

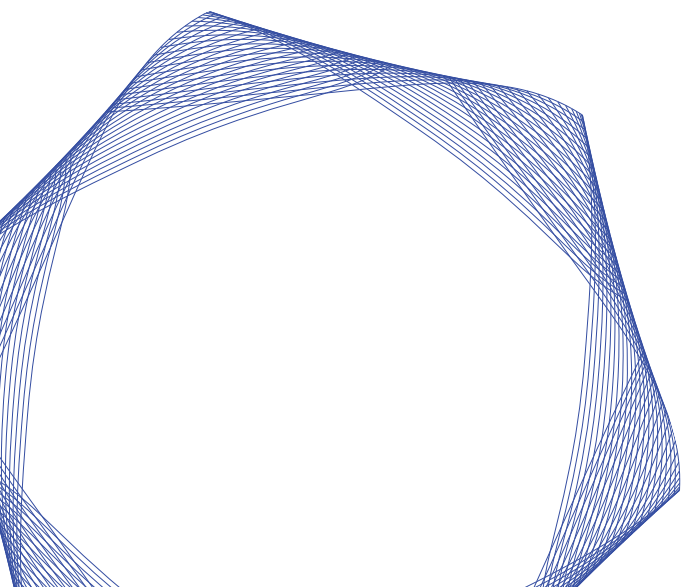
Endogenous UMIs as quantifiable reporter elements – validation studies & applications in rAAV vectorology

Patrick Wilmott

**Children's Medical Research Institute,
Faculty of Medicine and Health – The University of Sydney**

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

2023



Declaration

I declare that the work described in this thesis has not been submitted previously for any other degree at any university or institution. The contents of this thesis are the original work of the author except where due reference is made in the text.

Acknowledgements

Eagerly I dedicate this thesis's first original sentence to express my gratitude to the Australian National Measurement Institute's Bioanalysis team. All of us roll the dice when we attempt to build on human capabilities – us arguably more than others when we set upon my ITR project. The serum cortisol will halve time and again but the learning experiences, true to your assurances, are indelible. They will remain one of the most practically and intellectually rewarding experiences I have acquired in this doctorate. I owe my deepest thanks to Daniel Burke, Anna Baoutina, Jacob McLaughlin and Michael Forbes-Smith for their patience, mentorship, and quality company. These thanks I extend to the rest of the Bioanalysis Team for the genuinely welcoming and supportive atmosphere I have enjoyed at the NMI. I look forward to meeting again before long.

My principal supervisor, Assoc. Prof. Leszek Lisowski. I can't say what a privilege I've had taking this leap in the dark with you in my corner. I've had the opportunity in many a progress evaluation meeting to share what a positive experience it's been, but not to thank you for the freedom and support I've enjoyed under your supervision, to follow my nose in these formative years. It was not the easier option for either of us, but it was more important. Thank you.

To my auxiliary supervisor Prof. Ian Alexander, your endorsement more than anyone else's convinced me I should make a go of this. It will always benefit me to have entered science by way of your community, around your eight-cylinder brain and your embodiment of scientific humility.

As I scan the Translational Vectorology Research Unit in my mind I come across no one in our team who hasn't earned my gratitude in the last four years. The only thing you have deprived me of is the opportunity to complain about colleagues at dinner parties. You are also the reason I can look back on my doctoral research with nostalgia. Tasks involving animals are gratefully accredited to colleagues Cindy Zhu, Sophia Liao, Inna Navarro, and Matthieu Drouyer.

To my friends, who made me feel safe and okay whenever my PhD got the slightest bit difficult, you belong on this page. Thanks to you the worst of it was conquerable and the rest was more or less a joy.

My big sister, Clancy, that you've accomplished what you have without a Clancy is one my most enduring sources of inspiration. I don't know that I would have. That said, having an Emma can't hurt. My little brother, Joseph, I'm prouder of you than you are of me.

To my parents, Luke and Lee Wilmott, your unassailable love and support have meant everything. Nothing in these 240 pages was as difficult to put into words.

Abstract

In the creation of recombinant adeno-associated viral (rAAV) vectors, terminal DNA elements known ITRs (inverted terminal repeats) of the direct the intracellular synthesis and packaging of nonviral DNA. The need to clonally amplify ITR sequences in one form or another thereby underlies the existence of all rAAV clinical products and research materials worldwide. Their tendency to form strong nonduplex structures raises problems. The genetic precursors to rAAV vectors – typically prokaryotic plasmids – are known to possess heterogenous ITR sequences as a result of replicational instability, the effects of which on vector yield and efficacy are unclear and have not been systematically explored. To shed much-needed light on this decades-old problem, I utilised unique molecular identifiers (UMIs) as reporter elements for different rAAV plasmid preparations, so that massively parallel sequencing could be used to analyse their DNA and RNA derivatives through the course of production and in vivo gene transfer. The range of vector potencies observed, while not calamitous, definitively erases the notion that this problem can be further overlooked.

The success of this unconventional strategy proved to be an equally notable outcome, offering unprecedented insights into population kinetics, and achieving quantitative consistency between biological replicates comparable to q/dPCR measurement replicates of single samples. This triggered concerted efforts to formally investigate the capabilities of UMIs used in this fashion. The probabilistic principles underlying the technique were formalised and empirically validated, confirming precision capabilities akin if not superior to dPCR and qPCR at equivalent levels of stringency. Experiments also revealed a pattern of measurement bias with potentially adverse implications for other areas of count analysis including differential gene expression. These combined results urge us to build from the groundwork established herein to pursue a deeper technical and theoretical grip on this technique. The precision, robustness and strategic versatility showcased in this thesis make UMIs a strong complement to keystone techniques in bioanalysis when used as quantifiable reporter elements, not only to meet the coming decades’ vectorological challenges but as a tool for molecular biologists writ large.

Glossary ^{*}neologisms

	Unique molecular identifier (UMI)	A variable base sequence within an otherwise constant sequence, from which some aspect of an RNA/DNA molecule’s identity can be inferred
	Barcode	A common synonym for UMI
iii	*Reporter variant	A nucleic acid molecule containing a specific reporter barcode sequence, or the host entity to which that nucleic acid belongs
	*Reporter barcode	A UMI whose purpose is to allow the ratio of reporter variants in a

The Binomial Distribution

Term	context: binomial distribution	context: qRBA
BC_p	'Success', '1', one of two possible outcomes of a single Bernoulli trial	A specific reporter barcode
BC_q	'Failure', '0', the other of two possible outcomes of a single Bernoulli trial	A specific reporter barcode that isn't BC_p
c	Number of Bernoulli trials in a Bernoulli process	Readcount
p	P(success) for one Bernoulli trial,	Probability of an NGS read containing BC_p
q	P(failure) for one Bernoulli trial	Probability of an NGS read not containing BC_p
r	$p:q$	The ratio between two reporter variants (that physically exist). The measurand in the context of qRBA.
p_s	The <i>fraction</i> of Bernoulli trials from a Bernoulli process resulting in a success	The <i>fraction</i> of NGS reads containing BC_p
c_{ps}	The <i>number</i> of trials in a Bernoulli process that resulted in a success	The <i>number</i> of NGS reads containing BC_p
q_s	The <i>fraction</i> of Bernoulli trials in a Bernoulli process that did not result in a success	The <i>fraction</i> of NGS reads not containing BC_p
c_{qs}	The <i>number</i> of trials from a Bernoulli process not resulting in a success	The <i>number</i> of NGS reads not containing BC_p
r_s	The ratio of successes to failures in a Bernoulli process – i.e. $p_s:q_s$	The <i>ratio</i> of NGS reads containing BC_p and NGS reads containing BC_q
n	The number of independent Bernoulli <i>processes</i>	The number of qRBA measurement replicates

Mathematical Expressions

The binomial distribution

$P(p_s c, p)$	The probability of p_s being the fraction of Bernoulli trials resulting in a success. Calculated from the binomial probability mass function (right)	$\frac{c!}{(cp_s)!(cq_s)!} p^{cp_s} q^{cq_s}$
SD^2 or σ^2	The variance of p_s in a normal approximation of the binomial distribution	cpq
SD or σ	The standard deviation of p_s in a normal approximation of the binomial distribution	$\sqrt{cpq}$
ζ	('zeta') aka Sum.SD Hypothesised standard deviation of r_s according to the 'summing' method	$\left(\frac{p}{q}\right) \times \frac{1}{\sqrt{cpq}}$
ζ_s	The value of ζ inferred from outcome(s) of a Bernoulli process, where n is the number of independent Bernoulli processes.	$\frac{1}{n} \sum_{i=1}^n \left[\left(\frac{p_s}{q_s}\right) \times \frac{1}{\sqrt{c(p_s q_s)}} \right]_i$

Descriptive statistics

SD	Sample standard deviation (used unless otherwise specified)	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$
SD.P	Population standard deviation	$\sum_{i=1}^n (x_i - \bar{x})^2$
RSD	Relative standard deviation aka Coefficient of Variation (CV); standard deviation as a proportion of the average	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} / \bar{x}$
(RSD)_{product} (RSD)_{quotient}	Error estimated for the product or quotient of two measurands	$\sqrt{(RSD_1)^2 + (RSD_2)^2}$
S.E.Y	Standard error of the regression	$\frac{1}{n} \sum_{i=1}^n \left \frac{y_{pred.} - y_{obs.}}{y_{pred.}} \right _i$

List of Abbreviations

Species and strains

AAV	adeno-associated virus
Ad(5)	adenovirus serotype 5
FRG	Fah ^{-/-} Rag ^{-/-} Il2rg ^{-/-} mouse
HEK293	Human embryonic kidney 293
HEK293T	HEK293 derivative containing SV40 T antigen
HSV	herpes simplex virus
NHP	non-human primate
Sf9	Spodoptera frugiperda 9

Techniques & technologies

ATMP	advanced therapy medicinal product
CE	capillary electrophoresis
CRISPR	clustered regularly interspersed palindromic repeats
ddPCR	droplet digital polymerase chain reaction
DGE	differential gene expression
dPCR	digital PCR
ESI-MS	Electrospray ionisation mass spectrometry
HPLC	high performance liquid chromatography
LC-MS	liquid chromatography - mass spectrometry
MS	mass spectrometry
MWCO	molecular weight cut-off
NGS	next generation sequencing
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
SMRT	single-molecule real-time [sequencing]
UMI	unique molecular identifier

Chemicals & reagents

Amp	ampicillin
CsCl	caesium chloride
dH₂O	deionised H ₂ O
DMSO	dimethyl sulphoxide
EDTA	Ethylenediaminetetraacetic acid
EtBr	Ethidium bromide
NaOAc	sodium acetate
PEG	poly(ethylene glycol)
PEI	polyethylenimine
EtOH	ethanol
SDS	sodium dodecyl sulphate
TAE	tris base, acetic acid and EDTA
Tris	tris(hydroxymethyl)aminomethane
BSA	bovine serum albumin
DMEM	Dulbecco's Modified Eagle Medium
DPBS	Dulbecco's phosphate-buffered saline
FBS	foetal bovine serum
HEPES	N-2-hydroxyethylpiperazine-N-2-ethane sulfonic acid
LB	lysogeny broth
LB-Amp	lysogeny broth with ampicillin (1× = 100µg/mL)
PBS	phosphate-buffered saline
TAE	tris base, acetic acid and EDTA

Analysis

GoF	goodness-of-fit
R²	coefficient of determination
RFO	relative fold output
RSD	relative standard deviation
SD	standard deviation
SEG	standard error of the gradient
SEM	standard error of the mean
SEY	standard error of the regression
PMF	probability mass function

Cell & molecular biology

AAP	assembly activating protein
cDNA	complementary DNA
CNS	central nervous system
CV	coefficient of variation
DDR	DNA damage response
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
eGFP	enhanced green fluorescent protein
low-MW	low molecular weight
mRNA	messenger RNA
rAAV	recombinant adeno-associated virus
RHR	rolling hairpin replication
RNA	ribonucleic acid
scAAV	self-complementary AAV
ssAAV	single-stranded AAV
TLR-9	Toll-like receptor 9
WT	wild-type

Organisations

AEC	Animal Ethics Committee
CHW	Children's Hospital, Westmead
CMRI	Children's Medical Research Institute
EMA	European Medicines Agency
FDA	Food and Drug Administration
ISO	International Standards Organisation
NMI	National Measurement Institute

Units

bp	base pair
cp	copies
g	relative centrifugal units
kb	kilo base
kDa	kilo Dalton
nt	nucleotide
p.t.	post transfection
pmol	picomol
rpm	rotations per minute
rxn	reaction
V	volts
v/v	volume per volume
vg	vector genome
w/v	weight per volume
µg	micrograms
µL	microlitres
µM	micromolar
µmol	micromoles

Nucleic acid elements

AmpR	gene conferring ampicillin resistance	RBE	Rep binding element
bGH-polyA	bovine growth hormone polyadenylation signal	RBE'	Rep binding element (CTTTG)
cap	ORF encoding capsid monomers	rep	ORF encoded AAV Rep proteins
egfp	gene encoding enhanced green fluorescent protein	RRS	restriction recognition site
ITR	inverted terminal repeat	RSVp	Rous sarcoma virus promotor
ITR2	AAV serotype 2 ITR	trs	terminal resolution site
luc2	gene encoding luciferase 2	WPRE	woodchuck hepatitis virus post-transcriptional regulatory element
ORF	open-reading frame		
ori	origin of replication		

Table of Contents

CHAPTER 1 – Preface & Outline

CHAPTER 2 – Methodology

p.8	2.1	Reagents and Equipment
p.13	2.2	Standard protocols
p.18	2.3	Methods relating to Chapter 4
p.20	2.4	Methods relating to Chapter 6.1 (rAAV production)

CHAPTER 3 – Endogenous UMIs as quantitative reporter elements – what, why and how

p.33	3.1	Introduction
p.36	3.2	What are reporter barcodes, and what is qRBA?
p.42	3.3	The substantial analytical benefits of qRBA measurement
p.47	3.4	Relative Fold Output and Complete Binarisation
p.51	3.5	A binomial distribution model for qRBA count analysis
p.59	3.6	Conclusions

CHAPTER 4 – A metrological exploration of qRBA

p.62	4.1	Design and construction of experimental conjugate reporter plasmids
p.66	4.2	Testing ζ using in silico simulated Bernoulli processes
p.70	4.3	Empirically testing the binomial model of Illumina-derived qRBA measurement
p.86	4.4	Systematic error in qRBA measurement
p.107	4.5	Conclusions

CHAPTER 5 – Plasmid ITR instability in the rAAV vector system

p.109	5.1	Introduction
p.111	5.2	The AAV virus and its vector derivative – a snapshot
p.116	5.3	The biotechnological context of plasmid ITR instability
p.119	5.4	A deeper dive into AAV genome biology
p.127	5.5	The ITR instability problem
p.134	5.6	Conclusions

CHAPTER 6 – Exploring the impact of ITR replicative instability in the rAAV vector system

p.138	6.1	Plasmid ITR heterogeneity and rAAV production in adherent HEK293T cells
p.160	6.2	Truncations in plasmid 3'ITRs and gene transfer to murine liver and skeletal muscle
p.172	6.3	Conclusions

CHAPTER 7 – Discussion & Conclusions

p.175	7.1	Discussion
p.180	7.2	Conclusions
p.181	7.3	References

1

Preface & Outline

This thesis contains two seemingly disparate research themes, each independent in scope but directly synergistic. Though connected by compelling logic, it is difficult if not disingenuous to map their combined achievements to a unitary hypothesis or research aim. Initially, my doctoral research was driven wholly in response to unintended sequence heterogeneity in recombinant adeno-associated viral (rAAV) inverted terminal repeats (ITRs) and its implications for the vector technology. In that context of that work, the use of unique molecular identifiers (UMIs) in quantitative bioassays demonstrated outstanding potential, triggering efforts to systematically develop and validate this investigational strategy. Not only did these efforts greatly enhance my analysis of rAAV ITRs in the recombinant vector system, they strongly supported the broader adoption of UMIs for analysing the translocation, replication, expression, or modification of nucleic acids. These validation experiments were also able to identify and characterise systematic patterns of error in NGS count analysis. In addition to highlighting a clear next step to refining this technique, this error identified is highly relevant to the broader application of NGS count analysis including transcriptomics. Though widely applicable, the use of UMIs is conceptually distinct and has specific design requirements, calling for a distinct term. The term this thesis offers is *quantitative reporter barcode analysis*, or qRBA.

Aim 1

To demonstrate that UMIs can and should be used as explicitly quantitative research tools in molecular biology

- 1a** To determine whether random variation in UMI count ratios is governed by binomial statistics.
- 1b** To test the accuracy of qRBA using a set of plasmids containing two different qRBA target regions.

Aim 2

To explore the impact of plasmid ITR heterogeneity on rAAV production and performance (with qRBA as the core methodology)

- 2a** To determine whether plasmid ITR heterogeneity impacts the yield of packageable rAAV genomes and packaged rAAV vector capsids.
- 2b** To determine whether plasmid ITR deletions impacts the transfer and expression of exogenous transgenes in murine liver and skeletal muscle

An atypical structure (Fig. 1.1.1) is necessary to serve this thesis's two narrative cores. Each aim will have a review chapter followed by a research chapter. A final discussion chapter will reiterate and reflect on the combined significance of both Aims. The thesis (as its title would suggest) will lead with Aim 1, due to its greater breadth of impact. However, readers may find it useful to note that these aims are presented achronologically. 'Aim 1' was formulated directly in response to data generated earlier in 'Aim 2'. Using qRBA as the core investigational strategy into vector production and gene transfer, changes in the relative population size of rAAV transgene sequences could be observed with unprecedented clarity. Measurement consistency across groups of wholly replicated experimental samples were comparable with previously published q/dPCR precision with purified rAAV samples. Focus then shifted to qRBA for two reasons: (1) To formally validate the investigational approach and identify strengths and weaknesses, thereby disambiguating certain outcomes and vindicating novel observations. (2) To properly countenance the potential of qRBA that had been made apparent. Conversely, Aim 2 serves to demonstrate that qRBA has clear and immediate practical significance, both within and outside rAAV vector research.



Figure 1.1.1 A guide to the structure of this thesis and its two research themes| This thesis has two interlinked themes and is not well suited to a typical thesis structure with a single introductory review. Both themes are represented by an aim, a critical review, and a research chapter. Material and methodology used for all research is detailed in Chapter 2, and a final chapter will summarise the successes of both topic areas and reflect on their shared scientific significance.

Chapter 3

Chapter 3 has characteristics of a treatise, as well as a review. The core practical and conceptual elements of qRBA, though conceived in a novel way, are not original to thesis. This Chapter's original contribution is to establish qRBA as a distinct, valuable, and implementable application of unique molecular identifiers. To this end it undertakes a seemingly eclectic group of tasks. It makes clear the novelty of qRBA with respect to other applications with similar attributes and highlights the design constraints that suit UMIs to the task of relative measurements of a specific nucleic acid target. In doing so, it will provide context to qRBA's development and highlight the contribution of previous researchers, particularly in the rAAV vector field. It will systematically argue why and in what sense qRBA can routinely outperform qPCR and dPCR. It will also supply a prototype mathematical toolset in support of qRBA measurement and analysis. It describes *relative fold output* as a means of dealing with fundamentally relative information, and draws on binomial statistics to predict random measurement variation.

Chapter 4

Chapter 4 is an original research chapter that will test the precision and accuracy of qRBA measurement. It uses Illumina MiSeq to perform population sequencing on mixture of 150bp PCR products containing 10 possible UMI sequences in a specific location. The PCR target is located on an experimental plasmid, for which 10 UMI variants were cloned. Measurements were replicated from the point of PCR amplification and the point of population sequencing. In theory (as explained in Chapter 3), a binomial probability model can be used to predict random variation between measurement replicates. This theory was successful, thus supporting a formal model to explain the relationship between *measurand* and *measurement* – significantly elevating the credentials of qRBA in bioanalysis. The question of accuracy was addressed by performing the measurement on an alternative PCR target on the same set of plasmids. Unfortunately, this test was successful in identifying highly replicable, highly patternistic non-random error. While the orthogonal targets agreed in qualitative terms, disagreements approached two-fold at either end of the measured range. Ad-hoc analyses were able to exclude sequencing errors, coverage bias, contamination, or probabilistic skew as the cause. While the basis could not be determined, the mode of error was described

mathematically and conclusions were drawn as to how qRBA (and similar) assays may be interpreted, including those performed in my studies of rAAV ITRs.

Chapter 5

Chapter 5 is a review of rAAV ITRs and the problem of plasmid ITR heterogeneity. Recombinant adeno-associated viral (rAAV) vectors are tools for delivery of nonviral genetic material to patient tissues via the infectious pathways of adeno-associated virus. They consist of recombinant AAV capsids packaged with linear single-stranded DNA, which terminates on both ends in hammerhead-shaped 145nt elements called ITRs. The ITRs are the only viral DNA elements retained in the recombinant vector system, as they are required for processing and packaging of nonviral DNA within cultured producer cells. Prokaryotic plasmids have historically acted as the precursors to packageable rAAV genomes. Unfortunately, they are typically heterogenous in one or both ITR loci due to replicational instability. Despite the centrality of ITRs in (r)AAV genome biology, the impact of this heterogeneity on vector production and performance has not been systematically studied. This chapter will provide the requisite biology needed to support an understanding of Chapter 6, explain the significance of rAAV ITRs in the biotechnological landscape, explore the roles of the ITRs in the gene transfer process, and discuss the challenges hindering progress on this issue in spite of its clear importance.

Chapter 6

Chapter 6 seeks to explore the potential impact of plasmid ITR heterogeneity on rAAV production yields and/or the outcomes of rAAV-mediated gene transfer. A novel approach to this problem was to use qRBA as the core investigational strategy. Clones of an rAAV plasmid with various 3'ITR sequences (including a wild-type-3'ITR-derived prep) were prepared with clone-specific hexanucleotide UMIs inside a 150bp PCR target region. qRBA analysis of UMI composition was carried out on plasmid mixtures and their nucleic acid derivatives through successive stages of production and gene transfer. Highly replicable fold changes in UMI composition demonstrated clearly that deletions in the 3'ITR negatively impacted vector production, predominantly due to lowered efficiency converting ITR-flanked plasmid sequences into replicable ssDNA. Interestingly, despite these unambiguous internal population

changes, I was unable to detect a correlation between ITR integrity (as a continuous variable) and total vector yield. The strongest predictor of total yield was the amount of plasmid measured in producer cell nuclei 24h post-transfection. Two ancillary findings were also significant. Firstly, a negative correlation between transgene delivery and transgene expression was clearly indicated in the UMI sequence data. The wild-type-derived rAAV vector prep differed from mutant-derived preps in that the pattern was not tissue- or timepoint-dependent. The precise relationship displayed tissue- and timepoint-dependence in the mutant-ITR-derived rAAV vectors, but not in the wild-type-ITR-derived vectors. Secondly, subpopulations bearing different UMIs exhibited different efficiencies despite having been derived from the same ITR variant, and grown and extracted from the same culture of bacterial transformants. The collective results firmly establish that plasmid ITR heterogeneity, while non-fatal, is a variable of concern in the translation of rAAVs and support further research and improved methodologies to better understand the nature and extent of the problem.

Mathematics in this thesis: scope and accessibility

In my capacity as a student in molecular biology, I have endeavoured to investigate and use qRBA analysis in the most logical way possible. For better or worse, this ultimately called upon undergraduate statistical theories and various basic mathematical skills. Since I don't possess the relevant training, I have not tasked myself with making original contributions to either of these areas *per se*; their originality lies in their application to qRBA count analysis. The mathematical ideas expressed in this thesis are not expert perspectives (though their quality has been subject to external expert examination). My willingness to share them rested on their successful empirical validation (reported in Chapter 4). The goal of sharing them with a multi-disciplinary audience is to make clear the need for these sources of logic and the value in addressing those needs. The purpose of this model, as is true for most models, is to be improved.

Molecular biologists are fully capable of following these ideas but do not routinely engage with statistics. Where possible, I have refrained from referring such readers away from the text at hand to brush up on the necessary theory. To do so would treat molecular biologists as a secondary audience, where in actuality they are among most important readers I am attempting to persuade. By necessity, a considerable amount of text in Chapters 3 and 4 will feel overly discursive or expository to those who do possess such credentials. Extensive feedback has informed me that this excess is required to ensure my arguments are accessible to the widest

possible audience. Much of the expository content relating to the binomial distribution constitutes ‘common knowledge’ and has been written in consultation with peer-reviewed secondary sources authored by Rosner¹ and Wypij².

Work not included in this thesis

A significant portion of my doctoral research was not included in Chapter entries, in accordance with The University of Sydney Thesis and Examination of Higher Degrees by Research Policy 2015 Section 8(6) stipulation that only material serving the argument of the thesis be included in the main text. This regrettably excludes discontinued work that was undertaken with the Australian National Measurement Institute. With the support of scientists in that laboratory, a project was initiated to develop a methodological framework for measurement of ITR heterogeneity in plasmid DNA. An additional endeavour was to develop a qRBA technique grounded in mass spectrometry rather than DNA population sequencing. The decision to discontinue these projects was a difficult act of calculus rather than one clearly attributable to one or two results. Aside from specific hurdles that arose, the decision was affected by logistical factors, changes in the technological landscape, and complications arising from the COVID-19 pandemic. Synopses of these projects are appended to the thesis. These appendices are not submitted as direct contributions to my candidacy. Rather, they are included to informally share concepts and technical endeavours that arose as part of my doctorate. These appendices are light in technical detail in order to comfortably avoid disclosure of intellectual property, ensuring that access to this thesis by other researchers is unrestricted.

Image attributions

All figures in the manuscript were prepared by its author using Adobe Illustrator (versions 25.1 through 27.2) with the following exceptions: mouse silhouettes in Figures 3.1.1 and 6.2.1 (which were licensed for use without attribution), Figure 4.3.1 (prepared by modifying an image generated using SnapGene software – accredited in figure caption) and Figure 3.4.3 (prepared using an open-source online 3D plotting tool made available through by LibreTexts.org – URL in figure caption). Any resemblance between vector images in this thesis and those encountered elsewhere is due to the same physical objects being used as models (e.g. Eppendorf brand 1.5mL microfuge tubes).

2

Methodology

In the course of my doctoral research, I have prepared visualised protocols for my own ease of reference and for communicating methodologies to others. These have been included in this thesis and amended where necessary to ensure that details (alongside this Chapter's written portion) are sufficient to repeat the work and to interrogate whether the methodology supported the research aims. Most of these visualised protocols occupy full A4 pages, designed to be useful at a glance in the laboratory. It is preferable then that the majority are allocated to the end of the Chapter, otherwise, they are guaranteed to obstruct the flow of information significantly.

2.1 Reagents and Equipment

Tables 2.1.1 to 2.1.5 detail the resources used in this doctoral project and their manufacturers, PCR protocols, oligonucleotide sequences used throughout and a list of buffer constituents that are not otherwise specified in the text or figures of this Chapter.

Table 2.1.1 A list of equipment, disposables and reagents used in this doctoral project

Cat. No.	Manufacturer	Resource
SLGP033RS	MilliporeSigma	0.22um syringe filter
324000	Interpath Services	0.2mL 8-tube strips (capped)
CLS3745-1000EA	Sigma-Aldrich	0.2mL PCR tubes, capped
10787026	Thermo Fisher	1 kb plus DNA ladder
30125150	Eppendorf	1.5mL microfuge tubes
352051	Corning	14mL round-bottom aerated polystyrene test tubes
353025	Corning	150mm TC dishes
352095	Corning	15mL Polystyrene Conical Tube
NN 1838R	Terumo	18G needle
AM9530G	LifeTechnologies	1M MgCl ₂ (100mL)
15567-027	LifeTechnologies	1M Tris-HCl pH 7.5
15568-025	LifeTechnologies	1M Tris-HCl pH 8.0
SD8141	Astral	1M Tris-HCl pH 8.5
3131-00	Interpath Services	200µL 8 strip with 8 strip dome caps
CLS430776-102EA	Sigma-Aldrich	250mL centrifuge tubes
HSP3805	Bio-Rad	384-well qPCR plates
B1015-2	Beckmann-Coulter	3mL Inoculum water
352070	Corning	50mL Polystyrene Conical Tube
24710-030	LifeTechnologies	50x TAE Buffer DNA Typing Grade
SS+05L	Terumo	5mL syringe
353046	Corning	6-well plates
AS373	CSA Scientific	Agar bacteriological grade
BIO-41025	Bioline	Agarose
UFC910024	MilliporeSigma	Amicon® Ultra-15 Centrifugal Filter Unit
AB0028	Bio-Basic	Ampicillin (dry)
1.00983.2511	MilliporeSigma	Analytical EtOH
A94469	Beckmann-Coulter	Beckman Coulter Optima XPN ultracentrifuge
1.01695.0002	MilliporeSigma	Benzonase 250U/uL
172-5120	Bio-Rad	BioRad iTaq Universal SYBR Green Supermix
1225W62	Labconco	Blue anodized aluminium cool rack
5406000682	Eppendorf	Centrifuge 5425 R
1855484	Bio-Rad	CFX384 Touch Real-Time PCR Detection System
C0549	Thermo Fisher	Chloroform:isoamyl (24:1, v/v)
430776	Sigma-Aldrich	Corning 250mL Centrifuge Tubes
C29821	New England Biolabs	DH5-α
41966	Gibco	DMEM
D2650-5	Sigma-Aldrich	DMSO
D4008	Zymo Research	DNA Gel Recovery Kit
M0303	New England Biolabs	DNase I (RNase-free)
N0447	New England Biolabs	dNTPs
D8537	Sigma-Aldrich	DPBS
14040216	Thermo Fisher	DPBS (+Ca ²⁺ , +Mg ²⁺)
AM9262	Thermo Fisher	EDTA 0.5M

Table 2.1.1 (cont.)

Cat. No.	Manufacturer	Resource
12391	Qiagen	EndoFree Plasmid Giga Kit
E1385	Sigma-Aldrich	Ethidium bromide
12006C	Sigma-Aldrich	Fetal Bovine Serum (FBS)
UFC210024	Fisher Scientific	Gibco MWCO
15630130	Gibco	HEPES
71196-100ML	MilliporeSigma	MgCl ₂ 1M solution
S6546-1L	Sigma-Aldrich	NaCl 5M
S2889	Sigma-Aldrich	NaOAc 3M
74385	Sigma-Aldrich	Nonidet P-40 substitute
31985	Gibco	OptiMem media, 500 mL
D1556-250mL	Sigma-Aldrich	OptiPrep Density Gradient Medium
361625	Beckmann-Coulter	OptiSeal tubes
44I200005	Eppendorf	Orbital shaking incubator (Innova S44i)
BP2331-1KG	Fisher Scientific	PEG 8000
23966-2	Polysciences	PEI (polyethylenimine MW 25,000)
15140122	Gibco	Penicillin-Streptomycin 10,000U/mL
P0290	Sigma-Aldrich	Phenol red solution
15593031	Thermo Fisher	Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v)
24040-032	Gibco	Pluronic F-68 10% Solution
81268-5KG	Sigma-Aldrich	Poly(ethylene glycol) 8,000
352070 & 352096	Corning	Polypropylene conical centrifuge tubes
60135-250ML	Sigma-Aldrich	Potassium chloride solution BioUltra, ~3 M in H ₂ O
UN 12 19	Ajax Finechem	Propanol
A31231	Invitrogen	PureLink™ Expi Endotoxin-Free Maxi Plasmid Purification Kit
B7024	New England Biolabs	Purple running dye
M0491	New England Biolabs	Q5® High-Fidelity DNA Polymerase
27106	Qiagen	QIAprep Spin Miniprep Kit
28104	Qiagen	QIAquick® PCR Purification
(Various)	New England Biolabs	Restriction enzymes
337922	Beckmann-Coulter	Rotor 70Ti
746398	Sigma-Aldrich	Sodium chloride
S6546-1L	Sigma-Aldrich	Sodium chloride solution 5M
D6750	Sigma-Aldrich	Sodium Deoxycholate (DOC)
S6546	Sigma-Aldrich	Sodium chloride (NaCl) 5 M solution
18090200	Thermo Fisher	SuperScript IV Reverse Transcriptase
18091200	Thermo Fisher	SuperScript™ IV First-Strand Synthesis System
369702	Beckmann-Coulter	SX4750 Swinging bucket rotor
M0202	New England Biolabs	T4 DNA Ligase
M0302	New England Biolabs	T7 Endonuclease I
2232000083	Eppendorf	ThermoMixer C
T9424-100ML	Sigma-Aldrich	TRI Reagent
BIOSD8141	Astral	Tris-HCL Buffer
SD8127	Astral	Tris-HCl pH 8.0 1M

Table 2.1.1 (cont.)

Cat. No.	Manufacturer	Resource
SD8141	Astral	Tris-HCl pH 8.5 1M
12604021	Thermo Fisher	TrypLE dissociation reagent
12604-021	Gibco	TrypLE Express
16922-500G-F	Sigma-Aldrich	Tryptone (vegetable)
AM1907	Thermo Fisher	TURBO DNA-free™ Kit
P1379-100mL	Sigma-Aldrich	Tween-20 (100mL)
15553027	Thermo Fisher	UltraPure™ SDS Solution, 10%
500607	Du Pont	Virkon Tablets
B2000-2-E	Benchmark Scientific	Water bath
70161	Sigma-Aldrich	Yeast Extract
12604-021	Gibco	TrypLE Express
16922-500G-F	Sigma-Aldrich	Tryptone (vegetable)
AM1907	Thermo Fisher	TURBO DNA-free™ Kit
P1379-100mL	Sigma-Aldrich	Tween-20 (100mL)
15553027	Thermo Fisher	UltraPure™ SDS Solution, 10%
500607	Du Pont	Virkon Tablets
B2000-2-E	Benchmark Scientific	Water bath
70161	Sigma-Aldrich	Yeast Extract

Table 2.1.2 List of oligonucleotides used in this thesis

qRBA PCR		Sanger sequencing	
Target:	<i>egfp</i> reporter region	Target:	<i>egfp</i> reporter barcode
For	ggcatggacgagctgtacaag	Upstream	gagatcgtggactatgtggc
Rev	NNNNNNgcagcgtatccacatag	Downstream	gcattgtctgagtaggtgtc
Product:	150bp		
Target:	<i>egfp</i> reporter region	Target:	<i>luc2</i> reporter barcode
For	cgagattctcattaaggccaag	Upstream	tgagcaaagaccccaacgag
Rev	NNNNNNgatggctggcaactagaagg	Downstream	gggccacaactcctcataaag
Product:	150bp		
ITR Sanger sequencing		vector qPCR	
5'ITR	cgtatgcggtgtgaaatacc	Target	<i>egfp</i> gene
3'ITR	tggcacgacagggtttcc	For	tcaagatccgccacaacatc
		Rev	ttctcggtgggggtctttgtc
		Product:	148bp

Table 2.1.3 PCR reaction components used in this thesis

qRBA PCR		Vector qPCR	
Component	50µl rxn	Component	10µl rxn
Q5 Buffer	10.0	2x I-Taq SYBR Green	5.0
dNTPs	1.0	F primer (100µM)	0.08
For 100mM	0.25	R primer (100µM)	0.08
H2O	32.75	H2O	2.84
Q5® HiFi DNA Polymerase	0.5	Diluted template	2.0
Rev 10mM	2.5		
DNA	3.0		
		Total	10
Total	50		

Table 2.1.4 PCR thermocycling protocols used in this thesis

quantitative real-time PCR					
Cycles		Step	Temp	Duration	
1 ×		denaturation	95 °C	2 min	
39 ×		denaturation	95 °C	5 sec	
		annealing & extension	60 °C	15 sec	
		hold	12°C	∞	
qRBA PCR					
Cycles		Step	Temp (<i>egfp</i>)	Temp (<i>luc2</i>)	Duration
1 ×		denaturation	98°C	98°C	30 sec
25-32 ×		denaturation	94°C	94°C	10 sec
		annealing	68°C	67°C	15 sec
		extension	72°C	72°C	12 sec
1 ×		final extension	72°C	72°C	60 sec
		hold	12°C	12°C	∞

Table 2.1.5 List of Buffers and Constituents

Lysogeny broth (LB)	LB Agar
NaCl 10g/L, yeast extract 10g/L, tryptone 10g/L	NaCl 10g/L, yeast extract 10g/L, tryptone 10g/L, bacteriological agar 15g/L
AAV PCR alkaline digestion buffer	AAV PCR neutralization buffer
25mM NaOH, 0.2mM EDTA	40mM Tris-Hcl, 0.05% Tween20, pH 5.0
PBS-MgK	PBS-MgKNa
PBS with 1mM MgCl ₂ 2.5mM KCl	PBS with 1mM MgCl ₂ 2.5mM KCl 1M NaCl
Concentration Buffer	25% iodixanol – 30mL
PBS, 50mM NaCl, 0.001% Pluronic F68	17.5mL PBS-MgK 12.5 Optiprep 75µl phenol red
15% iodixanol – 30mL Total	60% iodixanol – 22mL
22.5mL PBS-MgKNa 7.5mL Optiprep	100% Optiprep 100µl phenol red
40% iodixanol – 22mL	
7.3mL PBS-MgK 14.7mL Optiprep	

2.2 Standard protocols

2.2.1 DNA Ligation

Ligations were performed using T4 DNA ligase following the suppliers instructions. Fragments to be ligated were fractionated using agarose gel electrophoresis and the DNA extracted from excised gel slices using a DNA Gel Recovery kit. Ligation reactions contained approximately 0.02 pmol of DNA fragments containing a bacterial ampicillin resistance gene and/or origin of replication ('backbone' fragments) and 0.06 pmol of other fragment(s). All ligation reactions reported in this thesis involved fragments with overhangs on both termini and the incubation took place at room temperature for one hour.

2.2.2 Restriction digestion

DNA samples were subject to preparative or analytical digestions using restriction endonucleases purchased from New England Biolabs, which were supplied with appropriate

reaction buffers. DNA quantity by default was 1µg though this varied in some instances according to how much of an extracted gel fragment was required for cloning. Incubation conditions varied according to the enzyme in accordance with the manufacturer's instructions.

2.2.3 Agarose gel electrophoresis

Ethidium-bromide-stained agarose gels were prepared at concentrations of 2% w/v for analysis of band lower than 200bp, 0.8% w/v for analysis of undigested DNA and 1% w/v unless otherwise specified. Where measurable, DNA was added in 1µg quantities in 1× purple loading dye (NEB) or 200ng for thin wells.

2.2.4 Establishing monoclonal plasmid preparations

Chemically competent DH5α cells were transformed via the heat-shock method. Ligation products (including no-backbone controls) (1µl) were added to 50µl volumes of thawed DH5α cells. The cell-DNA mixture was incubated on ice for 30mins and then subject to heat shock consisting of a 45 second incubation in a 42°C water bath. Cells were then returned to ice to recover for 20min before being transfer to 1.5mL microfuge tubes containing 200µl LB medium pre-warmed to 37°C. The tubes were shaken for 20min (37°C, 350rpm) in a benchtop orbital shaker (Eppendorf) and 50µl was transferred to 10cm plates of LB agar containing 100µg/mg ampicillin and incubated overnight at 37°C.

Monoclonal cultures of transformed bacteria were prepared by inoculating LB-Amp (3mL) with cells from single, well-separated colonies that were transferred to end of a disposable 10µl pipette tip. The inoculum was incubated overnight (37°C, 350rpm) in 14mL polypropylene aerated tubes.

Plasmids were prepared in intermediate quantities (~5-20µg – 'minipreps') before proceeding to 200-1000µg amounts ('maxipreps') of low-endotoxin plasmid necessary for tissue culture or other experimentation, so that the ligation product could be verified by restriction analysis and Sanger sequencing. Miniprep cultures were prepared using QIAprep Spin Miniprep Kits (Qiagen). Briefly, cultures were grown overnight in 3mL of LB-Amp. Cells from 1mL of the culture were pelleted and disrupted via alkaline lysis. Cell debris was pelleted and plasmid DNA was purified from the supernatant via hydroxy silica column purification and eluted in TE buffer.

Maxiprep cultures were prepared using PureLink™ Expi Endotoxin-Free Maxi Plasmid Purification Kits (Invitrogen). Briefly, cultures were allowed to enter the exponential growth phase in 3mL shaken volumes then transferred 50µl to sterile 120mL conical flasks for overnight incubation (37°C, 350rpm). After a maximum of 16h, cells were pelleted in 250mL centrifuge tubes disrupted via alkaline lysis. Cell debris was removed via spin filtration and the plasmids extracted from the clarified lysate via anion exchange chromatography. The plasmid was purified from the eluate via ethanol precipitation and the DNA pellet resuspended in TE buffer (unless otherwise specified).

2.2.5 Tissue DNA extraction

Harvested tissues were snap-frozen in liquid nitrogen and stored at -80°C until processing. For both DNA and RNA extractions, cells were disrupted in 1.5mL microfuge tubes using a handheld rotor-stator and sterile polypropylene pestles. Approximately 100mg of liver or skeletal muscle tissue were disrupted initially in 50µl Lysis Buffer, then with an additional 30µL. 300µg proteinase K was added and the samples were incubated overnight at 56°C in a benchtop orbital shaker (800rpm). DNA was separated from the bulk of proteins and lipids via phenol:chloroform extraction and two washes with chloroform (1:1 v/v) were performed to assist in excess phenol from the sample. A standard phenol:chloroform separation is visualised in a figure belonging to a later section (Fig. 2.4.5.2). The DNA in the aqueous phase was separated via EtOH precipitation and resuspended in TE buffer. Quantity and purity were assessed with a Nanodrop spectrophotometer.

2.2.6 Tissue RNA extraction

Approximately 30mg of liver or skeletal muscle tissue were disrupted in 1.5mL tubes. The tubes were held in an anodised aluminium cool rack equilibrated to -80°C resting on ice while tissues were disrupted. (I had the most success with RNA extraction using this ‘halfway’ method similar to grinding in liquid nitrogen. Staying semi-frozen allowed the cells to be efficiently ground, limited the activity of RNases and prevented the sample from over warming due to friction). Tissue was disrupted initially in 50µl TRIzol™ reagent, then in an additional 300µl. RNA was separated into an aqueous phase by adding 0.2× volumes of chloroform. RNA was extracted and from the aqueous phase via ethanol precipitation and the pellet washed twice with 80%

ethanol before resuspending in RNase-free water. Quantity and purity were assessed with a Nanodrop spectrophotometer and samples were stored at -80°C.

2.2.7 Construction of cDNA template from RNA samples

Construction of cDNA took place in two steps: DNase treatment and reverse-transcriptase-mediated synthesis of the cDNA strand. DNase treatment was carried out using TURBO DNA-free™ Kits (Invitrogen) and cDNA synthesis using SuperScript™ IV First-Strand Synthesis System (Thermo Fisher Scientific), with reagents included in both respective kits, according to the manufacturers instructions. Oligo(dT)₂₀ was used to prime complimentary strand synthesis. The protocol is visualised in Figure 2.2.1.

2.2.8 Tissue culture

HEK293 cells were maintained in DMEM containing 10% FBS and 1× penicillin-streptomycin and passaged at a ratio of 1:3 to 1:5 before reaching 80% confluence using TrypLE dissociation reagent. Cells were incubated at 37°C in humidified air containing 5% carbon dioxide. A 'PBS wash' consisted of warm DPBS buffer used to resuspend pelleted cells (10mL/dish), then aspirated away after repelleting. Cells pelleted for maintenance were spun at 100g for 5 minutes at ambient temperature.

2.2.9 Triple-transfection of HEK293 cells

Two hours prior to transfection, medium was replaced with a half volume of complete DMEM containing 25mM HEPES buffer. An Ad5 plasmid, a Rep-Cap plasmid and a vector plasmid (aka transfer plasmid) were mixed in a 3:1:1 ratio (22.5µg, 7.5µg and 7.5 µg per 150mm dish). To transfect, equal volumes of DNA and polyethylenimine diluted in Opti-MEM® (DNA:PEI at a 1:2 mass ratio) were combined to a final volume of 550µl per dish (or 33µl per 3cm well), including 10% overage. Complexation occurred by allowing the mixture to incubate at room temperature for 10 minutes before adding to the plate/well and swirling). Cells were harvested 72h post transfection unless otherwise specified, and processed as specified in later sections.

2.2.10 Measurement of vector concentration using qPCR

Vector concentrations in the crude samples were estimated using qPCR with NcoI-linearised pGL_ITR2.Ch6_ΔBssHII (quantified using a nanodrop spectrophotometer) as an external reference. The assay was performed using an iTaq Universal SYBR Green Supermix (Bio-Rad) on a CFX384 Touch Real-Time PCR Detection System instrument (Bio-Rad) using the protocol and primers specified in Tables 2.1.3-4. Iodixanol-purified samples were prepared as 1:40 dilutions for qPCR by adding 5µl to 47.5µL Alkaline Digestion Buffer and incubating at 95°C for 10mins then quickly allowing to chill on ice before addition of 47.5µL neutralisation buffer and 100µl of H₂O. Crude samples were prepared as 1:200 dilutions with the above steps, after initial treatment of 2µl in a 10µl DnaseI reaction in 1× DnaseI Buffer (New England Biolabs).

2.2.11 Data analysis and visualisation

Once read count data was obtained, all data handling and analysis was performed using Microsoft® Excel for Mac (Versions 16.45 through 16.70), with the exception of Shapiro-Wilks or D'Agostino-Pearson tests for normality, which were performed using the online resource StatsKingdom available through the URL <<https://www.statskingdom.com/shapiro-wilk-test-calculator.html>>. All data represented visually in this thesis were plotted in Excel and saved in Scalable Vector Graphic (SVG) format, then modified using Adobe Illustrator (versions 25.1 through 27.2). All figures (including all graphical assets other than typefaces) in the manuscript were prepared by the author with the following exceptions: mouse silhouettes in Figures 3.1.1 and 6.2.1 (for which use was permitted without attribution), Figure 2.3.1 (prepared by modifying an image generated using SnapGene software – accredited in figure caption) and Figure 3.4.3 (prepared using an open-source online 3D plotting tool made available through by LibreTexts.org – URL in figure caption). Any resemblance between vector images in this thesis and those encountered elsewhere is due to the same physical objects being used as models (e.g. Eppendorf brand 1.5mL microfuge tubes).

2.3 Methods relating to Chapter 4

2.3.1 Construction of plasmids containing conjugate reporter barcodes

The plasmid construct with the working label pCRB.Ch5 (pConjugate_Reporter_Barcodes) was prepared that contained two qRBA targets. Each 'qRBA target' consisted of a 6N reporter barcode within a PCR target region that produced a 150bp reporter amplicon. For this thesis, the barcodes were labelled according to their immediately adjacent (non-functional) genes *egfp* and *luc2*. A catalogue of *egfp*-adjacent barcodes were available in the laboratory, which have featured in a publication by Westhaus *et al.*³. Barcodes adjacent to *luc2* available were available from the pGL_ITR2.Ch6 plasmids featured in Chapter 6. These two constructs acted as donors of barcoded inserts for the final construct, with a wild-type AAV2 plasmid providing the recipient plasmid backbone.

A three-piece ligation strategy was devised making use of compatible PflMI and DraIII overhangs (Fig. 2.3.1). Three colonies were picked from each plate of transformants and gross structure was confirmed with analytical restriction digests using EtBr-stained agarose gel electrophoresis. Confirmed clones were submitted for Sanger sequencing, primed upstream and downstream from both reporter barcode loci using primers specified in Table 2.1.2. As a defensive measure, monoclonality in the final preparations was ensured by streaking the successful liquid cultures and inoculating LB medium with single colonies for preparation of maxipreps. Cell pellets were snap-frozen on dry ice and maxipreps proceeded upon receipt of Sanger sequencing confirmation of the intended barcode sequences. Plasmids were dissolved and stored in TE buffer.

2.3.2 qRBA measurement of reporter ratios in plasmid samples

Mixtures of pCRB.Ch5 reporter variants were prepared as targets for PCR amplification by a targeted fragmentation (a DraIII-HF restriction digest) and direct purification of the restriction fragments using a QIAquick® PCR Purification Kit (Qiagen). DraIII sites were the only convenient targets for this treatment but the enzyme DraIII-HF was not heat labile, necessitating separation of the DNA fragment. The *luc2*- and/or *egfp*-targeted PCR reactions were performed using 25 cycles and the protocol specified in Tables 2.1.3-4. 10µl of the PCR reaction mixture was reserved for gel electrophoresis to confirm the intended product, no

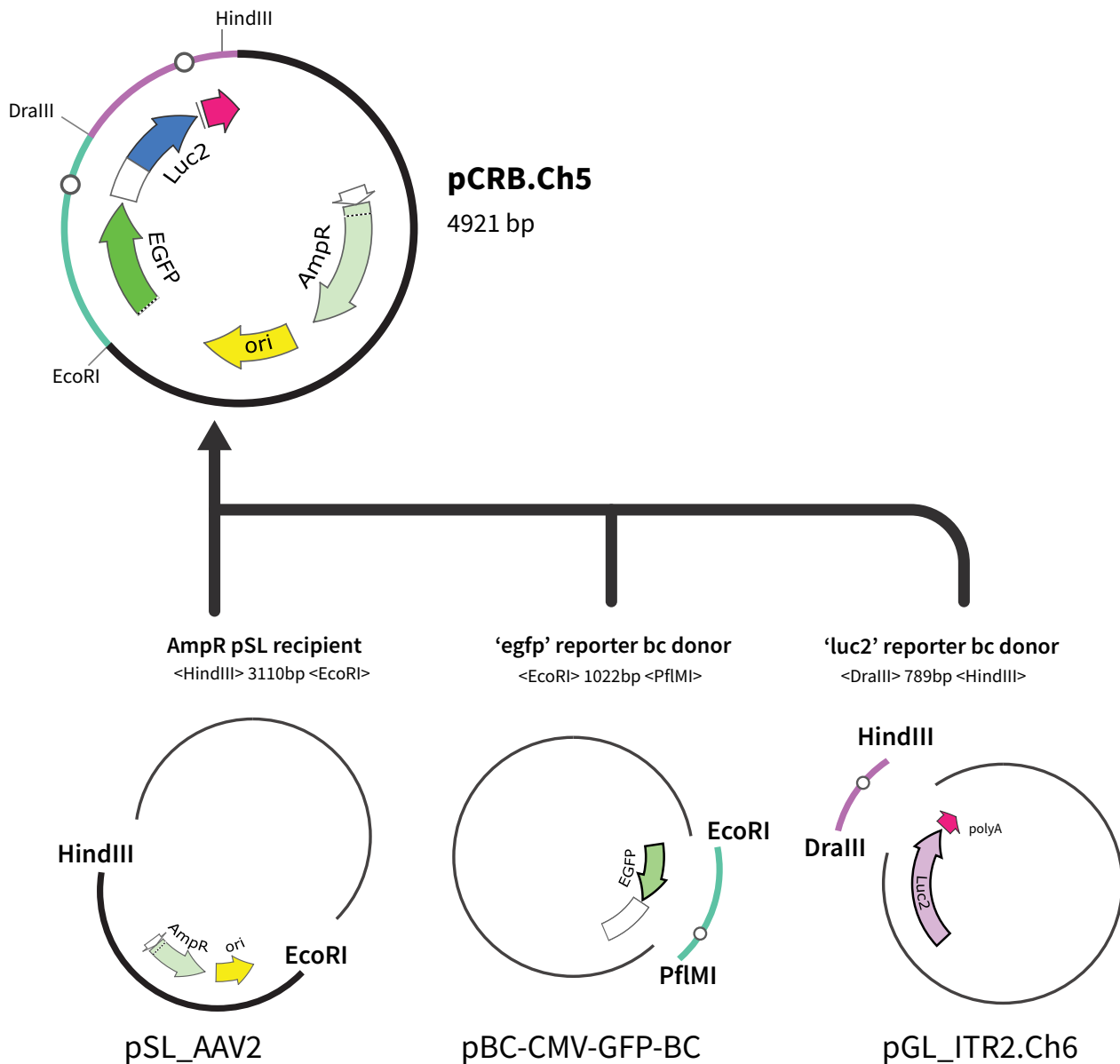


Figure 2.3.1 Cloning conjugate barcode variants pCRB.Ch5 | Plasmid constructs containing two independently targetable reporter barcodes ('conjugate reporter barcodes') was cloned using a three-piece ligation with compatible DralIII and PflMI sticky ends. Two fragments contained reporter barcode targets developed in prior projects in our laboratory, and a third provided a plasmid backbone recipient fragment allowing high-copy-number replication and conferring ampicillin resistance. Ten unique variants of pCRB.Ch6 were made, two of which had a luc2-adjacent barcode in common.

unwanted bands and no bands appearing in the no-template control. Upon confirmation, the PCR products were combined (in amounts described in Chapter 4) and purified using the QIAquick purification columns mentioned above. The 150bp product was eluted in EB buffer and 1µg samples were dried before submission for MiSeq 2× 150bp paired-end sequencing. The service provider (GENEWIZ Genomics, Suzhou) conducted NGS library preparation from the PCR amplicon sample, including ligation of the Illumina NGS adapters. The unpaired read data was received in FASTQ format. Pair merging was performed using the BBMerge tool⁴

and reporter barcodes tallied using an open-source barcode counting pipeline (ngs_barcodes v0.1.1 https://github.com/szsctt/ngs_barcodes/). Readers may refer to the GitHub repository for a detailed description. In short, the pipeline first identifies merged reads with the specified length and forward primer sequence and orients the base positions according to the position of the forward primer. Then, reads containing unique (either defined or undefined) UMI sequences in the specified base positions are enumerated and sorted according to index sequences. The qRBA assay process as a whole has been visually summarised in Figure 2.3.2.

2.4 Methods relating to Chapter 6.1 (rAAV production)

2.4.1 Subclonal isolation of 3'ITR plasmid variants

The reporter variants of the plasmid pSL-ITR2-RSV-GFP-Luc2-6nBC (referred to in this thesis under the alias pGL_ITR2.Ch6) were available in our laboratory from a prior project. To prepare barcoded 3'ITR variants of this plasmid, the variants were first subclonally isolated and then the barcodes were incorporated. Chemically competent DH5- α cells were transformed with ~1ng of the non-barcoded parent plasmid and 48 plated subclonal isolates (colonies) were grown as minipreps and screened for the presence of XmaI and AhdI restriction sites. Isolates that were wholly AhdI-negative (Δ AhdI), XmaI-negative (Δ XmaI), and negative for both sites (Δ BssHII) were identified. Subclones representing each of these profiles (three in total) were retained for downstream cloning, along with the wild-type-derived parent plasmid.

To insert unique barcodes into each 3'ITR variant, barcoded 2.5kb SacI-KpnI restriction fragments were used to replace the unbarcoded equivalent SacI-KpnI region using a two-piece ligation. To create α/β barcode pairs, both donor plasmids were mixed in the SacI+KpnI restriction digest mix and the 2.5kb fragments were co-extracted from an agarose gel slice following electrophoresis. Both α and β inserts were present in the ligation reaction and plasmid outgrowth proceeded directly from bacterial transformation with no intermediate colony selection. This method is summarised visually in later in the thesis, in Figure 6.1.1. Restriction analysis was used to reconfirm the plasmid identity and gross structure, and to verify the absence of unintended ligation products. Undigested plasmids were also analysed, to identify any viable ligation products that were equivocal in their restriction profile. Maxiprep cultures of the transformants were grown and the pellets snap frozen on dry ice and stored at -20°C until the ITRs were sequenced. For the pWT variant, lower plasmid yields are typical and

a larger quantity was required to create a dilution series (below). For this reason, a 3.5L culture was grown and a prep (yielding >1.5mg) was prepared using an EndoFree Plasmid Giga Kit (Qiagen).

A specialised ITR Sanger sequencing service (GENEWIZ Genetics, NJ) confirmed the clonally dominant sequences of these isolates in both the 3' and 5' ITR loci. Electropherogram traces are shared in Appendix I.

2.4.2 Preparation of variant mixes simulating plasmid ITR heterogeneity

Each 3'ITR plasmid variant was diluted to its mass equivalent of 100 fmol/ μ l (near to ~430ng/ μ l) using a NanodropTM spectrophotometer. This allowed for equal molar quantities to be transferred equivolumetrically, reducing dependence on accurate pipette calibration. A roughly 1:1:1 mix of the p Δ AhdI, p Δ XmaI and p Δ BssHII variants was prepared. After adding 800 μ l of the wild-type isolate to 200 μ l of the non-wild-type mix, two serial 1:1 dilutions were carried out in the non-wild-type mix. From qRBA measurements of wild-type barcode content (based on the sum of paired wild-type reporters), working labels were retroactively replaced with WT_82, WT_40 and WT_21 denoting 82% wild-type, 40% wild-type and 21% wild-type respectively.

2.4.3 Transfection of HEK293T cells in 6-well plates

To transfect, equal volumes of DNA and PEI diluted in Opti-MEM[®] (DNA:PEI at a 1:2 mass ratio) were combined to a final volume of 30 μ l per plate (+10% overage) and allowed to incubate at room temperature for 10 minutes before adding to each plate and gently swirling. Each well received ~100 fmol of mixed or individual transfer plasmids. A Rep2Cap(DJ) or Rep2-only helper plasmid in masses equivalent to 100fmol were assigned to their appropriate plates, along with 1.35 μ g of a pAd5 helper plasmid. 150mm dish preparations of reference cells were made in parallel.

The transfection mixtures (DNA in OptiMEM medium) were constituted using a multi-step mastermix strategy that minimised variation of shared elements due to liquid transfer. Other than the vector-plasmid-only transfections (containing only vector plasmid and OptiMEM), this mix strategy followed the hierarchy: pAd5 $\rightarrow$ pRep2/pRep2CapDJ $\rightarrow$ pWT/ Δ BssHII/p Δ AhdI/p Δ XmaI.

Five wells each of pWT, pWT_82, WT_40, pWT_21ΔBssHII, pΔAhdI, and pΔXmaI were transfected for all three sample types of interest.

2.4.4 Cell harvest and spiking

Reference cells were harvested ahead of experimental cells by dissociation (for nuclear plasmid and accumulated genome samples), addition of warm DPBS to 50mL, pelleting, one DPBS wash, and thorough resuspension in pre-warmed complete DMEM medium. The tube containing reference cells was intermittently agitated to maintain a homogenous suspension, especially prior to transfer into experimental wells.

Cells in nuclear plasmid and accumulated genome samples were not readily dissociable. To harvest these cells, medium was first transferred into labelled 15mL polypropylene tubes. To reduce variation due to shear and dissociation enzyme over-incubation, 500μl 0.5× TrypLE dissociation reagent in DPBS was added to the cell monolayer and allowed to incubate for 3-5 minutes at room temperature. Activity was then reduced with addition of 2mL complete medium while cells were spiked. At this stage, cells were readily dissociable with minimal shear. WT-mix wells were spiked with the addition of 1mL resuspended reference cells, mixed, and then transferred to the sample tube containing the spent culture medium. Each well was rinsed twice with 1mL complete medium and the volume added to the sample tube to recover residual cells. For the no-WT wells, resuspended reference cells were added to all four wells of a given replicate. The contents were then resuspended while being passed from well to well to ensure a homogenous recovery of cells. The rinse step was carried out as described above, with the medium passed from well to well in the same fashion.

2.4.5 Extraction of target nucleic acid species from harvested cells (and a novel cellular lysis technique)

The protocols for processing of harvested cells are visualised in Figures 2.4.5.1-3. Fractionation of cell nuclei 24 p.t. (Fig. 2.4.5.1) was carried out according to a published protocol⁵, in which cell membranes are lysed in 0.5% Nonidet™-P40 and nuclei separated via discontinuous sucrose gradient centrifugation. For this experiment, a nonconservative treatment could be applied (higher shear, longer incubations in the surfactant) since consistent retention of intact nuclei was not a priority, unlike removal of the cytosolic fraction.

Low-MW ‘Hirt’ DNA extraction (Fig. 2.4.5.2) was performed for transfected cell nuclei and accumulated vector genomes, using a previously published protocol⁶ with CsCl used in place of NaCl, following the advice of Arad⁷. My own modification was to first resuspend cells in Hirt buffer lysis containing 0% SDS and then lysed with the addition of an equal volume of chilled buffer containing well-precipitated 2× SDS (1.2%), against conventional wisdom. Addition of SDS in solution had led rapidly to lysis of some but not all cells, quickly releasing chromosomal DNA, forming a thick bolus that greatly hindered dispersal of the surfactant and lysis of remaining cells. By contrast, adding precipitated SDS to a chilled sample allowed the surfactant a brief opportunity to evenly disperse throughout the cell suspension before the onset of lysis. Rather than hindering SDS dissolution, the rapid loss of dissolved molecules to the lipid-bound compartment of the reaction led the SDS to dissolve much more quickly than suspensions of solid SDS allowed to warm in a water bath or incubator. Lysates free of turbidity (judged visually) were achieved in a dramatically shorter time in a lower volume of lysis buffer and with minimal mechanical assistance. Lower process volumes led to more concentrated low-MW DNA and avoided the need to split samples between multiple tubes in later steps. Proteinase K treatment was followed by a high-salt overnight 4°C incubation to precipitate chromosomal DNA. Chromosomal DNA and cell debris were pelleted by a 3h centrifugation at 15,000g and DNA was isolated from the supernatant via phenol:chloroform extraction and stored in TE buffer.

Crude vector isolates (Fig. 2.4.5.3) were obtained from the harvested cell pellets by first resuspending in benzonase buffer and lysing with three freeze-thaw cycles. The lysate was treated with benzonase (1h 37°C) and clarified centrifugally. Further cell debris was separated from the supernatant by an additional round of centrifugation after 1h incubation on ice in 1M CaCl₂.

2.4.6 Pre-PCR treatments and qRBA measurement

Preparations for PCR (including pre-PCR treatments) were carried out as depicted in Figure 2.4.6. Prior to PCR amplification of the barcoded region, samples (1µl per 20µl reaction) were treated with 20U MscI for one hour at 37°C. The MscI restriction recognition site was present in the AA’ region of sequenced ITR variants in both the linear or hairpin conformations. MscI could therefore be used to cleave away heterogenous termini (which may affect amplification efficiency⁸). In addition to MscI, accumulated genome samples were treated with 20U DpnI for

preferential amplification of replicationally active ssAAV genomes. A DpnI restriction recognition site is located in the barcode amplicon sequence, meaning that cleaved copies will not be amplifiable. Isolated vector samples (1µl) were treated with Dnase (20U 1h 37°C) and the enzyme was heat deactivated at 75°C for 15min. The temperature was raised to 95°C for an additional 5 minutes to denature capsid particles. The sample was then treated with MscI (as well as DpnI for consistency). These processes are summarised visually in Figure 2.4.6. PCR amplification for qRBA was carried out as per Table 2.1.3-4 and the Illumina sequencing process as previously described. Not however that unlike Chapter 4, the total volume of all indexed PCR samples were added to their respective libraries (minus 10µl for agarose gel analysis) and the NGS sequencing process itself was not replicated.

2.5 Methods relating to Chapter 6.2 (rAAV gene delivery)

2.5.1 Production and crude extraction of rAAV9 vectors from 3'ITR plasmid variants

To prepare a mixture of rAAV9 vector products barcoded according to plasmid 3'ITR variants-of-origin, sixty 150mm dishes of adherent HEK293T cells were subject to triple transfection. These comprised 12× pWT transfections, 16× ΔAhdI transfections, 16× ΔXmaI transfections and 32× ΔBssHII transfections. Separate crude preparations were performed according to a scaled-up protocol of that depicted in Fig. X, with the PEG precipitate of the media fraction added to the cell pellet after one round of freeze-thawing and sodium deoxycholate used in place of CaCl₂ (Fig. 2.5.1).

Following qPCR estimation of vector concentrations, crude preparations were combined into a single 50mL polypropylene conical tube and treated as a single preparation, in quantities that would achieve a roughly 1:1 ratio according to the crude qPCR result. This mixture was subject to discontinuous gradient ultracentrifugation and molecular weight cut-off filtration.

15%, 25%, 40%, and 60% preparations of iodixanol with alternating layers coloured using phenol red (Table 2.1.5) were prepared and layered into 32.4mL OptiSeal polypropylene tubes (Beckmann-Coulter). The crude vector mix was added and the tubes were spun at 58,000g for 2h. The vector fraction was extracted from the 25%/40% interface with an 18-gauge needle and 10mL syringe and passed through a 0.22µm syringe filter prior to MWCO filtration (100kDa).

Spin columns had 15mL filter units (MilliporeSigma), which were pre-wet with 15mL Concentration Buffer and emptied via centrifugation before the addition of the vector sample. The vector sample was raised to 15mL with Concentration buffer and spun at 3000g for 5mins at 18°C (or until 250µL remained) and this process was repeated three times. The deadlock retentate volume of 150µL was transferred to a sterile cryotube and the filter was rinsed twice with an additional 500µL prior to storage at 4°C. Vector concentration was measured as before without a DnaseI step, and the final injection mix was raised to 150µl per animal (with 10% overage) prior to handover.

2.5.2 Animal handling

Tasks involving animals were carried out in under AEC Project C389, which was approved by the Children's Medical Research Institute / Children's Hospital Westmead Animal Ethics Committee by circulation and approved for a period of 3 years from 28 April 2021 to 30 April 2024. In accordance with that ethics protocol, animal work was performed with my great thanks by colleagues Cindy Zhu, Sophia Liao, Inna Navarro, and Matthieu Drouyer, rather than myself. The rAAV vector mix was prepared in its final formulation by me and handover of the materials took place for animal administration (150µl tail-vein injections). Tissues were harvested and labelled at the specified time points and snap-frozen in liquid nitrogen. Work was resumed by me upon access to the frozen tissues. RNA and DNA samples were prepared as described in Section 2.1.5-6.

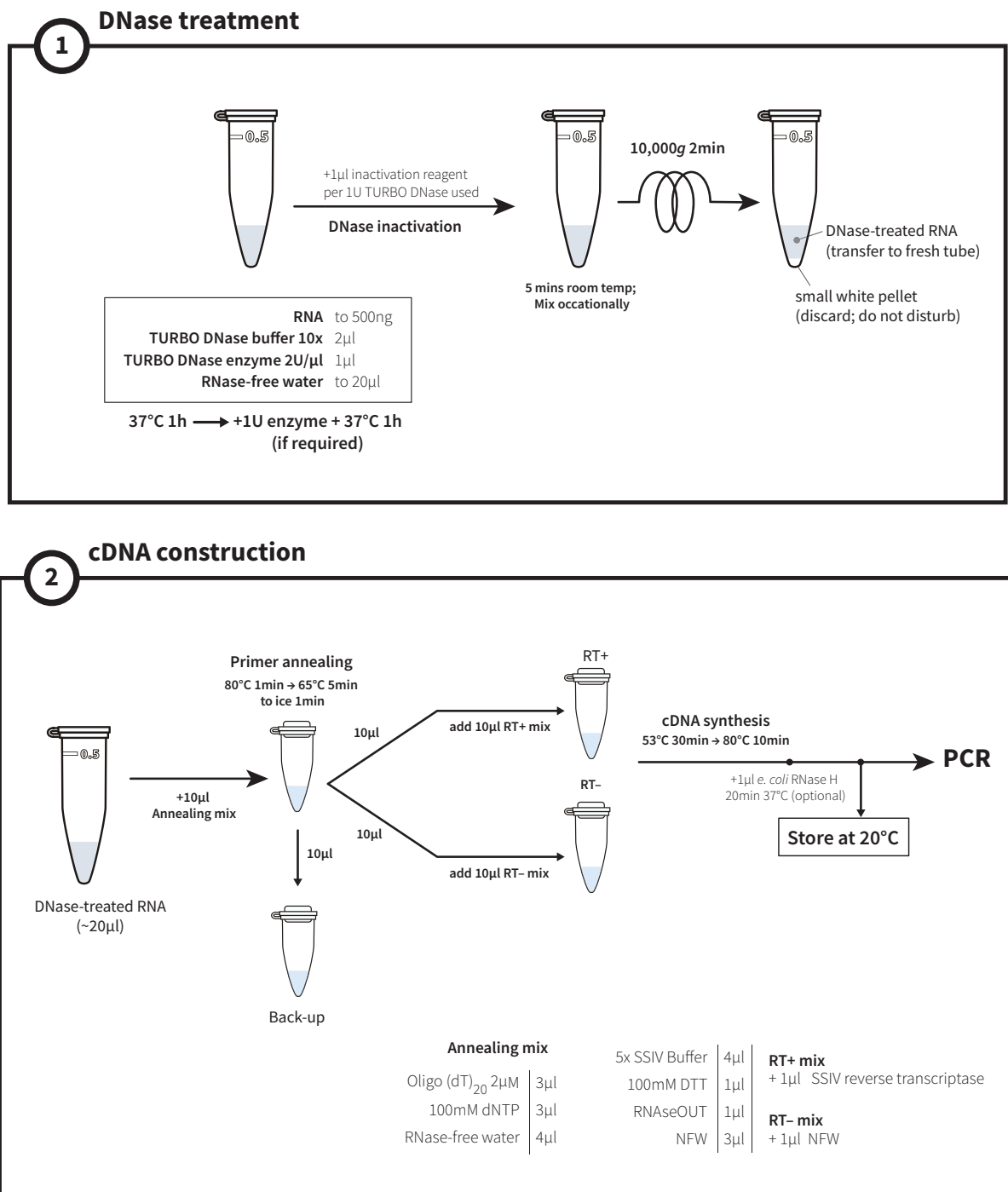


Figure 2.2.1 cDNA Preparation of cDNA templates for qRBA analysis of tissue mRNA extracts]
Reverse-transcriptase PCR amplification was used to generate a reporter amplicon for qRBA analysis. DNase treatment was applied first in order that only mRNA targets were amplified. The DNase-treated sample was then added to a primer annealing mix in which messenger RNA containing polyadenylation signals were primed (to the exclusion of other RNA subtypes) using (dT)₂₀ oligonucleotides. A reverse transcriptase synthesised a strand of DNA complementary to the barcoded rAAV vector transcript which would be targeted for PCR using conventional reagents. An RT- control contained no reverse transcriptase enzyme in the cDNA synthesis step, allowing amplification from DNA targets to be positively identified.

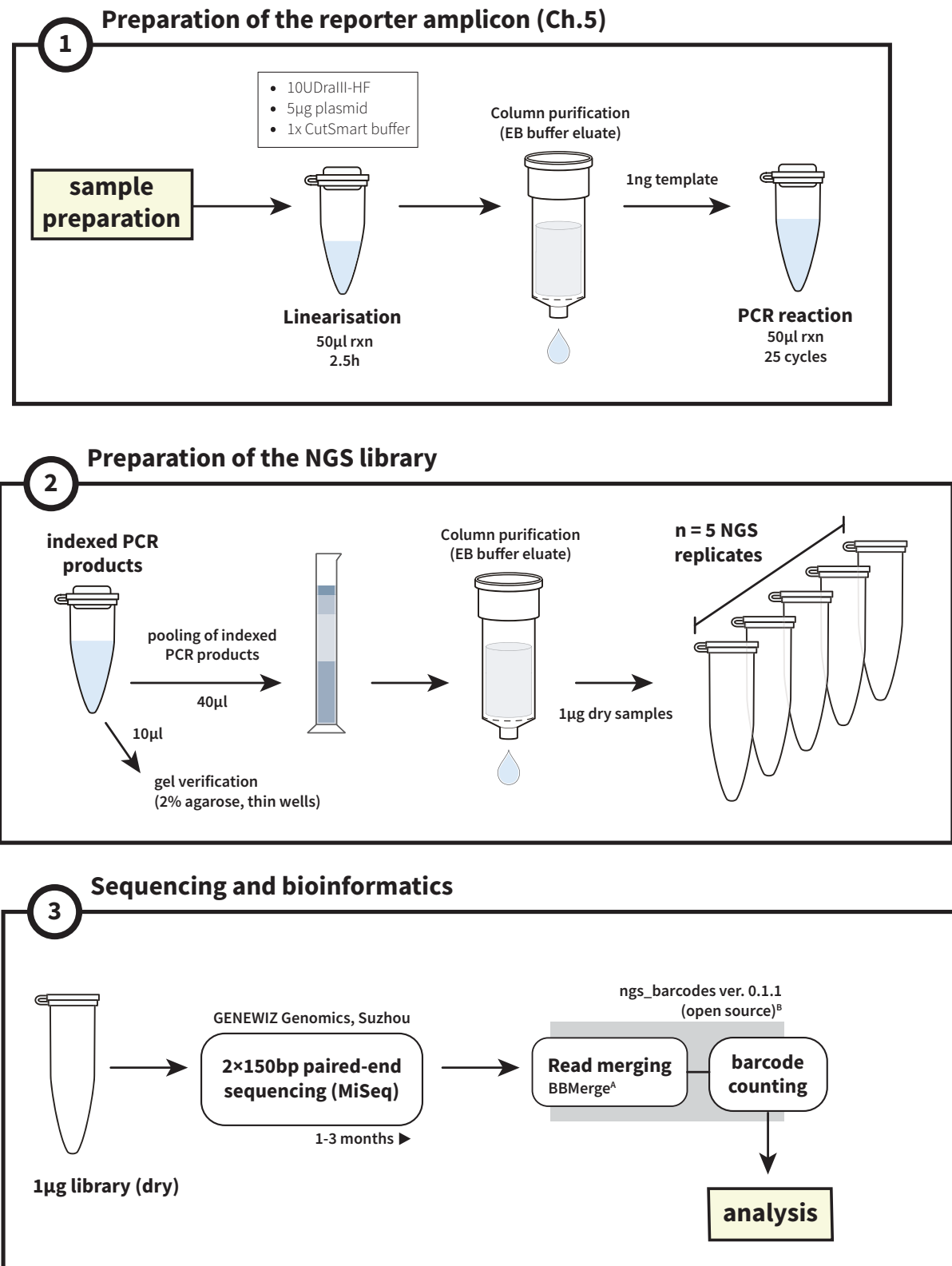


Figure 2.3.2 PCR amplification of qRBA samples and preparation for Illumina population sequencing | The above protocol is specific to Chapter 5. In Chapter 6, pre-PCR treatments differ depending on the sample type and context, PCR products were added in roughly similar volumes, and technical replication of the Illumina sequencing process itself was not replicated. The GitHub repository for ngs_barcode ver. 0.1.1 can be accessed at https://github.com/szscct/ngs_barcode/

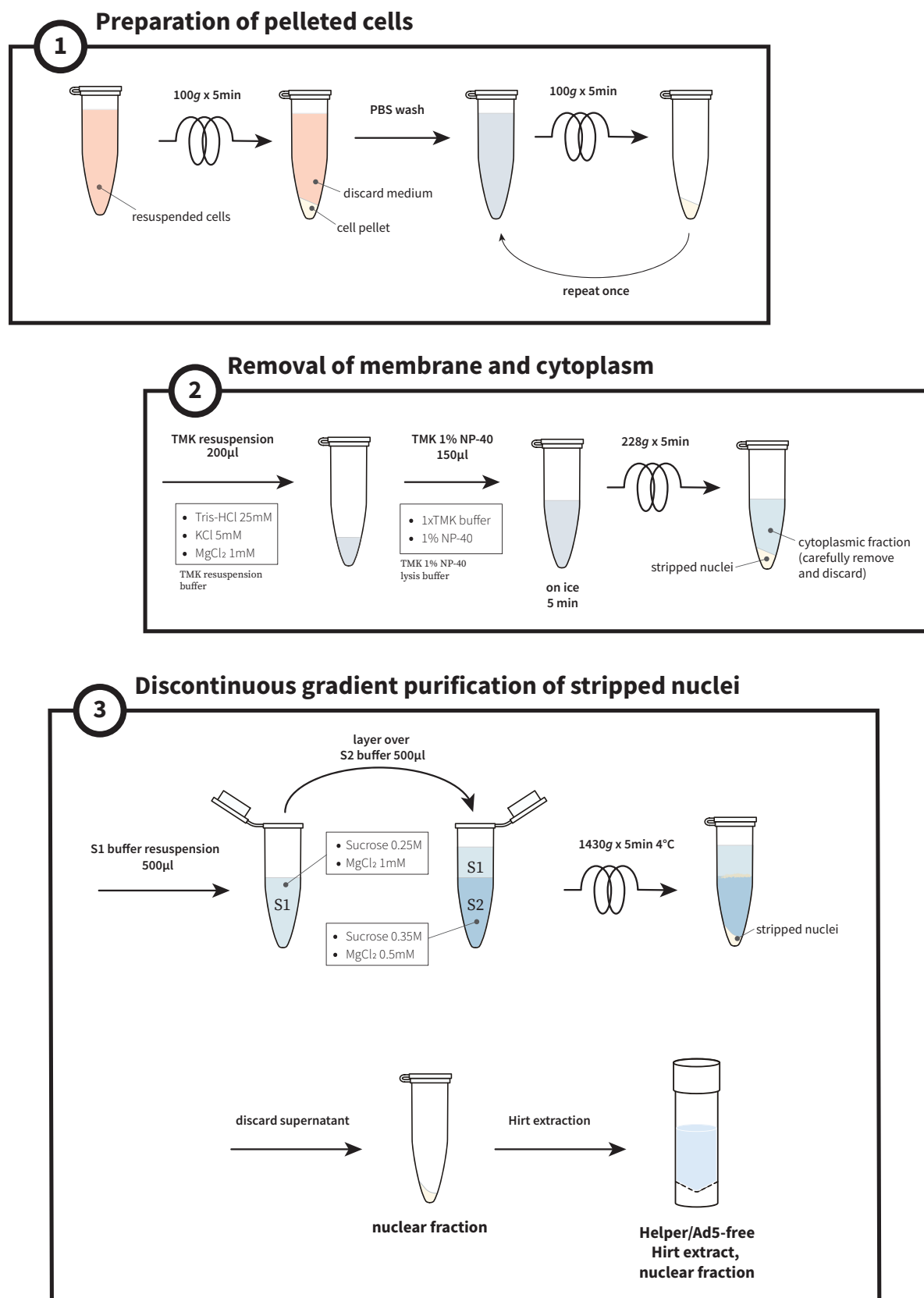


Figure 2.4.5.1 Selective purification of cellular nuclei | Cell nuclei were fractionated following a protocol published by Galvis (2017). Nonidet P-40 (or substitute) is used as a gentle lysis buffer that disrupts cellular membrane more efficiently than nuclear membranes. Intact nuclei are separated from the cytosolic lysate via sedimentation and further enriched using a discontinuous density gradient constructed from two solutions of sucrose, one 0.25M (S1) and one 0.35M (S2). Since ratios were the measurand in this experiment rather than total quantities, a nonconservative enrichment of nuclei could be carried out, with increased residence times and mechanical stress.

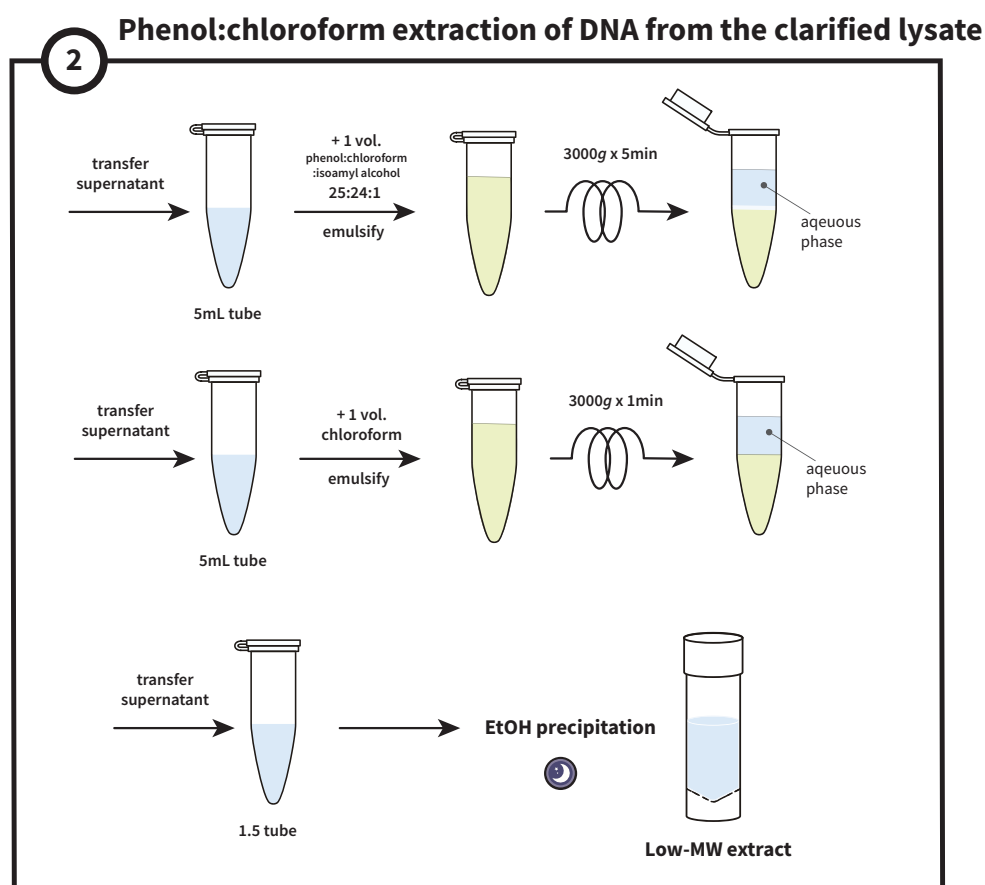
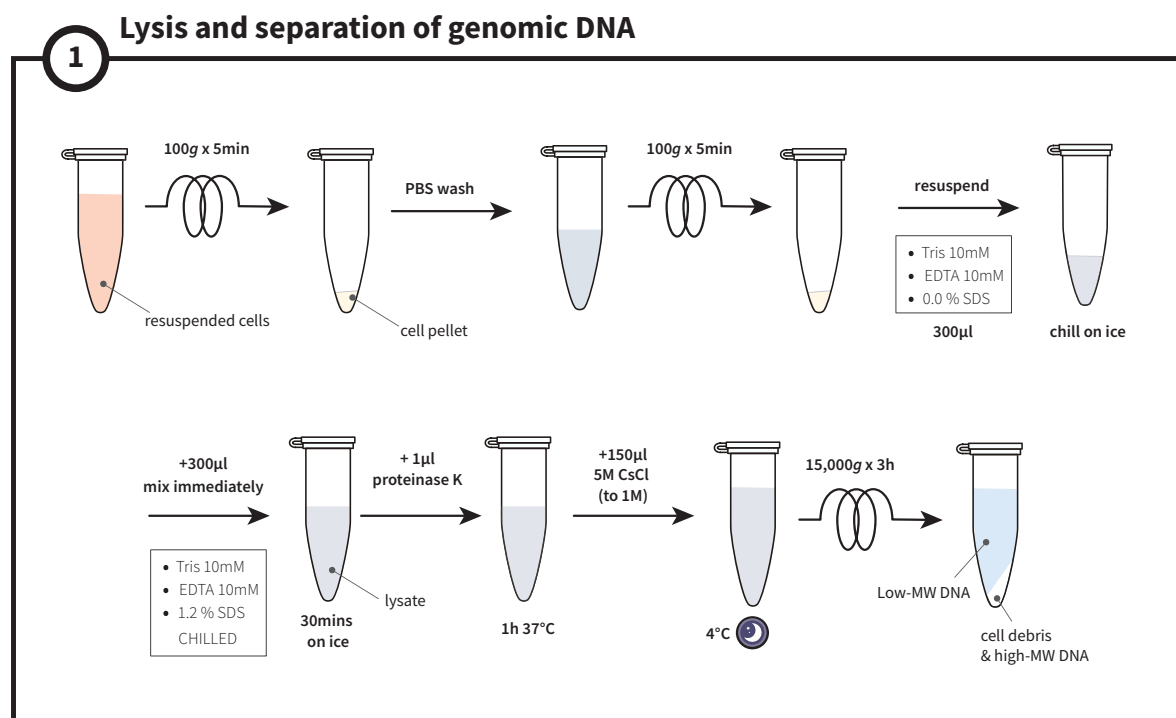


Figure 2.4.5.2 Low-MW DNA extraction and separation with phenol:chloroform | Low molecular weight DNA is selectively extracted by applying minimal mechanical stress to the sample after cellular lysis so that chromosomal DNA remains intact and is thereby co-pelletable with other cell debris following an overnight high-salt incubation. The supernatant is then enriched with smaller and more mobile DNA molecules. A published protocol by Bardelli and colleagues (2017) is cited in the main text. CsCl was used in place of NaCl, following the suggestion of Arad (1998). My own modification was to resuspend cells in a half-volume of Hirt 'lysis' buffer containing no SDS and then adding the other half volume containing chilled, well-precipitated 2× SDS (1.2%).

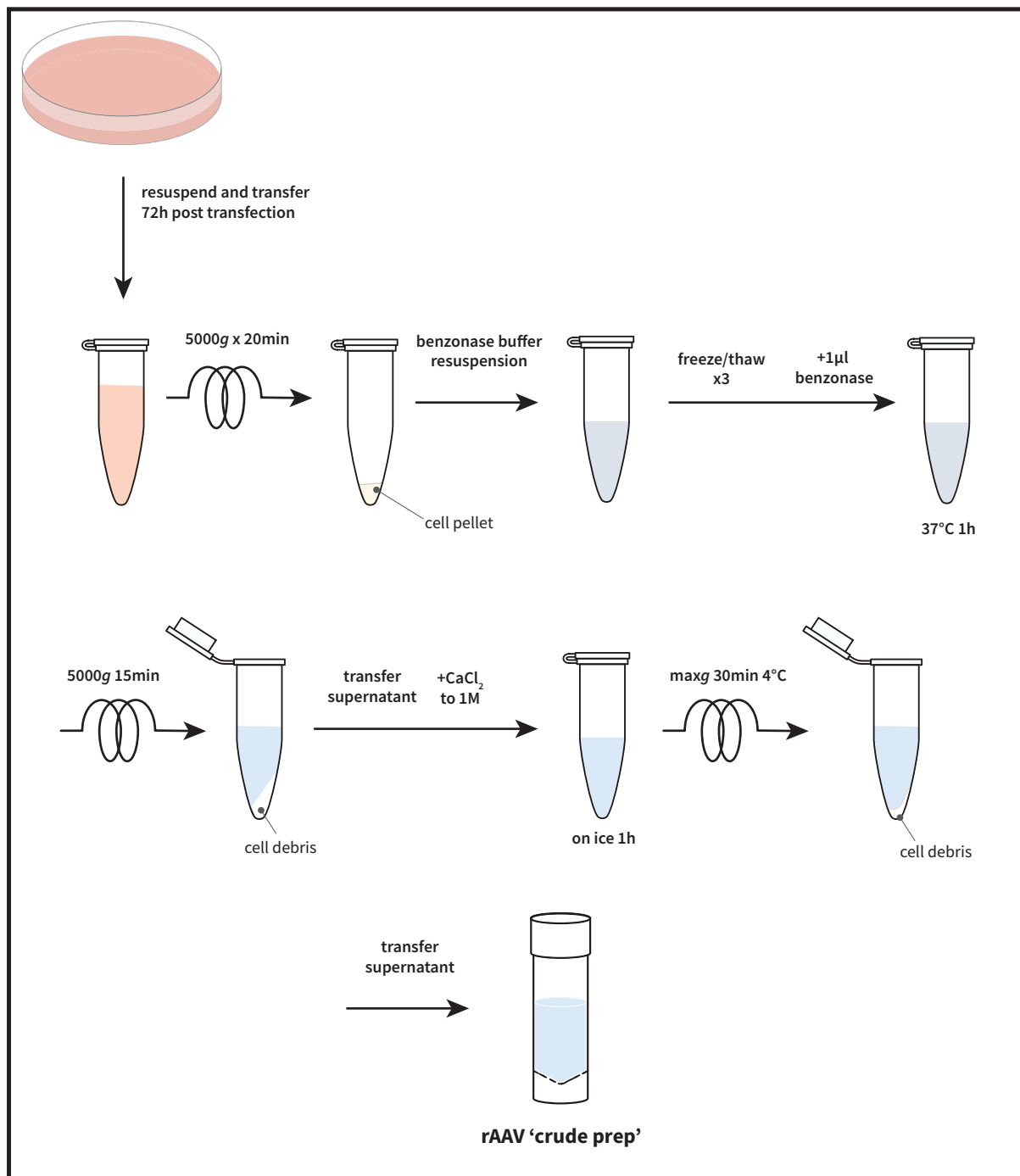


Figure 2.4.5.3 Preparation of crude vector samples from adherent producer cells in 3cm wells | rAAV vectors produced via triple-transfection of producer cells can be prepared for some forms of analysis by a simple isolation procedure (resulting in a 'crude' prep), which is otherwise intended to precede higher-intensity purification steps. This is well suited to generation of preparative PCR samples and avoids variation incurred from downstream purification. The cell pellet is resuspended in a buffer conducive to benzonase activity and cells are lysed via three freeze-thaw cycles. Benzonase is added to the lysate to digest genomic and unpackaged vector DNA. Cell debris is then cleared centrifugally and residual cell debris is removed with further centrifugation after cold incubation in calcium chloride.

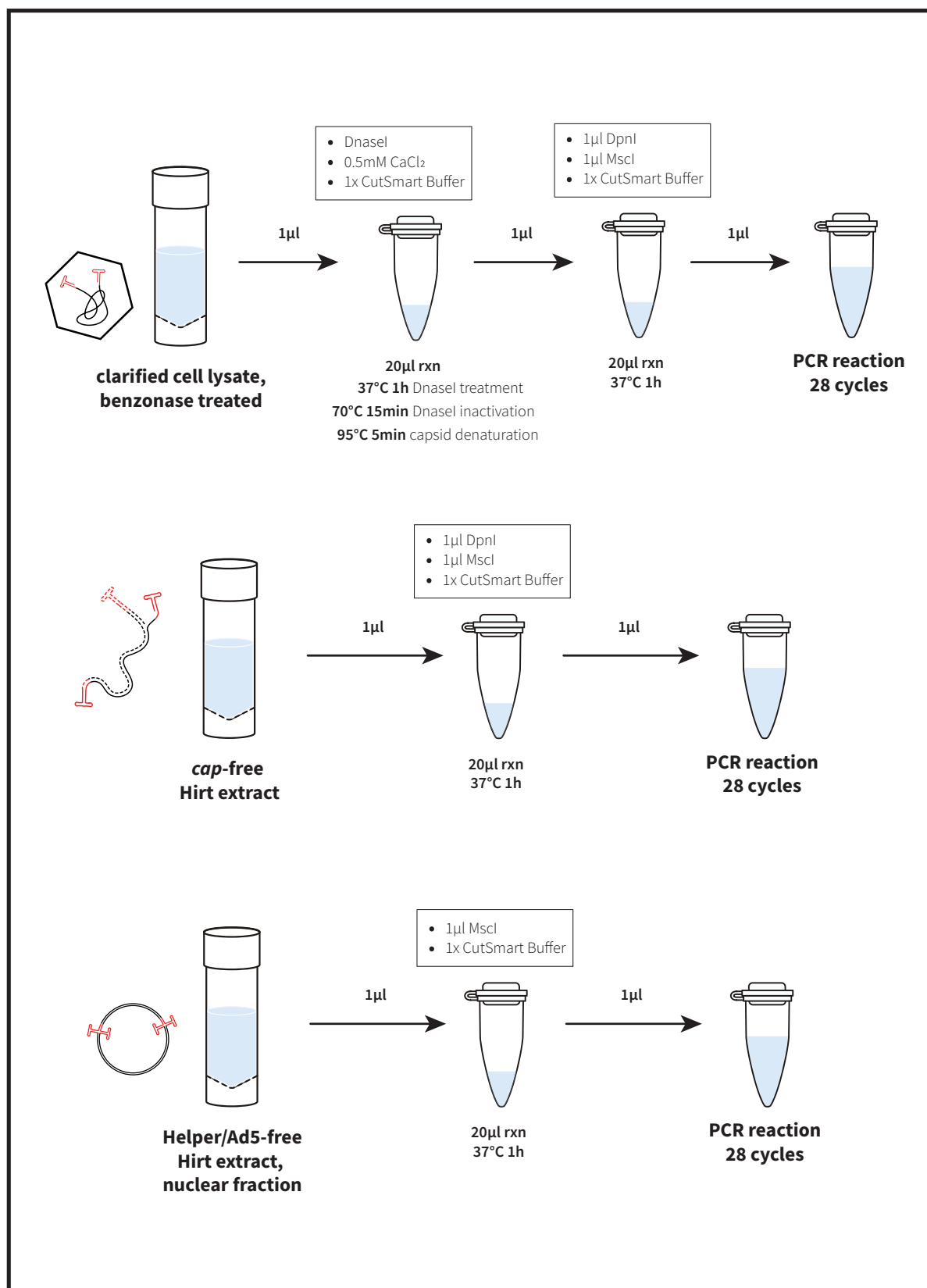


Figure 2.4.6 Pre-PCR preparation of samples for transfection products for qRBA analysis in Chapter 6.1 | Three nucleic acids of interest were targeted for analysis by strategic omission of helper elements, selective purification procedures and appropriate pre-PCR treatments. The methylation-specific restriction endonuclease DpnI was used to degrade vector DNA sequences of bacterial origin so that eukaryotically derived vector genomes could be analysed with a reduced plasmid signal. MscI targets the AA' sequence of both ssAAV and dsAAV, removing terminal variation that may influence PCR amplification efficiency



4.5.1 Crude preparation of vesicles from the media and cellular fraction

Review: Endogenous UMIs as quantitative reporter elements – what, why and how

In this Chapter:

- A conceptual and terminological framework for qRBA is established.
- The contributions of previous researchers to the development of qRBA are highlighted.
- The analytical strengths of qRBA are argued.
- Relative fold output (RFO), a tool for comparing relative measurements, is defined.
- The standard multinomial statistical model is reformulated for application to qRBA measurement and analysis.
- The term ζ is defined and hypothesised as the SD of measured ratios between UMI variants.

3.1 Introduction

In this thesis, *quantitative reporter barcode analysis* (qRBA) refers to the use of unique molecular identifiers (UMIs), endogenous to a nucleic acid target, used to measure relative quantities of that target. Various practical and conceptual elements of this technique can be

seen in microbial ecology, lineage tracking, biodistribution and even transcriptomics, but their power and novelty when coalesced in a fully quantitative setting have yet to be fully recognised.

This chapter might be better understood as a treatise than as a review. Its scope, for the most part, is conceptual. It will pass through a seemingly eclectic collection of subjects, each in service of a cohesive, overarching goal: to advance the view that UMIs can and should be exploited as explicitly quantitative bioanalytical tools. It will dissociate qRBA from the myriad of qualitative and semi-quantitative UMI applications and establish a terminology grounded in the concept of a *reporter*, so that ideas can be fluently exchanged. It will identify key advantages of qRBA over quantitative real-time PCR (qPCR) and digital PCR (dPCR) to make clear why it is worth pursuing in the face of technical challenges, and erode the notion that its strictly relative nature diminishes its usefulness. Readers can think of these areas as the *what* (Section 3.2) and the *why* (Section 3.3) of qRBA.

The *how* portion of this chapter (Sections 3.4 and 3.5) are practical contributions to qRBA with immediate relevance to Chapters 4 and 6 of this thesis. It refers not to protocols or technologies, but rather to a theoretical framework for conducting count analysis and predicting measurement uncertainty. Two components of this framework are described. Firstly, *relative fold output* (RFO) – a metric featuring heavily in Chapter 6 – will be defined and promoted as a tool for making scientifically valuable calculations based on relative information. The second is a prototype statistical model grounded multinomial distributions, from which we can infer the fundamental accuracy and precision of qRBA measurement. Though identical in substance to the standard multinomial model, this prototype “rephrases” the model in a manner best suited to qRBA measurement and the metrological thought process in general. In short, the model treats a multinomial result as a collection of binomial results, privileges the success/failure *ratio* as the dependent variable, and predicts ζ (defined in Section 3.5) to be the standard deviation of that ratio. The term ζ specifies the relationship between UMI ratios, total molecular counts, and variability of qRBA measurements, which will be tested empirically in Chapter 4.

3.1.1 The Illumina mechanism of (relative) molecular counting

Before opening the chapter, it would also be useful to summarise the technological underpinnings of DNA population sequencing. ‘Next generation sequencing’ (NGS) is a nonspecific market term applied to several unrelated sequencing technologies, the hallmark of

which is much higher throughput with lesser individual sequence confidence than Sanger methods. For NGS applications, the source of sequence quality is the ‘massively parallel’ alignment of overlapping results, which together contribute to a high-confidence consensus at each consecutive base position. While salient technologies such as ion semiconductor sequencing, single-molecule real-time sequencing and nanopore sequencing^{9,10} meet this description, this chapter will only elaborate on the clonal bridge amplification and reversible dye terminator sequencing method provided by Illumina, Inc., as it features heavily in this thesis and is currently the dominant means of conducting qRBA experiments. This focus on the Illumina platform is intended to support an understanding of the thesis content, rather than to imply that other technologies are less appropriate. On the contrary – a diversity of technologies with context-specific strengths and weaknesses for qRBA applications should be encouraged. Numerous bioanalyzer instruments have been developed by Illumina, Inc. with names such as MiSeq, NovaSeq and NextSeq. These are intended to service a range of potential output and stringency requirements, and all utilise the same basic sequencing mechanism described below.

Illumina sequencing technology is designed for computing long contiguous sequence information from thousands to millions of short overlapping sequence results termed *reads*. Each individual read is the biochemical sequencing result of a single DNA molecule close to 150nt in length. Reads are obtained through a four-colour sequencing-by-synthesis process that takes place on a planar surface (*flow cell*) underneath a camera. In preparation for the sequencing reaction, the DNA fragments of interest are first bookended with short adapter sequences, which hybridise to immobile antisense adapter oligonucleotides that cover the flow cell in a dense lawn. The immobile adapters serve both to prime and spatially constrain DNA amplification, leading to monoclonal amplicon clusters dispersed across the flow cell surface. All strands of one particular sense are then remobilised by specifically cleaving one of the two adapter sequences, reducing mixed-sense clusters to single sense clusters. Once established and uniform in sense, the clusters will produce the fluorescent signal intensity required for four-colour sequencing-by-synthesis. Key to this process is a removable strand-extension-blocking moiety conjugated to each free dNTP, which enables reversible chain termination. Base sequences are determined via successive cycles of base incorporation, fluorophore excitation/detection and removal of the blocking group^{10,11}.

The fidelity of Illumina sequencing is, of course, highly consequential to the efficacy of Illumina-based qRBA. Unfortunately it is not so easy to suggest a ‘typical’ level of raw sequence quality that will hold in a randomly selected scenario. It has, in fact, been recently reported that error rates for the same instrument model and sample type (the *Escherichia coli* genome) can be convincingly grouped according to the submitting laboratory¹². As a generalisation, the authors reported a median error (wrong base) rate of 0.47% with a standard deviation of 0.94% for Illumina MiSeq data generated from *E. coli* chromosomal DNA. Illumina technology utilises a Phred-like algorithm equivalent to Sanger sequencing and expresses read quality using Q scores¹³. Each read is attributed an overall Q score equal to the average for all bases, typically decreasing toward both termini. A base determination with a score of Q30 is implied to be incorrect in 10^{-3} (0.1%) of cases. A score of Q20 implies the same for 10^{-2} (1%) of cases. The effective read length and/or quality score for a DNA fragment is enhanced by a ‘paired-end’ sequencing mode wherein each cluster is sequenced twice – once forward then once in reverse. Provided there is some overlap in the forward and reverse read pairs, this allows users to exceed the technical limit of ~150bp, thereby strengthening the overall contiguity of aligned reads^{10,11}. For qRBA experiments presented in this thesis, paired-end sequencing of 150nt DNA strands acts purely to increase the quality of raw read data. In my own laboratory, Illumina MiSeq paired-end sequencing of a 150bp PCR product routinely results in >2 million reads per sample, approximately 95% of which have an average base score of Q30 or higher. In a later Chapter (Section 4.4) I will demonstrate unambiguously that Illumina-based qRBA assays can be performed with negligible impacts from amplification and/or miscall errors, and discuss why this finding should not surprise us despite the lower individual sequence quality characteristic of population sequencing technologies.

3.2 What are reporter barcodes, and what is qRBA?

One of this thesis’s main arguments is that qRBA can readily contend with q/dPCR when measuring relative quantities of a c/DNA target sequence. Given the boldness of this claim, it is important to establish a clear and common understanding of *what* the neologism ‘qRBA’ refers to. This is achievable, though not trivial. The technique is applicable to a wide range of molecular biological subject areas, across which terms like ‘barcode’ and ‘quantitative’^{14–16} will have greatly variable connotations. A short discussion of these terms with illustrative examples, aside from highlighting context and the contributions of previous researchers, will align the

attributes of reporter barcodes and qRBA and their implications coherently, so that they that can be understood and discussed by all relevant audiences.

A UMI is a base sequence, present on a nucleic acid molecule, from which some aspect of that molecule's identity can be inferred. UMIs themselves are so impressively versatile that 'UMI' hardly denotes a specific tool. The popular synonym 'barcode' is no more specific. On the contrary, it is used frequently both in common parlance and as a term-of-art in a multitude of incompatible senses. UMIs used for qRBA measurement are termed *reporter barcodes*. By analogy, a defining characteristic of a reporter *gene* is measurability, in addition to identifiability¹⁷. This makes *reporter* more specific than *identifier*. **The purpose of a reporter barcode, like a reporter gene, is to signify the presence and quantity of a nucleic acid. In lay terms, to say “Here I am; this is how much of me there is”.** The variation among barcode sequences is confined to a specific length and location, and is endogenous to the qRBA target. For quantitative applications, an additional criterion is that the mix of possible reporter barcodes be low in complexity (Fig. 3.2.1). These three attributes distinguish qRBA from a number of similar applications and are intended to support a quantitative mode of analysis, in which relative read counts are viewed as direct measurements, not statistical correlates, of subpopulation abundance.

1. **qRBA** is the use of molecular counting to measure the relative abundance of reporter variants
2. A **reporter variant** is a nucleic acid target (or its host entity) containing a specific reporter barcode sequence.
3. A **reporter barcode** is a UMI whose length and location are consistent to all reporter variants, and is endogenous to the reporter variants.
4. In qRBA, the number of unique reporter variants is low enough to support the desired measurement precision.
5. The **qRBA target** is almost exclusively (though not necessarily) a PCR target, whose PCR products are the direct analyte of qRBA measurement.
6. Molecular counting is almost exclusively (though not necessarily) achieved with DNA population sequencing, prominently Illumina sequencing.

Similar applications include the use of naturally occurring taxonomic barcodes, where the number of reads containing a specific barcode signifies the relative abundance of species (or higher taxonomic clades) in an ecological community¹⁸. This is distinct from qRBA in that variations will occur in uncontrolled location, introducing additional sources of miscall and/or coverage bias that are difficult to predict and distinguish from population variations. Endogenous barcodes are also a powerful ‘lineage tracking’ mechanism, providing viewing windows into growth and selection kinetics, strongly enhancing transposon mutagenesis studies¹⁹ (with illustrative examples in yeast²⁰ and in bacteria²¹), experimental evolution^{22,23} and

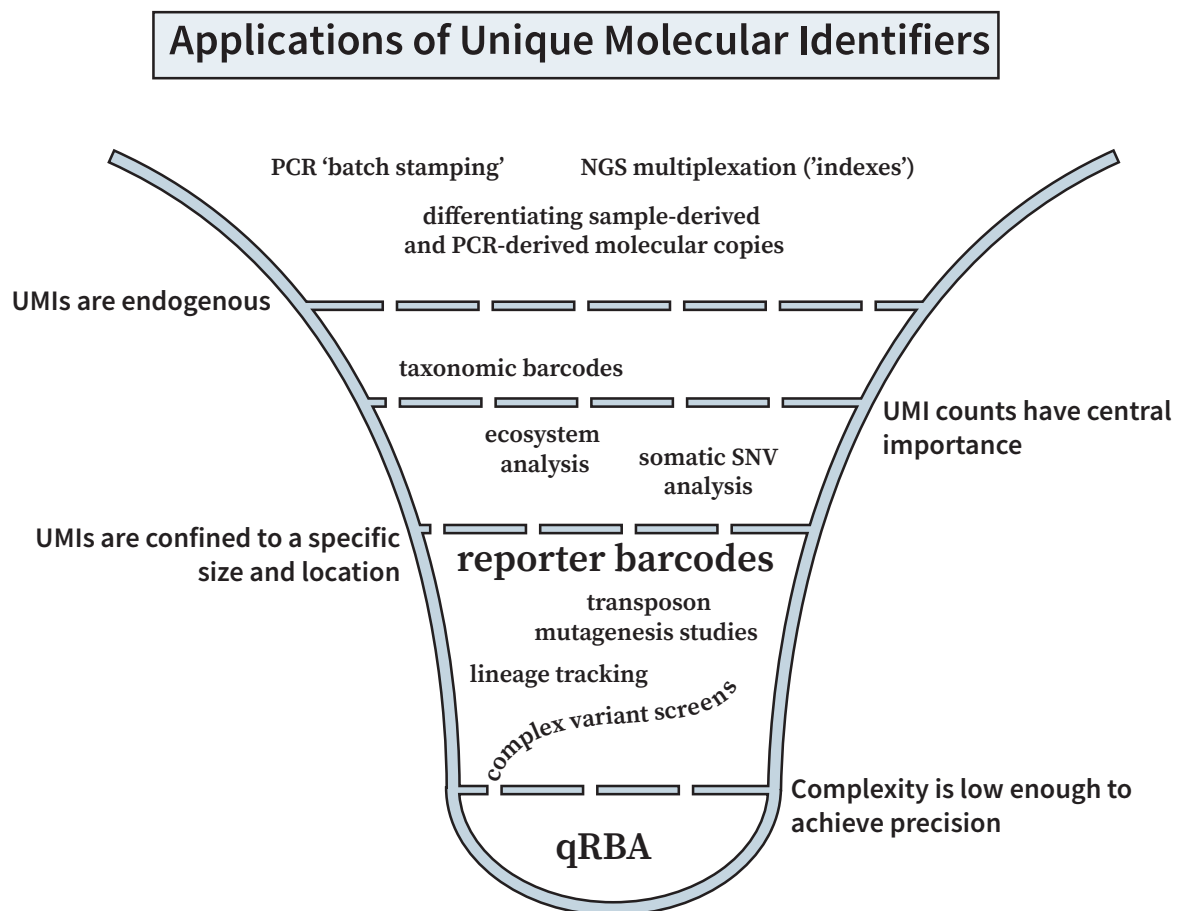


Figure 3.2.1 Specific elements of reporter barcodes and qRBA | Unique molecular identifiers are exploited across a range of biological disciplines, with an array of complexity, sources, designs, and means of incorporation. As a result, terms such as ‘barcode’ and ‘quantitative’ have developed connotations that may vary depending on the audience. This figure is a non-exhaustive map of UMI applications, with examples to be found referenced in the main text. A filtration metaphor is used to delineate what this thesis terms ‘reporter barcodes’ and ‘qRBA’. By avoiding the need to incorporate UMIs from an exogenous source, and by limiting the variability of a UMI (and its host sequence) to a specific length and location, complex sources of coverage and miscall bias can be avoided, making reporter barcodes better suited to quantitative applications than other types of UMIs. Increased complexity (the number of unique variants) detracts from the quantitateness of NGS reporter compositions by reducing the level of representation possible for each variant in a given readset. The difference then between qRBA and other uses of reporter barcodes, while not immune to semantics, is at least conceptually distinct in that the NGS reporter composition is viewed as a direct reflection of the sample rather than a statistical correlate. The specific determinants of quantitative precision will be discussed in Section 3.5.

development of designer rAAV capsid variants through directed evolution^{3,24}. Reporter barcodes used in this fashion typically meet the confinement criterion but are random (at least initially) on a copy-to-copy basis. Increasing the number of reporter variants intrinsically decreases the individual readcount of all variants²⁵, reducing quantitative precision and rendering them a lesser fit for qRBA.

Certain elements of transcriptomics are important to distinguish from qRBA, particularly where UMIs are used to measure the relative abundance of RNA transcripts. The RNA content of a sample, like a mixture of reporter variants, can be probed using population sequencing and count analysis. UMIs are exploited prominently in this area when reverse transcription PCR is required to generate sample quantities sufficient for analysis. In this setting, random UMIs are incorporated into the PCR template molecules and used to distinguish sample-derived from PCR-derived molecular duplicates^{26–28}. Elsewhere this has been proposed in analysis of copy-number variation (CNV)²⁹ and even as a method for estimating the absolute molar quantity of nucleic acids³⁰. At face value alone, UMIs used like this do not meet the description of a reporter barcode because (1) random from copy-to-copy and (2) exogenous to the target. They are also fundamentally different in that the number of sample-derived RNA copies is estimated from the *number of unique UMIs* associated with mappable NGS reads. Unlike reporter barcodes, the *number of times a specific UMI appears* is more or less irrelevant to the analysis. In principle, the RNA fragment as a whole might be viewed as a ‘UMI’ specific to the gene-of-origin, though this is the extreme opposite of the ‘confinement’ criterion.

For completeness, reporter barcodes can be readily distinguished from UMIs that function as metadata to an NGS read result. For example, an NGS *index* borne by molecules from a given sample enables multiple samples to be run on the same flow cell and re-separated bioinformatically³¹. (In this thesis, a ‘barcode’ meant in this sense will be referred to exclusively as an *index*). A similar concept has been proposed as a laboratory housekeeping technique, with UMIs acting as ‘batch stamps’ on PCR products to guarantee sample traceability³² and monitor cross-contamination.

One of the earliest examples of a reporter barcode system may be a repository of >953 unique *Saccharomyces cerevisiae* gene knockout strains, in which every variant was painstakingly engineered and cloned with a UMI-containing sequence in place of the knocked-out gene³³. The reporter barcodes for this repository were initially intended for measurement not by

population sequencing, but rather via UMI-specific microarray hybridisation. In 2009, Smith and colleagues³⁴ demonstrated that the exceedingly large quantity of individual sequence results offered by Illumina technology could reasonably reflect the relative abundance of each barcoded knockout strain, enabling simpler and more efficacious comparisons of strain composition in mixed cultures after exposure chemical challenges.

The use of reporter barcodes first appeared in the rAAV field in 2014, in a vector biodistribution experiment published by Adachi and colleagues³⁵, though the authors were clear not to assert that the technique should be used quantitatively. The authors developed large panels of unique rAAV capsid peptide variants (up to 92), each represented by multiple possible reporter barcodes. Along with benchmark variants, the pooled panel were administered to experimental mice. After a period, DNA was isolated from various tissues and bodily compartments and the reporter compositions of the barcoded PCR products were compared. The authors named this method “Barcode-Seq”. This strategy has since been adopted in the study of lipid nanoparticle-based nucleic acid transfer vectors^{36–41}, which have risen to public prominence in the global vaccination effort against SARS-CoV-2⁴². More so than quantitiveness, the authors highlighted the significant throughput capabilities and the alleviation of animal/tissue resources. They formally demonstrated that the differential reporter composition across various recipient tissues correlated reliably with testable gene delivery phenotypes.

Other early contributions arose out of Florida, the first of which was published in November of the same year⁴³. Having enriched a random capsid pool with favourable vector candidates through iterative selection, Marsic and colleagues prepared two final candidates and two benchmark variants (AAV2 and AAV8) with uniquely barcoded transgene constructs in order to compare their DNA transfer capabilities in a single animal via Ion-Torrent (not Illumina) NGS analysis of tissue DNA extracts. The technique was used in conjunction with qPCR to obtain a ‘total’ transgene copy number per diploid genome and the contribution of those four variants to that total. Notably, though earlier authors had emphasised advantages in throughput capacity, this paper applied the technique to the study of only four capsid variants. The lead and senior authors (along with colleague Dr H. Méndez-Gómez) published an additional report the following year showcasing the UMI-NGS capsid screening method as an explicitly quantitative tool for measuring gene delivery⁴⁴, in an obvious (though seemingly undeliberate) contrast with the views of Adachi and colleagues. Though the experimental process was demonstrated, the titular claim that the method was ‘highly accurate’ rested on

common sense reasoning rather than empirical evidence (only three mice were used, and no technical replicates were performed). Reporter composition was again coupled with qPCR measurements of tissue isolates to calculate viral genomes per ng of gDNA, however, the accompanying uncertainty was not reported in the main text. In the supplementary material, the standard deviation of vector genomes per ng of gDNA was on the order of 100%. Importantly though, these authors were able to demonstrate (albeit without a statistical test) that amplicon reporter composition was unaltered over 20 to 30 rounds of PCR.

In the literature, there is no clear original conceiver of what I retrospectively characterise as qRBA. Smith *et al.* and Adachi *et al.* both understood Illumina readsets to represent the relative abundance of reporter barcodes, though neither relied on this in a quantitative sense, and the latter expressly declined to assert that their technique was quantitative. Reports led by Dr Marsic from the same period recognised the data as quantitative and are therefore directly aligned with the concept of qRBA, particularly in their initial analysis of only four variants. Regrettably, those publications went no further than previous researchers in advancing that notion scientifically.

What is clearer, and has been reviewed^{45–47}, is the enthusiastic adoption of reporter barcodes and NGS sequencing in rAAV vector research (Fig. 3.2.2). Engineering novel rAAV variants through directed evolution was long underway⁴⁸ by the time UMIs arrived on the scene in the mid-2010s. UMIs provided researchers with lineage tracking capabilities, as well as a much-needed means of comparing the influx of novel capsid candidates generated from their iterative selection experiments^{45–47}. As a result, many rAAV researchers today are very well-practised in cloning, managing, implementing, and analysing reporter barcodes in their everyday research. Because of this, the rAAV research environment has become increasingly permissive to UMI-based analytical strategies and allowed researchers to veer into territory where throughput would appear less of a limitation, conducting studies for which q/dPCR would conventionally have been the tool of choice^{49,50}. Many of these constitute non-random capsid mutagenesis studies^{51–53}, though quantitative comparisons of promoter activity have also been conducted by comparing reporter composition of exogenous transgenes with that of their mRNA products⁵⁴. Unpublished experiments from my own laboratory include analyses of column affinity, subcellular localisation, vector production efficiency, plasmid propagation, mechanisms of DNA freeze-thaw damage and plasmid topology.

3.3 The substantial analytical benefits of qRBA measurement

Though the throughput benefits of reporter barcodes have long been emphasised^{34,35,55,56}, the ground-breaking quantitative benefits have yet to be fully argued. In Chapter 6.1 of this thesis, a qRBA-based assay will be applied to the study rAAV vector production in mammalian tissue culture. This assay achieves quantitative consistency between complex, multi-day experiments comparable to historical q/dPCR measurements of single, purified rAAV samples. **In this section I will explain how qRBA can routinely achieve high-quality, highly reproducible**

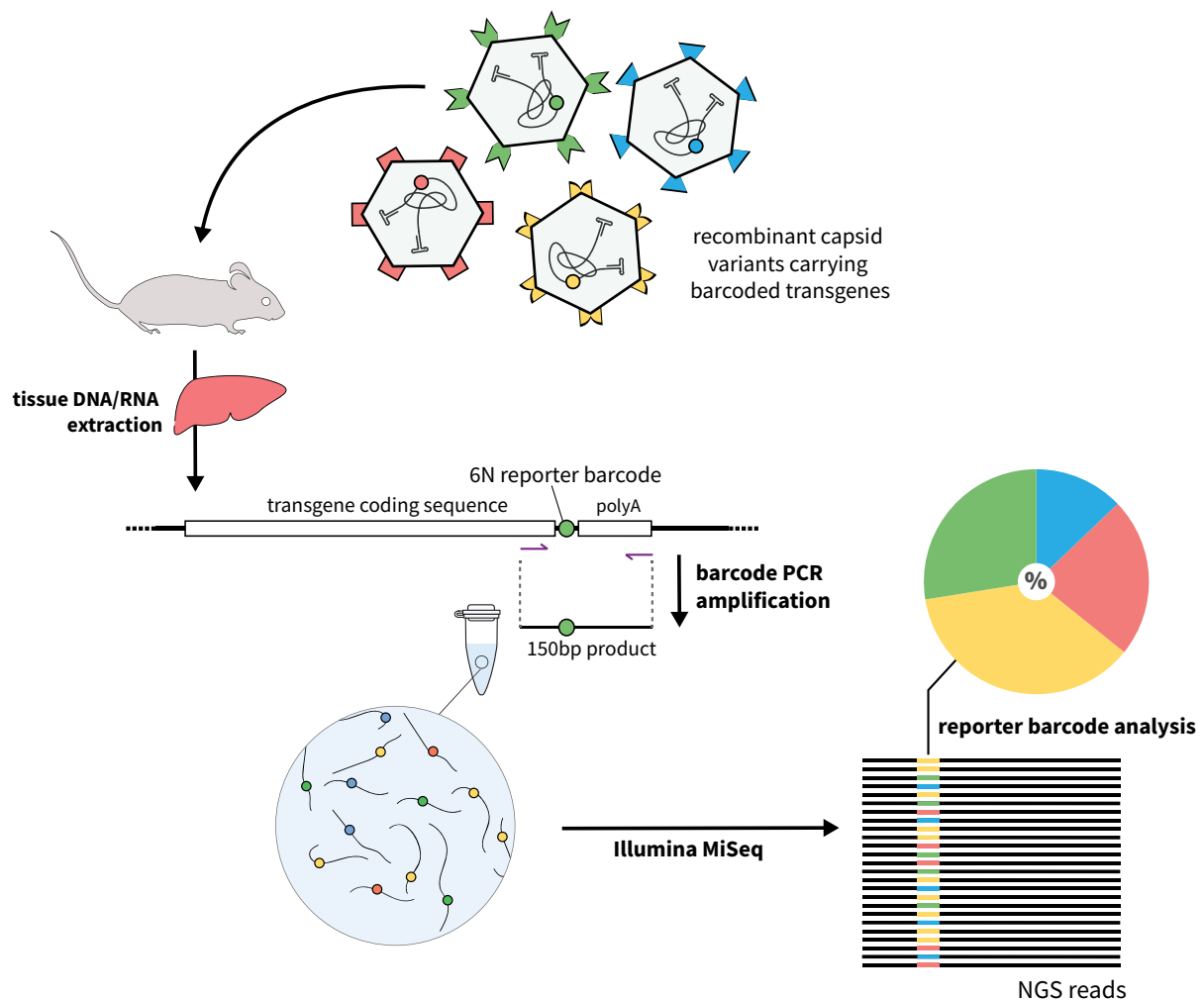


Figure 3.2.2 Application of endogenous UMIs and massively parallel sequencing to the study of rAAV vectors | 'Next generation' sequencing technology has been utilised in conjunction with unique molecular identifiers to produce a subpopulation profile of otherwise-identical DNA copies, which can be amplified from a PCR target. Population sequencing chemistries produce a large number of sequence results derived from single molecules. Read data can thereby be viewed as a multinomial representation of subpopulations with variable UMI sequences. UMIs can be used to signify groups of interest. In the example depicted above, UMIs (termed reporter barcodes in this thesis) allow the viral vector capsid variants responsible for delivering barcoded genes to target tissues to be quantified. This approach to biodistribution analysis has fast become a keystone technique in the study of rAAV vectors. Extending this technique to RNA extracts of the same tissues has enabled the transcriptional activity of exogenous transgenes to be measured.

measurements with an exceedingly less burdensome quality infrastructure than is required for q/dPCR. My argument will not be that qRBA is superior because it is 'less effort'. This would certainly disregard its power as a quantitative tool. More seriously though, it would understate the scale and complexity of systems required to maintain traceability from q/dPCR measurements to the International System (SI) of units, downstream from which lie the unavoidable inconsistencies between humans, equipment, laboratories, and times. Rather, my argument is that a lesser reliance on this infrastructure is a source of considerable robustness. This compounds two further advantages: (1) qRBA measurements are an intrinsically superior reflection of target quantities *in situ*, and (2) the precision of qRBA measurements is inherently comparable to qPCR or dPCR at equivalent degrees of stringency.

The advantages of qRBA over PCR-based measurements are numerous and synergistic. On the advice of friends and colleagues, I will open this argument with an analogy. Suppose ten people stood in a queue, behind which was an analytical balance. A pouch of sand is emptied into the carefully cupped hands of the frontmost member. To find out how much sand was in the pouch, the members of the queue must pass the pile from person to person until the last member passes it onto the balance. Now, suppose another ten people are similarly organised, only every single grain of sand in their pouch has one of ten distinct colours. At the back of this queue, in place of a balance, is a high-tech instrument that can sample over a million individual grains and report the colour of each one. In that first queue, try as they may, their last member will be passed a lower and less consistent amount of sand than their first. While the same is true for the second queue, the relative abundance of colours will be practically unassailable from beginning to end.

The advantage illustrated above is the stability of reporter variant ratios (once mixed) through upstream collection and preparation processes, compared to their absolute concentrations. The amplification kinetics of qPCR, and the droplet integrity of ddPCR, are both sensitive to the solute profile and colligative properties of the reaction mix. This necessitates numerous physical and biochemical processing steps to isolate and refine nucleic acids from a less permissive source, often comprising a long succession of manual and/or automated manipulations. Every manipulation unavoidably contributes to variation in template concentration of the eventual reaction mixture, widening the disconnect between an experimental sample and its readout and lowering our sensitivity to differences between

treatment groups. In a sense, this principle has already been recognised through the ubiquitous use of endogenous references in q/dPCR.

Some may correctly raise that because reporter variants are not categorically identical outside the reporter barcode sequence, there is room for non-negligible bias in their passage through sample preparation processes. For example, rAAV capsid variants may have variable thermal/chemical stability and exhibit nonequal denaturation kinetics leading to a bias in genome release prior to PCR. Factors like these certainly do bear relevance to experimental design, as will be seen in later chapters. This being true, it is no less true for q/dPCR applications. In fact, as will be seen in later chapters, the resilience of reporter ratios means heightened process intensities can be tolerated in order to reduce sources of potential bias.

Another distinct advantage is enjoyed during PCR amplification itself (albeit one more specific to qPCR). As with dPCR, the delicacies of amplification efficiency have a dramatically reduced influence on qRBA measurement. If PCR is required to produce a qRBA sample, the step is strictly preparative. More importantly, the many nuisance variables responsible for nonequal amplification efficiency are categorically irrelevant when all targets are amplified in the same reaction mixture. Even difficult culprits are avoided, such as nonequal heat distribution in a thermocycler plate⁵⁷, algorithms used to calculate C_t values in response to a photomultiplier signal, and even the thermoconductive properties of PCR reaction tubes⁵⁸. The same is greatly applicable to the sources of inconsistency arising from the sample and/or targets, such as GC content⁵⁹, secondary structures⁶⁰, primer affinity, nonspecific amplification⁶¹, and DNA/RNA storage and degradation^{62,63}. Many incorrectly intuit that reporter variants present in nonequal ratios will diverge as a result of exponential expansion. Though arithmetic differences in subpopulation size will indeed increase exponentially, proportions are unchanged if amplified with the same efficiency (see Appendix F). Endogenous references (aka ‘housekeeping genes’) would have little efficacy were this not true. If not clear mathematically, this was evident when tested directly by Marsic *et al.*⁴⁴ over 20-30 cycles and will continue to be evident in the consistency of measurements seen in later chapters despite variable degrees of amplification.

For dPCR, the efficiency of the PCR reaction is inconsequential so long as a distinguishable degree of fluorescence between empty and full partitions can be achieved⁶⁴. Therefore, the advantages of qRBA in this regard are less impactful but nonetheless worth raising. These are that a validated assay with a strong track record is less needed to predict the success of a novel

PCR reaction or of a sample with unspecified quality. In contrast to droplet dPCR (ddPCR) in particular, it is also free from dependence on volumetric consistency between droplets, which is assumed to be perfect both as a Poisson parameter and for calculating concentration⁶⁵.

There remains one key advantage yet to highlight, whose significance is less readily appreciable to biologists but no less important. For qRBA, the measurand is a unitless quotient. Both of these attributes are highly relevant to final measurement uncertainties (which we might simplify to mean ‘the correct size of error bars’). It means that once mixed, there is categorically zero importation of volumetric, fluorometric, kinetic, thermometric, barometric, or even gravimetric uncertainty^{58,66,67} to the ratio between reporter variants. This alleviates the infrastructure and human labour required to minimise and quantify uncertainty from these sources. Upstream from research laboratories, these error sources are characterised and curated to a highly stringent degree by dedicated metrological institutions under legally cognisable Quality Management Systems^{68–71}. Outside these strictures, such factors are seldom contemplated when reporting error in published research⁷². Certain volumetric factors, without which error bars will always be underestimated, are typically accepted (reasonably) as too onerous to incorporate despite their potential influence on raw data and its analysis. Examples include the effect of temperature and altitude on fluid density and the uneven distribution of precision/accuracy over the volumetric range of a calibrated laboratory pipette. For dPCR, volumetry is consequential in two regards: (1) volume is the denominator of concentration and (2) readout is strongly determined by the volume and consistency of partitions (droplets in the case of ddPCR^{73,74}). In its independence from prior measurements, qRBA bypasses many unavoidable error sources. The resulting uncertainty is lowered, less dependent on individual technique and intrinsically less complex to quantify.

Yet further, the fact that the measurand is itself a quotient gives qRBA a ‘head start’ toward observations of interest (often quotients in one form or another) without raising the uncertainty. This includes any instance where q/dPCR results need to be interpreted relative to another sample (i.e., a fold difference) or ‘per’ some other calculated quantity (e.g., per an endogenous reference or a unit of biomass). To be clear, this alone does not grant qRBA superior precision. Rather it highlights that uncertainties of individual q/dPCR measurements and individual qRBA measurements are not valid to compare. The relevant comparison is the combined uncertainty of at least two q/dPCR measurements and must include uncertainties introduced by appurtenant measurements.

On the advice of colleagues, I will briefly digress to make explicit why biologists might underappreciate these advantages or underestimate everyday uncertainty in d/qPCR measurement. Identifying and totalling measurement uncertainty often involves situation-specific methodological variations, a reasonable extent of which lie understandably outside the purview of the final analyst^{58,66,70}. However, the following fact is both universally true and heuristically accessible: **products and quotients of measured values propagate their uncertainties geometrically**^{75–79}. For example, when measuring the height and width of a rectangle to calculate its area. Since error analysis is often minimally detailed in biological reports^{72,80,81}, particularly in relation to qPCR⁸², it is difficult to estimate (though all too frequently obvious) how often a published set of q/dPCR-derived error bars fail to incorporate all fair instances that uncertainty was increased, despite their implications for downstream statistical tests⁸². Misapplication of descriptive statistics is also an area of concern^{82–84}. Ko and colleagues⁸⁵ determined that for the year 2011, 22.26% of research articles published in the *American Journal of Obstetrics & Gynecology* inappropriately used Standard Error of the Mean (S.E.M.) to express uncertainty in their data. Mathematically, S.E.M. is always (often unsubtly) lower than sample SD and is therefore guaranteed to underreport uncertainty when misapplied. More to my point, a lack of acuity among reviewers and editors is required for this issue to persist. In the same report, it was evident that the general audience tended to overlook the issue, as articles misapplying S.E.M. were cited more frequently than those reporting SD correctly ($p = 0.025$). This alarmingly suggests that a reasonably complete representation of experimental uncertainty, in addition to incurring effort, can incur academic penalty.

In summary, qRBA measurements are more direct reflections of target ratios *in situ*, are minimally influenced by PCR reaction conditions, and present direct relative information whose sources of random error are intrinsically simpler to manage and report. Where implementable, these factors make qRBA an extremely valuable asset to molecular biologists as either a primary or orthogonal investigational tool. While technical hurdles are likely to arise, these advantages justify concerted efforts to overcome them.

Before proceeding to the ‘how’ portion of this Chapter I will close this section by addressing two concerns that tend to arise in personal communication. Firstly, that the relationship between measurand and readout is probabilistic, not deterministic. This is however no less true for dPCR, where a stochastic partitioning process results in an empty-to-full partition ratio, relating to target concentration through a Poisson probability distribution. In fact, the same

can be said for the ratio of two m/z variants measured using mass spectrometry (sample complexity and resolution notwithstanding). In this scenario, the counts observed are subject to fundamental random variation, however small, being samples drawn from a larger population. Secondly, that the strictly relative nature of qRBA deprives us of information relevant to the interpretation of our data. Certainly, as is obligatory for qPCR, qRBA applications can be enhanced with the incorporation of references and calibrants. An example will be seen in a later Chapter (Section 6.1) with the use of reference samples added to a pool of experimental reporter variants. More importantly, the interpretation of absolute measurements is usually only possible in the context of other absolute measurements. This is true in a broad sense, but even more specifically true where experimental controls and benchmarks are concerned. If relative information can be obtained without absolute inputs, then absolute measurements often fulfil their role by giving a sense of magnitude, with little need for fine-tuning. **Whatever the case, there are plentiful scientific demands where quality relative information is more valuable than quality absolute information.**

3.4 Relative Fold Output and Complete Binarisation

To exploit qRBA fully, we must be comfortable designing studies and drawing conclusions from relative information. Suppose we have two populations, whose sizes v and w are altered by some process to V and W . In the context of qRBA, these represent two reporter variants.

$$v \rightarrow V \qquad w \rightarrow W$$

Arithmetically speaking, the metric I term ‘RFO’ is by no means novel. It refers simply to the output ratio $V:W$ divided by the input ratio $v:w$. While *fold difference* or *fold change* are more universally applicable than *fold output*, modelling two measurements as an input and output is applicable often enough and makes the ‘direction’ of the change clear. It also greatly simplifies verbal communication about the metric.

$$\text{RFO} = \frac{V:W}{v:w}$$

The usefulness of this calculation is that it tells us the fold increase or decrease of one population, relative to the other. A more intuitive calculation might involve four absolute measurements: v , V , w and W . The RFO provides the same answer from two relative

measurements, such as can be obtained using qRBA. **Though naturally it represents less overall information than four absolute measurements, as a standalone observation it is no less complete or accurate.** This is provable with rudimentary algebra (Fig. 3.4.1a). On the advice of colleagues, I have risked an overabundance of examples in Figure 3.4.1c to help others gain an intuition for the interpretation RFOs.

$$\text{Fold Output}_{[v]} = \frac{V}{v} \quad \text{Fold Output}_{[w]} = \frac{W}{w} \quad \text{Relative Fold Output} = \frac{V}{v} : \frac{W}{w}$$

$$\frac{V/v}{W/w} = \frac{Vw}{vW} = \frac{V/W}{v/w}$$

From its simplicity, a brief explanation in a methodology paragraph later in the thesis was also a realistic option. Space is dedicated to the RFO in this review chapter for two reasons. Firstly, there is a lack of harmony in current strategies for handling relative count data and comparing two groups with different reporter compositions. Relegating the concept to a methodology section would only contribute to this methodological mosaicism, and fail to make clear its advantages. Secondly, placing the RFO and the binomial model in the same neighbourhood makes clear their shared context in advancing an analytical infrastructure for qRBA and encouraging its adoption.

Naturally, most qRBA experiments will present more than two possible subpopulations in a set of count data. RFO analysis is made possible by appointing one variant (such as w) as a reference, then expressing all counts relative to that reference (Fig. 3.4.2). Two populations can then be compared by reporting the fold output of all variants relative to the reference variant. I refer to this approach tentatively as **complete binarisation** (having not found an existing term in the literature). This is especially useful where the reference variant has some special significance, such as a benchmark or control.

In NGS count analysis, the more common practice is to treat fractional abundances (in one form or another) as the objects of comparison between two readsets. Unfortunately, by viewing subpopulations this way, we are vulnerable to the assumption that changes to a subpopulation are wholly attributable to that subpopulation. In reality, the change in fractional abundance is arithmetically inseparable from all changes in all subpopulations. More complex still, measurement error in one subpopulation (both random and systematic) is inseparable from that of every subpopulation in the sample. We may note of course that a percentage, a fraction-

of-the-whole, and reads-per-million are fundamentally the same metric; 100 reads-per-million is simply 0.01%, 1×10^{-4} , or ‘one ten-thousandth’ of the population. Changes to reads-per-million (prevalent in transcriptomics⁸⁶) are therefore fully equivalent to changes in percentage-of-the-whole analysed commonly in rAAV reporter barcode experiments^{3,87}.

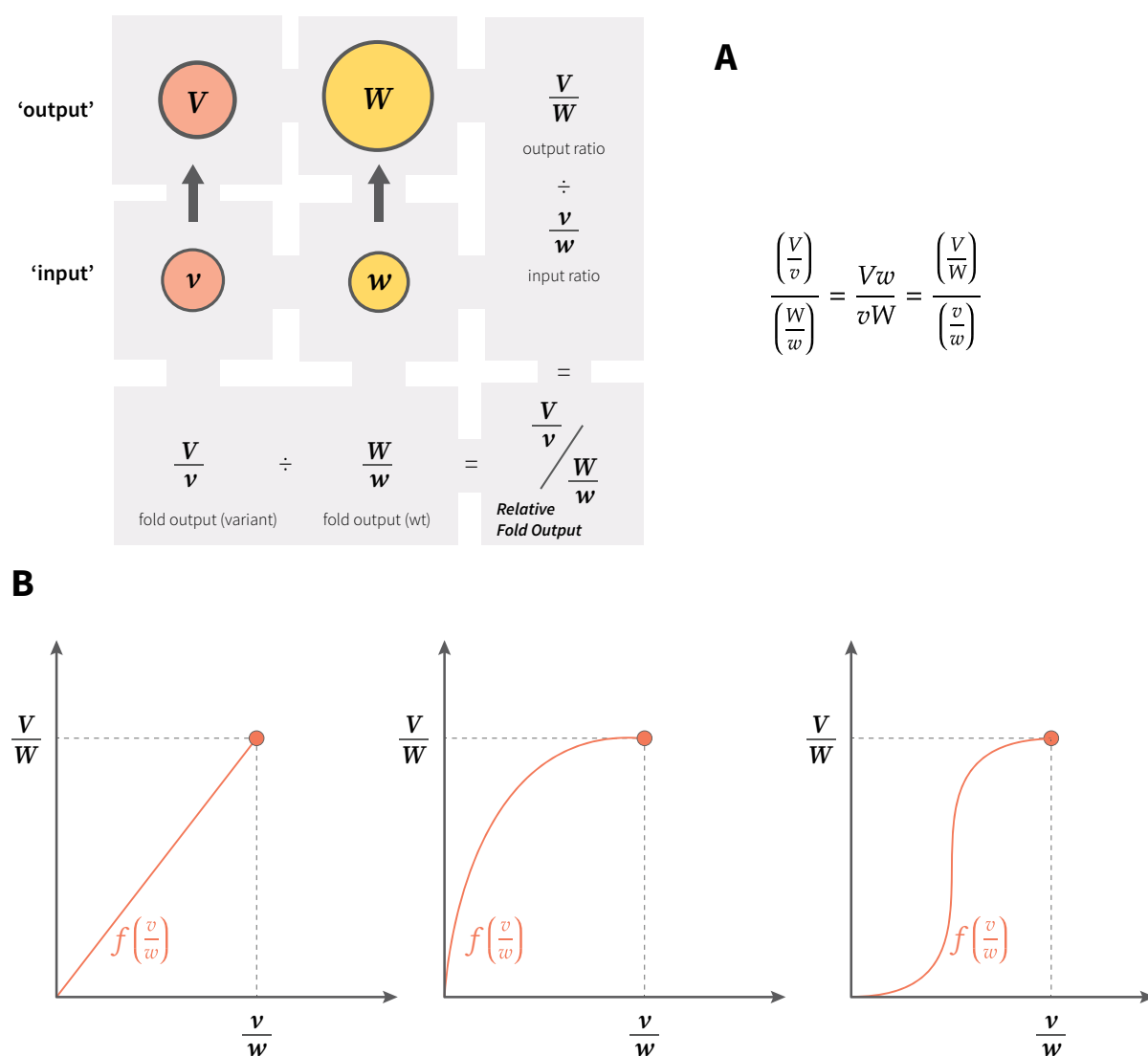


Figure 3.4.1 (1 of 2) 'Relative Fold Output' - a metric for comparing multinomial populations given relative information | **(A)** RFO is the output-to-input ratio of one group relative to that of another group. We can arrive at the RFO either by dividing the output-to-input ratio of one group by the other's or by dividing the output ratio of the two by their input ratio. Knowledge of or absolute individual quantities are therefore not required to compare the productivity of two groups in relative terms. Compared to other methods, the RFO makes obvious the relative nature of the measurement, avoids a multinomial error structure, and is easier to accurately interpret **(B)** The term ‘fold output’ also makes plain what is, and is not, being measured. It measures a fold change, but does not inform us of the underlying kinetics that led to that particular fold change. **(C)** Four examples of RFO measurements are provided to help readers build an intuition for this concept.

Mathematically, dividing %_[out] by %_[in] succeeds at stratifying fold changes of each subpopulation in relative terms. However, I argue that these strata are a more convoluted expression of RFOs. In fact, at the log scale, the vertical separation between RFOs is identical to that between $\frac{\%[\text{out}]}{\%[\text{in}]}$ values. This is because the sum of all reads acts as an ‘imaginary’ variant in the population, whose fold difference (decided mainly by the sequencing instrument) acts a point of reference. This fold difference is ultimately as arbitrary as that of a true reference

C

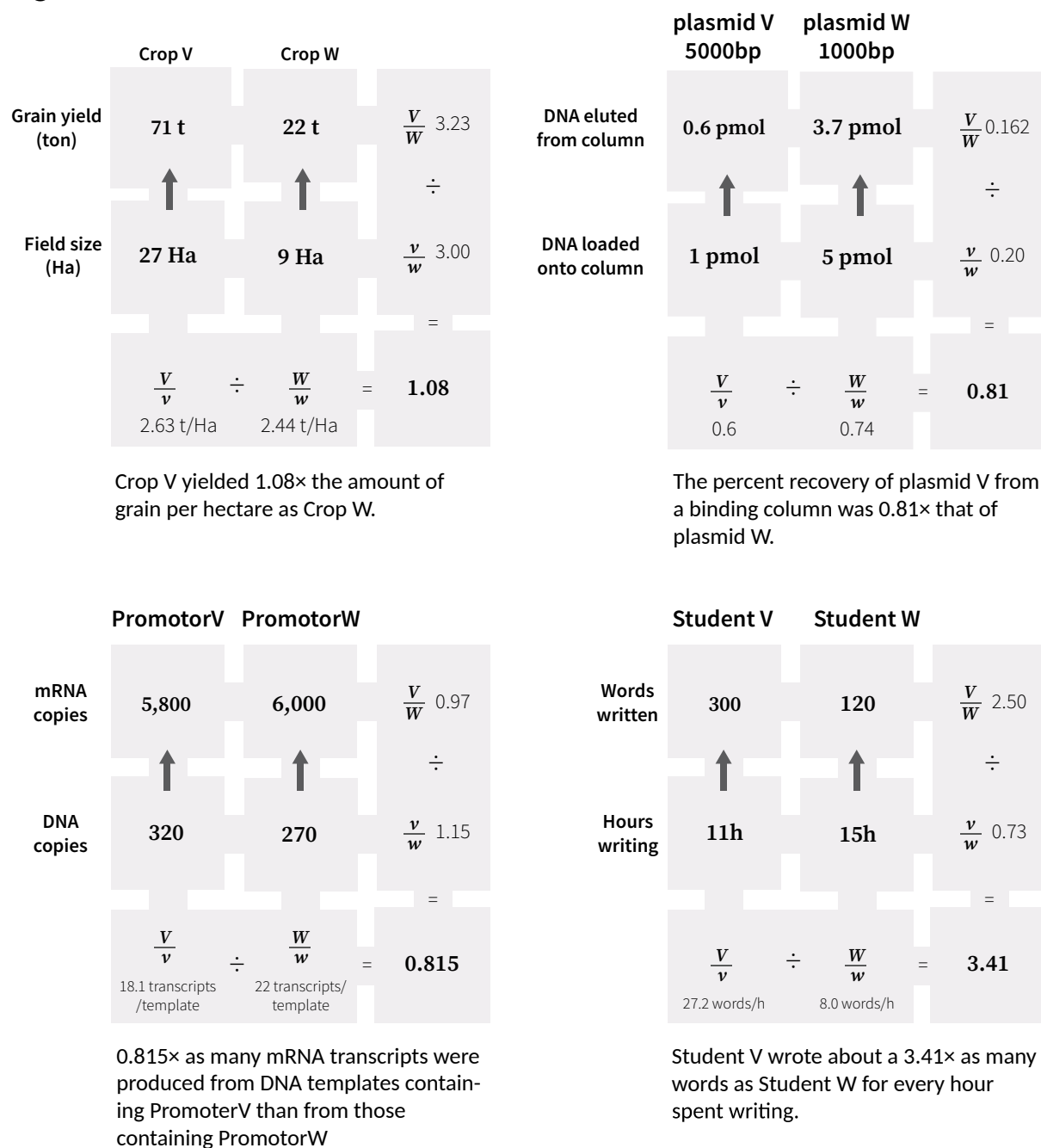


Figure 3.4.1 (2 of 2) 'Relative Fold Output' - a metric for comparing multinomial populations given relative information

variant, but unnecessarily abstract. The RFO is also clear in defining the *kind* of difference between two populations, i.e. their fold difference. This makes explicit a limitation inherent to both approaches – that this ratio does not tell us *what mode of change* resulted in the output values, and that a linear mode may or may not explain the data (Fig. 3.5.1b).

3.5 A binomial distribution model for qRBA count analysis

3.5.1 Preface

Non-experts in NGS count analysis have tended to rely informally on the notion that as sample size increases, a population can be more reliably represented, and that read counts achievable with Illumina MiSeq are adequate in this regard. To many molecular biologists, the notion of ‘a model’ in data analysis might feel somewhat esoteric. A model in the present context, stated simply, formally explains how measurand and readout are related. qRBA and dPCR⁶⁴ are alike in that that relationship is fundamentally probabilistic, not mechanical. This means that the ‘model’ will take the form of a statistical theory. The need for theories like this is perhaps more obvious with dPCR because the number of targets per partition cannot be ‘eyeballed’ from the number of empty and full partitions; it must be calculated using our knowledge of Poisson probability distributions. By contrast, the barcode composition of qRBA read data is the maximally likely composition of the sample². We can therefore infer the quantity of interest directly from the NGS barcode composition without resorting to undergraduate statistics.

However, as the International Bureau of Weights and Measures stresses⁷¹, a measurement reported without its degree of confidence has little meaning. We therefore cannot truly advance qRBA without a formal understanding of measurement uncertainty provided by a statistical model. As this thesis will show, a model like this offers more than just metrological credentials. It raises testable predictions regarding the relationship between UMI ratios, total molecular counts, and variability of qRBA measurements. It lets us judge the error contribution of its mechanical elements, i.e. protocols and instrumentation. It also allows analysts to make defensible guesses about data variability where numerous technical replicates are unfeasible.

Since I am not a statistics or bioinformatics specialist, I have not tasked myself with developing a consummate system. Here, as is usually true, the most important purpose of a model is to be improved.

3.5.2 Statement and basis of the model

As with the *negative binomial* model standard to transcriptomic count analysis²⁵, this model begins by treating the production of an NGS read sequence as a Bernoulli trial, and the category of that read (e.g. the gene-of-origin) as the outcome of that trial. It then diverges from the standard model to better fit the context and limitations of that application. Limitations to a multinomial model are thrown into sharp relief in transcriptomics, where a sample's sheer number of unique mRNA transcripts creates a dynamic range of alignable read counts, extending as low as 1 to 20 per million. The skew and variability of such read data, though understood to follow a multinomial distribution²⁵, renders a multinomial model (the assumption of normality in particular) unuseful, leading analysts to rely on Poisson-derived models. The standard negative binomial model (which will not be further discussed) is related to both of these models, though primarily because the degree of dispersion encountered in empirical data is incompatible²⁵. I note then that in transcriptomics, the negative binomial distribution is implemented not to estimate the relative abundance of cDNA transcripts *per se*, but rather to minimise the frequency of false hits when identifying differences between two or more transcriptomes.

This Chapter suggests we treat the production of an NGS read sequence as a Bernoulli trial and its reporter barcode sequence as the outcome. It deals with multinomial data on a pair-by-pair basis, 'restating' the multinomial process as a series of binomial processes. Pairs of interest can be analysed in isolation, or all variants can be paired with a common reference variant. The ratio between two reporters (r) is treated as the parameter to the binomial probability distribution, rather than fraction of one in a population consisting of both (p). The normal approximation has a mean of r and a standard deviation of ζ ,

$$\zeta = \left(\frac{p}{q}\right) \frac{1}{\sqrt{cpq}} \text{ or } \frac{r(1+r)}{\sqrt{cr}} \quad (3.1)$$

$$P(r_s|c, r) \approx \frac{1}{\sqrt{cpq}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{r_s-r}{\zeta}\right)^2} \quad (3.2)$$

NB: the normalising constant $\frac{1}{\sqrt{cpq}\sqrt{2\pi}}$ is suggested tentatively, since it results in equal estimates when $r_s = r$ as when $cp_s = cp$. The correctness of this normalising constant will not affect the spread of the hypothesized normal distribution.

Table 3.5.1 A summary of terms used in this thesis to express the binomial distribution

Term	binomial distribution	qRBA
BC_p	'Success', '1', one of two possible outcomes of a single Bernoulli trial	A specific reporter barcode
BC_q	'Failure', '0', the other of two possible outcomes of a single Bernoulli trial	A specific reporter barcode that isn't BC _p
c	Number of Bernoulli trials in a Bernoulli process	Readcount
p	P(success) for one Bernoulli trial,	Probability of an NGS read containing BC _p
q	P(failure) for one Bernoulli trial	Probability of an NGS read not containing BC _p
r	p:q	The ratio between two reporter variants (that physically exist). The measurand in the context of qRBA.
p_s	The <i>fraction</i> of Bernoulli trials from a Bernoulli process resulting in a success	The fraction of NGS reads containing BC _p
cp_s	The <i>number</i> of trials in a Bernoulli process that resulted in a success	The number of NGS reads containing BC _p
q_s	The <i>fraction</i> of Bernoulli trials in a Bernoulli process that did not result in a success	The fraction of NGS reads not containing BC _p
cq_s	The <i>number</i> of trials from a Bernoulli process not resulting in a success	The number of NGS reads not containing BC _p
r_s	The ratio of successes to failures in a Bernoulli process – i.e. p _s :q _s	The ratio of NGS reads containing BC _p and NGS reads containing BC _q
n	The number of independent Bernoulli processes	The number of qRBA measurement replicates

A summary of pronumerals was provided as front matter and is provided here again for ease-of-reference (Table 3.5.1). I note two deviations that experts will find unfamiliar: Firstly, the successful fraction of trials is assigned its own term p_s . The *absolute number* of successes in

a Bernoulli process (cp_s , denoted as k in Rosner¹ and Wypij²) is generally not referred to outside this Chapter. Since trials can number in the millions in qRBA, the *fraction* of successful trials will have clearer meaning more often than not. Secondly, the parameter conventionally expressed by n is expressed as c in this thesis to avoid later confusion between an NGS readcount (c) and the number of qRBA measurement replicates (n).

3.5.2 A primer on the binomial distribution

For the non-experts reading this chapter, consider the pouch of sand mentioned earlier, in which each grain had one of ten distinct colours. One of those colours is blue. Taking a single grain and noting its colour, the grain will be blue or not blue. Having two possible independent outcomes, this test can be referred to as a ***Bernoulli trial***. The two possible outcomes of a Bernoulli trial are canonically described as ‘success/failure’ or ‘1/0’. While arbitrary, the term ‘success’ is typically applied to a subjectively ‘positive’ result (e.g., blue rather than *not* blue). This will serve us well enough initially.

- The probability of selecting a blue grain is the proportion of blue grains in the entire pouch. Both of these amounts can be expressed as p .
- The probability of selecting a grain that isn’t blue is $1 - p$, or more simply q .

$$\begin{aligned} P(\text{grain is blue}) &= p \\ P(\text{grain is not blue}) &= (1 - p) = q \\ p + q &= 1 \end{aligned}$$

- When a Bernoulli trial is repeated a certain number of times, we will call this number c (analogous to *readcount* in qRBA).
- The fraction of these trials resulting in a ‘success’ will be called p_s .
- $q_s = 1 - p_s$ and $cp_s + cq_s = c$.

We can think of p_s as the fraction of the *sample* that were successful – distinct from p , the fraction of the *population that is being sampled*. When c repeated trials are performed, the possible number of ‘successful’ trials (cp_s) will be any number between 0 and c (assuming $p \neq 0$ or 1). **Repeating a Bernoulli trial c times and obtaining cp_s successes is called a *Bernoulli process*.**

A binomial distribution relates p and c to the probability of any possible outcome of a Bernoulli process (Fig. 3.5.1a). Given the parameters c and p , the *exact* probability of any possible outcome cp_s is calculated using the binomial probability mass function (PMF), which is shown below (though only for completeness). The cumulative probability of every possible cp_s is 1.

$$P(cp_s \text{ given } c \text{ and } p) = \frac{c!}{(cp_s)!(cq_s)!} \times p^{cp_s} q^{cq_s}$$

Probability is maximal when $cp_s = cp$, and probability decreases from the maximum in an *approximately* normal fashion. The normal curve that approximates the binomial PMF has a mean (μ) of cp and standard deviation (σ) of $\sqrt{cpq}$. (Non-experts can note for completeness: Euler's number exponentiated as below gives a normal curve, and the preceding term serves to reduce the area under that curve to 1, as is characteristic of a probability distribution).

$$P(cp_s \text{ given } c \text{ and } p) \approx \frac{1}{\sqrt{cpq}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{cp_s - cp}{\sqrt{cpq}}\right)^2}$$

A binomial distribution can be thought of as a normal distribution with at least three important corrections. Firstly, the range of possible cp_s values is limited by 0 and c , unlike the normal distribution, which in theory has infinitely long tails. Secondly, the binomial distribution is only perfectly symmetrical when $p = 0.5$. In an intuitive sense, we might see this as a consequence of the first correction; as p approaches 0 or 1 there is decreasing 'room' for complete symmetry. More formally, we could deduce from the binomial PMF that $P(cp + x)$ and $P(cp - x)$ are not completely equal unless $p = 0.5$. It is important to stress however that whatever the asymmetry, the probability is always maximal at cp , meaning that p_s is always the maximally likely true value of p^2 . Thirdly, the binomial distribution is fundamentally a discrete distribution, where cp_s is always an integer and p_s therefore has a discrete list of possible values. Binomial distributions lie on a spectrum, approximated by a Poisson distribution at low c and/or p further from 0.5, and approximated by a normal distribution at high c and/or p closer to 0.5^{2,88}. Transition boundaries suggested as rules-of-thumb¹ often provoke credible contention⁸⁹. For illustrative purposes however, authors Brown *et al.* have drawn attention to certain shortcomings of the normal approximation even at values as 'high' as $c = 100$ (specifically its ability to pinpoint a 95% confidence interval). Nuances of binomial confidence intervals aside, we can infer from that discourse that $c = 100$ is generally regarded as a high number of trials by statisticians. For

qRBA, read counts are typically order/s of magnitude higher than 100. That is, far higher than applicable in the usual discourse about limitations of the normal approximation.

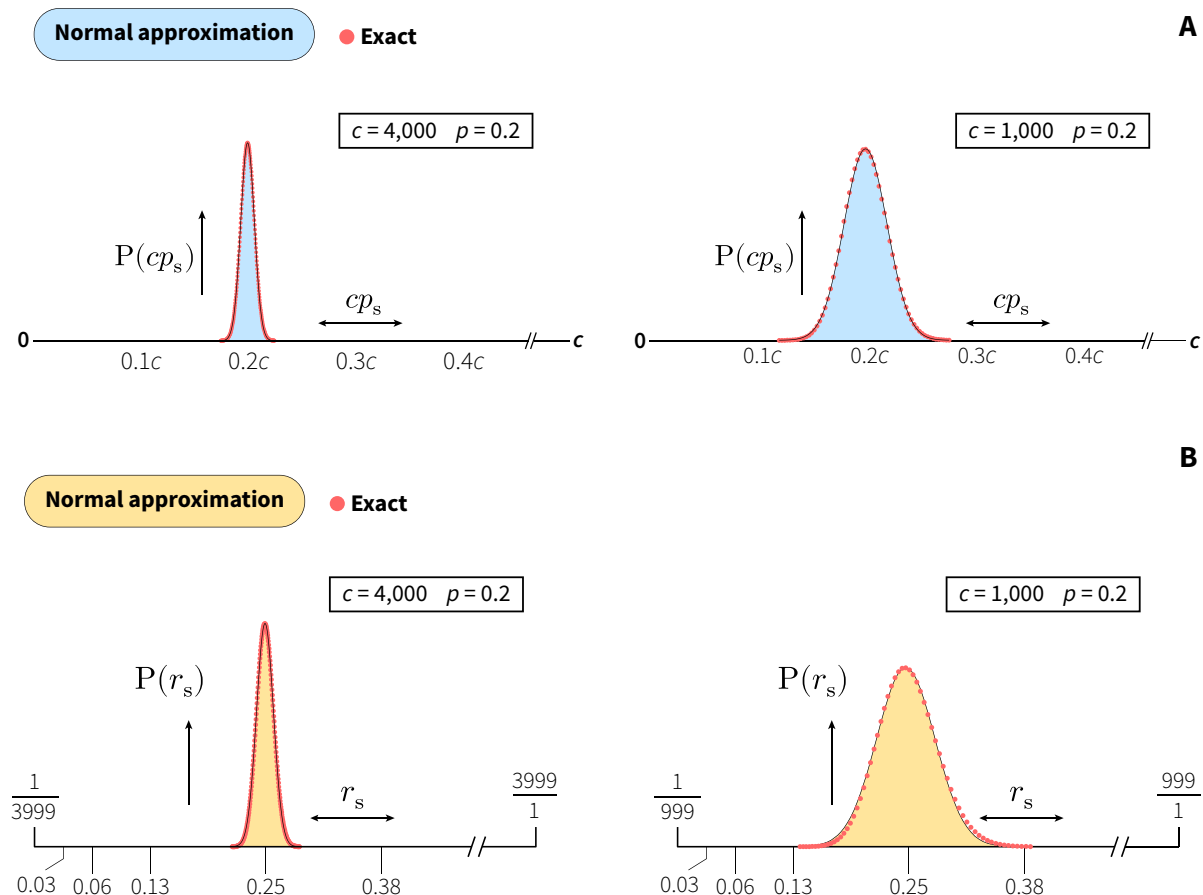


Figure 3.5.1 A binomial model for qRBA count analysis A Bernoulli trial is one whose outcomes have the same probabilities every time the trial is repeated. A Bernoulli process is a series of Bernoulli trials, some fraction (p_s) of which will result in a certain outcome, and the remainder will not. The probability of observing that outcome in any given trial is p . A binomial distribution applies when the number of different possible outcome is two. It describes the relationship between the number of trials (c), any possible value of p_s and the probability of p_s occurring. The maximally probable value of p_s will be p , and the probability overall will be distributed in an approximately normal manner. This approximation improves as p approaches 50% and/or as the number of trials in the Bernoulli process increases, and its standard deviation (in absolute terms) will be equal to $\sqrt{cpq}$. In relative terms, the spread decreases as c increases, as is shown above. **(A)** The exact probability distributions and normal approximations for the parameters indicated are superimposed, with p_s plotted on the x-axis as per convention. **(B)** The ratio of the two outcomes (r_s) is plotted as the independent variable in place of p_s and the corresponding exact probabilities shown as red spots. The normal approximation uses ζ (defined in the main text) to determine the standard deviation of this normal approximation.

3.5.3 Complete binarisation, use of the ratio, and the SD of the ratio

To better suit the binomial model to qRBA measurement, I suggest we ‘rephrase’ the model so that the $p:q$ ratio (r) is viewed as the main parameter, rather than the fraction p . Likewise, the independent variable of the binomial distribution would be the ratio of $p_s:q_s$ (r_s). The

parameter c , rather than shaping the outcome in absolute terms, affects only the spread of the distribution about the mean r . To analyse multinomial data in binary terms, we apply complete binarisation using a variant of interest (or an arbitrary variant) as the common denominator to all pairs. Analysing ratios of completely binarised count data has two advantages. Firstly, it harmonises the use of the binomial model with RFO analysis. As argued in Section 3.4, complete binarisation simplifies the error structure and clarifies the thought process by foregrounding the relativeness of count information. It also removes a layer of abstraction when two multinomial populations are compared. Secondly, it avoids a distribution whose *relative standard deviation* (RSD) – and therefore metrological precision – depends on which barcode is classed as a ‘success’.

This chapter has discussed measurement uncertainty extensively but not yet made mention of RSD aka the *coefficient of variation* (CV). Like variance and 95% confidence intervals, the RSD is an expression of metrological uncertainty. The RSD of a measurement is its standard deviation relative to its average. For example, 100 ± 10 , 44 ± 4.4 and 1 ± 0.1 all have an RSD of 0.1, or 10%. Crucially, the RSD of prior measurements is needed to determine the uncertainty of products or quotients of those measurements. Using a Bernoulli process to measure p , we run into a problem concerning RSD. Recall that the designation of ‘success’ (aka p) is arbitrary, and consider that any measurement of p is an equally valid measurement of q . Despite this, cp_s and cq_s have different RSDs. The more precisely we measure p , the less precisely we measure q , and vice versa.

$$\text{RSD}_{cp_s} = \frac{\sqrt{cpq}}{cp} \quad \text{RSD}_{cq_s} = \frac{\sqrt{cpq}}{cq}$$

While not truly paradoxical, it can easily lead to confusion and mishandling of uncertainty. A more ‘egalitarian’ measure of p and q is to take the $cp_s : cq_s$ ratio (simplified to $p_s : q_s$) as the result of the Bernoulli process, rather than cp_s as a fraction of c . Taken as a quotient, the RSDs of p_s and q_s are combined additively, the result of which is the same for $cp_s : cq_s$ and its reciprocal. That is to say:

$$\text{RSD}_{cp_s} \neq \text{RSD}_{cq_s} \text{ but } \text{RSD}_{\frac{cp_s}{cq_s}} = \text{RSD}_{\frac{cq_s}{cp_s}}$$

Unfortunately, lacking the relevant mathematical training, I cannot provide a solution to the RSD of r_s with an analytical proof. Derived in Appendix C however is a solution ζ (as given in Eq. 3.1) arrived at simply by adding the RSD cp_s and cq_s . This method is arguably crude, usually

taught only for its heuristic value^{76,77}, though apparently successful (Fig. 3.5.1b). International Standards Organisation guidance for testing and calibration laboratories (ISO/IEC 17025:2017[E]) indicates that the RSD of a product or quotient be calculated by treating it as the hypotenuse of a right triangle, whose other two sides are the RSDs of the two prior measurands. I note however that this method was unsuccessful and possibly inapplicable (not shown).

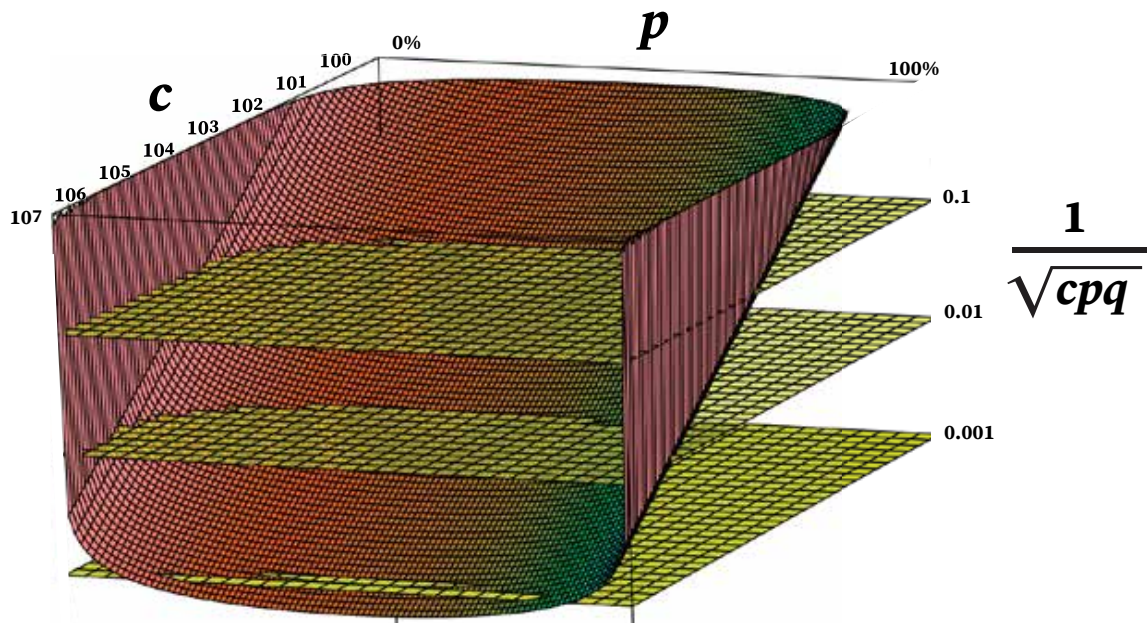


Figure 3.5.2 A hypothesised relationship between readcount, reporter pair ratios, and the theoretical precision of qRBA measurements | $1/\sqrt{cpq}$ is a hypothesised expression for the inherent variability of $p:q$ in Bernoulli processes, relative to their average. Taking the ratio of p to q rather than p as a fraction of the whole, both operands contribute to the variability equally and a symmetry is achieved that satisfies both possible ways that 'success' and 'failure' of a Bernoulli trial can be classified. Aside from offering a testable prediction of the model, it offers an intuitive sense of how readcount and pair ratios affect qRBA measurement precision. It implies that variability is lowest when p and q are equal in proportion, and rises parabolically as either operand approaches 0% or 100%. It also implies that variability is inversely proportional to the square of the readcount, such that a ten-fold increase in precision would require a 100-fold increase in readcount for any value of p . 3D plot generated using the CalcPlot3D online tool <<https://c3d.libretexts.org/CalcPlot3D/index.html>> accessed 10 Nov 2022.

The RSD of $p_s:q_s$ is a critical aspect of the binomial model as it is the means by which the boundaries of 'quantitativeness' can be inferred. Figure 3.5.2 illustrates the relationship between p , c and the RSD of $p_s:q_s$. In the sense that acceptable partition counts and target concentrations for dPCR can be inferred from a Poisson model⁹⁰, the binomial model might be used to infer acceptable ranges of c and p for qRBA. Aside from acting as a testable prediction, the model offers useful insights. For example, precision is highest when $p = 0.5$ (i.e., when $p = q$) and declines parabolically as p approaches 0 or 1. It also indicates that precision is inversely

proportional to the square root of c . This would imply that readcount must be increased by 4 ($=2^2$) in order to double the precision of an Illumina qRBA measurement.

With this binomial model, a basis for the precision and accuracy of qRBA measurement is proposed. The basis for precision is ζ . The basis for accuracy is that p_s is an unbiased estimator of the parameter p and by definition has the same discrete probability distribution as r_s . One of the features needed in the present model is a method for formally assigning confidence to estimations of p from p_s , and from multiple replicated p_s measurements. The binomial distribution predicts outcomes of Bernoulli processes given p , but the intended application is to infer p from p_s . Here it may be useful to note that the likelihood ratio of all possible $P(p \neq p_s)$ to $P(p = p_s)$ is distributed equally to $P(p_s)$ in relative terms, given that this true for any curve divided by a constant. That is to say, if the distribution of $P(p_s = p)$ is taken to be normal, then the distribution of $P(p \neq p_s) : P(p = p_s)$ is also normal and the empirical rule is equally applicable. This said, in this area I am at even greater risk of being unproductive without expert assistance. Like the rest of this model, it should be regarded as the start of a scientific conversation and not the end of one.

3.6 Conclusions

Scientists in the field of rAAV vector engineering have been responsible for numerous methodological innovations exploiting population sequencing technology in the analysis of large, mixed populations of nucleic acid copies with variable origins, made distinguishable and countable by unique molecular identifiers. Though highly versatile, the core methodological concept is unique in the UMI landscape and worthy of a formal synthesis so that it may be communicated, developed, and exploited by the molecular biological community at large. To this end, this review has refined the concept of a *reporter barcode* and laid the groundwork for the larger methodological concept of *quantitative reporter barcode analysis*. This dissociates the method from a cluster of techniques that align in part, but characteristically more stochastic and error prone. The chapter has identified and articulated qualities of qRBA that suit it outstandingly to the task of making replicable, precise, and accurate quantitative measurements. Its mechanism, in principle, makes it a valuable complement to dPCR and qPCR in the molecular biological landscape and in many contexts a superior option. It therefore calls resoundingly for systematic development, including the pursuit of metrological

credentials and a sustained effort to identify and manage limitations and hazards. To support this endeavour, this review has also suggested a statistical model based on the binomial distribution, as a means of formally handling the probabilistic connection between measurand and readout.

4

A metrological exploration of qRBA

Introduction & Outline

In Chapter 3, I formalised a methodological concept that has coalesced in recent years, wherein predefined, endogenous UMIs are used to distinguish and quantify experimental subpopulations in population sequence data. It was argued that this technique has considerable and underrealised robustness both at the point of measurement and upstream from measurement. In the interest of advancing qRBA, a model was developed based on the binomial distribution, which formally describes the probabilistic process relating measurand and readout. The model holds that reporter variants in a multinomial population can be measured as independent pairs, the measurand being their ratio and the readout being the sample ratio r_s . With r_s treated as the outcome rather than the more conventional variable cp_s , the standard deviation of the normal approximation distribution is hypothesised to be ζ . This chapter will test three hypotheses, each with the mutual goal of establishing the analytical credibility of qRBA for audiences within and outside rAAV vector research.

In service of these aims, I generated a set of ten plasmid reporter variants (**Section 4.1**), which would act as targets for qRBA measurement. Unlike the construct used in earlier rAAV experiments, each plasmid copy contained two independently targetable reporter barcodes in separate 150bp PCR target regions ('conjugate' reporter barcodes). This allowed me to mutually cross-examine qRBA measurements of plasmid ratios with a guarantee that the conjugate targets were equimolar. Section 4.1 is a technical prerequisite to later sections and is not accompanied by a hypothesis or a discussion.

Aim 1a To determine whether random variation in UMI count ratios is governed by binomial statistics.

Hypothesis 1 (Section 4.2) The standard deviation of r_s in in silico simulated Bernoulli processes will vary randomly and symmetrically from ζ

Hypothesis 2 (Section 4.3) In fully binarised multinomial qRBA data, the residual distribution between r_s and ζ_s will match that of Bernoulli processes simulated in silico under equivalent constraints.

$$\zeta = \left(\frac{p}{q}\right) \frac{1}{\sqrt{cpq}} \text{ or } \frac{r(1+r)}{\sqrt{cr}} \quad (3.1)$$

$$r_s \sim N(r, \zeta^2) \quad (3.2)$$

Aim 1b To test the accuracy of qRBA using a set of plasmids containing two different qRBA target regions.

Hypothesis 3 (Section 4.4) When plasmid ratios are measured from two possible qRBA targets, the differences between measured ratios will vary randomly and symmetrically about zero.

4.1 Design and construction of experimental conjugate reporter plasmids

A simple, accessible, manageable PCR template such as a plasmid is an ideal material to explore the precision and accuracy of Illumina NGS-based qRBA measurement. It is particularly appropriate where cross-contamination of reporter variants is a concern, since monoclonality can be used as a tool to help ensure homogeneity. This advantage applies in the construction process and later if cross-contamination is suspected. This would be impractical if premade amplicons were used as template material, though these would be more directly adjacent to the measurement process. The main drawback of a plasmid template is its variable topoisomer composition, which can affect amplification efficiency⁸. However, this problem can be obviated by cleaving the plasmid with a restriction endonuclease prior to PCR.

To investigate the accuracy of qRBA, I made a set of ten plasmid constructs, each containing two separate 150bp reporter PCR target regions. Where two reporter barcodes are present on a single molecule, I will apply the term *conjugate reporter barcodes*. These will allow me to measure plasmid ratios using two different qRBA assays with a guarantee that the alternative reporter PCR targets are equimolar. These experiments, by design, are unreliant on orthogonal measurements of individual plasmid variants. Such measurements, as accurate and/or precise as they may be, will not exceed our certainty that conjugate reporters are equimolar. Orthogonal measurements, including qPCR or ddPCR and their surrounding workflow (see Chapter 3) may in fact be less precise than qRBA measurement and are of course open to biases and errors. This means that orthogonal measurements (though potentially valuable in future experiments) risk introducing unnecessary ambiguities. *Conjugate* reporter barcodes are not to be confused with *partnered* reporter barcodes. Since both strategies are used in this thesis it is worth clarifying the terms proactively (Fig 4.1.1).

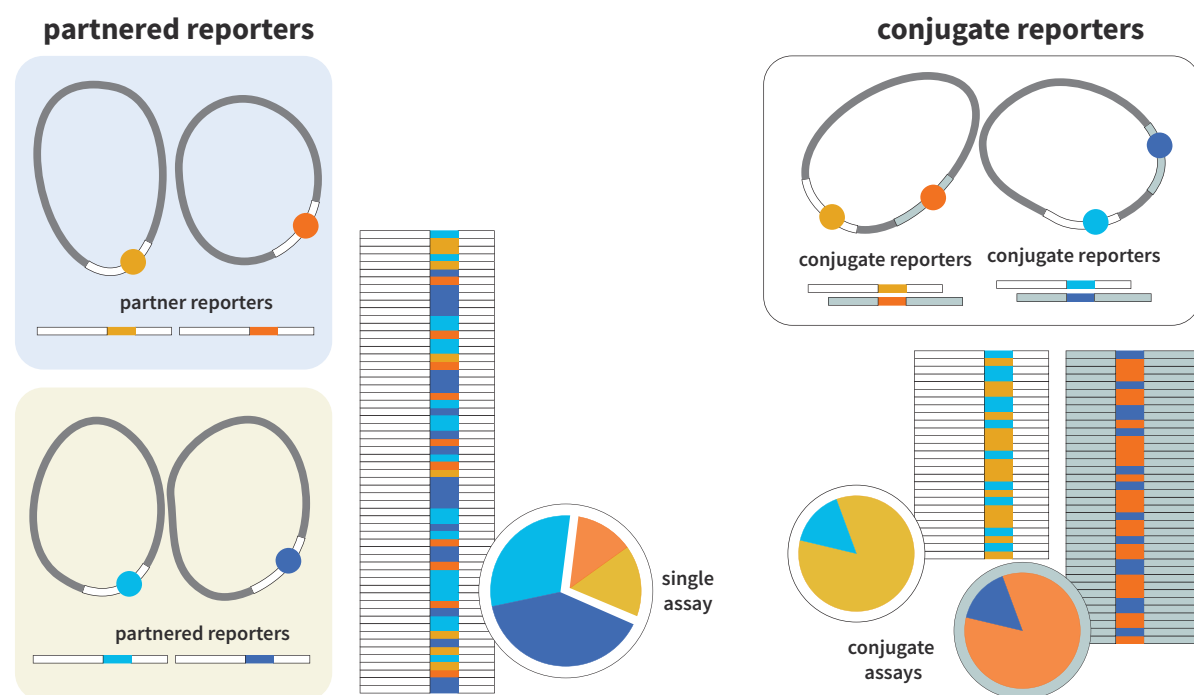


Figure 4.1.1 Distinction between 'partnered' and 'conjugate' reporter barcode strategies | Two unrelated but easy-to-confuse strategies involving a link between two different reporter barcodes are utilised in this thesis. Partnered reporters do not appear in this Chapter, but are prominent in Chapter 6. Conjugate reporter barcodes coexist in separate locations on a single molecule of D/RNA. They are amplified with different primer sets, producing two different reporter amplicon sequences, representing two completely separate multinomial populations. Partnered reporter barcodes are two UMI variants of the same PCR target sequence and do not coexist on one molecule – each copy in a paired population contains one of either barcode. A pair of reporter barcodes may be designed to encode the same subpopulation or treatment group, acting as experimental duplicates of that group or simply a control against read bias. Reads derived from conjugate reporters belong to completely distinct countable populations of NGS reads. Reads derived from partnered reporters are variants of the same reporter amplicon sequence and therefore part of the same multinomial population. Two 'conjugate qRBA assays' use different PCR targets of the same barcoded population to measure their ratios. They do not measure the ratio between one reporter barcode and its conjugate reporter barcode, which would always be 1:1.

These plasmids were constructed from two different sets of qRBA targets sequences available in my laboratory. The first was a set of *luc2*-adjacent reporter barcodes generated in collaboration with fellow doctoral candidate from my laboratory Kimberly L. Dilworth for applications in ITR-related experiments. The second was an unrelated *egfp*-adjacent reporter barcode constructed for use in a published rAAV capsid ‘kit’ study³. These two sets of plasmid reporter variants acted as restriction fragment donors. The final construct(s) was cloned via three-piece ligation into a pMB1-derived plasmid backbone (assisted by calf intestinal phosphatase treatment) and cloned via bacterial transformation and colony screening. The gross structure of the ligated product was verified via a series of analytical restriction digests and each reporter barcode was verified via Sanger sequencing from both directions. Cross contamination was minimised in the preparation stage by culturing and preparing the plasmids in ~1mg quantities (‘maxipreps’) in groups of three, three and four variants on separate days. The final construct is shown in Figure 4.1.2.

These ten constructs comprised ten “*egfp* reporter” variants and nine downstream “*luc2* reporter” variants – two constructs shared a *luc2* reporter sequence. There were a number of factors which on balance led me to pursue this design. Primarily, the time available to design and execute the study was constrained. Only nine *luc2* donor plasmids were available at the time the project was commenced, and due to a combination of logistical factors, obtaining the tenth would have incurred a nontrivial delay. The rationale behind proceeding with the nine available *luc2*-adjacent barcodes was that only one planned experiment (not shown) called for ten reporters by necessity. Having nine reporters in one of the regions, though it would lead to inconsistency in the sample structure, would not otherwise prevent the successful completion of this project. Another contributing reason was that having two *egfp* barcodes with a conjugate *luc2* barcode in common could provide insight since the binomial model predicts that the quantity of the identical *luc2* targets should be the sum of their *egfp* conjugates.

One pertinent design consideration was the Hamming distance between the reporter sequences of a given panel. In lay terms, this is the number of base mismatches required for one sequence to be read as the other by mistake. By analogy, the words READ and MEAD have a hamming distance of 1, READ and MEND have a Hamming distance of 2, and READ and MIND have a Hamming distance of 3. A Hamming distance greater than 1 allows us to mitigate the potential for sequencing bias by tolerating one mismatch. In my panel of ten conjugate-reporter plasmids, with 19 unique reporters in total, *egfp* barcodes G and D are the only two reporter

barcodes that can be mutually misidentified as a result of two specific mismatches. Note that the most likely result of two mismatches is exclusion of the read rather than misidentification, since only one of nine possible miscall combinations will match the alternative reporter.

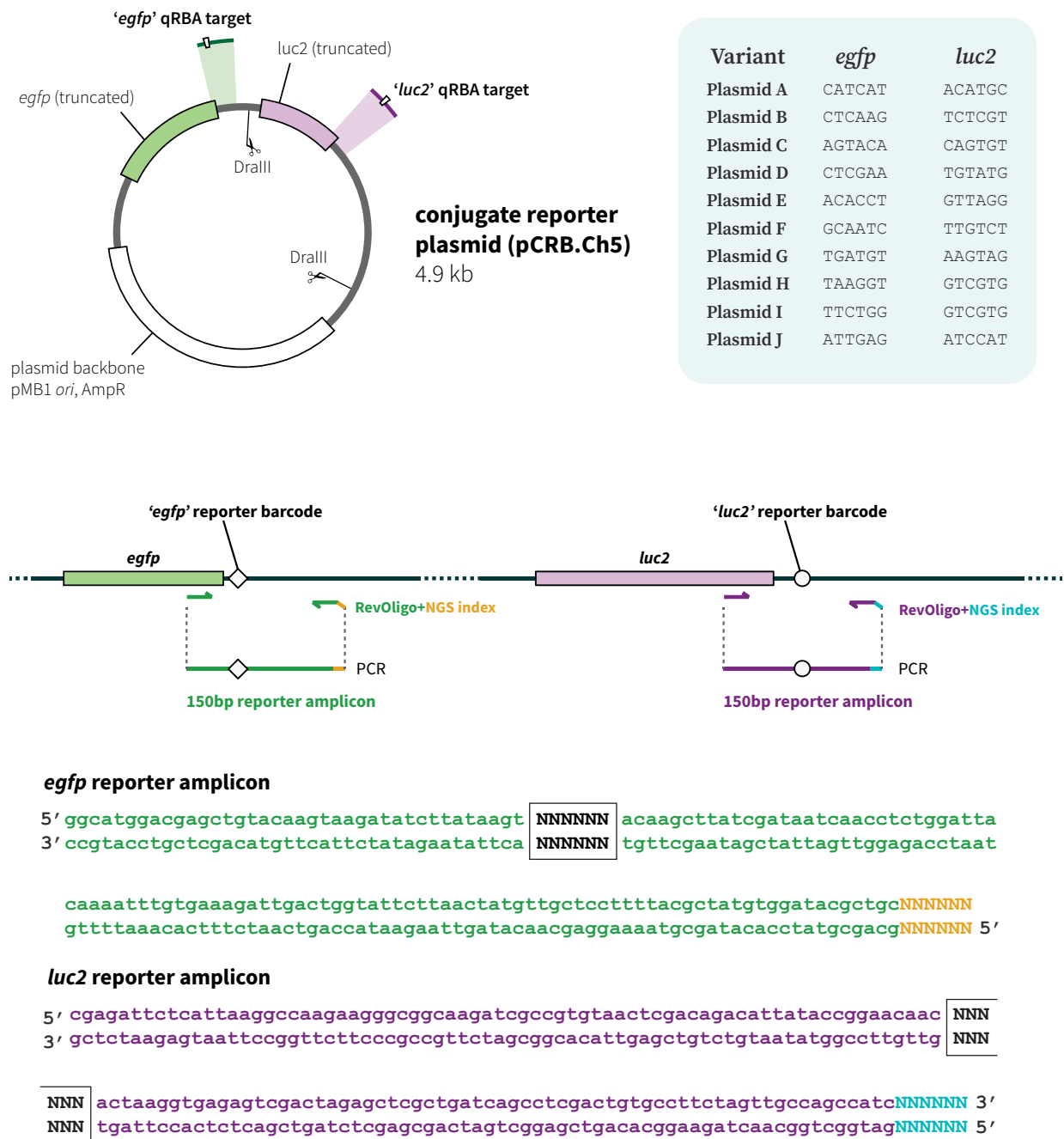


Figure 4.1.2 A plasmid construct with two independently measurable qRBA targets | For this study a plasmid was designed with conjugate reporter barcodes so that qRBA measurements could be made with two independent assays, which would produce the same result if accurate. The barcodes are adjacent to partial *egfp* and *luc2* sequences, which are vestiges of the donor plasmids used to clone the conjugate barcode construct. Throughout this section, these two alternative qRBA targets are termed with reference to their adjacent reporter gene. Ten variants of this construct were prepared, though variants H and I bore identical barcodes in the *luc2* assay. DralII sites indicate where the plasmid was cleaved prior to PCR amplification. This eliminated topological variation and split the qRBA targets between equimolar plasmid fragments.

To analyse either target with Illumina MiSeq, all plasmid topoisomers were reduced to a linear structure via DraIII restriction digestion and the endonuclease was inactivated by incubating at 80°C for 20 minutes. Approximately 1ng of template was then transferred from the inactivated digest mix to a PCR reaction mix and amplified using 25 cycles. Each separate reaction mixture contained a reverse primer with a 6N tail, which was incorporated into the amplicon as a terminal 6N index unique to each sample, enabling multiplexing (pooling and analysis of separate PCR samples as a mixture). Prior to pooling the PCR products were verified via gel electrophoresis of a 10µL sample in a narrow-welled 2% agarose gel. Indexed, gel-verified PCR products were consolidated into a single tube and column purified. This NGS-ready mix of PCR products is referred to as a **library**. 1µg of the column-purified PCR products were sent to the service provider (GENEWIZ Genomics, Suzhou) as dry samples in 1.5mL microfuge tubes.

4.2 Testing ζ using *in silico* simulated Bernoulli processes

The normal curve specified by parameters r and ζ fit satisfyingly to the exact binomial distribution calculated from the binomial PMF. However satisfying though, this is not a substitute for a rigorous mathematical proof, which has not been provided. To further support the appropriateness of ζ , Bernoulli processes were simulated *in silico* and ζ compared to the standard deviations of r_s that result. Each SD obtained from the simulation describes the result of 621 independent Bernoulli processes ($n = 621$). Parameters for p and c are given in Table 4.2.1. Note that p and r are interchangeable; $r = p/q$ and $n = r/(1 + r)$.

To simulate a Bernoulli trial in Microsoft Excel, I prepared a simulated ‘Urn’ (WB5.1 WS1) represented by a column of 100 cells, each containing either the character p or the character q . To ‘draw from the urn’ I used the INDEX and RANDBETWEEN functions to return the content of a random Urn cell into a ‘Bernoulli cell’ (WB5.1 WS2). The probability of returning a p was controlled by the number of cells in the Urn containing “ p ”. Carrying this process out in c Bernoulli cells simultaneously constituted a Bernoulli process with c trials. The fraction of those cells returning a “ p ” was p_s and the fraction returning “ q ” was q_s . To carry this out more efficiently, each row in the Urn contained a string of 13 characters rather than a single “ p ” or the letter “ q ”, with variable probabilities of “ p ” specific to each character position. This allowed

Table 4.2.1 list of parameters applied in *in silico* simulated Bernoulli processes

<i>p</i> =			<i>c</i> =	<i>n</i> =
	0.20		1,000	
0.01		0.90	2,000	
	0.40		5,000	
0.02		0.95	7,000	621
	0.50		10,000	
0.05		0.98	20,000	
	0.60		50,000	
0.10		0.99		
	0.80			

each Bernoulli cell to perform 13 Bernoulli trials every turn. To test multiple parameters of *c* simultaneously, the COUNTIF function was used to count Bernoulli cells containing “*p*” in specified numbers of Bernoulli cells. To compensate for the ‘pseudorandom’ number generation used by Excel and most other software, the column of Urn cells was randomly rearranged between each turn. This meant that the contents of all Urn cells and the Urn cell that was selected were both pseudorandom. This Bernoulli process was repeated 621 times and the average (not shown) and standard deviation of r_s (SDr_s) was compared to ζ (WB5.1 WS4-5).

The average, sample standard deviation, and relative sample standard deviation of 621 independent Bernoulli processes were obtained for 91 unique pairs of (*c*, *p*) parameters. These parameters (listed in Table 4.2.1) were designed to extend into domains where the normal approximation of the binomial distribution was decreasingly applicable.

4.2.1 Results

Viewed at a log scale in the simulated data conformed approximately to the line $y = x$ when ζ and SDr_s were plotted as (*x*, *y*) coordinates respectively, though a negative bias was apparent as *x* increased (Fig. 4.2.1). *x*-dependent biases were clearer when residuals were plotted on the *y*-axis (Fig. 4.2.2). For added context, residual analysis was also performed on the relationship between the more conventional variable cp_s (often elsewhere expressed *k*) and its established SD of $(cpq)^{0.5}$. Least-squares regression analysis on log-log data returned the model $\text{Log}(SDr_s) = 1.000 \text{ Log}(\zeta) + 0.01$, which differed from the observed data by an average of 1.1%.

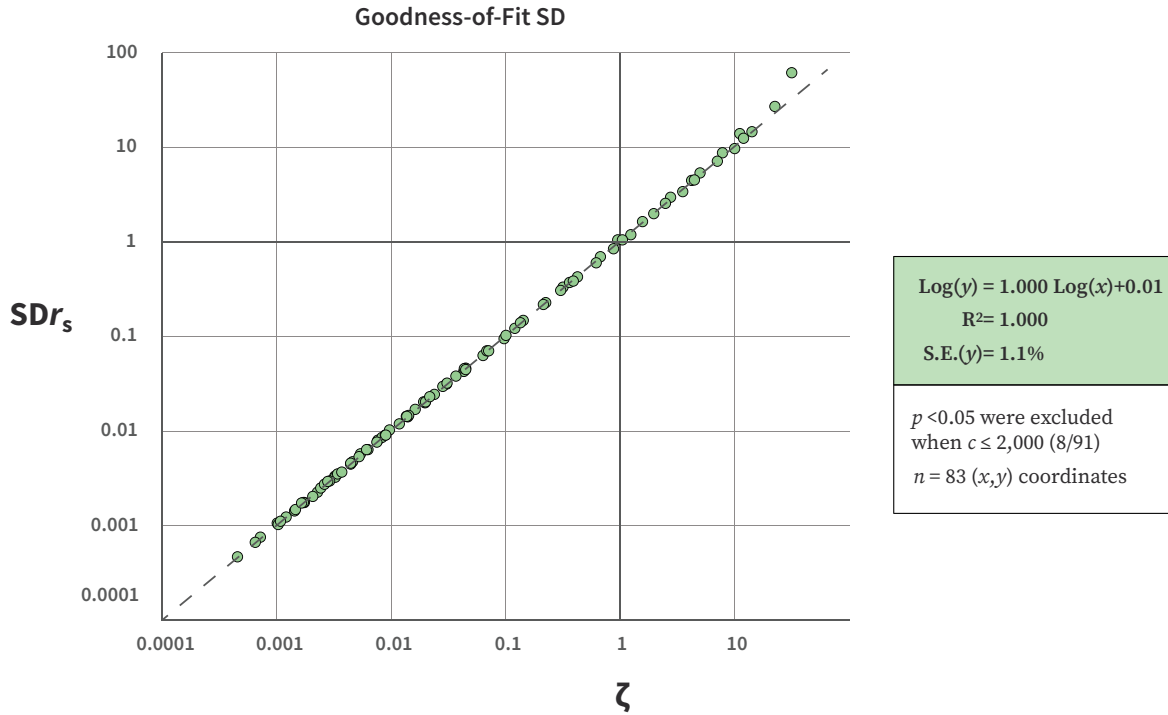


Figure 4.2.1 Relationship between ζ and variation in simulated Bernoulli processes | While $\sqrt{cpq}$ is the accepted expression for the SD of cp_s , an expression for the count ratio r_s could not be identified in the literature. The term ζ was Proposed in Chapter 3 and tested in *in silico* in this section. Each circle in this plot represents the SD r_s taken from 621 replicated Bernoulli process. The plot contains 91 green circles, representing 91 unique c, p parameters where c ranged from 1,000 to 50,000 and p ranged from 0.01 to 0.99. Predicted and observed values were plotted as x, y coordinates.

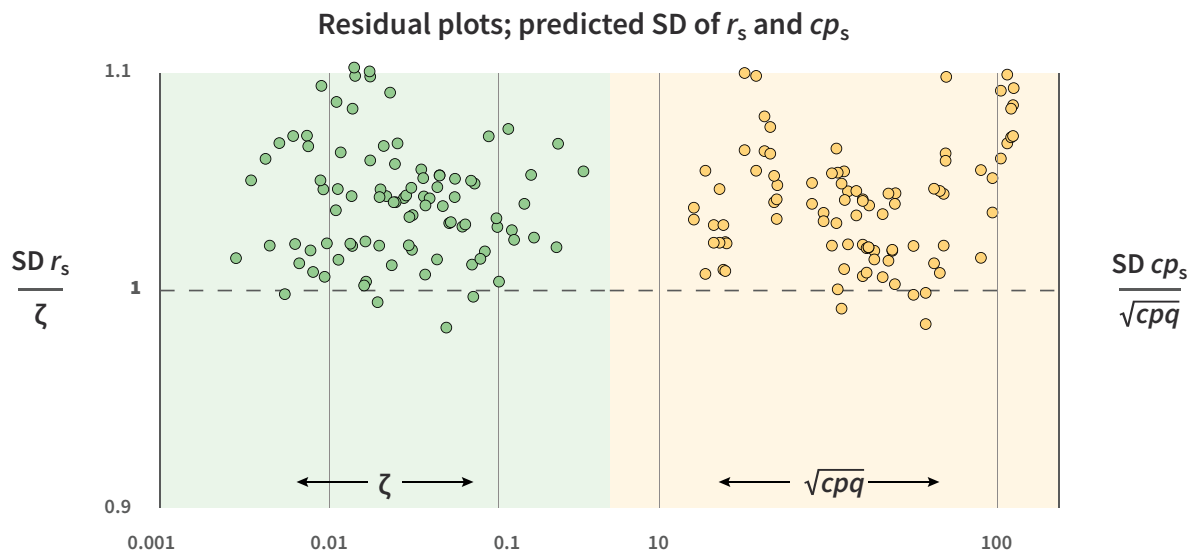


Figure 4.2.2 Residual analysis of predicted and simulated variation | According to a log-log linear regression, the relationship between observed and predicted variation was well approximated by $y=x$, though residual analysis highlighted a pattern of x -dependent negative bias in both predictors. Both the datasets above exclude processes where $c \leq 2,000$ and $p < 0.05$.

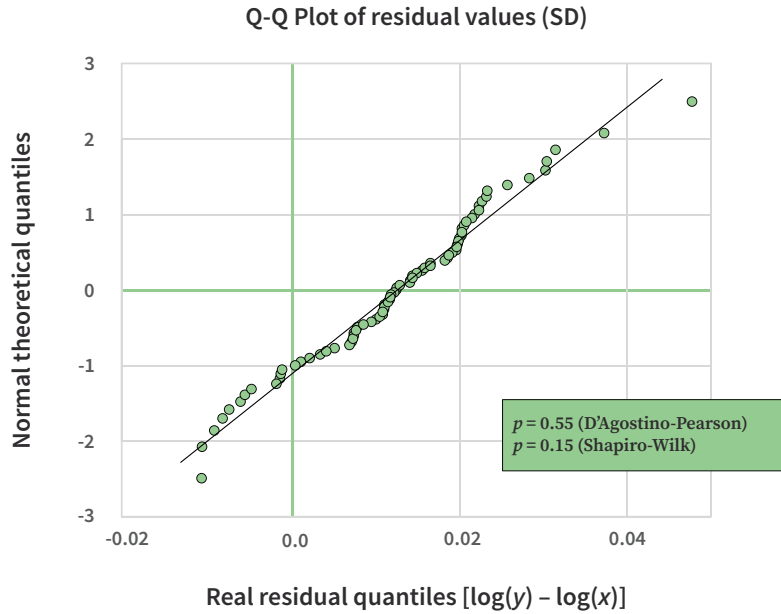


Figure 4.2.3 A comparison of the residual distribution with a theoretical normal distribution | A Q-Q plot is used to characterise the distribution of a set of data by comparing it to another (real or hypothetical) distribution, most often a normal distribution. In the present example, the set of data is a list of residuals (in log form) between the predicted variation and the actual variation observed in simulated Bernoulli processes. A straight on this plot line supports to the notion that the data is normally distributed, which would indicate that disagreements between the predictor and the observation were random (albeit not whether predictor and the observation have the same central tendency).

SDr_s and ζ diverged as a normal distribution became decreasingly valid approximate for the binomial distribution. Residual outliers (Tukey fence; $k = 1.5$) were identified exclusively from among the Poisson-associated range: parameters being $(c, p) = (1K, 0.99), (1K, 0.98), (1K, 0.90), (2K, 0.98)$. For $c = 1K$ and $2K$, p values below 0.05 and above 0.95, the tendency to diverge was well explained by the limits of the normal approximation under these parameters, which justified their exclusion from residual plots, regression analysis and tests for normality. In the normally-approximated range, the residual distribution still reflected a degree of x -dependence, where negative bias increased as the ζ approached ~ 0.01 and decreased as x moved away in either direction. Notably, this was observed for both ζ and the conventional model. The residual distribution was not distinguishable from normally distributed data by either D'Agostino-Pearson or Shapiro-Wilkes tests (Fig. 4.2.3).

4.2.2 Discussion of ζ and r_s as expressions of the binomial distribution

This experiment was successful at falsifying Hypothesis 1, since differences between ζ and the observed SD of r_s were not symmetrically distributed about zero. Fortunately, this appears not to have precluded ζ as a useful approximant, since discrepancies were limited to $\leq 10\%$ and

averaged ~3.5%, and comparable to that of the canonical model. Whether this resulted from idiosyncrasies of binomial probability theory or artefacts of the simulation was not determined.

The initial hypothesis for Section 4.3 was that the SD of empirical qRBA data would fit to $y = x$ regression model when plotted against ζ , and that the residuals would be normally and symmetrically distributed about zero. The fact that this was not the case for *in silico* Bernoulli processes means that the exact hypothesised residual distribution is less clear. Though not unambiguous, ζ (or a number very closely approximated by ζ) appears to be the central component of the SD of Bernoulli processes, with respect r_s . With additional controls in place, ζ will serve adequately to explore the validity of the binomial distribution model to NGS-based qRBA analysis.

4.3 Empirically testing the binomial model of Illumina-derived qRBA measurement

A qRBA sample mix containing ten plasmid reporter variants was subject to replicated measurements using Illumina MiSeq (Fig. 4.3.1) to test whether $p_s : q_s$ varied according to the binomial model proposed in chapter three. The ten plasmids were first mixed in pairs at in a roughly 1:1 ratio. These five mixes were then added to a final qRBA sample in relative quantities ranging from 70% to below 1%. Twenty sets of indexed PCR products were produced from this sample, which were added to the final library in relative quantities ranging roughly from 1% to 15%. Only the *egfp* reporter target (See Fig. 4.1.2) was analysed, since testing accuracy was not the objective and would have detracted from the range of readcounts testable.

Measurement replication was initiated from two stages of the Illumina qRBA process: PCR amplification of the reporter target ($4 \times n_{\text{pcr}} = 5$), and analysis of the library with Illumina MiSeq amplicon population sequencing ($n_{\text{ngs}} = 5$). The twenty sets of PCR products were grouped into four ‘tranches’, consisting of five indexed PCR replicates mixed equivolumetrically. Each tranche was added to the final library in variable quantities, establishing four grouped tiers of total readcounts. This divided pentaplicate approach was implemented in order to test a wider range of c than was possible with an $n_{\text{pcr}} = 20$ design. 1 μg ea. of this qRBA library was added to five microfuge tubes and each tube was submitted as an independent sample for 2×150bp paired-end Illumina MiSeq analysis ($n_{\text{ngs}} = 5$). In this manner, each uniquely indexed PCR

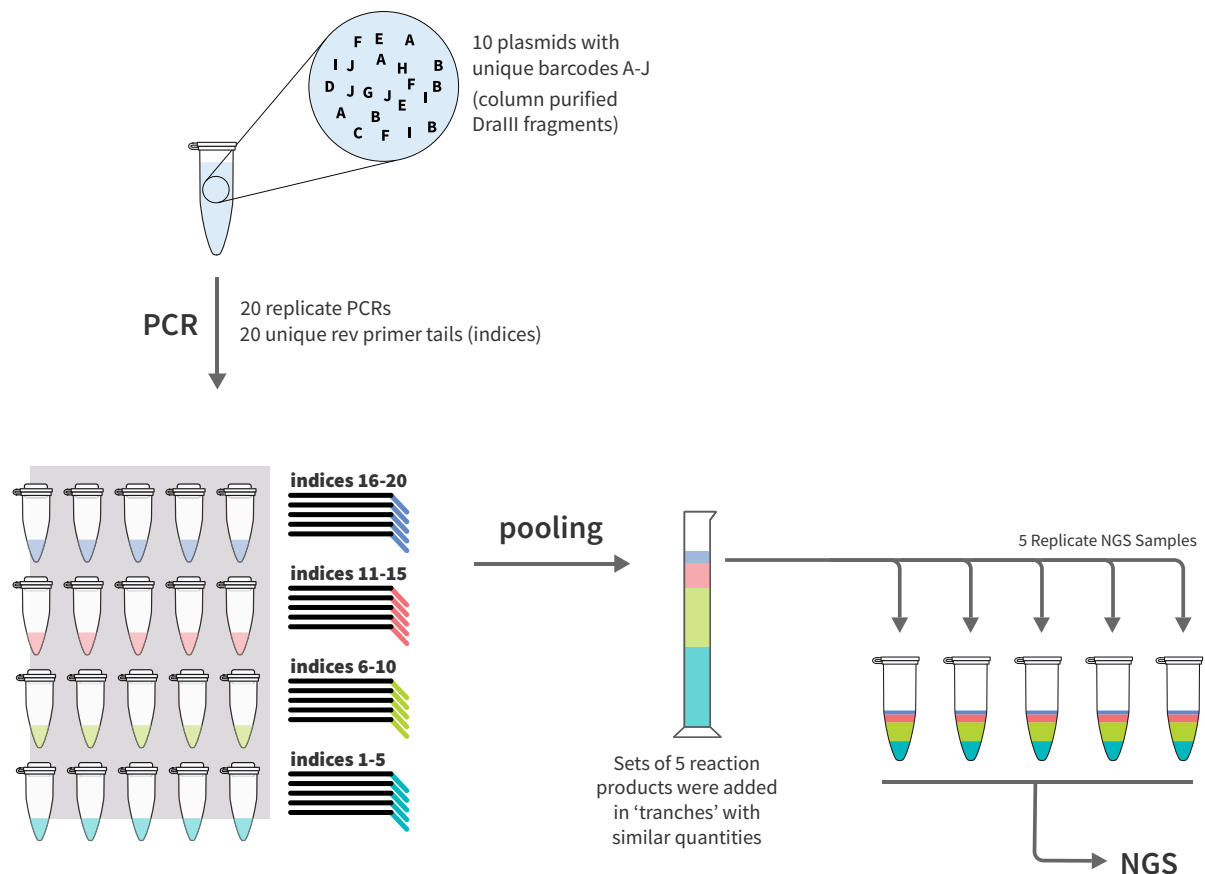


Figure 4.3.1 Generation of replicated qRBA data from a mix of 10 barcoded plasmid variants | A single mix of barcoded plasmid variants was PCR amplified twenty times, each with a unique NGS index sequence incorporated in the reverse primer. This set of twenty was divided into four 'tranches' (represented with a shared colour), each of which contained five samples. Samples were combined into their tranche equivolumetrically from the completed PCR reaction mixes and each tranche was added to the final NGS sample (the NGS 'library') in a distinct quantity so as to diversify the range of readcounts in the experiment. The mixed PCR products were then column purified and the library submitted for five independent Illumina sequencing runs using the MiSeq 2× 150bp paired-end sequencing mode.

sample would each be analysed five times. Upon receipt of the NGS read data, the reporter ratio was taken from every possible pair combination (excluding reciprocals and self-pairs) of the ten barcodes in each indexed PCR sample. For ten barcodes, this was 45 unique pairs. Each datum contributed to two different standard deviations – one taken from the tranche of five PCR samples, and one taken from the same PCR sample in each of five NGS replicates. These two data groupings will be referred to as the n_{pcr} group and the n_{ngs} group. One of the major strengths of this experiment was that the sample under study was immune to cross contamination; it could not 'self-contaminate', either in our laboratory or that of the NGS service provider. The exact proportions of each reporter variant in the PCR template mix were not critical in the experimental design, and any plasmid that may arrived in the PCR template mix unintentionally was a valid member of the population. The use of NGS index sequences

meant that the PCR products themselves were not cross-contaminable since the index would identify the PCR reaction of origin for all successfully sequenced molecules.

To mitigate key limitations of this experiment, more *in silico* data was prepared (WB5.3.1) with parameters equivalent to those encountered in the empirical data, including the number of independent Bernoulli processes ($n = 5$). Economic and practical limitations constrained the number of technical replicates below the number required to directly characterise the distribution of random error, which would be closer to twenty-five. With n as low as five, the degrees of freedom are fewer and conformity of data to the model less predictable. The *in silico* version of the experiment would therefore act to inform us how ideal binomial data should behave under the same constraints. It could also potentially assist by exhibiting any idiosyncrasies of the binomial distribution not obvious from the introductory model.

Another limitation was that unlike the simulated experiment, it was not feasible to fully predetermine the parameters p and c . Where p is concerned, we are limited to the parameter inferred from five averaged p_s values. As mentioned, the likelihood of an unknown p equalling a known p_s has not been incorporated into my model, as I judged myself unreliable without expert assistance. While a very close approximation might be inferred from the totality of reads, this presupposes the validity of the model being evaluated and would still not constitute a 'known' p . With simulated data in hand, we have a means of analysing variability where p is an estimate. **For this I will introduce the additional term ζ_s (zeta-s).** Note that in this report, ζ_s signifies the average estimate of ζ made from individual (p_s, q_s, c) results. It is not the ζ inferred from the averaged values of p_s, q_s and c .

$$\zeta = \left(\frac{p}{q}\right) \frac{1}{\sqrt{cpq}} \quad \zeta_s = \frac{\sum_{i=1}^n \left[\left(\frac{p_s}{q_s}\right) \times \frac{1}{\sqrt{c(p_s q_s)}} \right]_i}{n}$$

The question of c is potentially more complex; while the parameter c is perfectly discoverable upon delivery, it would not be consistent between replicates.

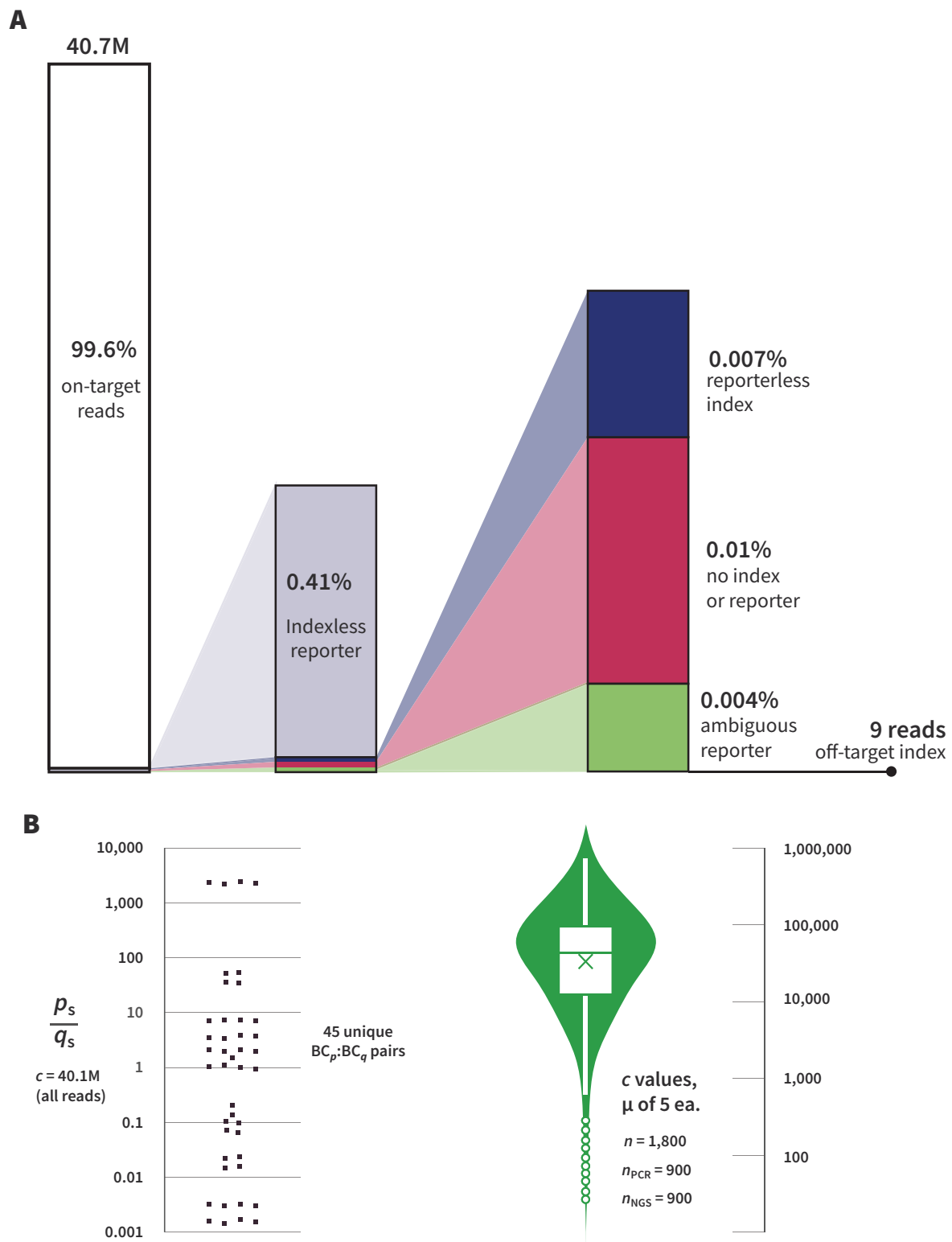


Figure 4.3.2 Summary of qRBA read data | (A) Of the ~41 million raw data returned from the Illumina sequence provider, 99.6% possessed the correct length and forward primer sequence, and contained one of the 10 prescribed barcodes 24 prescribed indices in the proper base positions. Of the 0.43% unused reads, 0.41% were reads with valid barcodes but no valid index sequence. Only nine reads in total contained one of the four index sequences that existed in our laboratory, but were not used in the experiment. **(B)** Among the grand total of reads generated, ratios between barcodes exhibited a dynamic range from just above 0.001:1 to just over 1000:1, though there were no ratios in the 100-1000:1 range. The c parameter of a Bernoulli process signifies how many reads were counted that resulted in the measured barcode ratio.

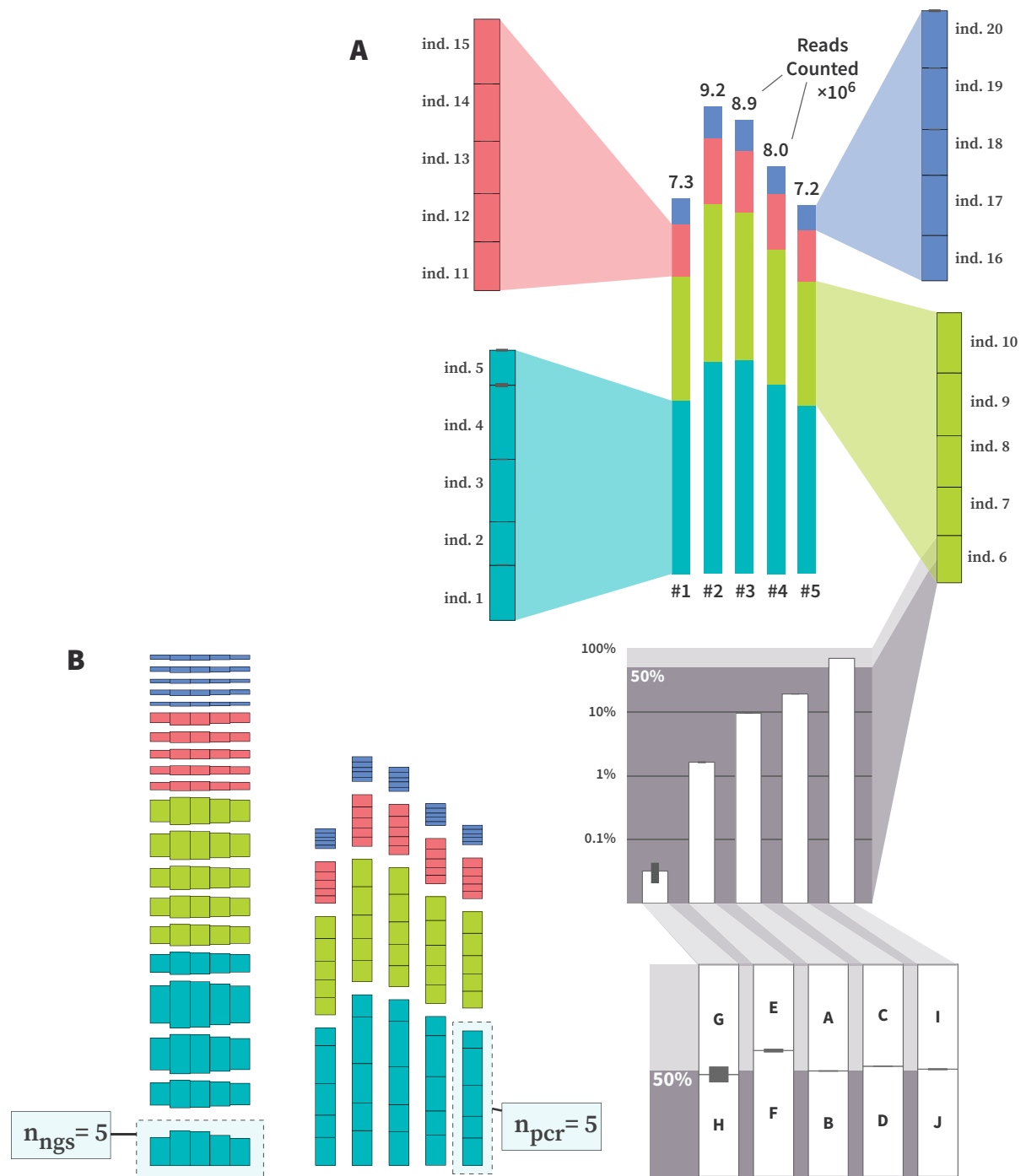


Figure 4.3.3 Summary of indexed and barcoded read counts | (A) The five independent NGS runs returned readcounts between roughly 7 and 9 million. Coloured blocks represent four 'tranches' of PCR replicates, which each contained five uniquely indexed PCR products of the same template mix. Total tranche size in a given readset varied intentionally in order to create a diversity of readcounts for the qRBA measurements. The coloured blocks projecting from the five stacked columns indicate the percentage abundance of each indexed PCR sample. Error bars are included to show the sample standard deviation, though these may not be visualisable in print. The ten reporter barcodes were incorporated into the mix as five sets of two with approximately 1:1 ratios, to ensure that the parameter $p = 50\%$ could be evaluated over a wide range of readcounts. The percentage abundance of each reporter variant was averaged from all 100 indexed PCR samples. **(B)** The standard deviations analysed in this experiment were the results of five replicated qRBA measurements. Each individual qRBA result was part of two pentaplicate groups: five PCR samples in a tranche and five independent NGS runs of the same indexed sample. These two groupings are referred to as the n_{ngs} group and the n_{pcr} group. The 1,800 SD values analysed comprised 45 unique barcode pairings with twenty sets of five replicates and two modes of replication ($45 \times 20 \times 2$).

4.3.1 Breakdown of count data

Between five replicated NGS runs and 20 indexed PCR reactions, 1000 individual barcode counts were obtained. With a complete matrix of possible reporter pairs (excluding reciprocals and self-pairing), a total of 4500 individual qRBA measurements were made, amounting to 900 PCR replicates ($n_{\text{pcr}} = 5$) and 900 NGS replicates ($n_{\text{ngs}} = 5$). The two following figures (Figs. 4.3.2 and 4.3.3) are summaries of these data.

Of the total ~41 million reads returned, 99.6% were useable (Fig. 4.3.2a). The majority of unused reads lacked an identifiable index (0.004%). The counting script searched for all 24 indices in the batch of reverse primers ordered for the project, of which only Index1 to Index20 were utilised in this study. Nine reads were attributed to unused indices (Index21 to Index24) among the grand total of ~41 million reads. Taken from the grand total, the five reporter couples amounted to approximately 70%, 20%, 10%, 2% and 0.03% of all reads. Within coupled barcodes, the first reporter ranged from 40.4% to 51.7% (Fig. 4.3.3). The equivalent $p_s:q_s$ ratios exhibited a dynamic range from over 1000.0 to just above 0.001, albeit with a paucity of data between 100.0 and 1000.0 (Fig. 4.3.2b).

1664 sets of pentaplicated Bernoulli processes in total were simulated. These comprised 416 unique c, p parameter combinations (listed in Table 4.3.1) simulated in pentaplicate, four times. That is, 416 unique parameters ($n = 5$ ea.) $\times$ four repeats = 1664 values of ζ_s and SD to analyse as (x, y) coordinates respectively.

Table 4.3.1 list of parameters applied in additional *in silico* simulated Bernoulli processes

<i>r</i>		<i>c</i>				<i>n</i>
0.01		50	1K	10K	45K	5
	1.50	100	2K	12K	50K	
0.02		200	3K	15K	70K	
	4.00	300	4K	20K	100K	
0.05		400	5K	25K	120K	
	9.00	500	6K	30K	150K	
0.11		800	7K	35K	200K	
	19.00	900	8K	40K	250K	
0.25						
	49.00					
0.67						
	99.00					
1.00						

4.3.2 Analysis of SDRs and its relationship with ζ_s

All 1800 ζ_s values were set on the x -axis and the observed sample standard deviation for the same five replicates was set on the y -axis. These plots are shown in Figure 4.3.4a at the linear and log scales. At a glance, a linear relationship between y and x is depicted at the log scale, approaching the locus with a gradient near to one.

Within some small ranges, values of ζ_s occurred that uncharacteristically underestimated $SDp_s:q_s$ compared to the rest of the data. Ad-hoc tests (results not shown) could not identify a specific tranche, barcode, index or mode of replication that correlated consistently with this divergence. Residual analysis revealed that this tendency was strongly related to the range (max – min) of associated readcounts, which appeared to take effect at a difference of 50,000 reads (Fig. 4.3.4b). A Q-Q plot corroborated the impression that data in this range diverged from the linear model in a distinct mode (evidenced by the two apparent gradients – Fig. 4.3.4c). This effect was further visualised by grouping the data into deciles and isolating the upper three (Fig 4.3.5). Given that variability in the c parameter is an undesired limitation of this particular experiment, this effect justified excluding the upper two deciles from further analysis. Out of conservatism, the upper third decile was retained. 201 n_{ngs} samples and 151 n_{pcr} samples were excluded. This was seen as a less biased method for excluding nuisance data than a cut-off based on directly on residual values.

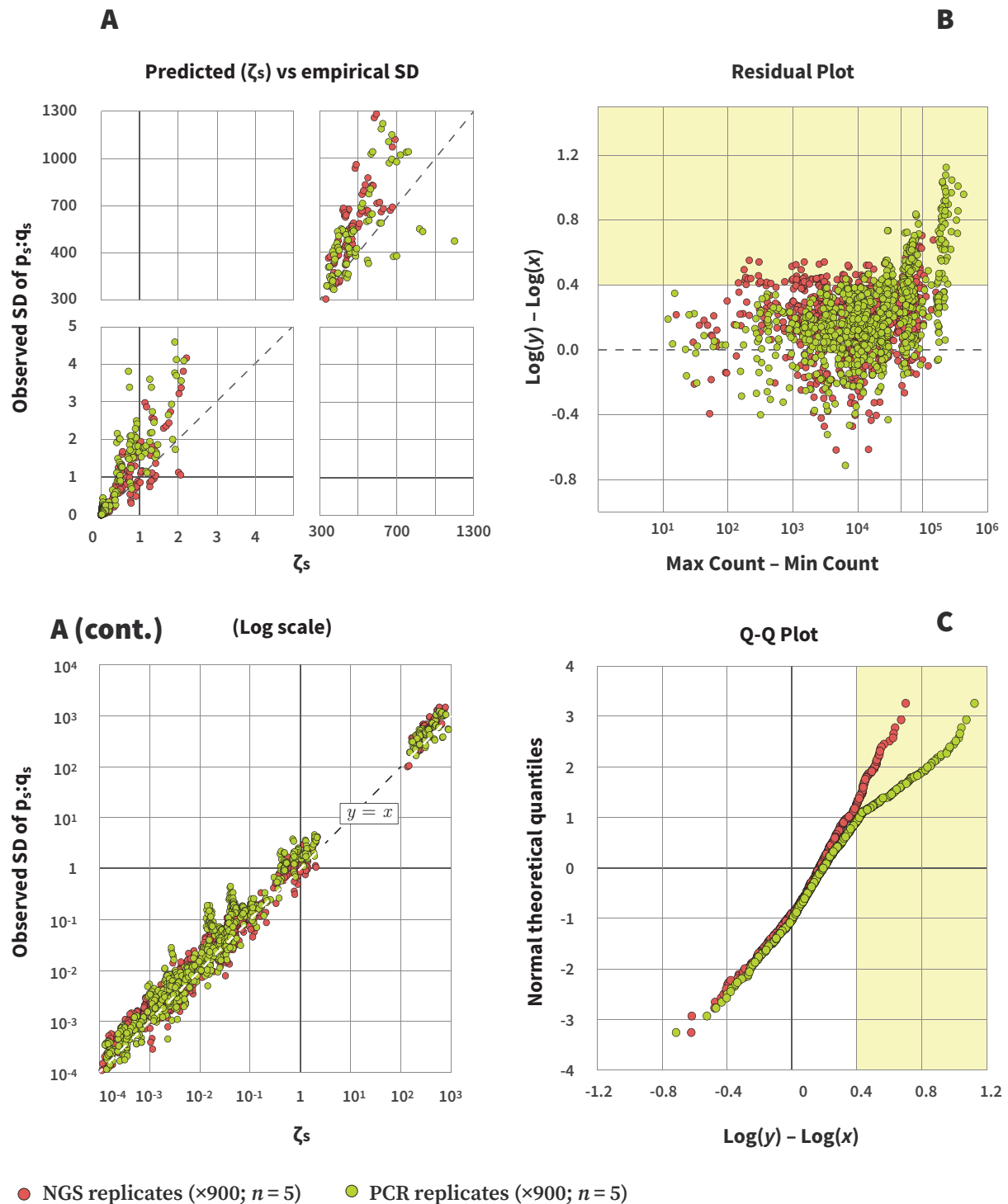


Figure 4.3.4 Relationship between observed standard deviations and the predictor ζ_s | (A) The empirically observed and predicted standard deviations were plotted as x,y coordinates respectively in order to assess the fit of the data to the model $y=x$. As with the simulated data in Section 5.1, the model $y=x$ was a pleasing visual approximate of the data aside from an increase in negative bias observable at the higher ζ_s values. Isolated instances are observed where SD was uncharacteristically underestimated. **(B)** The tendency to underestimate empirical SD increased with the absolute range of underlying readcounts. This appeared to take effect more strongly from a difference of roughly 50,000 reads. **(C)** A Q-Q plot supported the view that this tendency was divergent from the bulk of the data. A yellow field indicates the approximate residual range where a change in gradient can be observed (note that this was not the 'cut-off' range for exclusion of data).

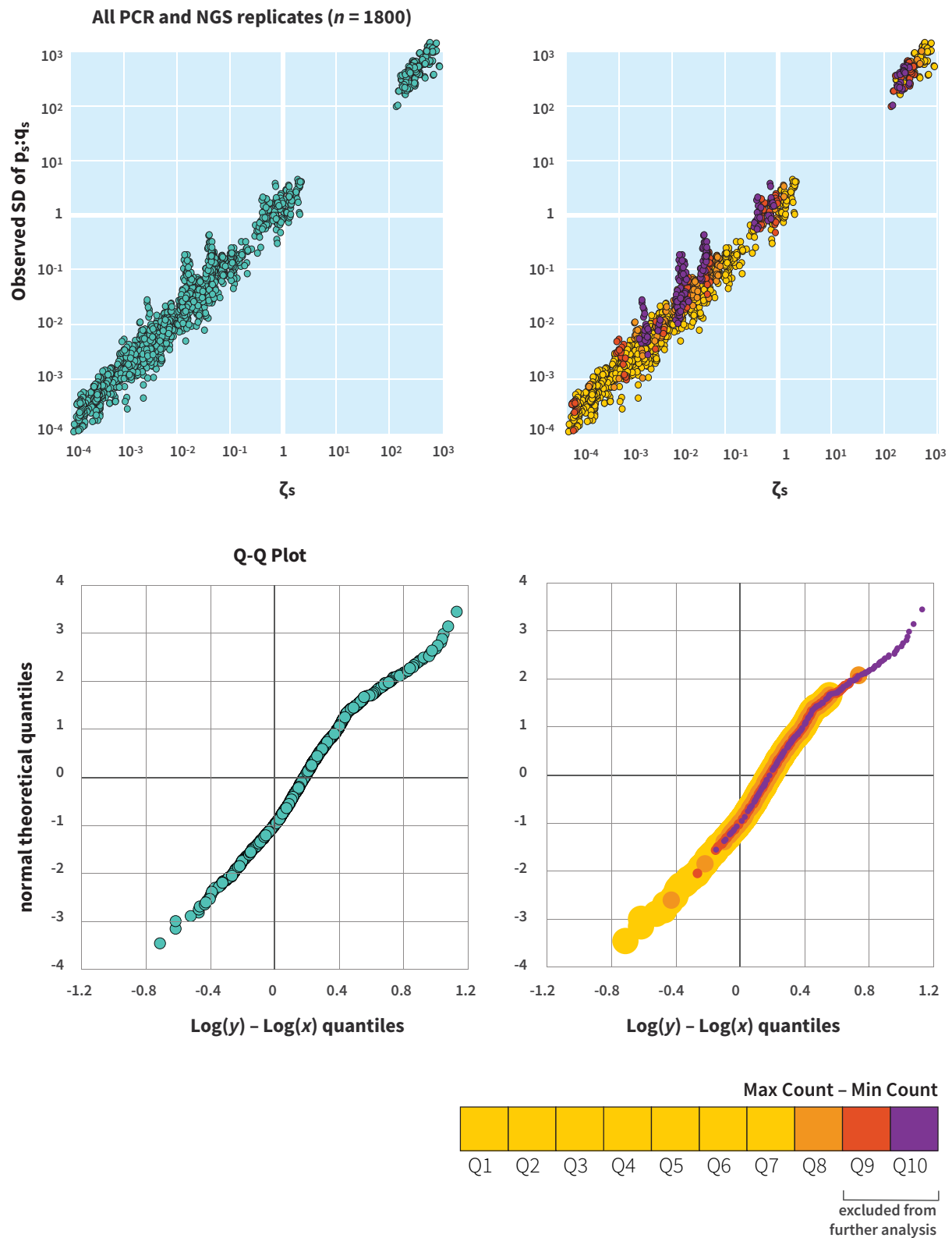


Figure 4.3.5 Larger differences in readcount affected the validity of ζ_s | Where ζ is used to predicted the SD of replicated Bernoulli processes, it is assumed that the replicates share the same readcount. This was not possible in the empirical setting, meaning inconsistent readcounts were a confounding factor. In Figure 5.3.4, fold residuals were seen to increase with the range between maximum and minimum readcounts. This relationship was further examined by organising residuals into deciles and highlighting data belonging to the eighth, ninth, and tenth deciles. This revealed that a particularly high negative bias seen in some data was a characteristic of the upper deciles. Excluding the upper two deciles from further analysis was justified, since the reason for this divergent behaviour was attributable to a known confounding factor.

Figures 4.3.6a-e are distributed across two pages and together constitute a comparison of *in silico* and empirical Bernoulli processes and their relationship with ζ_s . The empirical data were plotted on a log-log chart as in Figure 4.3.4, alongside data obtained from *in silico* Bernoulli trials (Fig. 4.3.6a). As observed in Section 4.2, random divergence from $y = x$ was heteroscedastic and increased in absolute terms as the predicted SD increased. A least-squares regression of the log-log data (Table 4.3.2) resulted in exponents ranging from 0.99 to 1.00, indicating that the relationship between predicted and observed variability was approximately linear. Power relationships inferred from these regressions would indicate that ζ_s tended to underestimate SDr_s by approximately 35% for empirical replicates and overestimate SDr_s by 15% for simulated replicates.

Table 4.3.2 Least-squares regression models for predicted vs empirical SDs

	n_{ngs}	n_{pcr}	Sim. ζ_s	Sim. ζ
log-log regression	$y = 0.99x + 0.12$	$y = 1.00x + 0.13$	$y = 0.99x - 0.07$	$y = 0.99x - 0.07$
R²	0.99	0.99	0.98	0.98
S.E.Y.	19%	19%	17%	18%
power relationship	$y = 1.32x^{0.99}$	$y = 1.36x^{1.00}$	$y = 0.85x^{0.99}$	$y = 0.86x^{0.99}$

Table 4.3.3 Tests for normality in log residual distributions ($\log y - \log x$)

		Effect size (all data)		Outliers excluded ($k=1.5$)		
y axis	(D'AP & SW) $p =$	D'AP	SW	D'AP	SW	Population - outliers
n_{ngs}	$<10^{-5}$	0.265	0.064	0.037	0.08	97.4%
n_{pcr}	$<10^{-5}$	0.267	0.049	0.028	0.12	97.8%
Sim. ζ_s	$<10^{-5}$	0.340	0.075	0.049	0.15	97.2%
Sim. ζ	$<10^{-5}$	0.299	0.067	0.047	0.15	97.1%

D'AP = D'Agostino-Pearson test for normality; SW=Shapiro-Wilks test for normality; k =Tukey fence size; Q=quartile; IQR = interquartile range | Tests performed using the online resource *StatsKingdom* <<https://www.statskingdom.com/>>

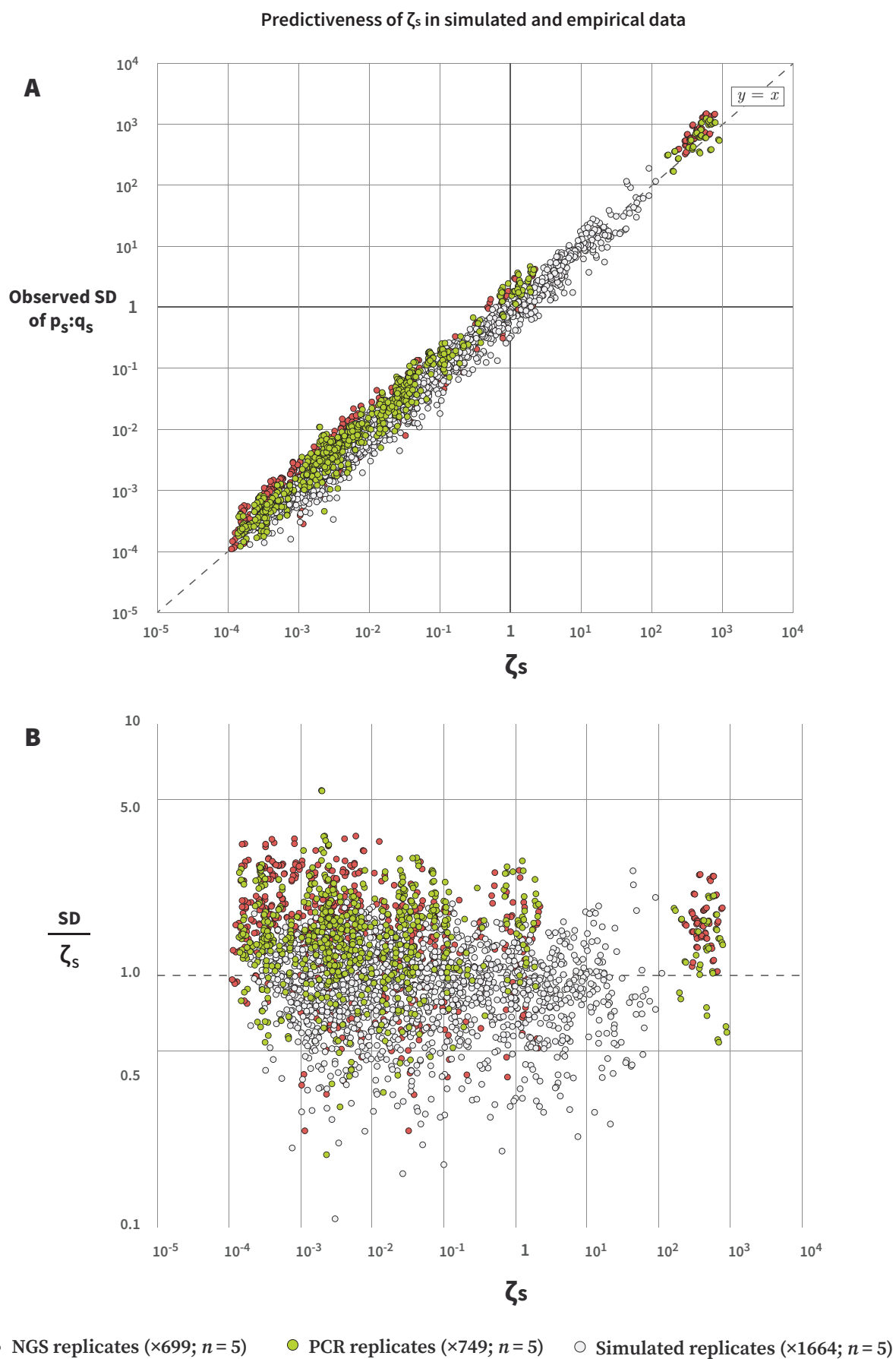


Figure 4.3.6 (1 of 2) Comparing the predictive value of ζ_s for empirical data with that of simulated data

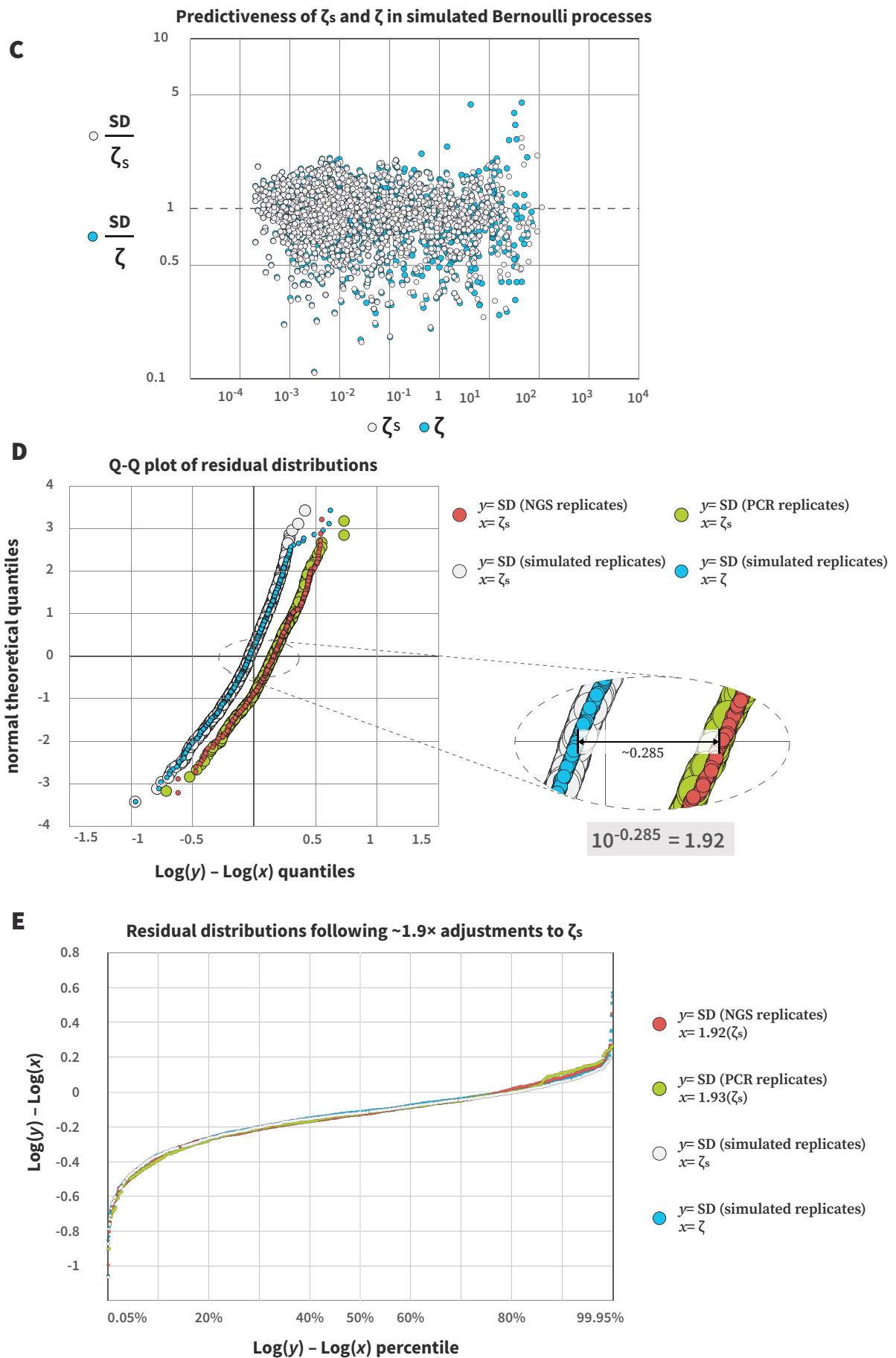


Figure 4.3.6 (2 of 2) Comparing the predictive value of ζ_s for empirical data with that of simulated data

Figure 4.3.6 (cont.) Comparing the predictive value of ζ_s for empirical data with that of simulated data (A) Having excluded data with unacceptable differences in readcount, NGS and PCR replicates were plotted as in Figure 5.3.4 alongside those of *in silico* simulated Bernoulli processes. (B) the fold-discrepancy between SD and ζ_s was compared to ζ_s , revealing some degree of x-dependence, as was the case in Section 5.2. We may note for residual analysis that $\log(y)-\log(x)$ is equal to $\log(y/x)$. The fold discrepancy is therefore a valid representation, and makes it easier to intuit the level of disagreement between the predicted and observed outcomes. From (A) and (B) we can observe that the data share similar profiles, though ζ_s had less of a negative bias in the *in silico* setting than the empirical setting. (C) For this experiment, the SD of simulated data was predicted from inferred values of p , despite the real value of p being known. This was to reflect the same limitation in the empirical data. For interest, the SD calculated from the known parameter (blue) was compared to the observed SD, and its predictive value was compared to ζ_s . (D) The residual distribution of each of these datasets was shown to be approximately normal, albeit with some small non-normal component. A coefficient was given to ζ_s that would translate the q-q plot in the negative direction, such that the NGS replicates, PCR replicates and simulated replicates would share an x-intercept approximately. (E) When $1.98\zeta_s$ (n_{ngs}) or $1.99\zeta_s$ (n_{pcr}) were used as predictors of variation, the residual distribution was indistinguishable from that of simulated data ($p=0.99$ and $p=0.99$). To visualise this similarity, the percentile of each residual and its actual value were plotted as x,y coordinates for each of the four groups, including simulated data where known ζ was used as a predictor. At the risk of limiting visibility, very small data markers were plotted in order to convey that variation in one group was not being visually occluded by another.

This tendency to over/underestimate was reflected in the asymmetrical residual distributions. These are shown in Figure (Fig. 4.3.6b), which compares ζ_s with the fold difference between ζ_s and observed SD. Since the data was available, an equivalent residual plot was constructed for simulated data alone, to compare the (x, y) relationship when based on either inferred p or known p (Fig. 4.3.6c). To assess normality of residuals in these datasets, Q-Q plots (Fig. 4.3.6d) were constructed and analysed in conjunction with D'Agostino-Pearson and Shapiro-Wilk tests for normality (Table 4.3.3). Each of these tests returned p -values lower than 10^{-5} , indicating strongly that the log residuals were not normally distributed. Although highly confident in their non-normality, the magnitude of difference between their distributions and a normal distribution was minor, with the effect size suggested to be 'small' by the statistical resource used or 'very small' by when data outside $1.5\times$ the interquartile range were excluded, which represented $<3\%$ of data in all cases. This result permitted the use of Student's T-test to explore the distinction between each population of residuals. The log residuals of NGS and PCR replicates could not be distinguished by Student's T-test ($p = 0.82$) and neither could those of simulated data when ζ was predicted from known parameters or inferred from the data themselves ($p = 0.33$).

The contours of the Q-Q plots for empirical and simulated data appear extremely similar, indicating that although the 'midpoint' for residuals differed, the distribution of data across quantiles set was closely comparable. To test this formally, the horizontal distance (at $y = 0$)

between the (inferred) simulated and empirical datasets was used to estimate a coefficient for ζ_s that would align the groups, which were then compared using Student's two-tailed T-test with equal variances assumed. When $1.92\zeta_s$ and $1.93\zeta_s$ were used as predictors of PCR and NGS replicates respectively, residuals were indistinguishable from simulated data ($p = 0.99$ and 0.99 respectively). Altering either these coefficients by 0.01 resulted in p values of 0.80 and 0.79 respectively. To illustrate the similarity, residuals from all four of these groups were arranged into their respective percentiles and plotted on a shared y -axis (Fig. 4.3.6e).

4.3.3 Discussion of qRBA precision

The results obtained indicate strongly that qRBA measurement is governed by a binomial probability model and encourage its further refinement. As in section 4.2, ζ_s was an incomplete predictor of measurement SD, given that the residuals were asymmetrically distributed and showed some degree of x -dependence. However, as before, ζ_s (or a quantity very closely approximated by ζ_s) was a powerful predictive component of variability in binomial measurement. Further, the applicability of the Binomial distribution was not judged solely by the predictive value of ζ – such a result would have been limited by an inadequate number of technical replicates and the need to infer the parameter p from the data itself. Much credibility derives from the *in silico* simulated Bernoulli processes, whose behaviour under the same limitations was strongly recapitulated by the empirical qRBA measurements.

The 1,645 pairs of predicted and observed SDs returned residuals whose distribution aligned exceedingly well with the ~1500 pairs obtained empirically. This meant that the difference between the empirical result and the ideal 'mistake-free' process can be described in very specific terms: when a qRBA sample was PCR-amplified five times and measured once with Illumina MiSeq, the (log) differences between their standard deviations and ζ_s were 1.92-fold higher than those attributable to 'pure chance'. For a sample PCR-amplified once and measured five times with Illumina MiSeq, this factor was 1.93. When $1.92 \times \zeta_s$ and $1.93 \times \zeta_s$ respectively were used to predict the standard deviation of five replicates, differences between predicted and observed values were statistically indistinguishable from those of *in silico* Bernoulli processes predicted by $1 \times \zeta_s$ ($p = 0.99$ and $p = 0.99$ respectively).

In this experiment, the contribution of PCR to random measurement error was negligible. If random error had been increased by PCR amplification, this would have resulted in higher variability within a tranche of PCR replicates than between NGS replicates of a single PCR

sample. Instead, the two paths of technical replication could not be distinguished ($p = 0.82$). This would indicate that random error in the experiment was attributable mainly if not wholly to MiSeq massively parallel sequencing. This itself encompasses numerous processes including adapter ligation, hybridisation to the flow cell, clonal bridge amplification and sequencing-by-synthesis. As with any analytical process, these steps are expected to increase random error, though it is not clear why they increased variance measurably whereas the PCR process did not. Some of this increase may be an artefact of the variable readcounts, which were not a feature of the simulated data. However, care should be exercised in assigning blame to unequal readcounts, as it was possible to identify and exclude data where this had clearly influenced the regression. It is possible that nonequal readcounts had minimal impact because c factors into $(cpq)^{-0.5}$ as $c^{-0.5}$, and any range on the number line will be shortened when given the exponentiated by -0.5 .

Another noteworthy observation was that residuals were indistinguishable in the simulated Bernoulli processes whether ζ was calculated using known parameters or from parameters estimated from the data itself. As stated in Chapter 3, I am at risk of wasting time by analysing this observation too deeply without assistance from researchers familiar with Bayesian statistics. However, I do draw encouragement from this finding and the potential of $p_s:q_s$ to act as a minimally biased estimator of $p:q$, which is a critical aspect of this measurement model.

These results vindicated the complete binarisation approach to NGS count data. Though the readset on the whole was multinomial, individual pairs behaved according to a binomial distribution when analysed in isolation. From one perspective this could be seen as a trivial or predictable consequence of the model. On the other hand, there is often an inertia to established practises and a perceived risk in adopting alternatives, meaning that at times even predictable validations can have outsized demonstrative value. This is aside from the need to verify our assumptions for the sake of scientific rigour. The principle of binarisation may be greatly underutilised in more complex readsets. Many transcriptomic targets represent a small and highly variable proportion of reads when measured against the readset as a ‘total’. These will however be close enough in proportion to other less abundant targets to be measured and analysed with far higher precision according to binomial model. This was illustrated by the G and H reporter pair, which averaged one per 5,000 reads in each total sample (median 56 and 49 reads per sample), but whose pair ratio was $1.10 \pm 26\%$ across all 100 samples and whose variability conformed well to the patterns observed in the higher-count populations.

The data encourages redoubled efforts to affirm these findings with a practicably higher number of technical replicates, and more importantly, independent experiments to canvas which aspects of the results are generalisable across different personnel, NGS service providers and reporter cassettes. As this Chapter will soon address, it is critically important to validate the accuracy of qRBA measurements, which this particular experiment was not designed to test. One concern this experiment strategically avoided was the effect of so-called ‘Index hopping’, which is when nonzero cross-contamination between index sequences leads to the misattribution of reads to the wrong sample. Here, this would have had no effect because each index sampled the same population. It was encouraging to note that only nine reads belonging to the 4 (of 24) unused indices were detected among the 41 million reads returned. While unimpeachable manual technique is my preferred explanation, it is also feasible that complete complementarity with the primer binding site is a competitive advantage in successive rounds of PCR amplification, which would act to enrich the majority index.

There is more that can be analysed in the existing dataset, though unfortunately not within the time constraints of this doctorate. A more formal analysis of nonequal c and its impact on the regression would add clarity to the results. Another approach may be to model Bernoulli processes *in silico* with variable count sizes. Grouping of multiple indexed populations would expand the range of ζ_s values analysed. Examining whether ζ_s and ζ are less similar given fewer technical replicates would be required before suggesting that ζ can be inferred from fewer than five samples.

In order to better define the precision capabilities of qRBA, refinements to ζ are required that can achieve a fit to observed variance with residuals normally and symmetrically distributed above and below zero. In this experiment, as in 4.2, ζ displayed neither of these attributes (albeit the effect of the non-normal component was small). This is critical because the precision of qRBA measurement can only be evaluated with respect to parameters c and p . The most precise 10% of empirical measurements had RSDs below 1.00%. In a sense this capability is worthwhile showcasing, but also relevant is the fact that all these were supported by p values between 0.1 and 0.9, which are proportions favourable for keeping $(cpq)^{-0.5}$ low. The dynamic range of qRBA is determined directly by expectable readcounts.

This experiment has contributed greatly to the credibility of qRBA in bioanalysis by validating the logical basis of the measurement process. It has demonstrated that qRBA data behave highly

similarly to *in silico* simulated Bernoulli trials, strongly supporting the binomial model. By verifying this fundamental probability structure, the impact of this study may well extend to other areas of NGS count analysis.

4.4 Systematic error in qRBA measurement

4.4.1 Introduction

A measurement's precision and its accuracy are two separate and highly distinct questions. Precision, however impressive, conveys no direct information regarding accuracy. Precision is conceptually simple to discover; perform valid replicates of the measurement process and observe their variability. To demonstrate that replicated measurements are accurate, we need evidence that the 'true' physical value is their probabilistic centre of gravity. Categorical proofs are evasive. For Illumina qRBA to be properly quantitative, we are relying on a minor, localised sequence variation within the reporter amplicon sequence (6N or 4% in the present case) not to act as a source of biased amplification, read production, or read detection.

This section will test the hypothesis that qRBA measurements are accurate. More formally, it will test the hypothesis that the differences in the plasmid ratio measured from two qRBA targets within the plasmid will be randomly and symmetrically distributed about zero. The known 1:1 stoichiometry of conjugate reporter targets makes this hypothesis falsifiable, because tendencies other than $y = x$ can only be attributed to measurement bias. For clarity – the measurand in this experiment is the ratio between barcoded plasmids, not the ratio between one reporter and its upstream/downstream conjugate. The hypothesis is *not* that measurements will centre on 1.0, but that the $p_s : q_s$ ratio measured from the *egfp*-adjacent target and *luc2*-adjacent target will match.

The ten reporter plasmids constructed in Section 4.1 were combined haphazardly into six separate mixtures (*Mix1-6*). Each Mix was used to produce three uniquely indexed sets of PCR products, creating a total of eighteen indexed samples in the NGS library. Rather than analysing each mix in triplicate, each of the three PCR samples were added to the NGS library in variable quantities so that the effects of read count could be interrogated. All samples were prepared for Illumina MiSeq analysis using the same qRBA assay methodology as in Section 4.3. In this section, the library contained 36 sets of indexed plasmid PCR products: 18 amplified from the

egfp-adjacent reporter region and 18 equivalent samples amplified from the *luc2*-adjacent reporter region (Figure 4.4.1). As with Section 4.3, the final library was submitted for five independent rounds of 2×150bp paired-end Illumina MiSeq analysis. The ratio r was taken for every possible barcode pair, excluding reciprocals and self-pairs. As explained in 4.1, two plasmid variants (H and I) had unique barcodes within the *egfp*-adjacent PCR target sequence, but an identical barcode adjacent to the *luc2* sequence. For plasmids H and I, the two *egfp* barcode counts were added and treated as the conjugate result to the corresponding *luc2* barcode count.

The results of this section revealed a strongly systematic pattern of misalignment between conjugate qRBA measurements. In response to this outcome, a series of *ad hoc* hypotheses were explored in addition to those originally planned. Through these analyses I was able to unambiguously exclude the possibility that sequencing infidelity, index-hopping or probabilistic skew as sources of this error. Modelling $r_{[egfp]}$ and $r_{[luc2]}$ as composite power functions of r enabled me to clarify the thought process surrounding reciprocals and p -assignments. Using this model I was also able to highlight subtle but strongly Mix-dependent patterns of error, and show that this error did not directly resemble PCR amplification bias. While they did not provide a catharsis, these analyses constitute a useful body of work that will constrain possible explanations for the error, inform appropriate next steps, and assist in evaluating data produced in Chapter 6.

qRBA measurements of r will be denoted according to which reporter target was assayed: $r_{[luc2]}$ or $r_{[egfp]}$. $r_{[egfp]}$ will be plotted on the x -axis by default and $r_{[luc2]}$ on the y -axis. However, it is noted that this is not ultimately interpreted as a ‘predicted vs observed’ relationship.

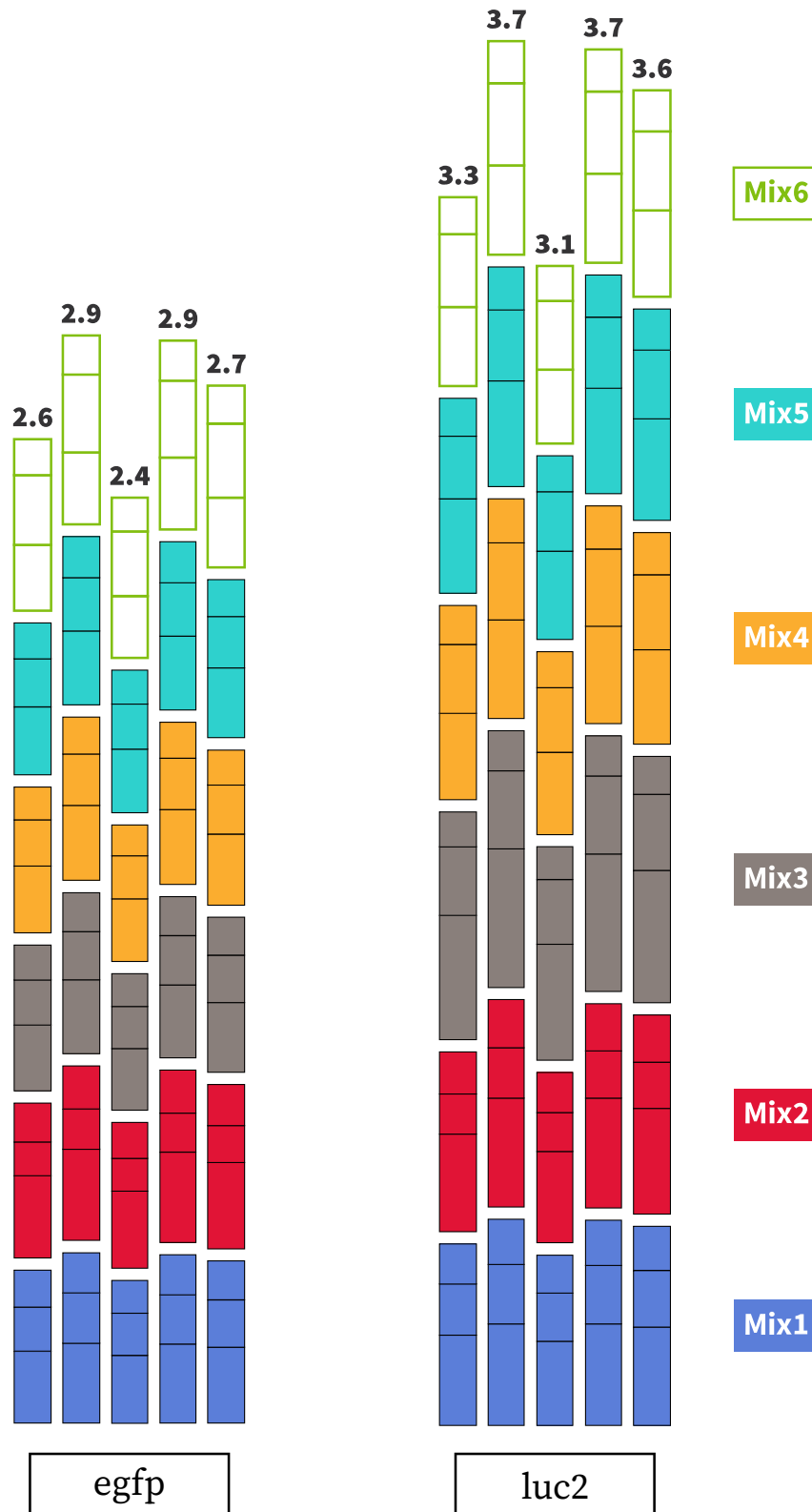


Figure 4.4.1 Samples tested for 4.4 Similarly to Figure 4.3.3, this figure summarises the samples that were generated for analysis of qRBA accuracy. Six 'Mixes' were made by haphazardly combining reporter plasmids. Mix1 to Mix6 were each PCR amplified in three separate indexed reactions, giving 15 unique measurements from 5 replicate NGS runs. The fill in Mix6 is inverted for the benefit of readers with colour vision deficiencies and has no special significance. Values above each column are the number of reads returned (million).

4.4.2 Result of conjugate qRBA measurements and *ad hoc* analyses

From five NGS runs, 13,533,088 *egfp* barcodes and 17,403,237 *luc2* barcodes were counted. As a cursory analysis, the whole NGS library was analysed as a single sample measured via five replicated Illumina MiSeq runs (Fig. 4.4.2). Reporter ratios from this single large dataset showed a dynamic range from $r = 0.00179$ to 561.2 (equivalent to $p\% = 0.179\%$ and 99.8% respectively) and so are shown at the log scale. At a glance, the $r_{[egfp]}$ and $r_{[luc2]}$ measurements conform satisfyingly to the line $y = x$. However, a divergence between $r_{[luc2]}$ and $r_{[egfp]}$ was apparent as $r_{[egfp]}$ diverged from $\sim 1:1$, with $r_{[egfp]}$ showing a clear and increasing tendency to underestimate $r_{[luc2]}$ above $\sim 1:1$ and overestimate $r_{[luc2]}$ below $\sim 1:1$. The consequence was disagreements between $r_{[luc2]}$ and $r_{[egfp]}$ measurements approaching two-fold as $r_{[egfp]}$ approached 0.01 or 100. The sample standard deviation for each coordinate may not be visualisable in print, which reflecting a strong consensus between NGS replicates underlying this result.

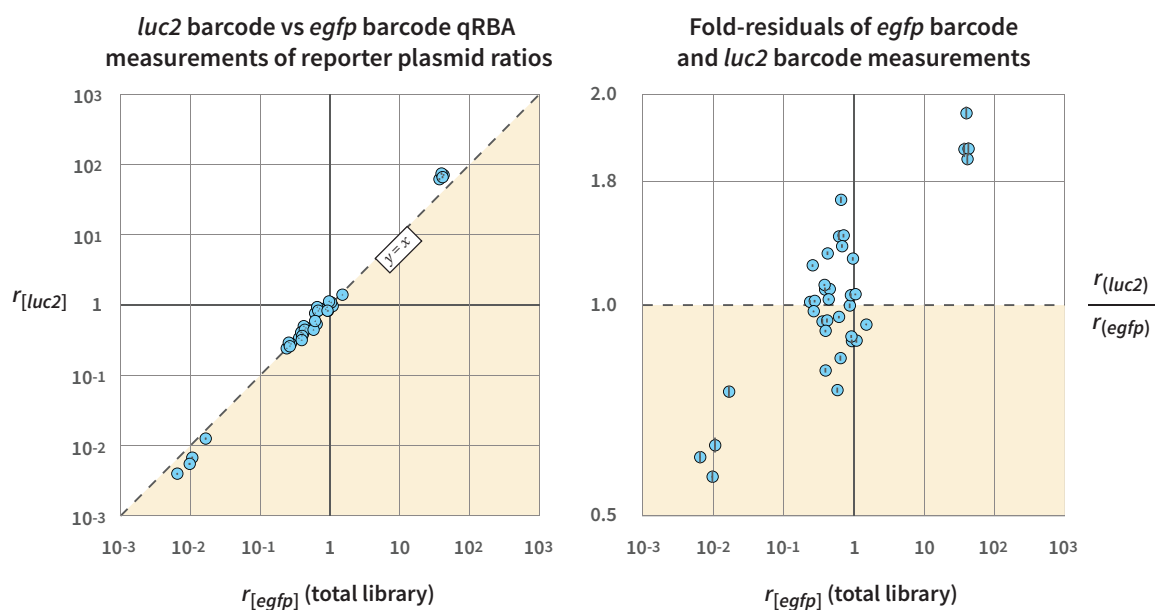


Figure 4.5.2 Barcode ratios across the total NGS library Eighteen indexed PCR samples (derived from six plasmid mixtures) were submitted as a single library for five independent rounds of Illumina MiSeq population sequencing. Shown above is the ratio of all possible reporter pairs, taken as a measurement of the total library. The fold discrepancies between *luc2* and *egfp* measurements approached two-fold in either direction as they approached ~ 0.01 or ~ 100 . A pattern of bias was observed increasing negatively as r decreased from 1 and increasing positively as r increased from 1. That is, when p approached 0 or 1. Horizontal and vertical error bars at the centre of each blue circle represent the sample standard deviation, though may not be visible in print. This reflects a strong consensus underlying each discrepant measurement, and shows that each conjugate result tended centrally toward contradictory readouts.

An equivalent plot was prepared for each $p:q$ pair in all eighteen indexed PCR samples, with averages and sample standard deviations calculated from the five replicated Illumina MiSeq runs (Fig. 4.4.3). Included in the figure are the results for the reciprocal of each barcode pair, in order to illustrate that the effect of measuring barcodes as $[Q]:[P]$ as well as $[P]:[Q]$. Collectively, a least-squares regression analysis of $\log[r_{(luc2)}]$ and $\log[r_{(egfp)}]$ produced the model $y=1.11x+0.08$ (the exact same regression resulted from the reciprocals, but with a change in sign in the y -intercept). This model had an R^2 of 0.992 and an S.E.G. of 0.004, or 0.4% in relative terms. Regression of individual reporter barcode pairs will be addressed in later in the results.

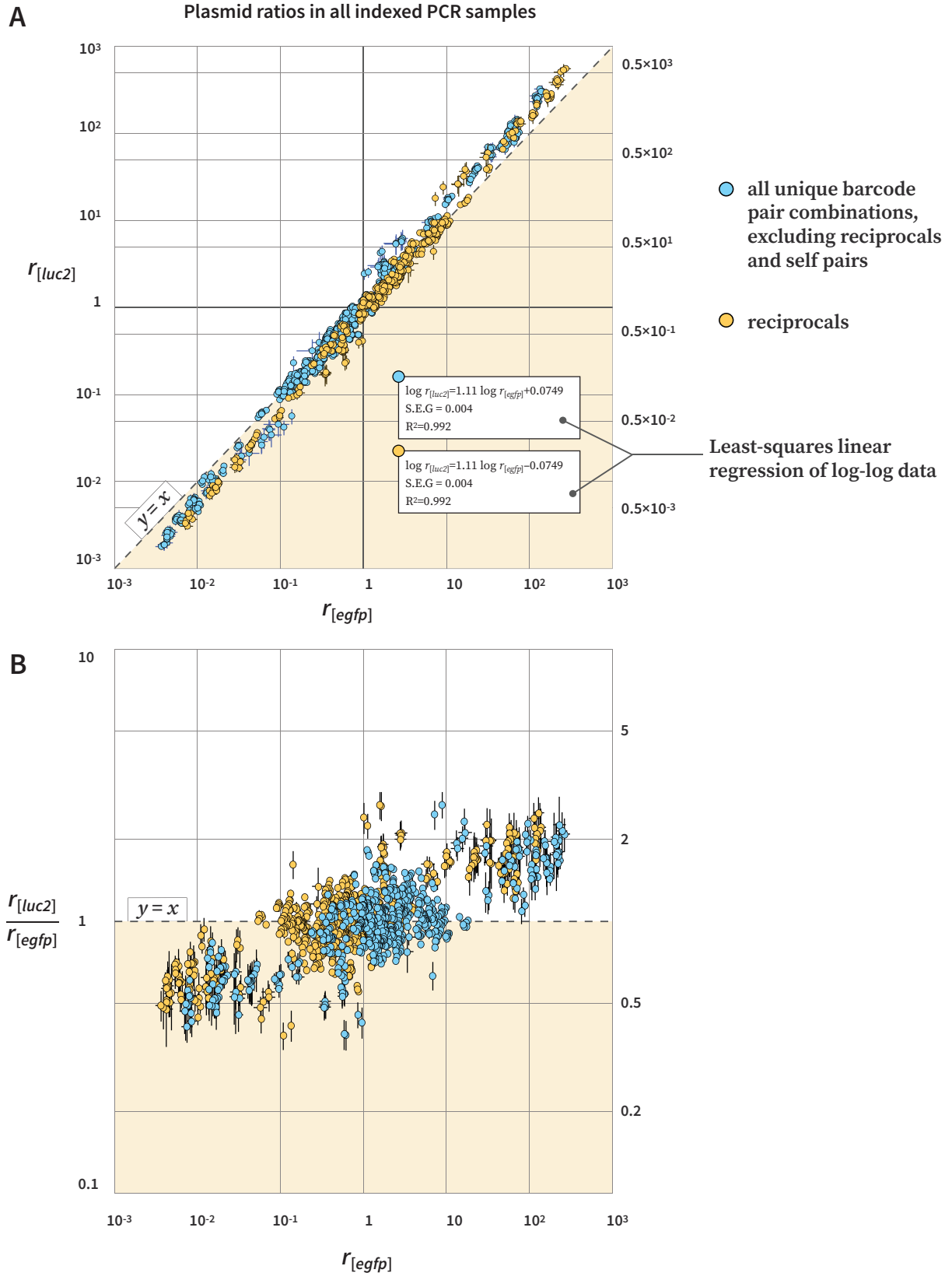


Figure 4.4.3 Comparing conjugate qRBA measurements in all indexed PCR samples (A) Observations in Figure 4.4.2 were also observed when the same analysis was applied separately to all eighteen PCR samples. A linear least-squares regression was performed on log-log data to characterise the error in general terms. Included in yellow are the reciprocals of each unique barcode pair, which serve to demonstrate that the same least-squares regression resulted, only with a change in sign in the intercept. **(B)** The fold discrepancies (shown at a log scale) exceeded two-fold.

While investigating the data, two groups with apparently distinct behaviours arose recurrently. I was unable to determine whether these were in alignment with unidentified trends, the result of some nuisance error source, or the result of unique mode of error taking effect under unidentified conditions. Since these groups are at times conspicuous, it will be useful to define them so that they can be distinguished in visualisations and analysed in isolation where appropriate. To borrow a term from the field of cognitive psychology, these will be referred to as ‘oddball’ groups (‘OB’), since they appear distinct, though the objective basis for their distinctness is not specified. These groups are defined below and highlighted in Figure 4.4.4.

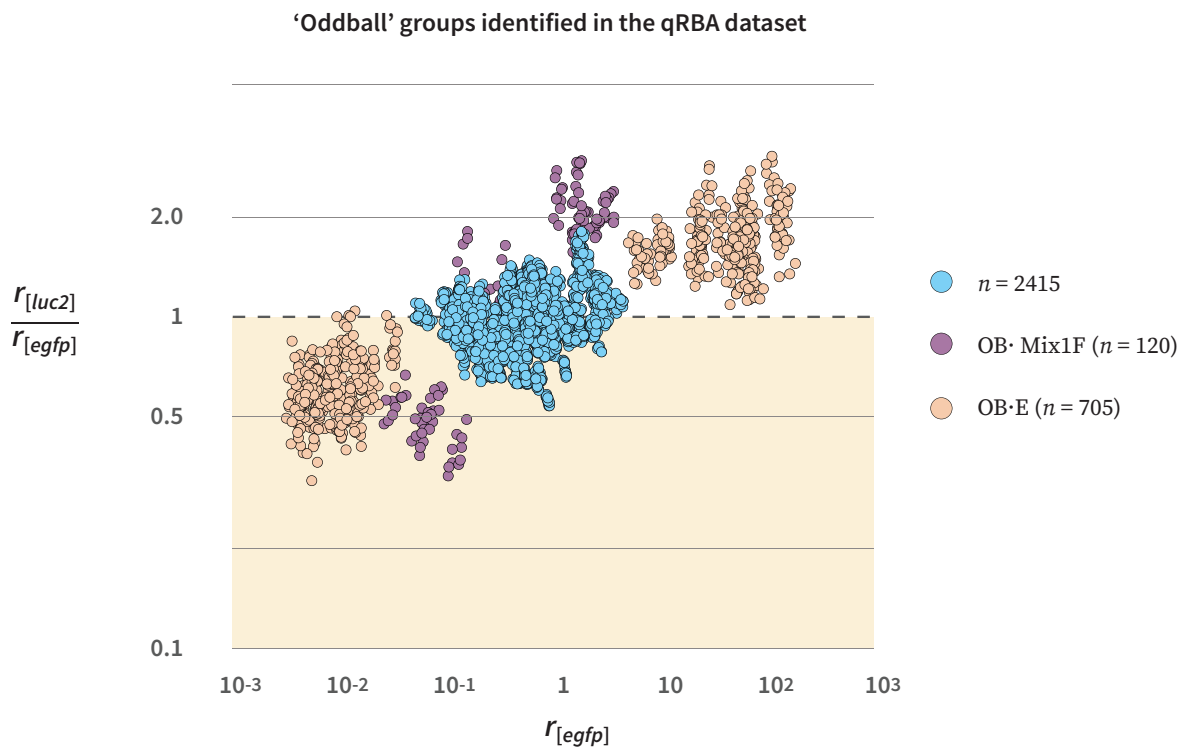


Figure 4.4.4 Demarcation for ‘oddball’ groups with apparently distinct behaviours Some of the more erratic behaviour observed in some analyses was attributable to two groups of data, which have been labelled OB·Mix1F and OB·E. OB·Mix1F consisted of measurements involving Plasmid F where Mix1 was the PCR template. OB·E consisted of measurements involving Plasmid E in any mix, which also represented all reporter ratios measured furthest from 1. The groups have not been systematically demarcated, since it is not known whether the distinct behaviours are divergent or consistent with some more general trend that was not identified. In the course of this report, it will be useful on occasion to delineate these groups from the bulk of the data. The term ‘oddball’ has been borrowed from the field of cognitive psychology, and emphasises the fact that their distinctness is not asserted to be objective.

The first group, OB-E, are reporter plasmid pairs that include the plasmid E. Because E was the least abundant reporter in all mixes, the farthest measurements from 1:1 all belong to OB-E. Unfortunately, there is a paucity of $p:q$ measurements between those in OB-E and those closer to 1:1. With more information in this transitional range, it may have been possible to discern whether the oddball tendencies of OB-E result from a unique mode of error or were part of a continuous trend. A least-squares regression of log-log data in the OB-E group indicated a relationship $y = 1.12x + 0.04$ ($R^2 = 0.999$; average residual = 2.67%) compared to the no-oddball ('none') group $y = 1.04x + 0.01$ ($R^2 = 0.959$; average residual = 215.4%).

The second group, OB-Mix1F, are reporter plasmid pairs involving the plasmid [F] in PCR samples derived from Mix1. These samples had two shared attributes, though their significance is uncertain. (1) each of these mixes was made by recycling the plasmid tested in section 4.3 and (2) the three reverse primers with Mix1 indices resided (in 200 μ L PCR tubes) in a shared column of the 96-well rack used in their storage/use. A least-squares regression of log-log data in the OB-Mix1F group indicated a relationship of $y = 1.39x + 0.22$ compared to the no-oddball ('none') group $y = 1.04x + 0.01$ ($R^2 = 0.959$; average residual = 215.4%).

To characterise the data further, the standard deviation of each qRBA measurement ($n = 5$ NGS replicates) was compared to ζ_s , as in 4.3 (Fig. 4.4.5). The data in this experiment was distinct from that of Section 4.3, in that the OB-Mix1F and non-OB groups exhibited an x -dependent divergence (up to 10-fold discrepancy) from ζ_s in the positive direction as ζ_s increased or decreased from 0.01. Standard deviations in the OB-E group conformed more consistently to ζ_s , though possible signs of positive divergence are apparent in the range of $0.1 \leq \zeta_s \leq 1$. The same trend was observed in the $r_{[egfp]}$ results as in the $r_{[luc2]}$ results.

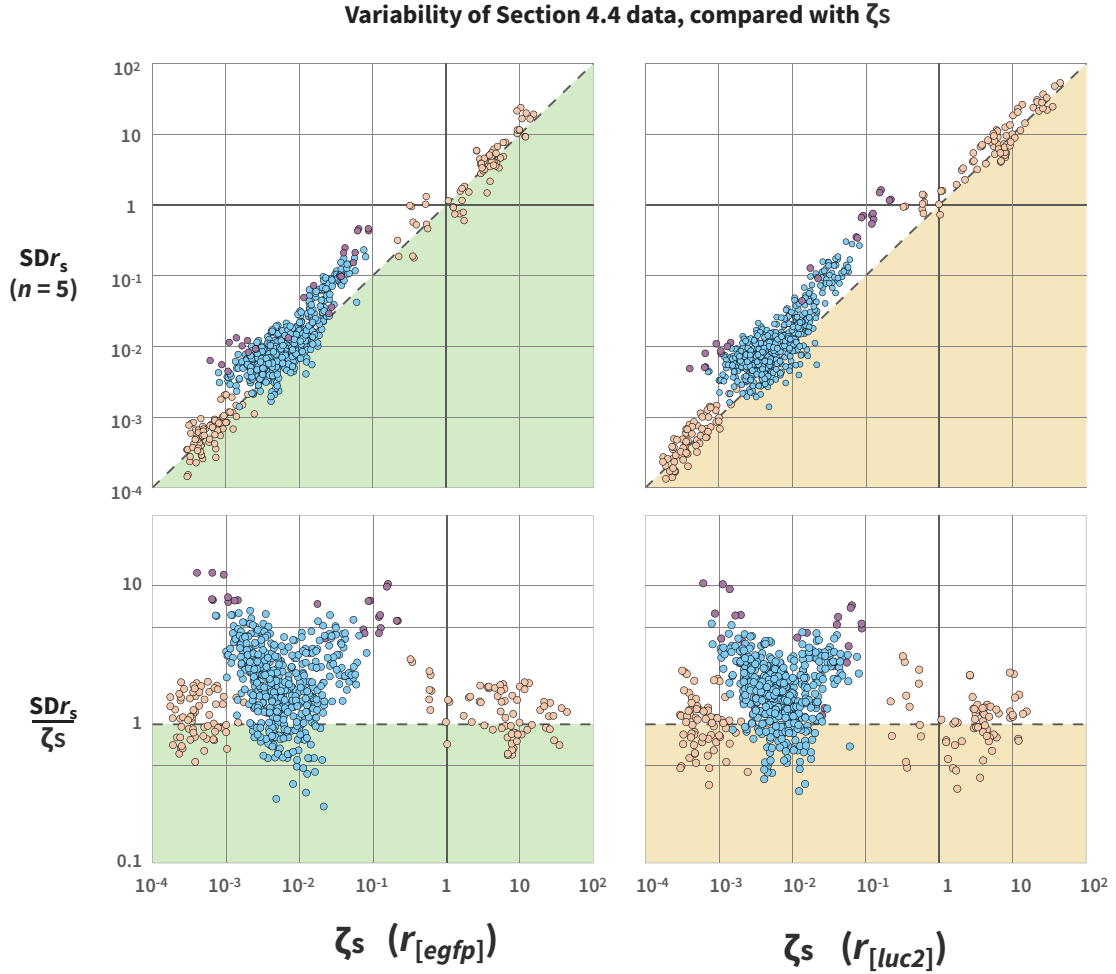


Figure 4.4.5 Comparison of measurement variability ζ_s To explore the data further, the sample standard deviation was compared to the predictor ζ_s , as was done in Section 5.3. With the possible exception of OB-E, the data displayed a markedly different residual pattern from that previous set of experimental replicates, with discrepancies approaching 10-fold in an x-dependent manner. In contrast, the residual distribution in the egfp and luc2 conjugate assays were similar in pattern and magnitude.

The reporter barcodes used for this experiment were designed such that only 5 of 6 bases were necessary for unambiguous identification. This meant that a single-base mismatch would not result in exclusion of a read. To test whether reporter-specific miscall rates could explain the qRBA misalignments observed, the same qRBA readsets were counted with a mismatch tolerance of zero and the fold discrepancies between $r_{[egfp]}$ and $r_{[luc2]}$ were compared to those presented in Figure 4.4.2. Misalignments between $r_{[egfp]}$ and $r_{[luc2]}$ differed negligibly when zero mismatches were tolerated, as opposed to ≤ 1 . Figure 4.4.6a shows these two sets of results as (x, y) coordinates respectively. For reporter pairs including Plasmid E, the $r_{[luc2]}/r_{[egfp]}$ misalignment with 0 mismatches permitted averaged 0.9998 ± 0.01 of that when 1 mismatch was tolerated. For pairs not including Plasmid E, this figure was 1.0001 ± 0.001 . The effect of 1nt

mismatches was also explored by comparing the relative and absolute differences in readcount as c increased, when mismatch tolerance was set to 0 or ≤ 1 (Fig. 4.4.6b). This revealed that as c increased, total incidence of 1nt miscalls increased for any given barcode but decreased in relative terms to a limit of approximately 0.1-0.2%.

Readers may note that the barcode counting pipeline will register and count any read with the specified length, index, leading 5' sequence (i.e. forward primer) and reporter barcode sequence, and does not reject reads based on Q-scores. Note that if two unmerged paired reads disagree on a specific base call, the merged read will contain either the base with the higher Q score, or an N if the Q scores are similar⁴ (N is classified as a mismatch). A miscall in one of two Illumina paired reads will not result in rejection.

Another potential source of error in Illumina sequencing is 'index hopping', which appears to have impacted this study negligibly (as was the case in Section 4.3). This term refers to contamination of indexed primers by other indexed primers, a certain extent to which is unavoidable in standard oligonucleotide production. The contaminating index will cause a read to be misattributed in the demultiplexing process and erroneously counted. Recall, a design advantage of Section 4.3 was to circumvent this problem by analysing only one reporter mix, since it could not self-contaminate either physically or via index hopping. In this section however, we have opened ourselves up to this problem. The reverse primers used in this project were ordered as a batch of 24, 18 of which (per qRBA target) were used in the present experiment. Taken together, *luc2* and *egfp* reporter amplicon sequences containing the predicted indices numbered 30,936,325, while the number of reads containing any of the six unused indices was 26 in total, or one per ~1.2 million reads.

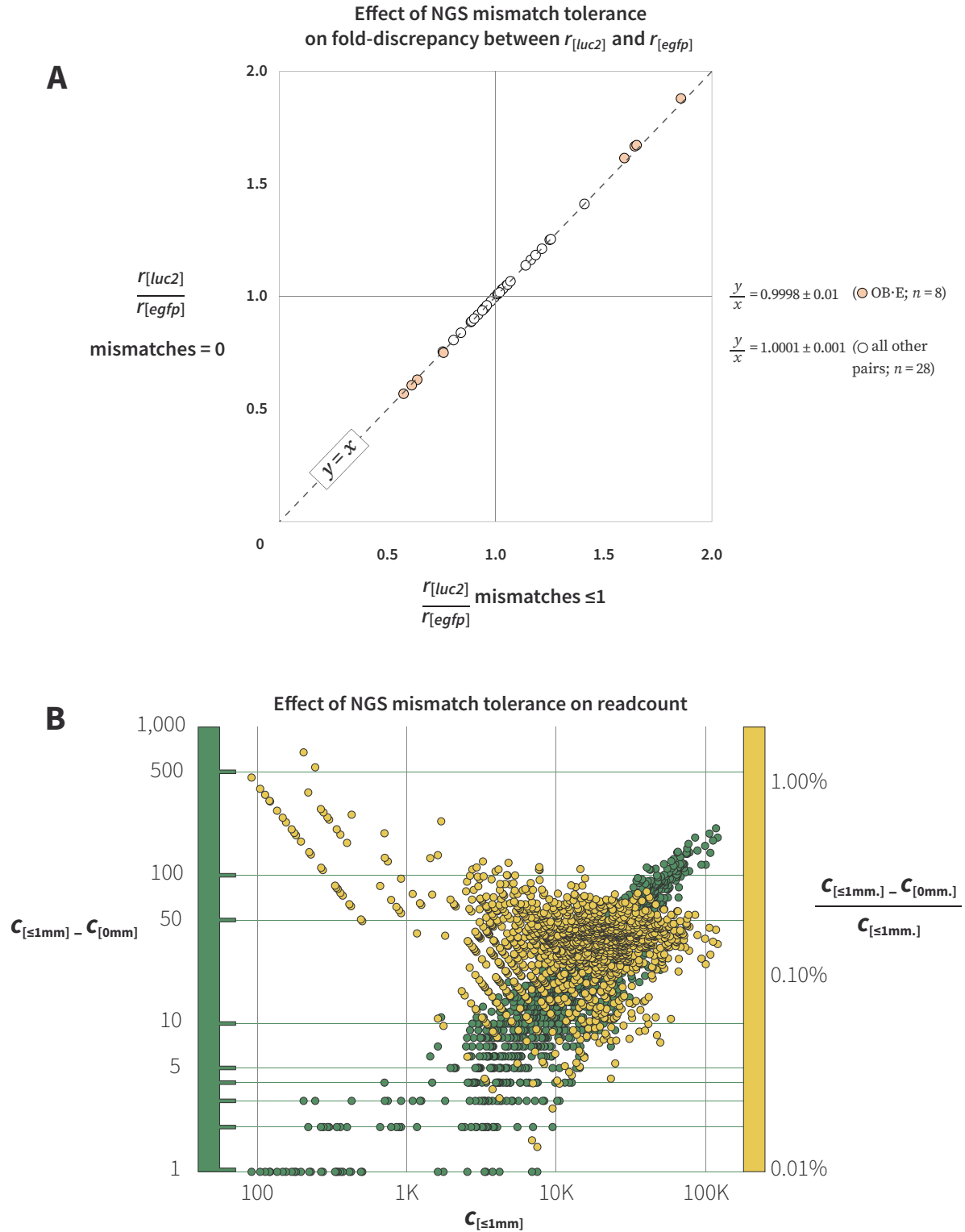


Figure 4.4.6 The effect of sequence infidelity on agreement between conjugate qRBA assays Each reporter barcode sequence in this experiment was positively identifiable by five of six bases, allowing reads to be correctly counted whether the sequence was fully accurate or contained one mismatch. To test whether sequencing infidelity could appreciably effect agreement conjugate qRBA measurements, the barcode counting pipeline was rerun with a mismatch tolerance of zero. $C_{[\leq 1mm]}$ is the reporter barcode count when 1 mismatch is tolerated, $C_{[0mm]}$ is the reporter barcode count when zero mismatches are tolerated. **(A)** The fold differences between luc2 and egfp results are plotted against the equivalent result with zero permissible sequence errors. The average difference in measurement disagreements is shown. **(B)** The difference in readcount for all barcodes in all PCR samples tended towards 0.1-0.2%. At the lower readcount range the impact of miscalls was higher in relative terms but reached a maximum of roughly one percent. Note that (A) concerns reporter count ratios while (B) concern absolute readcounts.

At the suggestion of other researchers, analyses were performed to explore the hypothesis that the measurement disparities between these qRBA assays were purely probabilistic. Two lines of investigation were carried out. The first (Fig. 4.4.7) was a case study of the plasmid pair with the least symmetrical likelihood distributions, as evaluated by the skewness formula $(1-2p)(cpq)^{-0.5}$. This was identified to be plasmids B and E in Mix4. A binomial distribution for this pair was calculated using the $r_{[egfp]}$ and $r_{[luc2]}$ ratios indicated by the grand total of reads in the experiment, with the minimum individual readcount used as the c parameter. Alongside with these PMFs, the figure shows each of the 15 individual qRBA results (5 NGS replicates $\times$ 3 PCR samples) for the *egfp* and *luc2* reporters. Both PMFs showed generally symmetrical profiles with nonoverlapping $\mu \pm 3\sigma$ boundaries. The conjugate qRBA measurements had overlapping ranges of c but separate ranges $r_{[egfp]}$ and $r_{[luc2]}$.

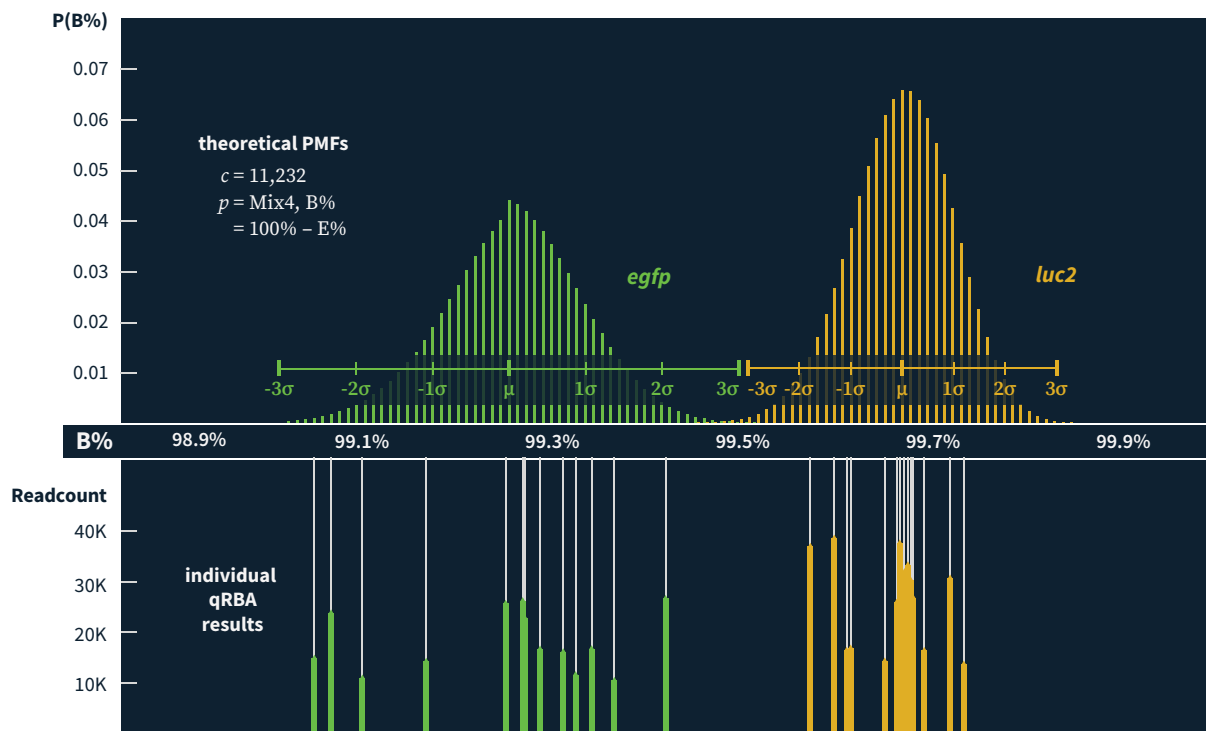


Figure 4.4.7 A case study of qRBA disagreement in the range most vulnerable to binomial skew To illustrate the resilience of Bernoulli stresses against skew at high values of c , the reporter pair with the highest predicted degree of skew was selected for focused analysis. The grand total of reads attributable to the reporter pair B and E in Mix4 were used collectively to measure the $[B]:[E]$ ratio according to both assays. This value was inferred as the p parameter and the PMF was calculated, with the lowest encountered readcount (11,232) used as the c parameter so as to assume the highest degree of skew. Contrary to intuition, the distributions were well resolved and generally symmetrical despite $[B]$ comprising between 99% and 100% of the read counts. In the lower panel, the result of all fifteen qRBA measurements of plasmid B and E in Mix4 are shown, with the coloured vertical bars indicating their readcounts. As the PMFs would predict, the *egfp* and *luc2* measurements fell within clearly separate ranges, and did not overlap on the alternative range even where read counts were similar equivalent.

The second (Fig. 4.4.8) was to assess the probability of obtaining $r_{[egfp]}$ given its equivalent readcount, if $r_{[luc2]}$ was taken to be the accurate value of r . Except for the change in c , this is akin to locating $r_{[egfp]}$ in the likelihood distribution of $r_{[luc2]}$. The location of the observed $r_{[egfp]}$ value in the probability distribution was expressed in the number of standard deviations from the mean. The red zone in Figure 4.4.8 indicates a range of $\pm 5\sigma$, which contains $(1 - 4.6 \times 10^{-7})$ of the normally distributed probability according to the empirical rule. An obvious majority of $r_{[egfp]}$ and $r_{[luc2]}$ measurements disagreed by more than 5σ , which is extraordinarily improbable according to the binomial distribution model. Readers may note that there are no binomial parameters in these readsets with less symmetrical distributions than those shown in Figure 4.4.7.

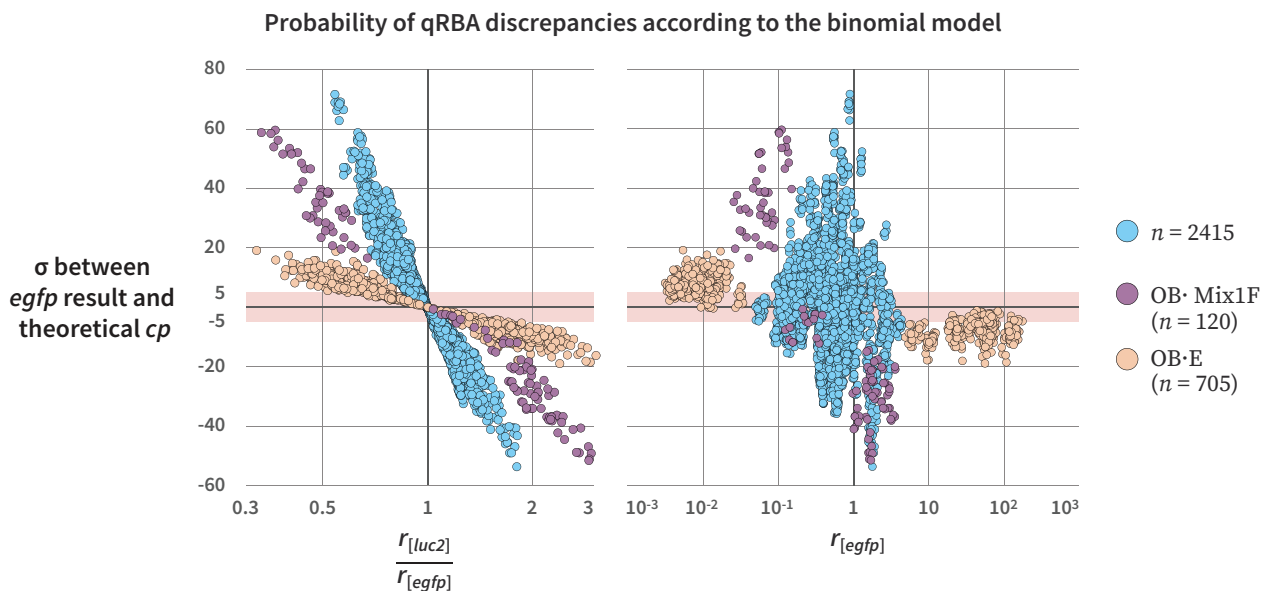


Figure 4.4.8. Probability of egfp measurements, given luc2 measurements] (A) The distance between conjugate egfp and luc2 measurements in probabilistic terms was analysed. The probability of obtaining an egfp measurement from the same number of reads was calculated with the assumption that luc2 measurements were accurate. Effectively, this expresses the likelihood of an $r_{[egfp]}$ result given the conjugate $r_{[luc2]}$ result. The difference between egfp and luc2 results was expressed in the number of theoretical standard deviations calculated according to the binomial distribution. The red zone indicates range of $\pm 5\sigma$, outside of which less than one ten-millionth of the probability mass exists. A majority of the data existed outside this range. This is extremely unlikely to occur randomly.

4.4.3 Ad hoc analyses involving composite functions

From this qRBA dataset we are limited in that we have no view of how $r_{[luc2]}$ or $r_{[egfp]}$ relate to r only how $r_{[luc2]}$ relates to $r_{[egfp]}$ (or vice versa). Although this limitation is unavoidable, further insights can be made if $r_{[luc2]}$ and $r_{[egfp]}$ are expressed in composite with respect to the more

generic variable x . At the request of examiners, the meaning, method and purpose of expressing the data this way will be explained in Appendix D rather than this Chapter. This Appendix D will highlight certain attributes of the data are consistent with a power relationship between r and our conjugate qRBA measurements.

$$r_{[egfp]} = f(r) = ar^b \quad (4.1)$$

$$r_{[luc2]} = g(r) = cr^d \quad (4.2)$$

With $f(r)$ and $g(r)$ plotted as (x, y) coordinates respectively, the plot can be described according to the function below:

$$y = g \circ f^{-1}(x) = c \left(\frac{x}{a} \right)^{\frac{d}{b}} \quad (4.3)$$

A log-log transformation of the data gives us a linear plot with gradient $\frac{d}{b}$ (which we can simplify to m) and y -intercept (ca^{-m}) .

$$\ln(y) = m \ln x + \ln(ca^{-m}) \quad (4.4)$$

The fact that gradients varied from 1.0 implied that although some functions may be approximated as linear, a source of error was present whose effect is to exponentiate r . Appendix D will also highlight that this model constrains the component coefficients (c and a) and exponents (d and b) to a limited (albeit indefinite) range of solutions. It also addresses how the error structure can be assessed irrespective of which barcode is classified as p and which qRBA target is represented on the x -axis.

Unsurprisingly, $(r_{[egfp]}, r_{[luc2]})$ coordinates derived from the same plasmid mixture tended to cluster within a single location, which was not conducive to linear regression analysis. However, clear Mix-specific trends could be resolved by plotting the data in the form of hyperbolae, whose horizontal and vertical asymptotes were m (the ratio of exponents in Eqs 4.1 and 4.2) and $\ln(ca^{-m})$ respectively. The method for producing these hyperbolae are shown in Appendix E. Though regression analysis was outside my skillset, the results are included – with illustrative examples in Figure 4.4.9 – to highlight the obvious mix-specific differences in the m and/or $\ln(ca^{-m})$. For ease of visualisation, the axes in Figure 4.4.10b are drawn along the lines $x \equiv 0$ and $y \equiv 1$. Presented alongside are the average $r_{[egfp]}$ ratios taken from each NGS replicate and PCR sample derived from a given plasmid mixture ($n = 15$). At least one tentative must be

highlighted: the position of the asymptotes for a given barcode pair often correspond, at least in rank terms, with the rank in $r_{[egfp]}$ ratio specific to a Mix. This would imply that a composite power function cannot be used to describe the error of a given barcode pair generally, since it does not allow variation in slope (let alone x -dependent variation in slope) in its log-log plot. Another observation is that in all cases, the position of vertical asymptote (signifying $\chi_{[y]}$) was positive where $r_{[egfp]} > 1$ and negative where $r_{[egfp]} < 1$. In Mix1, the $r_{[egfp]}$ result for the pair [C][D] was measured to be $1.00 \pm 1.3\%$, and the hyperbola had a vertical asymptote near to zero. As will be discussed (and explained in detail in Appendix E), these observations are not directly consistent with uneven PCR amplification efficiency.

Figure 4.4.9 (cont.) Regression to a hyperbolic model resolves Mix-specific trends | Six plasmid template mixtures were prepared for this section, each of which was amplified in three uniquely indexed PCR reactions. Each individual plasmid ratio was therefore measured fifteen times - with three PCR replicates and five NGS replicates. These fifteen replicates tended, predictably, to cluster in a single location on the x-y plane. Subtle differences in gradient were therefore poorly detectable using linear regression, and differences in the y-intercept wholly unachievable. **(A)** A hyperbolic regression was achieved using the x,y metrics specified in the upper panel, which produced the regression model shown in blue. Hyperbolae with this form will have a horizontal asymptote at m and a vertical asymptote at $\chi_{[y]}$ since both of these values tended toward 1.0, much of our empirical data will fall within a close proximity in the hyperbolic model, which also be a region of relatively high curvature. These factors make it possible to resolve qualitative differences in the error pattern of datasets that were difficult to distinguish via linear regression. **(B)** For each unique reporter pair, all six Mixes were analysed under a hyperbolic regression model. Since all functions pass through (0,1), these coordinates are where the axis lines in these plots will cross. To the side are shown the average p:q result according to the egfp assay. These data are stacked by rank to help illustrate the tendency for asymptotes in these regressions to correlate to the p:q ratio. White fill is used for distinction of Mix6 data by readers with colour vision deficiencies and has no other significance. Plots for all other reporter pairs can be found in Appendix E.

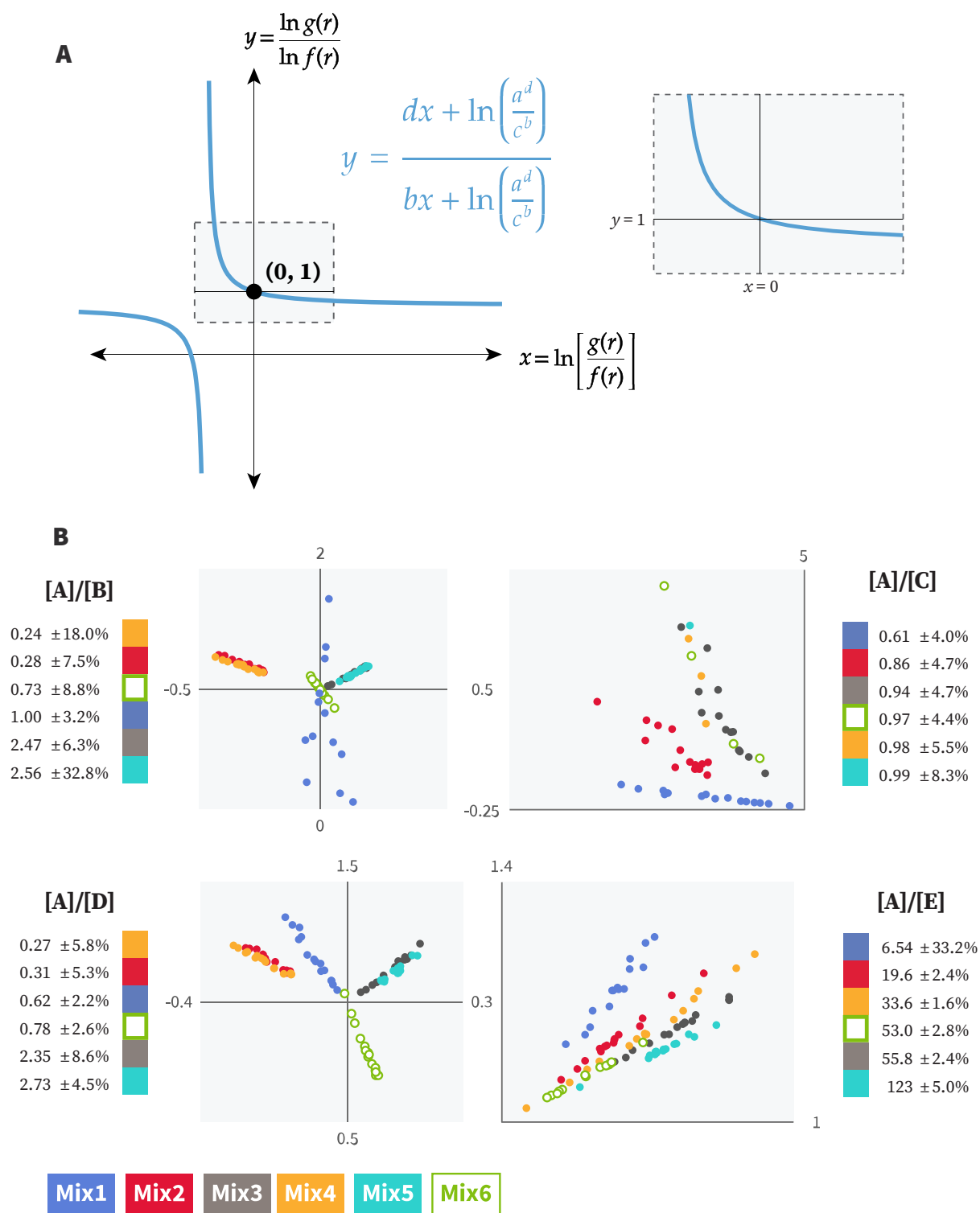


Figure 4.4.9 (cont.) Regression to a hyperbolic model resolves Mix-specific trends

4.4.4 Discussion – sources of qRBA measurement error

This experiment utilised plasmids with two 150bp PCR target regions, one adjacent to a truncated *egfp* gene and one adjacent to a truncated *luc2* gene. The ratio of *egfp* variants ($r_{[egfp]}$) and the ratio of *luc2* variants ($r_{[luc2]}$) was guaranteed stoichiometrically to be equal, and measurements of these ratios could be compared to test the accuracy of these two qRBA assays. Plotting $r_{[luc2]}$ measurements against $r_{[egfp]}$ measurements, data conformed satisfyingly to the model relationship $y = x$ at first glance. As is often the case, $y = x$ was a pleasing visual approximate for the data, compared to all possible outcomes. However, as is not always the case, formal analyses were carried out to test this impression more rigorously. Rather than varying from $r_{[egfp]}$ randomly, $r_{[luc2]}$ was seen to diverge from $r_{[egfp]}$ as it departed from ~ 1 in either direction, measuring approximately two-fold lower than $r_{[egfp]}$ as $r_{[egfp]}$ approached 0.01 and approximately two-fold higher than $r_{[egfp]}$ as $r_{[egfp]}$ approached 100. Collectively, the unique reporter pairs measured behaved similarly enough that a convincing linear regression model could be determined (even more strongly if ‘oddball’ groups were separated), though a hyperbolic model would show that individual reporter pairs display distinct divergence patterns dependent on their Mix of origin. The parameters of these error patterns appear to correlate with underlying reporter ratio, though not in a manner directly predicted by nonequal PCR amplification efficiency (as will be discussed).

In response to these results, *ad hoc* analyses were carried out to investigate two salient possible explanations – sequencing infidelity and binomial skewness – both of which were unambiguously excluded. To address sequencing infidelity, I supposed that 6/6 correct calls in the reporter barcode sequence were necessary for inclusion in the NGS readset and compared this to the result when one mismatch was permitted (as was in fact the case). Imposing this limitation, discrepancies between $r_{[luc2]}$ and $r_{[egfp]}$ averaged 1.0001 ± 0.001 that of the original dataset. For barcode pairs involving Plasmid E specifically, this figure was 0.9998 ± 0.01 . Readers should note that these figures in fact overstate the impact of miscall bias on these conjugate qRBA assays. Since all barcodes in this experiment were unambiguously identifiable from 5/6 barcodes, *two* miscalls were required to exclude a read from the final count, which probability dictates is an order of magnitude lower in incidence. To many this result may feel unexpected, given that NGS sequencing techniques are characteristically lower in individual sequence quality than Sanger methods. Figure 4.4.6b helps us to rationalise this result. It depicts clearly

the limited effect of miscall frequency as total readcounts increase and incidence stabilises in relative terms. It is also worth reiterating that miscall frequencies are specific to individual reporter counts, affecting reporter *count ratios* indirectly and with lesser arithmetic impact. As was the case in Section 4.3, the effect of ‘index hopping’ was trivial to exclude, since unused indices (6 from a batch of 24) were counted on the order of <1 read per million.

To experts in binomial statistics, investigations into skew may seem to be *a propos* of nothing. In my more local research community however, the inherent skewness of binomial distributions has proved an enduring explanation and it is prescient to acknowledge why. Given the patternistic nature of the error it is reasonable to contemplate whether the divergence of $r_{[luc2]}$ and $r_{[egfp]}$ is parametric in origin. Skew is a parametric phenomenon, which, like the divergence of $r_{[luc2]}$ and $r_{[egfp]}$, increases as p approaches 0 or 1 (that is, when $p:q$ departs from 1:1 in either direction). Skew, therefore, is guaranteed to correlate with fold-disparity in this dataset. Furthermore, both of the phenomena take effect in a domain where random variations exert an exaggerated effect on $r_{[luc2]}$ and $r_{[egfp]}$. This is because as either p or q approach zero, linear increments correspond to increasingly wide leaps in the $p:q$ ratio. In this region we can intuit that even subtle biases could be magnified by the outsized influence of the smaller operand. The data also suggests a basis for the direction of the bias – that the readcount underlying $r_{[luc2]}$ was consistently higher than $r_{[egfp]}$. It stands to reason that because of this, to whichever extent $p_s:q_s$ affected by skew, $r_{[egfp]}$ is more affected than $r_{[luc2]}$. This latter suggestion at least is simple to address, as fold disparity did not correlate with fold differences ($R = 0.0012$) or absolute differences ($R = -0.15$) in readcount.

Two aspects of the binomial distribution are critical to note, both of which relate directly to the fact that cp is always the maximally probable outcome of a Bernoulli process. Firstly, the maximally probable value of p_s is unaffected by the number of Bernoulli trials (i.e. the readcount). The number of trials limits the alignment between cp_s and $P(\max.)$ only insofar as it affects the list of discreet values between $cp_s = 0$ and $cp_s = c$ that are possible. Without a difference in $p:q$, the predicted $p_s:q_s$ of two Bernoulli processes will not differ even if their counts do. Secondly, given the outcome of a Bernoulli process, it is not possible to infer a more accurate value of p by ‘adjusting for skew’ – not because the solution is indefinite or challenging to calculate, but because any hypothetical value of p that isn’t p_s is less likely than p_s .

Ad hoc analysis have illustrated that measurement discrepancies between $r_{[luc2]}$ and $r_{[egfp]}$ do not accord with binomial probability theory. I first considered whether such discrepancies were probable for the sample ratio with the least symmetrical distributions predicted by the binomial model, in which one reporter was estimated to lie between 99-100% of the overall pair count (Fig. 4.4.7). Their exact probability distributions (equal proportionally to their likelihood distributions) contradicted our intuition that parameters in this range would exhibit visualisable asymmetry. Further, superimposing individual measurements illustrated that $r_{[egfp]}$ and $r_{[luc2]}$ results did not belong to single p_s distribution with variable readcounts. A more extensive approach was to ask: if $p = p_{s[luc2]}$ and $readcount_{[egfp]} = c$, what is the probability of the *egfp* result observed? Applying this question to all conjugate measurements in the dataset confirmed that the majority were misaligned in probabilistic terms by $>5\sigma$. The probability of this occurring in >3000 conjugate measurements is far below the range of human intuition. It also showed that this misalignment was not limited to the most extreme binomial parameters, such as for the oddball group OB-E. In fact, OB-E contained the least misaligned data in probabilistic terms, and misalignments in the remaining data increased as $r_{[egfp]}$ approached 1:1 (Fig. 4.4.8).

To assist in this and future studies, a composite power model was suggested which related $r_{[egfp]}$ and $r_{[luc2]}$ data, measurements plotted as (x, y) coordinated, to the more generic variable x . Aside from limiting error terms to a limited range of solutions, this model helps greatly to alleviate confusion surrounding how the choice of qRBA target on the x -coordinate and/or reciprocation of r affect the analysis (Appendix D). In short, some simple geometry can provide us with the same description of the error regardless of how it is arranged.

A hyperbolic representation of the data (Appendix E) demonstrated that relationships between $r_{[egfp]}$ and $r_{[luc2]}$ were not predicted by pair of reporter barcodes being analysed. It showed instead that a given barcode pair exhibited in highly deterministic, and clearly distinct hyperbolic regressions depending on which of the six reporter Mixes they derived from. We could also observe that the parameters m and ca^{-m} (horizontal and vertical asymptotes respectively) tended to correlate with r_s as measured from the *egfp* target. A formal analysis of these plots is counted among those unanticipated tasks that, while potentially fruitful, required specialised know-how and needed to be deferred decisively in the interest of a timely submission.

However, it is notable that this observation does not directly align with the error structure predicted if PCR amplification bias was responsible. As explained in Appendix F, this would lead to differences in the vertical asymptote, but not in the horizontal asymptote. While far from exhaustive, this approach is a sensible first step in exploring the PCR hypothesis. It does not exclude that aspects of the PCR process may lead to error, but it indicates that biased amplification is either an incomplete or an inaccurate explanation for the results of this experiment. It is worth considering this possibility in the context of endogenous references used in qPCR. For example, quantifying viral genomes relative to a host chromosomal target or an mRNA rt-qPCR target relative to a ‘housekeeping’ genes. In the present study, the PCR amplification bias being hypothesised is due to six consecutive nucleotides worth of variation, located away from primer binding sites, representing 4% of the overall nucleobase content. If we are willing to attribute significant amplification bias to this confined variation, we should not expect relative measurements of wholly separate PCR targets to be achievable.

Potentially one of the most relevant observations in these analyses was the variabilities of $r_{[egfp]}$ and $r_{[luc2]}$ and their relationships to ζ_s . The residuals and their distribution differed markedly from the data generated in Section 4.3, with the possible exception of the OB-E group. This strongly indicates the presence of some error source in these conjugate assays that was somehow avoided in Section 4.3. Despite this, the predictiveness of ζ_s in $r_{[egfp]}$ data and $r_{[luc2]}$ data was hardly distinguishable either in profile or magnitude. In other words, despite being an unreliable predictor, ζ_s was more or less equally predictive of $r_{[luc2]}$ variation as it was $r_{[egfp]}$ variation. This could indicate that both readouts were accurate measurements of the adapter-ligated reporter amplicon variants that were loaded onto the Illumina flow cell – if there were a disconnect between readout and the physical ratios in the adaptered library, we may expect a disconnect between the value of ζ_s inferred from the measurement and the empirical variation. Instead, the ζ_s value inferred from the readout appeared equally predictive in both assays.

Data for this analysis was initially provided by experiments that were intended to showcase the concept of qRBA by applying it to conceptually accessible research questions. These experiments utilised the conjugate reporter barcode plasmid(s) pCRB.Ch5, which would concomitantly supply the data for this stoichiometric test. As a response to this initial result, this dedicated dataset was produced to test those observations over a more diverse range of

parameters. The present data recapitulated the initial observations and illustrated them more comprehensively. Further independent repeats would therefore have questionable value.

Identifying a strategic path forward in this thesis would also have questionable value without the input of researchers practiced in bioinformatics and mathematical modelling. This said, small suggestions can still be made. There are straightforward experimental designs that would clarify whether PCR is a major source of qRBA bias. A 300bp PCR target could be designed containing a reporter barcode in both halves of the sequence. If a blunt-ended restriction site were included in the centre, cleavage of the 300bp PCR product would result in two blunt-ended 150bp products whose equimolarity would be categorically unaffected by the process of PCR amplification. A plasmid construct containing two such 300bp targets would be an even more robust experimental tool to explore sources of qRBA inaccuracy. If no expert insight can be gleaned from the existing data, this may justify the incorporation of orthogonal measurements, so that the accuracy or inaccuracy of qRBA measurements could be assessed with respect to a reliable reference, which itself would be a consensus between different measurement techniques. While the intended benefit of this strategy is obvious, we should not assume that achieving consensus between different measurement techniques is a trivial task. We should bear in mind that the most prominently misaligned qRBA measurements observed in this experiment were associated with p_s values that differed replicably by less than one percent. The scope of orthogonal measurements must reflect degree of precision achievable and the degree of precision required to be useful.

4.4.5 Discussion – how should existing count data be interpreted?

Identifying the error observed is significant not just for qRBA applications, but because of its adverse implications in the wider area of NGS-based count analysis. The reporter amplicons studied in this experiment could, from another field, be interpreted as an experimental control against most if not all feasible sources of read bias. There is no fragmentation involved in generating the adapter recipients, the termini are identical, the 144 bases outside the UMI sequence are identical, and there are no competing sources of alignment as might be seen in genome sequencing. The singular difference between one reporter target and another was six adjacent nucleotide variants, which contained no trimers and no more than 50% guanines and cytosines. What's more, the validity of these 'controls' could be directly observed, as the outcome of the experiment was unchanged whether 0 or ≤ 1 mismatch was permitted in

counted reads. The considerable disagreements between qRBA measurements in this experiment may therefore represent an as-yet unidentified source of inaccuracy affecting studies of differential gene expression.

Unless the mechanisms of error can be identified and eradicated, we cannot assume that multinomial reporter barcode data is a metrologically accurate representation of reporter variant populations. Fortunately, the highly systematic divergences between r and r_s offer certain boundaries within which qRBA data may be reliably interpreted. Present evidence indicates that the actual values may differ on the order of two-fold, with a lesser impact in the range of $p:q = 0.1$ to 10 but potentially dependent on the barcode. Replicate measurements were extremely precise (despite exceeding ζ_s), and the data overall regressed strongly to a model in which increases in the physical barcode ratio resulted in increases to the qRBA measurement.

Significantly for Chapter 6, this implies that changes in a measured reporter ratio can be confidently attributed to changes in the reporter population, rather than systematic error. While alternative qRBA targets may disagree on the quantity, the direction and magnitude of the difference are most likely trustworthy. The readout resulting from a given barcode ratio appears to be governed by a power relationship, meaning that the discrepancy increases geometrically as the ratio diverges from 1:1. However, the exponent of this function tended to vary in close proximity to 1, which gave the regression a steady pseudo-linear trajectory over a certain range. What this implies is that although misalignment increases as the ratio diverges from 1:1, the increase is generally not dramatic if the underlying ratio is in a moderate range of ~ 0.1 to 10 (corresponding to a $p\%$ between 10 and 90%), and even less dramatic within a range of ~ 0.5 to 2 ($p\% = 33\%$ to 67%). Note that this applies to changes in a specific pair of reporters in different populations, not differences between a different pair of reporters in the same population.

4.5 Conclusions

The precision observed in Section 4.3, and to follow in Chapter 6, greatly affirmed the metrological potential for qRBA argued in Chapter 3. With this potential set as a goal, the agreement between these two qRBA assays was unsatisfactory and did not support the ‘fully quantitative’ conception of reporter barcode analysis. This was judged not from the size of the

discrepancies (which was not trivial), but the patternicity and degree of consensus underlying them. These factors suggest that the measurement discrepancies were reached deterministically as a result of some systematic source(s) of error and would only have been reached with increased confidence had a larger number of replicates been implemented. While the source of inaccuracy could not be identified, the fact that skew, NGS artefacts and reporter-specific PCR bias were not supported may be seen as a desirable outcome, since these three issues would be very difficult to mitigate.

Critically, the outcome of this experiment does not deter efforts to pursue qRBA as a powerful and widely adoptable methodology. In its wider context, this Chapter represents only the first concerted efforts to explore its metrological validity. Further investigation stands either to reveal avoidable sources of inaccuracy or elucidate unavoidable ones. In the latter case, we would be empowered to work within known confines exploit the potential of qRBA made clear in Section 4.3 and in Chapter 6.

5

Review — Plasmid ITR instability in the rAAV vector system

5.1 Introduction

Recombinant AAV (rAAV) gene delivery vectors are currently the subjects of a surge in research and commercial activities – a state that contrasts palpably with the field’s previous decades of slow and steady conceptual development⁹¹. This rapid translational development presents a unique set of challenges, at least two of which call our attention to the terminal, hammerhead-shaped elements of the vector genome, the inverted terminal repeats (ITRs) (Fig. 5.1.1). One is achieving an adequate and economically tenable supply of the vector to support preclinical and clinical applications. The other is developing and validating a quality control infrastructure fit for the regulatory standards of human medicine. Interactions of the ITRs with host and viral proteins underly numerous crucial genome processing stages in the viral lifecycle, which are recreated in cultured tissue to produce the recombinant vector. They are how otherwise-nonviral gene cargo is converted from a dsDNA precursor into a replicationally active, encapsidatable ssDNA genome. The ITR-flanked genome initially exists as an integrated dsDNA sequence structurally equivalent to the dormant AAV virus. In the vast majority of cases present and historical, this sequence is supplied on a plasmid DNA construct produced in bacterial culture. The problem central to my doctorate is that AAV ITR sequences are prone to large truncations in course of bacterial propagation. Consequently, the genetic inputs to rAAV production are heterogenous in one or both ITR loci. Despite their fundamental role in (r)AAV biology, there has been limited motivation to understand the implications of this problem for vector production and performance.

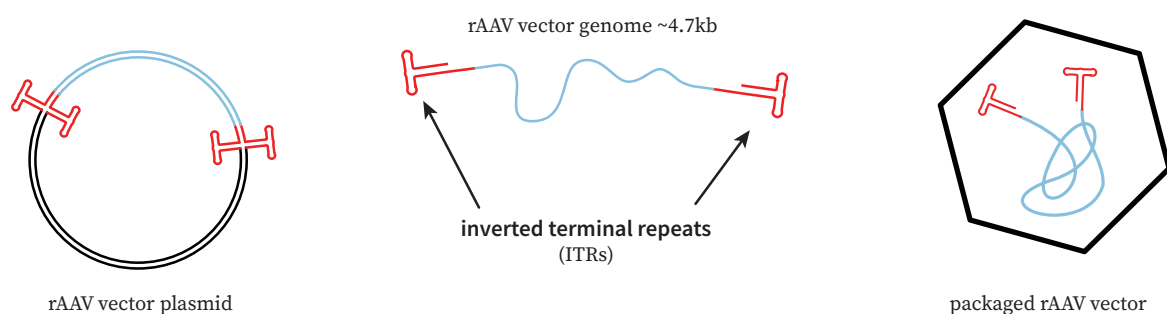


Figure 5.1.1 Shape and location of the inverted terminal repeats of on the (r)AAV genome | The genetic cargo carried by an AAV virus or rAAV vector is ~4.7 kilobases worth of single-stranded DNA, flanked on both ends by 144nt secondary-structure-forming elements called inverted terminal repeats (ITRs). In the recombinant vector context, the same genome sequence including both ITRs also exists in a double-stranded form, usually integrated within a plasmid molecule, which is propagated in bacterial culture.

Given their central importance to (r)AAV biology, it may seem remarkable that unquantified ITR heterogeneity has been viewed as acceptable by regulatory bodies. The explanation for this is far from straightforward, though it relates strongly to the fact that creation of rAAV vectors is not strictly contingent on both ITRs being intact. As will be discussed, the process can tolerate significant deletions in one of either ITR sequence. This is of course highly fortunate. Were it not true, it is questionable whether rAAV vectors would presently be a realised clinical technology. For regulators however, the question is not if or how ITR truncations preclude rAAV production – it is whether or not the proportion of plasmids with two intact ITRs affects the quality, and hence the clinical safety and efficacy, of the clinical rAAV vector produced. If not, what evidence exists to validate it? If so, how much variation in intact ITR content should be tolerated? Limitations to the yield and consistency of rAAV vectors are currently one of the most important limitations to the vitality of rAAV-mediated gene therapy. To stakeholders in rAAV therapies (both direct and indirect), the consequences of plasmid ITR heterogeneity, be they large or small, are immediate.

The ITRs exist on either end of an ssDNA genome, borne by adeno-associated viral virions, whose capsids are utilised as a gene therapy vector. Each piece of biology and their and technological significance must all be generally understood for the issue of ITR integrity to be properly appreciated. To address these areas, the Chapter will open with a brief description of an rAAV vector, including the canonical functions of its ITRs. With our main character introduced, Section 5.3 will describe the increasing demand for rAAV vector products and the strains in upstream bioprocessing that hinder their translation. Once we are aware of this

context, we will explore the grounds for concern regarding plasmid ITR heterogeneity in Section 5.4 by reviewing role of the ITR in rAAV biology more deeply. A broad synthesis of (r)AAV genome processing will be an important task because direct empirical research into the ITR problem is scarce in the current literature. The final section of this chapter will discuss the nature of plasmid ITR instability and its implications for rAAV vector research. It will reflect on the lack of inertia behind understanding and addressing the problem, and changes in the methodological landscape that could support a change to this trend.

This Chapter includes a small number of original experimental results, for illustrative purposes. These are not submitted as part of the research content of this thesis. Rather, they are instances where a modest amount of labour in the laboratory would effectively support arguments or concepts raised.

5.2 The AAV virus and its vector derivative – a snapshot

Viruses, more so than cellular forms of life, are an ontological oddity⁹². That said, the infectious particle (the virion) is conventionally a virus's appointed representative. Broadly speaking, it consists of a protein-based capsule (capsid), possibly lipid-coated, containing nucleic acids that encode the structural and non-structural products necessary to synthesise and fabricate a new generation of virions with the assistance of an invaded cell. The virion is capable of transmission between organisms, trafficking to specific tissues, entry into cells and delivery of the viral genome to the required subcellular compartment (in eukaryotes often the nucleus)⁹³. Viral vectors are powerful gene delivery vehicles that exploit sophisticated infectious pathways necessary for trafficking and entry into target cells and expression of their genetic cargo⁹⁴. To make an AAV virion that carries nonviral genes, the virulent products necessary to produce new progeny can be encoded from outside encapsidated gene cargo, and their presence confined to the production environment (Fig. 5.2.1). In most preclinical settings at least, they are split between separate plasmids⁹⁵. The packaged capsid thereby lacks the gene content necessary for infectious reproduction (or other undesired activities) and is described as *replication incompetent*. Both virologists and gene therapists agree that the word 'virus' or 'virion' is therefore no longer appropriate. We instead refer to this entity as a *vector particle*. Rather than 'infection', the word 'transduction' applies when we use such a particle to transfer DNA or RNA cargo into a cell via a viral invasion pathway. As of 2018, rAAV-delivered transgenes have been

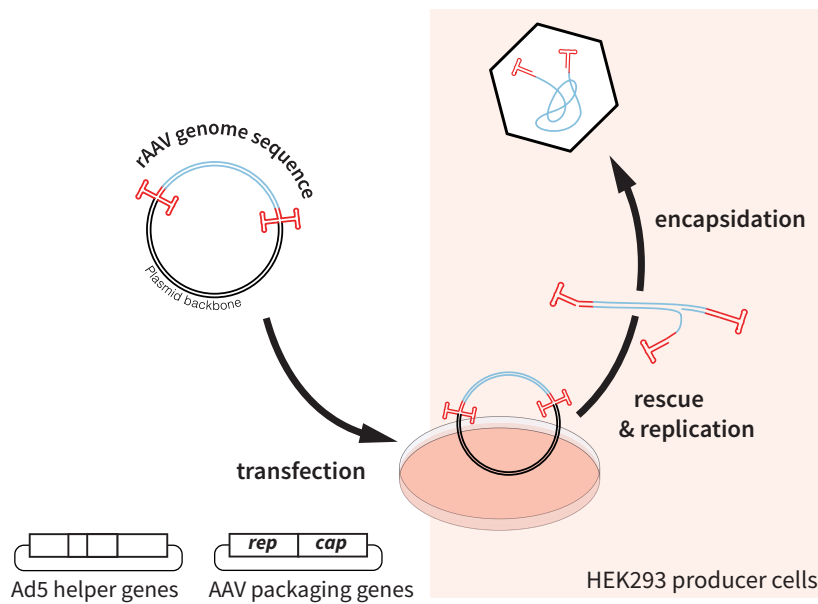


Figure 5.2.1 Generation of rAAV vectors from plasmid precursors in mammalian producer cells | The AAV virus has the natural ability to derive replicable, packageable ssDNA genomes from integrated, double-stranded precursors. In nature this constitutes the transition from a dormant ('latent') mode of cellular residence to a state of active lifecycle progression consisting of genome propagation and encapsidation of new genome copies. This process can be recapitulated in a tissue culture environment by transfecting cells with plasmids containing the necessary viral gene content, and a separate plasmid containing the 'integrated' sequence intended for packaging. In both contexts, the ITRs are the mechanism by which flanked DNA is targeted for each of the critical viral processing stages necessary for generation of packaged capsids particles. Essentially, the ITRs give an otherwise nonspecific length of DNA the structure of an rAAV genome (latent or lytic) and allow it to be processed in the same manner. Ad5 helper genes promote lytic drive in producer cells, which will be discussed in more detail in Section 5.4.

observed to remain transcriptionally active for over eight years in humans⁹⁶ and fifteen years in non-human primates⁹⁷, which are the latest timepoints measured at the time of writing.

An AAV virion consists of a ~4.7kb ssDNA genome (either sense or antisense) packaged within an unenveloped (lipidless) icosahedral capsid approximately 26nm in diameter. The wild-type AAV genome contains two open reading frames (ORFs)⁹⁸. The *rep* ORF encodes non-structural enzymes that regulate and catalyse AAV genome expression, replication, and translocation into unoccupied capsids. Four *rep* ORF products are encoded using two splice variations and two transcription initiation sites – Rep78, Rep68, Rep52 and Rep40 – often collectively referred to as 'Rep'. The *cap* ORF encodes the virion's structural components (capsid monomers)⁹⁸. Other genes have been identified within the AAV genome including assembly activating protein (AAP)⁹⁹ and 'X'¹⁰⁰ but are not greatly pertinent to this review. In the case of rAAV manufacture, the complete composite of the AAV genome is present within the producer cell (most prominently HEK293 cells) but split between two or more loci – the ITRs flank the DNA to be packaged, while *rep* and *cap* ORFs are located on a separate construct⁹⁵. In nature, superinfection with a 'helper virus' (canonically adenovirus) drives the AAV lifecycle forward,

as will later be discussed. This effect is recreated in rAAV production by supplying the producer cells with a set of adenoviral genes sufficient to promote lifecycle activity.

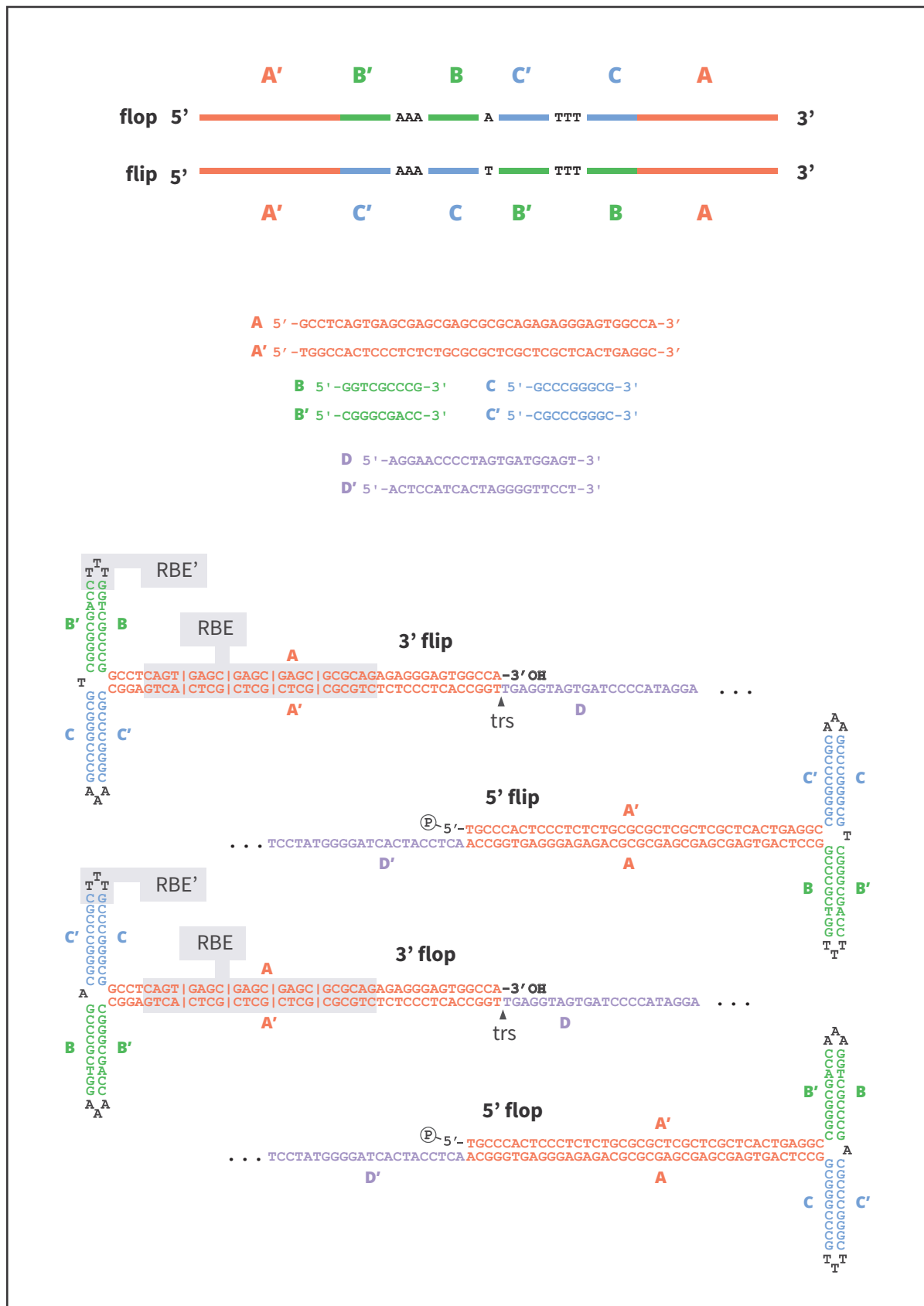
Adeno-associated virus encompasses considerable diversity of variants. Antibody reactivity has historically acted as the distinguishing factor between various natural AAV variants, leading to the term ‘serotype’, though AAVs display a remarkable diversity in tropism and transduction characteristics depending on tissue, species and the *in vitro/in vivo* context^{101–103}. Currently over 100 variants have been identified in humans and NHPs¹⁰¹, though a much smaller number (<30 or so) are likely familiar to any given vectorologist. This natural biodiversity can be leveraged to target a broad range of clinically important tissues and has even been instrumentalised in the engineering of novel AAV variants with improved vectorological attributes¹⁰⁴. The first serotype isolated, AAV2, holds canonical status and is by far the most extensively studied^{105,106}. Since the capsid is the primary determinant of AAV tissue tropism, it is unsurprising that the capsids have been the primary interest in vector bioengineering. In addition to cell binding and tropism, the capsid also directly influences intracellular trafficking and ultimately nuclear delivery, as well as interactions with the immune system, which have been reviewed elsewhere^{107–109}.

The AAV ITRs are 145nt elements on either end of the genome, consisting of inverted repeat sequences that give rise to an iconic ‘T-shaped’ structure possessing a paired guanine/cytosine content of ~80%. These inverted repeats (the ITRs) are also present as double-stranded sequences in a larger dsDNA molecule at various junctures within the AAV lifecycle. The ITRs fundamentally are what give a strand of DNA the structure of an AAV genome. This underlies our ability to use viral replication systems to create capsid particles containing DNA of our own design. Their roles in production can be separated under the general headings ‘rescue’^{110,111}, ‘replication’^{112–114} and ‘packaging’^{115–117}, each of which relies on interactions of unique sets of host and viral proteins with the ITRs¹¹⁸. ‘Rescue’ refers to the process by which replicationally active, encapsidatable genome sequences are obtained from dsDNA sequences present in plasmids, chromosomes or episomes¹¹⁰. Replication of the genome takes place through a modified rolling-circle process termed rolling hairpin replication, which the ITRs mediate by self-priming the genome for second-strand synthesis and recruiting the necessary effectors¹¹⁹. The ITRs also contain the signalling element necessary for translocation of the vector genome into unoccupied capsids¹¹⁶, referred to as encapsidation, though the term ‘packaging’ is often used synonymously. The influence of the ITR on the fate of ssAAV vector genomes in transduced

cells is not well understood compared to rAAV production. This will be discussed in greater depth in Section 5.4.

Four unique inverted repeat sequences exist in a packaged AAV genome, three of which appear twice (once per ITR – Fig. 5.2.2). The BB' and CC' repeats are adjacent to one another and flanked by the AA' repeat. When self-annealed, the BB' and CC' repeats create a three-way junction in place of a simple hairpin turn. The genome contains only one copy of the fourth repeat – the DD' repeat. Its upstream and downstream components are directly inset from the hammerhead-forming repeats and separated by the rest of the viral genome. Each (r)AAV genome terminus can be best described therefore as comprising three-and-a-half inverted repeats. Seven unpaired bases also exist within each terminal hammerhead. A single adenine or thymine separates BB' sequence from the CC' sequence, and both of those inverted repeats contain either a triadenine or trithymine base at their centres, which exist as short hairpin loops at the end of each 'arm' of the T structure. Note that although the 'T' shape is more iconic, a 'Y' shape is predicted *in silico* and more likely accurate¹²⁰. The forward or reverse complementary sequence of the terminal 125nt hammerhead can be found on either terminus, and do not correlate to the opposite ITR or the sense of the genome. Observed in 1977¹²¹ and confirmed in 1980¹²², the two apparent orientations of the ITR hammerhead were whimsically termed 'flip' and 'flop'.

Understanding the primary sequence of AAV's ITRs can be a demanding spatial reasoning task given their nested layout and variable internal configurations. Further, the packaged genome as a whole exists in either sense, is represented in a counterintuitive 3'-to-5' direction, and exists in a multitude of single- and double-stranded structures outside the virion stage of the lifecycle. An explainer (described by one examiner a 'a joy to read') can be found in Appendix G. A subsequent explainer, Appendix H, will explain the topological determinants of ITR secondary structure in dsDNA. This theory underlies AAV genome biology globally and bears relevance, direct and indirect at times, to the subject matter of this thesis.



5.2.2 Arrangement of inverted repeat sequences at both possible termini of the ssAAV genome | Shown above are reference sequences for 'flip' or 'flop' ITRs at both ends of the genome, including sequence labels that appear in this thesis. RBE(') = replicase binding element; trs = terminal resolution site

5.3 The biotechnological context of plasmid ITR instability

Technologies (virus-derived or otherwise) to add to or alter a cell's nucleic acid content, formerly confined to a research setting, have advanced to become the basis of clinically approved tools in human medicine. Their technological complexity, therapeutic mechanism and range of targetable health problems contrast greatly with what the European Council once referred to in passing as 'the traditional pharmaceutical field'. Such was the phrasing used in *Regulation (EC) No. 1394/2007*¹²³, which in 2007 established the new regulatory category of *advanced therapy medicinal products* (ATMPs), recognising the special legal challenges associated with their 'novelty, complexity and technical specificity'. In the year immediately preceding the global SARS-CoV-2 pandemic, the US Food and Drug Administration (FDA) announced an expansion in reviewing capacity for gene therapeutics to accommodate the anticipated 10-20 gene therapy approvals by 2025 and 200 investigational new drug (IND) submissions expected annually from the 2020^{124,125}. In that year, rAAV vector applications constituted approximately 65% of first-time-in-human gene therapy trials and 43% of IND trials outside of oncology¹²⁶. In March 2021 the number of active rAAV clinical trials was estimated to be 146¹²⁷. Since that year, pharmaceutical companies including Biogen, Pfizer, Novartis and uniQure have initiated expansions to their internal clinical-grade rAAV manufacturing capabilities, and new manufacturing sites have been secured by contract manufacturing organisations (CMOs), such as Lonza and Thermo Fisher¹²⁸.

Recombinant AAVs are one component of a larger viral vector technology sector, whose uses span each of the EU's four ATMP sub-classifications¹²⁹ (gene therapy, somatic cell therapy, tissue-engineered therapies, and 'combined advanced' therapies). The rAAV sector is smaller than other viral vector platforms (notably retroviral and adenoviral vectors) but leads in *in vivo* gene therapy applications¹³⁰. Many of the most difficult barriers to efficacy can be alleviated by administering viral vectors to a patient's cells *ex vivo* and re-engrafting the successfully modified autologous tissue. Such barriers include systemic dosing, tissue specificity, immune responses (adverse to the vector or its recipient) and the inability to screen away undesired outcomes such as unintended integrations of the genetic cargo. *In vivo* applications thrive despite these challenges in order to service poorly autograftable, clinically important targets such as the CNS, retina, skeletal muscle, and adult liver. The *in vivo* niche is shared largely (though not exclusively) with adenoviral vectors (ADV)¹³¹. The diversity and engineerability of

ADV tropism are comparable with rAAVs, and they can exceed the rAAV packaging limit of 4.7kb. They are also substantially less costly to produce and boast a much more favourable rate of gene delivery per vector particle than AAVs. Despite these factors, the two vector systems are not in direct competition for the most part. This is largely because, unlike ADV, rAAV can more consistently achieve non-transient transgene expression and has exhibited a much lower potential to provoke adverse inflammatory reactions^{132,133}.

Interest in rAAV vectors has been intensified by a series of gene therapeutic landmarks that have been increasing in frequency in only the last five years. The very first gene therapy to be awarded market approval by US or European regulators was an AAV-serotype-1-derived viral vector, *Alipogene tiparvovec* (trade name Glybera), approved by the European Medicines Agency (EMA) in 2012 for the treatment of an inherited dyslipidaemic disease via delivery of a functioning lipoprotein lipase gene to skeletal muscle tissue¹³⁴. At the time of writing, another five rAAV-mediated therapies have since been approved in one or both of the aforementioned jurisdictions. In 2017, *voretigene neparvovec* (LUXTURNA®), became the first gene therapy drug approved by the FDA. It is an AAV serotype 2 vector used to deliver a functional copy of the RPE65 gene to human retina¹³⁵. Mutations in this gene lead to dystrophic retinoid cycling resulting in blindness for individuals with biallelic mutations. *Onasemnogene abeparvovec* (Zolgensma®), an AAV serotype 9 vector treatment for type 1 spinal muscular atrophy supplying infant motor neurons with functional SMN1, saw FDA approval in 2019¹³⁶. The remaining three therapies received market authorisation less than twelve months before this thesis's submission. Two of these are haemophilia treatments – a haemophilia B treatment *etranacogene dezaparvovec* (Hemgenix™) in the USA¹³⁷ and a haemophilia A treatment *valoctocogene roxaparvovec*¹³⁸ (Roctavian™) in the EU. The third, *eladocogene exuparvovec*¹³⁹ (Upstaza™), is indicated for AADC deficiency and administered via direct injection to the putamen of the brain.

In the EU and the US, the first approved gene therapies set the unfortunate record for most expensive drug treatments ever to have been marketed in their respective jurisdictions. Glybera (priced at US \$1 million per treatment), historic for being the first gene therapy approved in the US or European markets, was historic again for having been withdrawn due to commercial inviability¹⁴⁰. LUXTURNA was brought to market at the price of US \$850,000 per eye (which fell short of speculation¹⁴¹), only to be overtaken by Zolgensma, priced at US \$2.1 million per treatment and again by Hemgenix at US \$3.5 million. Newly approved rAAV therapies have

been met with both celebration and predictable community upset in response to their high price tags. Of course, the concept of “one and done” treatments for rare genetic diseases aligns questionably with the venture capitalist model of medical advancement^{142,143}. The greater prevalence of targets such as Parkinson’s disease or macular degeneration presumably avoids this limitation, but the wider demand only calls further into question whether healthcare systems are prepared to pay prices driven to such heights by manufacture alone^{128,144–146}.

The biotechnological complexities of ATMPs like rAAV, revolutionary as they are, are also the basis of two hefty and interrelated obstacles incomparable with the ‘traditional pharmaceutical field’ – manufacturing and quality control. Within the wider viral vector cohort, rAAVs are associated with particularly low production yields and high proportions of empty vector particles. The most widely employed production system in basic research and early trials is adherent human embryonic kidney (HEK293-derived) cells transiently transfected with multiple plasmids containing the vector genome and necessary viral genes. Adherent HEK293 cultures are capable of servicing basic research needs and scaled-out systems may be sufficient for early-phase clinical trials, but they transition unreliably into suspension culture systems, where batch volumes from 250-2000L are necessary to supply latter-stage clinical trials and the commercial market^{128,147,148}. Alternatives to mammalian cells and/or plasmid transfection are a highly active research area. While not reviewed here, they attest to the seriousness of the problem and, like prototypical HEK293 systems, have thus far resulted in suboptimal production or vector performance outcomes^{128,149}. The cost and resource intensity of mammalian tissue culture is of course difficult to understate, though perhaps not fully appreciated are the economic burdens created by its delicate, convoluted supply chain and the complexity of biological raw materials¹⁵⁰. As manufacturing scale and regulatory oversight escalate, pressure to guarantee batch success and product consistency increases the risk associated with financing these bioprocesses¹⁵¹. The dose-per-treatment (and thus the doses yielded per batch) relate directly to gene transfer efficacy. Upstream-process-related determinants of vector efficacy are multifarious, production-system-specific, and poorly understood at a mechanistic level. The need to analyse, understand and manage them is inextricable from process inefficiency in limiting the growth of rAAV-based medicine^{128,147,152}.

Of course, were we to eliminate the problem of plasmid ITR instability in this very thesis, we would not expect to upturn this situation. What these points illustrate is the significance of plasmid ITR integrity to the broader project of translating rAAV vector therapies. In 2023, the

‘real world’ constraints of production and quality assurance are a central and urgent focus for the field, and must be viewed as tangible barriers to access for those living with targetable diseases and their communities. The impactful and unimpactful upstream-process-related determinants of vector performance must be understood so that effectual regulations can be implemented and non-concerns can be alleviated. Unless profound and unanticipated changes to the core biology of the vector occur, the ITRs will remain a universal component of its production and clinical application. Understanding the impacts of plasmid ITR instability is therefore a clear necessity.

5.4 A deeper dive into AAV genome biology

The species adeno-associated virus is a member of a family with indeterminate class, order and phylum named parvoviridae (subfamily parvovirinae), in the genus dependoparvovirus. The compact two-ORF structural/non-structural coding strategy of AAV is typical of all known parvoviruses and has been well-reviewed on numerous occasions by Drs Cotmore and Tattersall^{153–155}. Parvoviral genomes consist of linear single-stranded DNA (type II of the Baltimore classification), whose termini are self-annealed in various secondary configurations due to the presence of inverted repeat sequences, referred to in the above-cited reviews as ‘hairpin telomeres’. Although DNA sequences are the ultimate decider of parvoviral taxonomy, telomeric secondary structure is remarkably distinct between most characterised species. From a parvovirological perspective, ITRs are the hairpin telomeres of adeno-associated virus¹²⁰. Two monophyletic sets of genera currently constitute the vertebrate-targeting subfamily parvovirinae (desnovirinae – parvoviruses that infect invertebrates, is the other known subfamily). The set containing adeno-associated virus (as well as erythroparvovirus, copiparvovirus, tetraparvovirus) are homoteleric, possessing the same telomere at each terminus. The other set is heteroteleric. Telomericity is a useful indicator of the virus’s genome reproduction mechanism and the polarity of packaged genomes; as discussed later, homoteleric spp. replicate through a modified rolling-circle replication mechanism known as rolling hairpin replication (RHR) that generates both sense and antisense packaged genomes, with approximately equal frequency¹¹³. Heteroteleric parvoviruses have a rather more complex replication mechanism and package only a single polarity.

AAV exhibits a ‘biphasic’ lifestyle that can diverge and re-join through distinct potential paths¹⁵⁶. The most direct path is the replicationally active ‘lytic’ phase characterised by unsuppressed *rep* and *cap* transcription and the emergence of newly packaged virions. The ‘latent’ pathway can be thought of as a detour in the lifecycle, in which the virus enters a state of dormancy as a transcriptionally inactive, fully double-stranded provirus capable of reinitiating lifecycle progression at a later juncture^{156,157}. The eponymous characteristic of dependoparvovirus is the dependence on coinfection with an additional virus (termed a ‘helper virus’ in this context) for lytic reproduction to take place, either upon arrival in a host cell or at some stage during latency. Adenovirus (Ad) is by far the best-represented example of a helper virus and is rAAV’s partner-in-production in most non-experimental settings, either as a live virus, or, more commonly, as a defined set of plasmid-encoded Ad proteins capable of instigating lytic lifecycle activity¹⁵⁸. Other well-documented helper-viruses are herpes simplex virus 1 (HSV-1), human herpesvirus 6 (HHV-6)¹⁵⁹, vaccinia virus¹⁶⁰ and human papillomavirus(HPV)¹⁶¹. The helper effect can also be recapitulated by applying genotoxic stresses including UV irradiation and hydroxyurea^{156,162–166}.

Canonically, the AAV2 provirus is chromosomally integrated via a Rep-dependent process, specifically in the AAVS1 site (19q13.42)^{167,168}. It is surprisingly seldom noted that noncanonical, episomal structures are a significant portion, if not a clear majority of latent AAV viral genomes detected in *in vivo* studies. These exhibit a similar extrachromosomal structure to those responsible for long-term transgene expression in stably transduced humans¹⁶⁹ and non-human primates^{170–172}. One early study was unable to detect AAV-chromosome junctions at all¹⁷³ and a later study identified such a junction in only one of seven primary human tissue samples (sampled from tonsil-adenoid, spleen, or lung) with the remainder existing as low- and high-MW circular dsDNA concatemers of the AAV genome¹⁷⁴. It has been speculated that site-specific integration may even be an artefact of continuous cell culture¹⁷⁴, though a recent large-scale study of primary human samples (n > 1000 individuals) has identified both integrated and episomal latent structures in non-tumorous liver tissue¹⁷⁵. Chromatinisation is another under-acknowledged aspect of proviral AAV, seeing as latent wild-type AAV2 has exhibited hallmarks of chromatinisation (i.e. histone association) in multiple *in vitro* studies^{176–178}. Recently, AAV histone trimethylation has been discovered as a significant source of helper-mediatable transcriptional regulation in HEK293T and HeLa cells, via a KAP1/PP1/HLys6 pathway¹⁷⁹. Transcriptionally active copies of recombinant AAV genomes have been observed

in the skeletal of non-human primates 22 months post-administration predominantly as chromatinised episomal monomers and concatemers¹⁷¹. The same structures were observed in human skeletal muscle studied 3 and 12 months post-administration¹⁶⁹, though their chromatin structure was not analysed.

A key event in the latent-to-lytic transition the generation of a replicable, single-stranded AAV genome from the duplex proviral precursor. This reaction is termed ‘rescue’. Importantly, viral rescue is recapitulated in an *in vitro* environment to produce rAAV vectors whose vector cargo will mirror ITR-flanked genomes ‘integrated’ within double-stranded plasmids. Historically, plasmid-encoded (r)AAV genomes have acted as a model to study the rescue process, due to its structural similarity to the chromosomally integrated proviruses. The exact rescue mechanism remains unknown and may involve nicking at the ITR terminal resolution site, cruciform-specific interactions, *de novo* synthesis or some combination thereof, and almost certainly involves the combined activities of host and viral proteins^{110,111,164,180,181}. However, plasmid transfection can offer only a limited understanding of wild-type reactivation and its inefficiencies *in vitro*. Obviously, the natural time course of Rep (and Cap) expression (and thereby Rep-mediated activities) following rescue and replication of a limited number of proviral genomes differs profoundly from the first-order relationship between expression and plasmid copy number in *in vitro* transfection. There are also known *cis* regulatory mechanisms at play in wild-type reproduction that do not occur during plasmid rescue since they involve viral loci between ITRs^{182–184}.

Although many well-controlled experiments have identified key mechanisms by which Ad or its specific products can exert helper effects, our understanding of lytic drive in (r)AAV production remains limited. It is understood that none of the most commonly transfected Ad helper components are either necessary or sufficient to support lytic reproduction¹⁵⁸. Indeed, the special relationship between Ad and its eponymously associated dependoparvovirus has not been indicated by direct studies. By contrast, human herpesvirus 6 (HHV-6), which encodes an AAV2 *rep* homologue, has been more closely associated with AAV than Ad in two recent studies^{175,185}. Smith-Moore et al.¹⁷⁹ have argued that the multimodality of lytic suppression/activation signifies a tiered network of lifecycle regulation incorporating host, AAV and helper-viral factors. These tiers include nucleosome modification¹⁷⁹, transcriptional suppression/activation^{184,186} and genome processing (rescue and replication). Host proteins observed to influence AAV lytic activation (such as FKPB52^{187,188}, the MRN complex^{189–191}, DNA-

dependent protein kinase^{192–194}, Artemis¹⁹⁵ and Ku proteins^{196,197}) are usually members of DNA-damage response (DDR) pathways. Their intranuclear effects on AAV can differ drastically depending on dividing/nondividing status due to the predominance of either homologous recombination (HR) or non-homologous end joining (NHEJ) repair programs^{155,192,193}. Helper-specific effects have also been observed, with the Mre11/Rad50/Nbs1 complex downregulating AAV replication in the absence of Ad but promoting it in the presence of HSV-1¹⁹⁸.

The rolling hairpin replication mechanism (RHR) of rAAV genome propagation has been described as a modified form of rolling circle replication. The process (adapted from Ward 2005¹¹³) is depicted in Figure 5.4.1, though it omits the formation of unit dimers and trimers that occur depending on whether terminal resolution occurs before or after the cycle is completed⁶. The ITR has three canonical roles in this process – self-priming of the genome, recruitment of necessary Rep proteins, and hosting the terminal resolution site (trs) where a single-stranded nick separates parent strand from the *de novo* strand. AAV genome replication requires the cooperation of highly conserved host proteins including replication factor C (RFC), proliferating cell nuclear antigen (PCNA), polymerase δ and the minichromosome maintenance complex (MCM)¹⁹⁷. Of the proteins involved in facilitating RHR, the interaction between the ITRs and viral *rep* products (predominantly Rep68/78) has perhaps been best characterised. *rep* products are required to mediate terminal resolution, which occurs via helicase-mediated extrusion of the A'D junction into a 'nicking hairpin' stabilised by an imperfect repeat on either side of the nick site. The ITR contains a 22bp sequence within the AA' stem containing a [GAGC]₃ motif (the Rep binding element or RBE) and a CTTTG sequence denoted as RBE'. The RBE' sequence is always present on the hairpin loop closest to the 3' terminus (see Fig. 5.4.2). Both of these elements interact with the large Rep proteins to bring about this activity, though the RBE is less tolerant to primary sequence variation¹¹². It has been conjectured¹⁹⁹ that the invariant position of the CTTTG in space may help to account for the symmetrical hammerhead structure of the ITR. The spatial arrangement of the hairpin arms and RBE may indeed underly the well-documented replicative incompatibility between the ITRs and Rep proteins of AAV serotypes 2 and 5, lending credence to this theory, though I note that hairpin loop complementarity is not a conserved feature in known AAV ITR sequences^{200–204}. The Rep68/78 proteins possess the site-specific helicase activity needed for hammerhead reformation, which is required to initiate second-strand synthesis from the daughter strand¹⁹⁷ ('reinitiation') though this has not yet been demonstrated empirically.

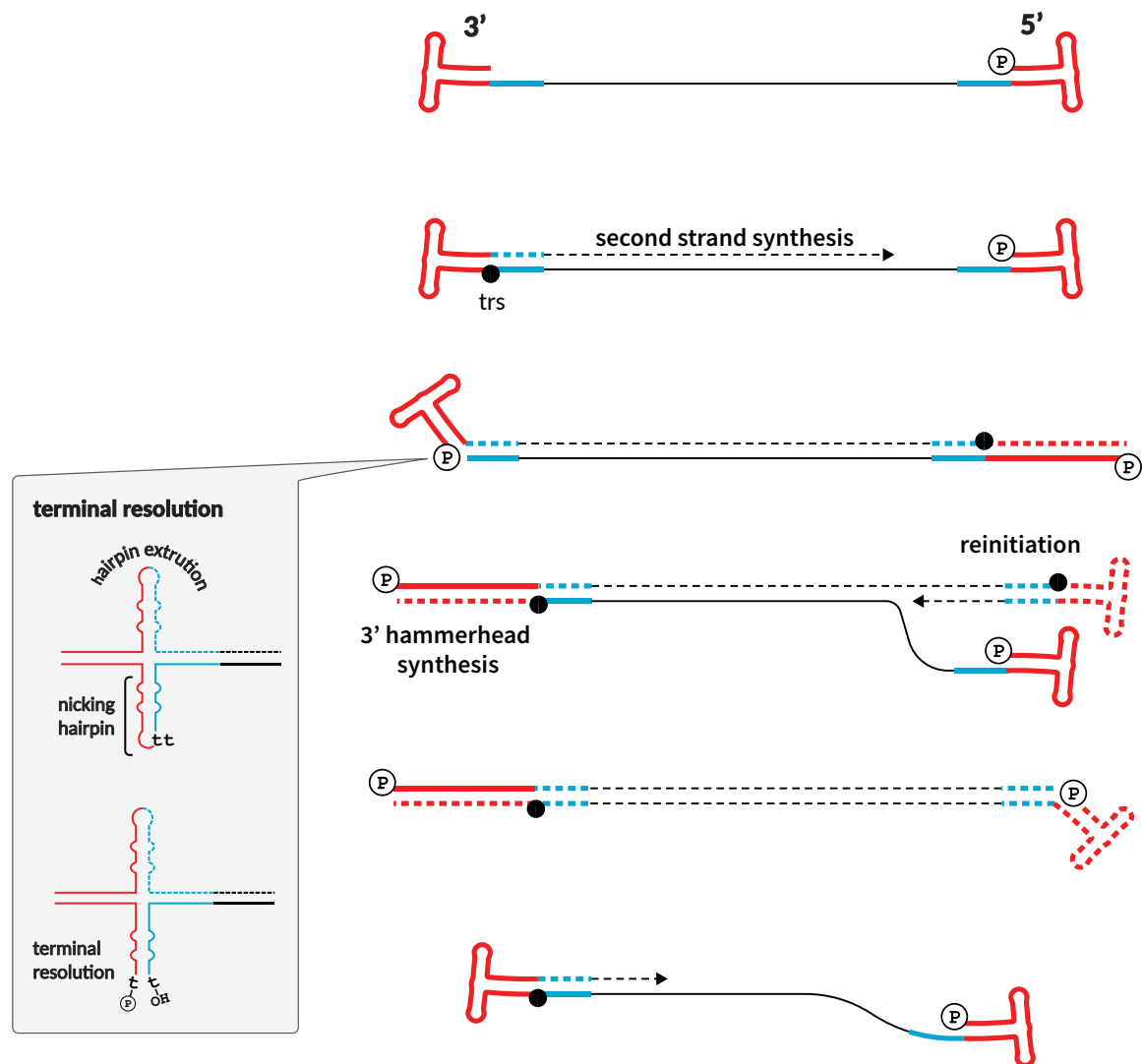


Figure 5.4.1 Replication of (r)AAV genomes via RHR | Homotelomeric parvoviruses such as AAV replicate via a unidirectional mechanism known as rolling hairpin replication (RHR). The self-annealed 3' terminal hammerhead acts to 'self-prime' the genome for second-strand synthesis. The detachment of the daughter strand occurs at a site at the A'D junction called the *terminal resolution site* (trs), and is dependent on interactions of the large AAV replicases (Rep68/78) with the hammerhead stem of the 3' ITR, further facilitated by additional interactions with the 3'-most hairpin sequence. Helicase activity is required on the on the 3' end of the *de novo* strand in order to separate the open-ended ITR duplex and reinitiate the process (an activity likely provided by the large AAV replicases).

The small *rep* products Rep40 and Rep52 are important in the encapsidation of rAAV genomes, where their helicase activity serves as a molecular motor to translocate the ssAAV strand into unoccupied capsids²⁰⁵. The D sequence of the ITR (but not the D' sequence, which is adjacent to the 5' ITR hammerhead) acts as packaging signals in this process, as evident in the ability to package genomes of a single polarity by ablating the packaging sequence of one ITR^{115,206}.

There is a small amount of direct research, a reasonable amount of indirect research, and very little clarity on the role of ITR sequences in rAAV-mediated transduction. It is tempting to

conceive of ITRs either as an inert vestige of the parental virus or the molecular ringleader of Rep-deficient latent infection pathways. In my own experience at least, neither of these notions are shortcuts to insight. It is notable, as Chen et al. have remarked²⁰⁷, that naked DNA has the potential to multimerise without the involvement of viruses, and that (r)AAV is far from the only virus known to persist latently in a chromatinised form¹⁷¹. The word ‘transduction’ itself is generally overburdened, which can make a synthesis of related research challenging. Its implications can depend distinctly what is meant and how (and when) it is measured. To non-experts, it may seem as though the desired phenotype is all but guaranteed once the genome is released into a host nucleus. Even more pervasive is the assumption that the phenotype observed in the days post administration will persist stably in the following months. In reality, both of these outcomes depend on post-delivery processes whose mechanisms we have little knowledge of, let alone knowledge generalisable between tissue types or vector genomes of variable quality. The remainder of this section will focus on paths toward the stable new phenotype that occur after release from the vector capsid (though should not imply that the capsid ceases to influence the process at the point of genome release^{208,209}).

Depending on the capsid, vector genomes delivered to the nucleus are steadily uncoated (for up to two weeks)^{210,211} producing a detectable transgene signal within 24h of administration, which rises slowly and reaches a steady state level on the order of 5-8 weeks in mice^{211,212}, concomitant with the appearance of highly stable double-stranded, predominantly circular (1 to ~12)mers of the vector genome. Linear dimers and circular monomers (products of a single joining event) are also detectable at the 24h timepoint, as judged by restriction analysis of low-MW DNA extracts²¹³ and join-dependent eGFP expression²¹⁴. I have identified only one study that has examined chromatinisation *in vivo* (in skeletal muscle of non-human primates 22 months post administration)¹⁷⁸, which found that the vector genomes were associated with histones. Those authors also found that expression of the transgene product was not unaffected during a three-month daily HDAC inhibitor treatment (initiated 5.5 months p.a.), contrary to observations in immortalised cell lines¹⁷⁸. To the authors, this indicated that unlike in immortalised cell lines, the episomal chromatin structure present in NHP skeletal muscle was constitutively permissive to transgene expression.

It is accepted that dsDNA is a prerequisite for transcriptional activity, though there has been debate (c.f. ^{206,215,216}) over whether transcriptionally competent vector genomes are generated via hybridisation of opposite-sense strands or *de novo* second-strand synthesis primed by the 3’ITR

after genome delivery. The latter is generally accepted since there is no clear difference in performance between single-sense and mixed-sense vector batches, or in their response to inhibition of factors known to reduce rAAV second-strand synthesis²⁰⁶, and since scAAVs result exhibit higher expression kinetics²¹⁷. This evidence is certainly compelling though to my knowledge the two possibilities have not been investigated with a directly comparable experiment, and confirmation of opposite-sense duplexes acquired by methylation analysis and restriction recognition site mismatch detection²¹⁶ have not been fully reconciled with the prevailing view. It is possible that opposite-sense strands that are free to anneal will do so to some extent, but that their contribution is either minor or context-specific.

Far less assured is if or how either process places vector genomes on a path toward long-term residence and transcriptional activity. Successful transduction is limited in the early stages by poorly understood processes of vector genome stabilisation (or avoidance of genome degradation) and is not determined solely by generation of dsDNA. In the murine liver, hepatocytes confirmed to have received exogenous genomes and converted them into a transcriptionally active form can substantially outnumber those that go on to measurably express the transgene eight weeks after administration²¹⁰. It is also the case that ssAAVs and scAAVs alike are susceptible to MRN-mediated failure to transduce, despite the latter arriving in the cell already in duplex form¹⁸⁹.

As already mentioned, (r)AAV genomes have a close relationship with proteins associated with cell-specific DDR pathways^{119,218} – a fact common to parvoviruses in general¹⁵⁵. DDR pathways are relevant to the persistence of rAAV genomes in transduced cells, in addition to their effects on lytic drive in vector production^{119,190}. There are strong indications that the presence of ITR termini strongly influence paths of multimerisation in a way that would be uncharacteristic of DNA without ITRs. In 2005 it was demonstrated that closed ITR hammerheads and open (blunt-ended) ITR duplexes formed by complete second-strand synthesis are distinct recombination substrates, and that intermolecular joins between closed-closed, closed-open and open-open ends of two linear molecules were distinct processes with varying efficiencies and susceptibility to topoisomerase I inhibition²¹⁴, and which require different host enzymes depending on dividing and nondividing cell status¹⁹⁰.

The hammerhead structure of the ITR, as opposed to a simple U-shaped hairpin, is also important in the involvement of ataxia telangiectasia mutated (ATM)-dependent DDR

pathways in vector genome processing²¹⁹. The hammerhead is required for the Holliday resolvase-like activity of the protein Artemis¹⁹⁵. In mice this protein cleaves the hairpin at diametrically opposed sites of the three-way junction – an activity apparently required in skeletal muscle, but not in the liver, to prevent the accumulation of unit-length “no-end” genomes with closed hammerheads at each terminus¹⁹⁵. Artemis shares an upstream PI3-K family transducer (DNA-dependent protein kinase) with Ku proteins, which are involved in non-homologous end-joining and are known to interact with Rep78¹⁹⁶ and partially substitute for the helicase activity of the MCM complex in AAV genome replication¹⁹⁷. Ku proteins have been shown to interact specifically with ITRs in the hammerhead conformation (as opposed to the linear duplex) to allow replication of rAAV genomes in producer cells¹⁹⁶.

It is sensible to assume that primary sequence variation within the ITR, which has accompanied the evolutionary divergence of rep and cap variants, confers a fitness benefit within a relevant physiological/zoological niche. One example is that RFX family transcription factors decrease rAAV transduction through an interaction with the single-stranded D sequence – an interaction with specific nucleotide requirements met by ITRs of serotype 2 and serotype 1, but not by serotypes 3,4 or 5²²⁰. The D sequence also contains a site with homology to a cellular NF- κ B-repressing factor (NRF) binding site, which, when removed, leads to improved transgenic GFP expression in mice²²¹. The D sequence is host to a number of other observed protein interactions including FKBP52, which exerts phosphorylation-dependent inhibition of second-strand synthesis in the early stages of transduction^{187,222,223}. (In the context of plasmid ITR integrity, we may note that the D sequence appears less vulnerable to truncation than other ITR subregions, and that D-deficient genomes may be at reduced risk of affecting vector performance due to their poorer packageability. Nonetheless, these features warrant mention).

Unfortunately, we are less able to judge the impact of primary sequence variations on transduction due to a near singular reliance on the canonical AAV2 ITR in most of AAV vectorology (though notably the MRN complex also binds to a snake ITR and inhibits transcription¹⁸⁹). Empirical evidence can be seen in the variable expression kinetics observed in a range of *in vitro* tissues dependent on ITR serotype^{103,224}. One recent study observed 0.25-1.25 fold variations (vs ITR2) in luciferase expression from promoterless transgenes flanked by ITRs from AAV serotypes 1 to 7 (sans AAV5) 48h post-transduction in a variety of adherent cell lines, with ITR7 exhibiting a degree of tissue specificity *in vitro* at this timepoint²²⁵. The differences were attributed to variations in the inherent (albeit modest) promoter activity of the ITRs²²⁶,

though this should be viewed as conjecture as the mechanism was not investigated. Differences in long-term transduction observed between ITRs from AAV1 through AAV6 have been less pronounced *in vivo*²⁰⁴ though the 2-4 fold differences nonetheless on display are substantial in a clinical setting. The significance of the two hairpin primary sequences (BB' and CC') is thus far difficult to judge in isolation though their significance in the production setting is clearer at present since a motif in the BB' hairpin is required for Rep40-mediated genome encapsidation²²⁷. The successful use of heterogenous ITRs to guide the orientation of split vector recombination²²⁸ demonstrates that the sequence of the ITRs plays a role in genome processing, at least insofar as homology between ITRs can drive paths of intermolecular recombination.

5.5 The ITR instability problem

5.5.1 What is the ITR instability problem?

The need to clonally amplify rAAV genome sequences to produce transfection material underlies all rAAV vectorological research and applications. Historically, prokaryotic plasmids containing the ITR-flanked dsAAV vector sequence have acted as surrogate latent AAV genomes in mammalian producer cells. During propagation in prokaryotic cloning strains, one or both plasmid ITR sequences are prone to *de novo* deletions and subclonal proliferation of the truncated variants. The consequence is that plasmids used to produce rAAV vectors are almost always heterogeneous in one or both ITR loci (Fig. 5.5.1). Conveniently, each inverted repeat of the AAV2 ITR contains a unique restriction recognition site, allowing us to observe the consequences of this phenomenon in rAAV plasmids' heterogeneous restriction products. Figure 5.5.2 shows three plasmid 3'ITR variants clonally derived from copies lacking different elements of the ITR sequence. Products like these are simple to identify and exclude via colony screening. The much more complex problem is represented in the leftmost lane. The mix of 3'ITR variants includes the intact WT sequence (confirmed via sequencing in Chapter 6), indicating that the original copy was itself intact (or else it arose spontaneously) while other subclonal variants are products of sequence instability. An AhdI or XmaI test can confirm the presence or absence of the intact ITR sequence in the plasmid preparation, as can Sanger sequencing, but it is clearly not the only sequence present. The fraction of intact ITRs in intact-ITR-derived populations ('ITR integrity') is a continuous variable. Thus far no methods for

reliably quantifying this variable have been reported, meaning it has never been included in published or unpublished research.

A small number of experiments have been reported assessing whether the ITR ‘intactness’ may influence outcomes of vector production and performance. Zhou and colleagues²²⁹ have experimented a number of different rAAV transfer constructs lacking BB’ and CC’ arms in both ITR loci of the plasmid (forming U-shaped AA’ hairpins). Compared to a plasmid with two wild-type ITRs (clonally speaking), This ITR variant resulted in ~75% reductions of packaged rAAV-DJ vectors from adherent HEK293 productions, as judged by qPCR quantification and DNA dot blots. Interestingly vectors prepared from double- Δ BB’/CC ITR plasmids lead to a higher response in a lower dose range than wild-type-ITR-derived vectors in several adherent cell lines (%GFP-positive cells) and in C57BL/6 J mice (blood luciferase over 100 days). Savy and colleagues²³⁰ have observed that restoration of 14bp in the AA’ sequence and switching the internal orientation of the BB’ and CC’ hairpins (characteristic of pSUB201-derived rAAV plasmids²³¹) led to an increase the ultracentrifugal peak ratio of full:empty capsids in an insect cell production system.

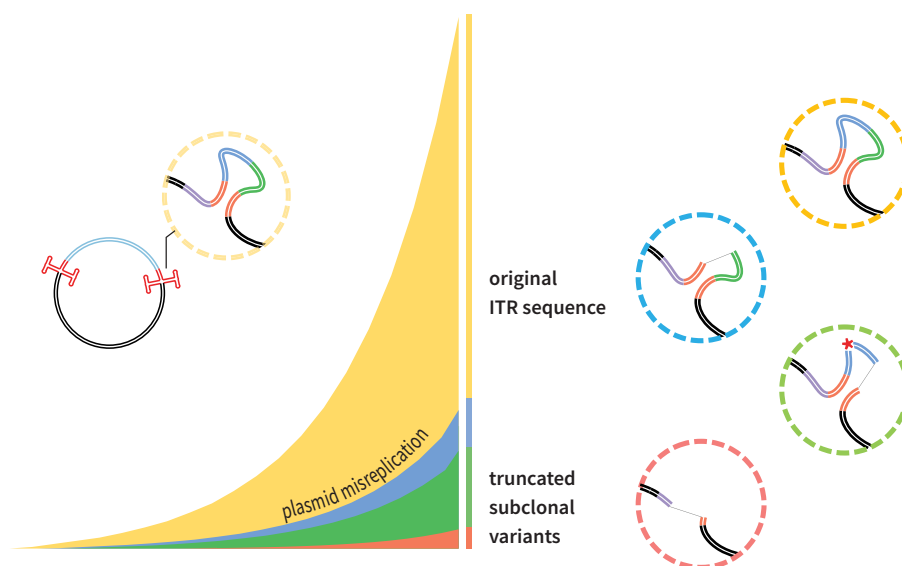


Figure 5.5.1 Emergence of plasmid ITR heterogeneity in the course of replication | Plasmid ITR sequences are propagated in prokaryotic culture in the preparation of transfection materials used for production off rAAV vectors. ITRs are prone to sequence instability in bacterial replication, leading to sequence deletions and expansion of truncated 3'ITR variants in the plasmid preparation. The result is plasmid preparations that are heterogenous in one or both ITR termini.

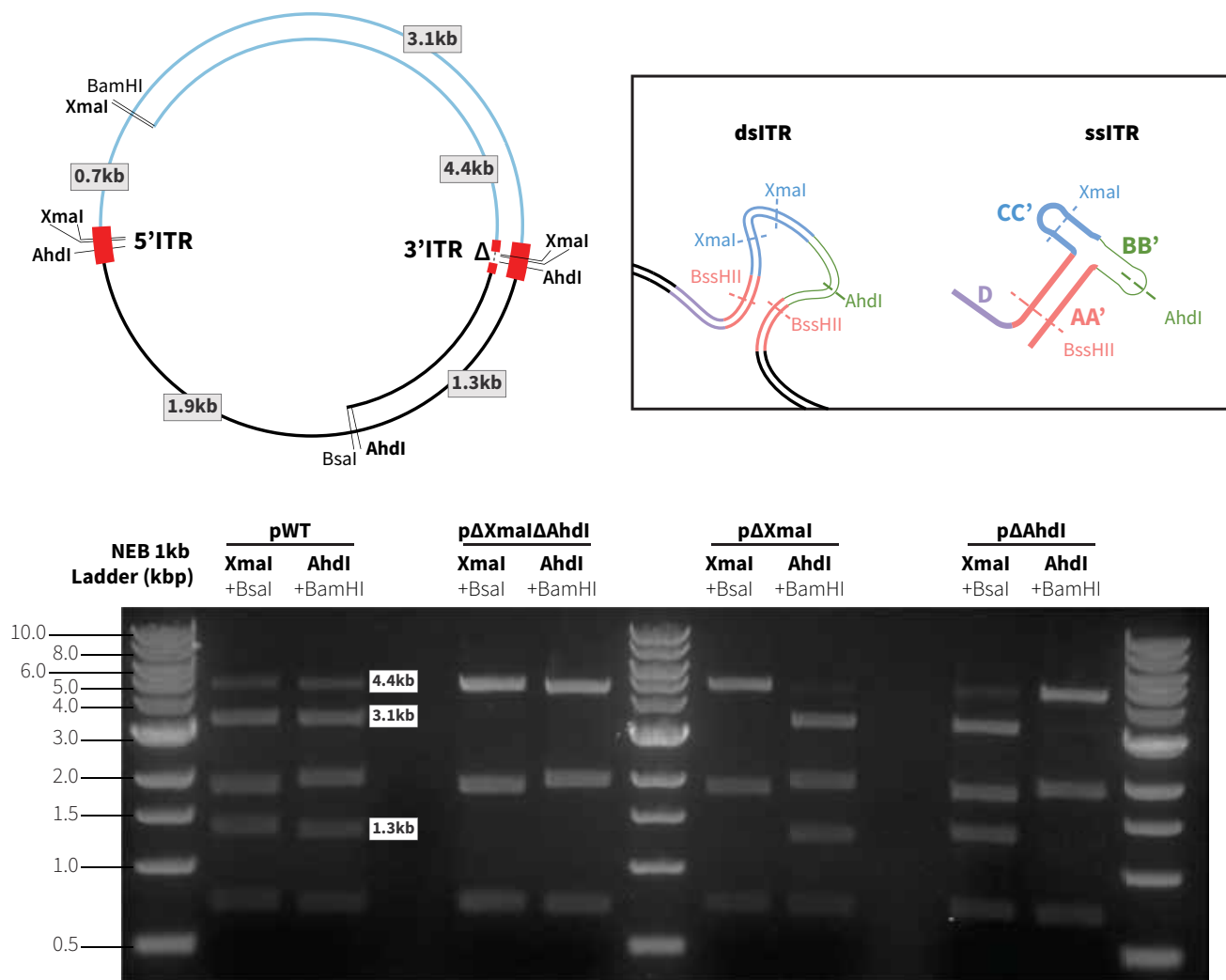


Figure 5.5.2 A routine analysis of plasmid 3'ITR heterogeneity using restriction recognition sites within the ITR

This gel image was generated in the experiments reported in Chapter 3 and is shown here for ease of reference. A lighter stroke weight is used for distinction between the AA' and BB' hairpins by readers with colour vision deficiencies and has no other significance. Heterogeneity in plasmid ITR sequences is observable in partial or total failure of restriction enzymes such as XmaI or AhdI to cleave the ITR locus. In the present case, failure of the 3'ITR to cleave results in a 4.4kb band (uppermost) and failure of the 5'ITR to cleave results in a 2.6kb band (not observed). The leftmost band contains a partial population of AhdI-positive and XmaI-positive copies most likely indicative of an intact ITR, indicating that the colony on the whole was derived from an intact copy.

The mechanisms of dsITR replicative instability remain unclear, and none have yet been suggested that concur with all observed phenomena. The problem is usually attributed to nonduplex-forming elements^{118,232}. This is very likely correct, given that long inverted repeats in *Escherichia coli* have been closely associated with replicative and non-replicative sequence instability^{233–235}, are susceptible to both protein-dependent (notably SbcCD and RecA) and protein-independent recombination²³⁶, and are known to be sites of structure-specific protein interactions in general^{237,238}. Importantly however, neither the inverted repeats of plasmid ITR sequences, nor their compound hairpin structure, nor their GC content, are sufficient to prevent their faithful replication in standard *E. coli*-derived cloning strains such as DH5α. This

is evident in the tendency for one of the two plasmid ITR sequences to be replicated with no less apparent stability than the rest of the plasmid (typified in Figure 5.5.2). It also does not account for the high degree of stability observed in the four-way (A-D') junction sequence relative to the BB' and CC' hairpins, at least in the present thesis. Clone-dependent emergence of certain truncated ITR variants is also possible, meaning that results observed in one outgrowth may not be representative of other outgrowths. This is illustrated in Figure 5.5.3, where two clones of an rAAV plasmid plated from the same mix of plasmid transformants exhibited a 'preferred' stability in different hairpin sequences of the 3'ITR. All too recently, both plasmid co-transformation and clonal mosaicism in outgrowth have been demonstrated in bacterial cloning²³⁹, calling on scientists to revisit the simpler model of monoclonal colony expansion suggested in Figure 5.5.1. This phenomenon may help to make sense of the result shown in Figure 5.5.3 and add a dimension of complexity to the question of plasmid ITR instability, assuming that mutations, like co-transformation, can act as a source of mosaic heterogeneity.

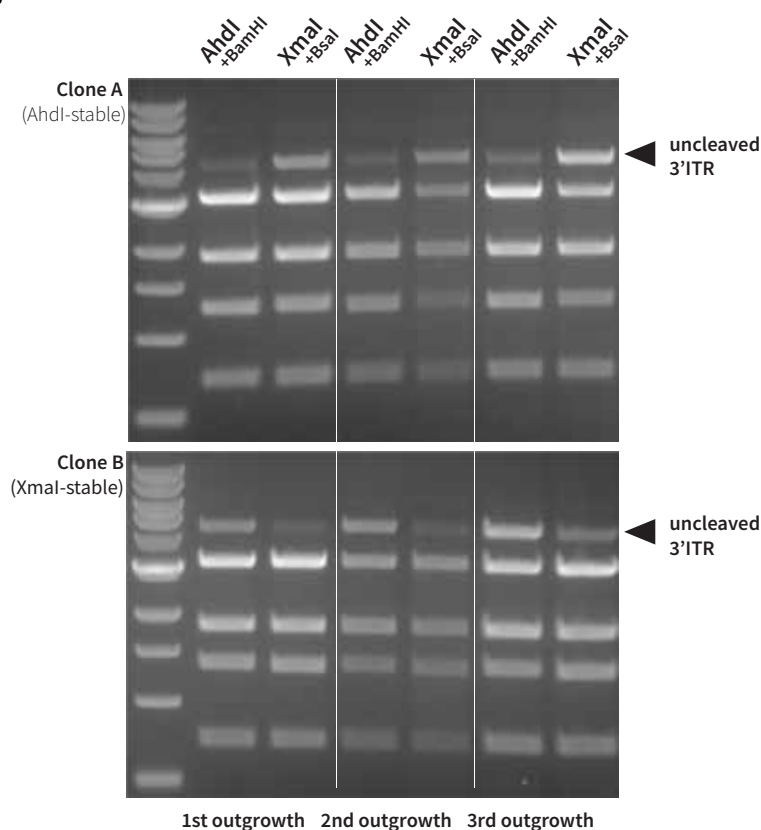


Figure 5.5.3 Differences in the apparent stability CC' or BB' arms between clonal outgrowths DH5- α cells were transformed with a wild-type-derived rAAV plasmid (as shown in Figure 2.5.2) and plated colonies were expanded in liquid culture and purified for restriction analysis. Clone A and Clone B showed dominant tendencies in retention of the BB' hairpin sequence or CC' hairpin sequence respectively and were subjected to serial outgrowths in 250mL liquid culture, through which the clone-specific trends were observed to persist. This illustrates the potential for interclonal variation in ITR truncation pathways, which adds a layer of complexity to the question of why plasmid ITRs sequences are challenging to stably propagate.

5.5.2 Dull, elusive, and non-fatal

It is reasonable to assume that all scientists with any prolonged experience in (r)AAV development are familiar with issue of plasmid ITR integrity. Despite this shared awareness, a scholarly review of the subject is greatly limited as there are few published experiences of the phenomenon to speak of. By and large, the discourse around this topic is idling in the realms of anecdote, community memory, and unshared commercial knowledge. For the duration of this doctorate, I have not become aware of any academic articles that include restriction analysis of plasmid ITR heterogeneity out of concern for reproducibility, other than to state that it was performed and that the wild-type ITR sequence was detected. I suggest three interrelated factors to help make sense of this. The first is conjectural and won't be discussed at length – that the 'dry' technical and analytical hurdles necessary to make progress do not serve the goals and predilections of most scientists who are drawn to the field of gene therapy. Second, those with an interest in tackling the problem have not been technologically or methodologically equipped. The third is an implicit reliance on 'ITR repair'.

We currently lack the methodological infrastructure necessary to tackle the problem in earnest, since its component techniques are either lacking or only presently emerging. Sanger sequencing services optimised for rAAV ITRs have become available in recent years from GenScript Biotech and Azenta Life Sciences (formerly GENEWIZ). This development is sorely needed and extremely useful but reports only the DNA sequence with a large enough majority to produce a coherent electropherogram trace. Its readout, like restriction analysis, is essentially categorical. Encouragingly this situation is quickly shifting with the emergence of 'next-generation' technologies capable of providing copy-by-copy sequence information and thus fulfilling the population analysis requirement. Specifically, long-read sequencing platforms such as single-molecule real-time (SMRT) sequencing provided by Pacific Biosciences Inc.^{240,241} and nanopore sequencing provided by Oxford Nanopore Technologies Ltd^{242,243} have shown promise in the characterisation of packaged rAAV genomes, which in numerous respects are a far more challenging analytical target than plasmid DNA. A metrological evaluation of these platforms for ITR analysis is particularly important. This is the true in a generic sense in the interest of scientific and clinical standards²⁴⁴ and made even more true by the challenges intact ITR sequences pose to enzyme processivity. Bridge amplification and sequencing-by-synthesis chemistry (provided by Illumina Inc.) has been utilised but is very poorly suited for population analysis of ITR-containing DNA fragments^{118,232,245–247}.

Aforementioned population-sequencing based approaches are conceptually well aligned to the problem and promising candidates for formal validation, though quantitative analysis of packaged ITRs remains an outstanding challenge. The Illumina format, as mentioned, is poorly suited to ITR-containing sequences. The need to sequence ITRs via strand extension in preparation of the library is problematic in and of itself, but the greater issues are the processivity of ITR regions in the flow-cell stage and the fact that ITRs present nested levels of short-read alignment ambiguity (see Appendix G). SMRT sequencing, a long-read format that utilises a virally-derived DNA polymerase, has demonstrated promising credentials for sequencing ITRs, particularly in self-annealed²⁴⁰ and close-ended genomes²⁴⁸. A challenge for this format with direct implications for ITR analysis is the requirement that ssAAV genomes be hybridised to generate blunt-ended termini compatible with the SMRT system. Intrinsic to the problem of ITR instability is the potential for heterogenous ITRs in ssAAV genomes (and thus termini with mismatched lengths and/or sequences), which could feasibly affect the possibility of being detected.

To understand the impact of ITR heterogeneity, there is an equally relevant demand for robust analysis of outputs as well as inputs. That is, products generated downstream from plasmid transfection, as well as the plasmids themselves. A salient example is our ability to quantify packaged vector particles with PCR-based methods (either dPCR or qPCR), which remains an area of active research and has thus far achieved unfavourable precision compared to simpler targets like chromosomal DNA^{8,249–255}. Other important attributes in rAAV production are full-to-empty capsid ratios and homogeneity of packaged DNA. Prospects for measuring full-to-empty ratios here are more favourable, though the most reliable techniques are dependent on scanning electron microscopy or analytical ultracentrifugation, to which access can prove an impediment for the literature on the whole²⁵⁶. Compounding this need are contextual factors previously discussed in this review. One is the degree of process variability within and between laboratories^{95,257,258}, and the fact that the capsid variants, their genetic cargo and their production systems are highly heterogenous. With techniques that are unstandardised, imprecise or yet to be fully validated, timely and durable contributions to the question of ITR will be a highly challenging. Another factor is the strain that current manufacturing techniques place on the supply of rAAV medicines. In this context, incremental gains or losses can be clinically impactful, especially where ITRs are concerned, given they are the basis of all current systems.

The phenomenon of ‘ITR repair’ has limited the urgency, real and perceived, behind the ITR problem. Since the 1980s it has been known from acrylamide gel electrophoresis experiments that the full-length ITR content of packaged rAAV vectors can be higher if not total compared to the plasmid population, including preps with no full-length ITRs in one loci^{111,259,260}. The term ‘ITR repair’ was used to describe this phenomenon and persists to the present, though AAV researchers should be fully aware that a concerted sequence restoration process is not implied. The ‘repairable’ deletions observed indicate strongly that the process is dependent on the mode of genome rescue and the presence/accessibility of an unmutated template elsewhere on the strand. Another possible factor is the kinetic disadvantage of viable-but-defective ITR mutants, which would have a desirable homogenising effect on the packaged vector population.

While ITR repair is certainly helpful to users of rAAV, is it far from a panacea for plasmid ITR instability. Perhaps most obviously, there has never been an implication that ITR instability would not affect production yields. Without a firm sense of context, readers could easily infer otherwise from the language used in keystone from the 1980s. At that time, vector yields five-fold and lower than intact plasmid transfections were judged to be “quite high”¹⁸⁰, obviously not in manufacturing setting, but in a context where the phenomenon per se was new and remarkable knowledge. The emergence of intact vector ITRs from plasmids with truncated ITRs, rather than something to be scrutinised quantitatively, was unambiguously welcome news for a vector system whose potential as a human medicine had not yet solidified.

The possibility of replicating and encapsidating incomplete ITR sequences has not been excluded, theoretically or experimentally. On the contrary, a recent study utilising SMRT long-read sequencing observed that not only are mutant ITRs packageable, but their packaging can differ between producer systems (having occurred more prominently in an insect cell producer system than in HEK293 producer cells²⁴⁸). Further to this, mutant ITR sequences were observed most prominently in incomplete genomes (with sedimentation coefficients between those of ‘empty’ and ‘full’ capsids in analytical ultracentrifugation) as well as in highly truncated genomes that sedimented within the ‘empty’ capsid fraction. There are multiple, significant implications to these observations. This speaks firstly to the different impacts that ITR instability might impart depending on the production system. As the authors also highlight, the truncated genomes packaged alongside mutant ITRs are capable of eliciting TLR-9-mediated immune targeting of transduced cells, but incapable of contributing to transgene expression. In terms of analysis, the presence of partial vector genomes implies a potential disconnect

between q/dPCR titres and the performance of the administered vector, assuming that at least some partial genomes contain the PCR target sequence. Aside from the producer cell system, this error source would depend on the purification system. Isopycnic fractionation can (i.e. CsCl ultracentrifugation) can separate full and empty capsids due to differences in their sedimentation coefficients, but is not the case for other highly utilised methods such as iodixanol-fractionation and tangential-flow-filtration.

5.6 Conclusions

DNA plasmids bearing rAAV vector sequences are the molecular precursors to ssDNA vector genomes for the vast majority of clinical and laboratory vector production. The inverted terminal repeats located at each terminus of the rAAV genome sequence are necessary for the production of packaged vector capsids from plasmid precursors, as they are central to numerous intracellular genome processing activities including rescue, replication and encapsidation of the rAAV genome sequence. To be generated in useable amounts, rAAV plasmids are propagated in prokaryotic hosts, typically derivatives of *Escherichia coli* such as DH5- α . Due to replicative instability, this process results in *de novo* deletions and subclonal expansion of plasmids bearing incomplete ITR sequences. Mutations in the plasmid 3'ITR appear most frequently to involve loss of secondary-structure-forming elements, such as the 21bp BB' or CC' hairpin sequences. Vector production systems have demonstrated a tolerance to such mutations, and the seeming intractability of the problem has led to the accepted use of heterogenous ITR sequences rAAV plasmid preparations. As yet there has been no systematic investigation into plasmid ITR heterogeneity as an upstream determinant of vector yield or vector performance. This is partially because we lack the well-developed toolkit needed to measure ITR heterogeneity, the yield of packaged vectors, and salient vector quality attributes.

Therapies based on rAAV vectors have well and truly expanded beyond the laboratory and into the landscape of human medicine, where their potential to thrive is limited predominantly by difficulties generating large, consistent, and affordable volumes of clinical-grade products. In 2023, technological hurdles are the only legitimate impediment to addressing the poorly understood question of ITR heterogeneity. The task ahead is challenging and from its outset must take into consideration the variability of production systems and their outcomes, and the significance of stepwise gains and losses in vector yield and quality.

Exploring the impact of ITR replicative instability in the rAAV vector system

Aim 2a To assess whether plasmid ITR heterogeneity impacts the yield of rAAV vector products.

Hypothesis 1 (Section 6.1) Mixtures of plasmid 3'ITR reporter variants and their derivatives in HEK293 cell transient transfection will have the same barcode compositions.

Hypothesis 2 (Section 6.1) When the fraction of intact plasmid 3'ITRs are varied, fold outputs of the total batch will be equal relative to an exogenous reference.

Aim 2b To assess whether plasmid ITR deletions impacts the transfer and expression of exogenous transgenes in murine liver and skeletal muscle

Hypothesis 3 (Section 6.2) The reporter composition of an rAAV9 vector derived from barcoded 3'ITR plasmid variants will be unchanged between the injected mixture and nucleic acids extracted after one and eight weeks.

In this chapter, a novel qRBA-based assay of rAAV production efficiency is employed to study the influence of ITR integrity. Core to this study are four subclonally isolated 3'ITR variants of rAAV vector plasmid, each derived either from an intact wild-type copy or one of three unique 3'ITR mutants. Each of these plasmid preparations contained hexanucleotide reporter barcodes

in a 150bp PCR target region, allowing their relative abundance in mixed samples to be measured using amplicon population sequencing (qRBA). Samples analysed by qRBA include the plasmid inputs themselves, plasmids in transfected producer cell nuclei, products of rescue and replication, packaged vector genomes, vector DNA in murine tissues and vector RNA in murine tissues. The latter two sample types were analysed one week and eight weeks post-administration. These sample types were modelled as inputs and outputs of production and gene transfer processes and changes to the composition of barcoded ITR variants allowed their relative fold outputs (RFOs – see Chapter 3) to be measured.

A pair of possible reporter barcodes were prepared for each 3'ITR variant, with the intention of mitigating barcode-specific measurement error. These partnered barcodes were ligated into to the recipient plasmid, transformed, expanded, and processed in the manner of single plasmid preparations. This represented a *partnered* reporter system (see Fig. 4.1.1), in which each plasmid copy contains one of either partner (not both). Partnered reporter barcodes for each 3'ITR variant were designated *a priori* as α or β . A fifth version of the rAAV plasmid was cloned and used an exogenous reference in Section 6.1. This plasmid was used to transfect cells in parallel with the experimental transfections. Experimental samples were then spiked with resuspended 'reference cells' (of the equivalent sample type) at the point of harvest, providing a point of mutual comparison for the total size of each experimental reporter mix.

The scope of this project reflects the present context in which our understanding of ITR integrity is underdeveloped, and updates to the consensus are made difficult by heterogenous production systems, variable outputs, and weaknesses in the overall analytical infrastructure. This study is intended to produce robust evidence that will inform rAAV researchers whether ITR heterogeneity can act as an upstream determinant of vector yield and performance.

Before proceeding to the main report, I will pre-emptively outline how certain terms will be used in this context and specify the vector genome sequence under study. '*ITR integrity*' will refer specifically to the proportion of rAAV plasmids containing the intended ITR sequence (in this case the wild-type) in both loci. '*Wild-type*' refers to the untruncated ITR sequence of AAV serotype 2. DNA or RNA molecules **barcoded as 'WT' (two capital letters only)** share a plasmid-subclone-of-origin. Though monoclonally derived from an intact copy, this plasmid preparation (and the others) are understood to be heterogenous, as per the nature of the research problem. Due to the phenomenon of ITR repair, the downstream products of the

non-WT plasmids may possess two wild-type ITRs, even if none existed in the plasmid. The experimental design is agnostic to the presence or proportion of wild-type ITR sequences in the products of HEK293 transfection. '**Accumulated**' **vector genomes (vgs)** will refer to the producer-cell-derived copies of the rAAV genome generated by intracellular processes upstream from encapsidation. The methodology used here does not permit a distinction between genomes propagated via rolling hairpin replication and the immediate (or intermediate) products of genome rescue.

This report calls on us at times to distinguish between a single plasmid molecule and a picomolar quantity of that molecule (1s to 100s of μg). Both of these could be referred to as '**a plasmid**' in common parlance. To avoid ambiguity where necessary, the latter may be explicitly referred to as a **plasmid preparation**, a **preparation** or more simply a '**prep**'.

Like 'infection', the word '**transduction**' is non-specific and can be broken down into a highly complex succession of steps, only one of which is the release of the packaged genome into a host nucleus. A lexicon that clearly delineates separate stages of transduction has not yet been established. For this Chapter, rAAV vector-mediated '**gene transfer**' will be taken to mean the non-transient residence of exogenous genes in cell nuclei, resulting from their physical transport into the nuclei by an rAAV vector.

The rAAV vector plasmid (aka transfer plasmid) used in this study, **pGL_ITR2.Ch6** contains an ampicillin resistance cassette and a pUC high-copy-number origin of replication. The vector genome sequence is cloned with a flip 5'ITR and a flop 3'ITR. The transgene cassette contains **tandem eGFP-Luc2 open reading frames** separated by a T2A ribosomal skipping sequence such that the two proteins are translated from the same mRNA copy. The genes are driven by a Rous sarcoma virus (RSV) promoter and **the reporter barcode is located between the Luc2 open reading frame and a bovine growth hormone polyadenylation (bGH-polyA) signal**, meaning that mature messenger RNA will contain the barcode. Note that because transcription of the *egfp-T2A-luc2*-barcode-bGH-polyA RNA sequence is driven by a single RSV promoter, we can infer that transcripts containing the reporter barcode originate from rAAV genomes with a complete sequence from RSV to polyA. **The proteins eGFP and Luc2 have no role in this experiment.** Sections 6.1 and 6.2 utilise the capsid variants AAV-DJ and AAV9 respectively.

rAAV sequence (flip_5'ITR)-RSV-egfp-T2A-luc2-barcode-bGH-polyA-(flop_3'ITR)

plasmid backbone ampicillin resistance; high-copy-number pUC ori

6.1 Plasmid ITR heterogeneity and rAAV production in adherent HEK293T cells

This study sought to compare outcomes of rAAV production where the wild-type 3'ITR content of an rAAV plasmid was varied. As the nature of the research problem would suggest, the plasmids used in this study are understood not to contain homogenous ITR sequences. The fraction of plasmids containing wild-type 3'ITRs, though not directly measured, would be directly proportional to the fraction of plasmids barcoded as 'WT'. By diluting the WT prep in a mixture of non-WT 3'ITR variants, the ITR integrity of the plasmid mixture would vary by a measurable factor. Three such mixtures were prepared containing approximately 82%, 40% and 21% WT-coded plasmids according to qRBA measurements. Non-wild-type 3'ITR plasmid variants were identified through restriction screening of plated transformants and were XmaI-negative, AhdI-negative or negative for both restriction sites in the 3'ITR locus. After Sanger sequencing of the 5'ITR and 3'ITR, barcodes were incorporated via simple two-piece ligation using existing barcoded constructs as donors. Each subclonal isolate was coded by two possible 'partnered' reporter barcodes designated *a priori* as α or β (Fig. 6.1.1). Reporter partners were cloned in the manner of a single prep, by mixing the barcoded donor constructs and ligating them to the recipient 3'ITR variant fragments in a single reaction. Bacterial transformants were expanded directly and homogeneity (as to gross structure) of the ligation product was confirmed by agarose gel electrophoresis. With this cloning strategy, partnered reporter plasmids could avoid prep-to-prep variations in ITR composition, topoisomer composition and process-related contaminants such as ethanol or bacterial endotoxins.

Aside from the plasmid inputs, this study analysed three DNA sample types purified from adherent HEK293T cells cultured in 6-well plates: (1) rAAV plasmids in transfected cell nuclei 24h p.t. transfected with no AAV or Ad5 helper genes (2) eukaryote-derived vector sequences 72h p.t. in the absence of a *cap* gene ('accumulated vgs') and eukaryote-derived vector genomes packaged into an AAV-DJ capsid 72h p.t. following a standard transient triple transfection.

These sample types were modelled as inputs and outputs of intracellular rAAV-DJ production, though it is understood the two are not connected physically by the processes in question because they were purified from separate cell samples. Different sample types were targeted for analysis via strategic omission of viral helper elements, selective purification and/or enzymatic treatments such as DpnI (which cleaves bacterially methylated DNA within the PCR target region) and nonspecific nucleases (which preserves only enzymatically inaccessible, encapsidated vector DNA). A pre-PCR MscI digest step was also incorporated to cleave the ITR region from the barcoded PCR target site, to ensure that terminal heterogeneity did not influence PCR amplification efficiency. Transfections were performed in pentaplicate for packaged and accumulated vector genomes, and triplicate for helper-free nuclear samples.

A fifth pair of the barcoded rAAV construct was made and utilised as an exogenous reference in qRBA analysis. The same transfections were carried out in parallel using this plasmid to produce 'reference cells' containing rAAV products equivalent to experimental cells. At harvest, experimental cells were spiked with a defined volume of resuspended reference cells (from an equivalent transfection), providing an external quantity against which separate qRBA samples could be normalised.

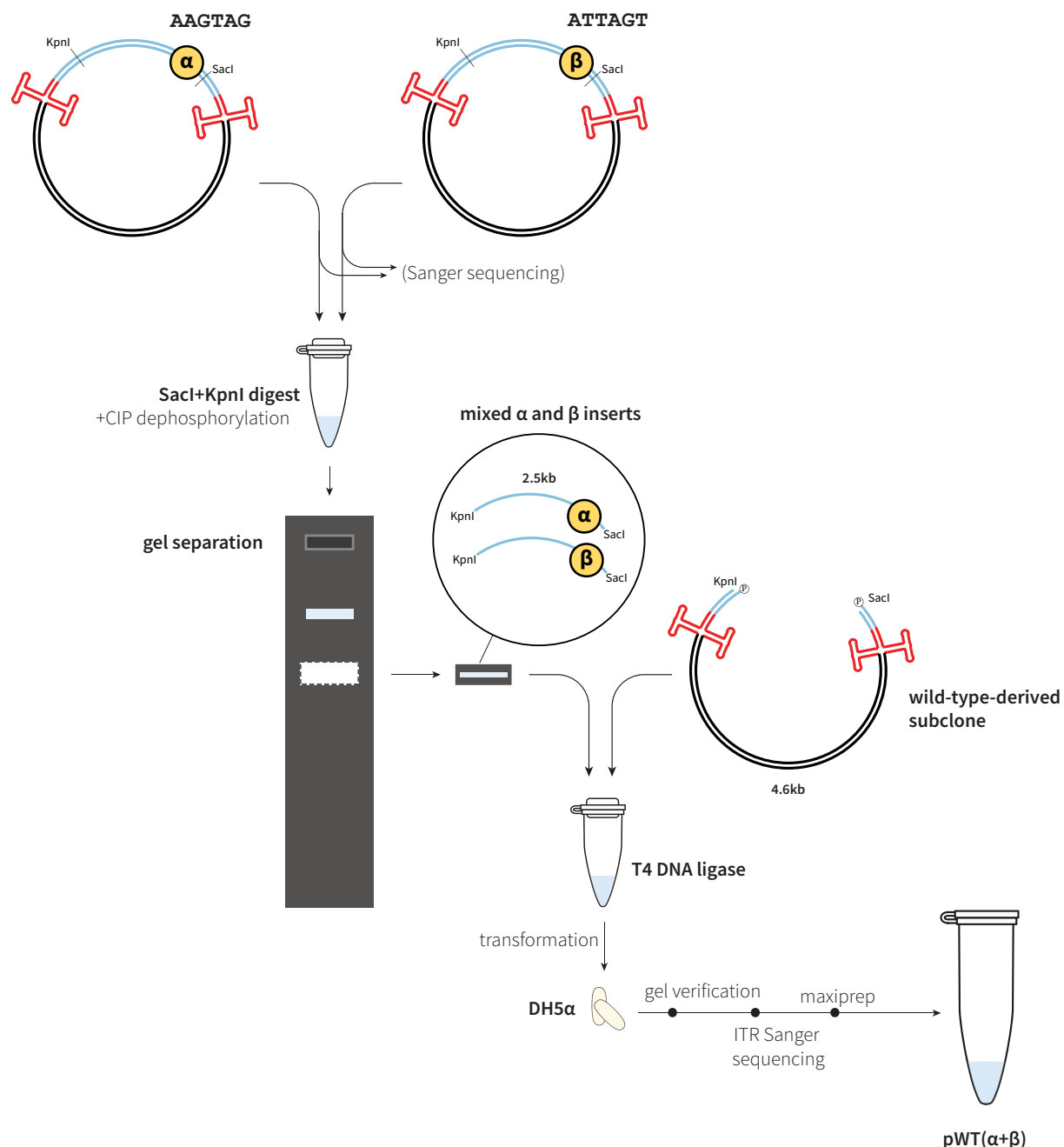


Figure 6.1.1 Cloning of $\alpha+\beta$ reporter barcode pairs into plasmid 3'ITR subclonal variants | Each of the ten reporter barcodes used in this experiment existed in the laboratory of the lead author within identical eGFP-Luc2 rAAV transfer plasmid constructs. The barcode sequence existed within a 2.5kb KpnI-SacI region and could be 'swapped' using KpnI-SacI cloning. To maximise the physical and genetic equivalence between α - and β -coded plasmids with the same subclonal origin, mixed α and β inserts were prepared by combining barcoded donor plasmids in the same KpnI-SacI restriction digest and co-purifying them from the same gel slice. Recipient fragments exhibiting 3'ITR variation were selected based on restriction colony screening of the 3'ITR-wildtype-derived eGFP-Luc2 plasmid. Recipient fragments were combined with the mixed, dephosphorylated α - and β -coded inserts and transformants in liquid LB were cultured directly without plated colony selection. The predominant 3'ITR and 5'ITR sequences were then determined via specialised ITR Sanger sequencing service (Genewiz LLC, NJ) and the gross structure of the plasmid verified by agarose gel analysis.

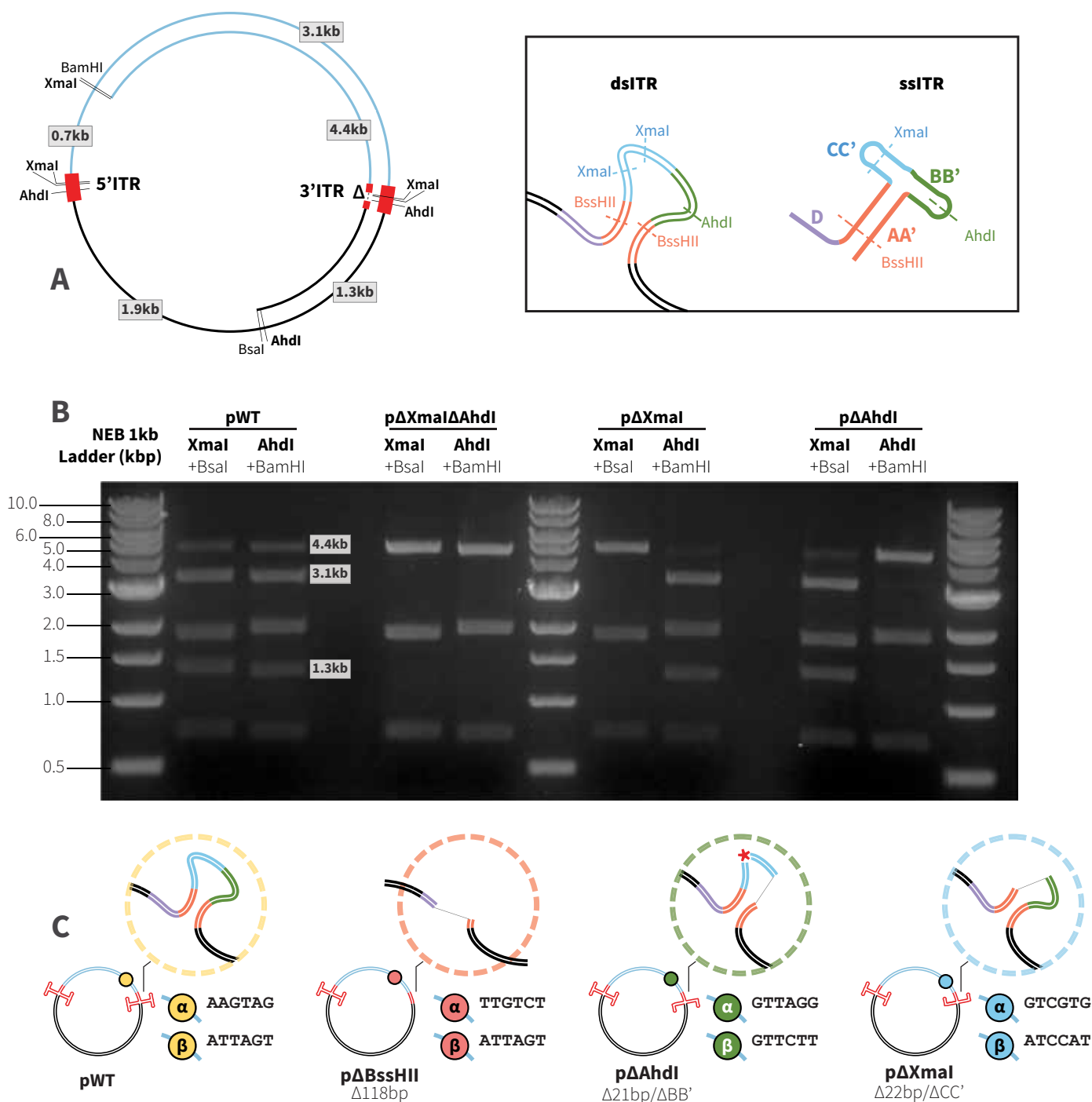


Figure 6.1.2 Subclonal isolation and characterisation of 3'ITR variant plasmids | (A) Restriction sites within specific regions of the ITR allow different truncations to be identified. For clarity, both a dsITR sequence and a corresponding ssAAV vector terminus are depicted. A ~4.4kb fragment of the plasmid can be cleaved into one 1.3kb and one 3.1kb fragment if the restriction recognition site is present, and will otherwise persist as an uncleaved band in the agarose gel. (B) Restriction analysis demonstrates partial populations of AhdI-negative and XmaI-negative 3'ITR variants in the wild-type-derived colony, observable as uncleaved 4.4kb bands. No variants or evidence of heterogeneity was observed in the 5'ITR locus. The ΔAhdI and ΔXmaI subclones exhibit heterogeneity, albeit lesser than the wild-type-derived prep. (C) Specialised ITR Sanger sequencing (Genewiz, NJ) was used to confirm the clonally dominant 5' and 3'ITR sequence in both the wild-type and restriction variants selected from plated colonies. The subclone ΔBssHII was labelled retrospectively based on Sanger sequencing results. pΔBssHII lacked 118bp including the site of resolution between parent and daughter strands in viral DNA replication (the terminal resolution site or trs) and all secondary structure-forming elements. Complete loss of a hairpin sequence occurred in the ΔAhdI and ΔXmaI variants. The red asterisk in the ΔAhdI variant indicates an inversion of the triadenine hairpin loop sequence from AAA to TTT. Each plasmid copy derived from a given subclone was barcoded with one of two possible reporter barcodes of an 'α' and 'β' pair.

6.1.1 Results

6.1.1.1 Subclonal isolation of 3'ITR plasmid variants and preparation of rAAV plasmid mixtures

Restriction screening of 45 plated transformants allowed for the successful isolation of subclones containing no wild-type ITR sequence in the 3' position (Fig. 6.1.2). An AhdI-negative clone was shown via ITR Sanger sequencing (GENEWIZ Genomics, NJ) to contain a 22bp deletion encompassing the complete 21bp BB' hairpin sequence and the central unpaired base. Likewise, the XmaI-negative clone contained a 22bp deletion encompassing the complete 21bp CC' hairpin sequence and the central unpaired base. It also exhibited an inversion of the triadenine hairpin loop sequence at the centre of the BB' sequence (that is, B'-aaa-B to B'-ttt-B). The BB' hairpin in this variant is therefore reverse complementary to the original, indicating that a recombination event occurred in conjunction with the CC' deletion. The clone missing both AhdI and XmaI sites contained a 118bp deletion encompassing both CC' and BB' repeats and the terminal resolution site. Remaining were the 3'-most 11bp of the A sequence and the 5'-most 15bp of the D sequence. No restriction variation was visible in the 5'ITR locus, and ITR Sanger sequencing demonstrated that the complete 145bp wild-type sequence was present (Appendix I). Though the clone was not screened with the restriction enzyme BssHII, the site was missing and the subclone was therefore called 'ΔBssHII' (with the alias 'Δ118bp').

Three mixtures of the 3'ITR variants were made by diluting the wild-type-derived plasmid in an approximately equimolar ($\pm \sim 10\%$) mix of the non-wild-type variants. Their molar compositions were measured in triplicate using qRBA and were indicated to have WT fractions of $82 \pm 1.2\%$, $40 \pm 0.3\%$ and $21 \pm 0.7\%$ (Fig. 6.1.3). These mixtures were given the retrospective labels pWT_82, pWT_40 and pWT_21.

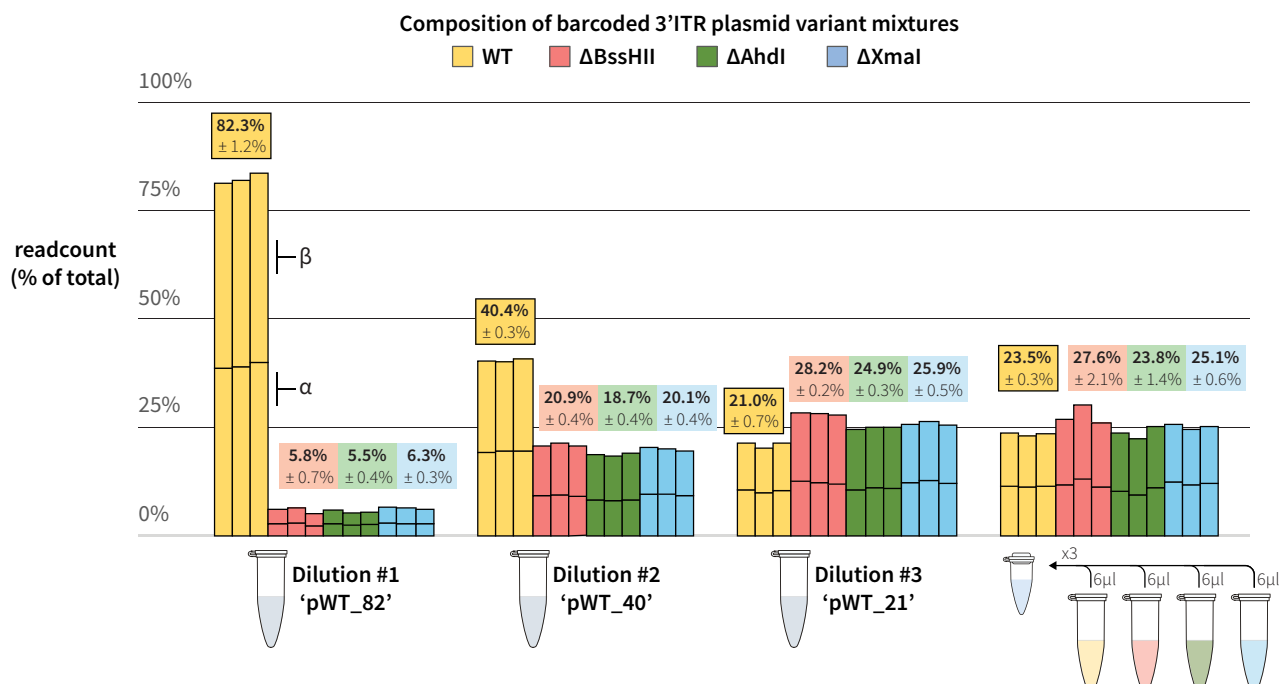


Figure 6.1.3 Determination of % wild-type-coded plasmid content in mixtures of subclonal 3'ITR variants | Each plasmid preparation was diluted to an approximately equimolar concentration using a Nanodrop spectrophotometer, with mass equivalents near to 430ng/μL. The three non-WT variants were then combined equivolumetrically in a 1.5mL microfuge tube and this mixture was used as a diluent for the WT-coded prep. 800μL of the WT-coded prep was added to 200μL of the non-WT mixture. Two further dilutions were performed with a dilution factor of 2. These dilutions were retrospectively labelled WT_82, WT_40 and WT_21 based on reporter barcode analysis results. Plasmids were digested with MscI prior to pre-NGS PCR amplification, which precluded 3'ITR variation from influencing PCR amplification efficiency. The rightmost set of data (23.5%WT) was the result of separate ~430ng/μL plasmid solutions combined equivolumetrically in order to determine their relative molar concentrations. 6μL from each prep was combined, which the volume used to transfect HEK293T cells in 3cm wells (of a 6-well plate).

6.1.1.2 Productivity of plasmid 3'ITR variants co-transfected with pWT

A global view of read data resulting from this experiment is provided in Figure 6.1.4. As this experimental design is unfamiliar to most, this figure was prepared to help readers understand the dataset, at the risk of presenting the same data more than once (which is generally not advised).



Figure 6.1.4 A global view of qRBA readcount data obtained in this experiment

Figure 6.1.4 A global view of qRBA readcount data obtained in this experiment | Three subclonal 3'ITR variants of an rAAV transfer plasmid were generated to study the impact of 3'ITR heterogeneity on rAAV vector production. These variants (and a wild-type-derived prep) each contained subclone-specific hexanucleotide barcode sequences in a 150bp PCR target region that would allow their relative abundance to be measured bioinformatically following Illumina amplicon sequencing. Each subclonal copy contained one of two possible unique reporter barcodes in an α - β pair. Plasmid 3'ITR heterogeneity was simulated by mixing these variants with the wild-type in various proportions and co-transfecting HEK293T cells in 6-well plates. WT_81, WT_40 and WT_21 were 82%, 40% and 21% WT-coded plasmids. Each variant was also studied in isolation by transfecting separate wells and pooling resuspended cells at the point of harvest. Analysis of different molecular production stages was achieved by omitting certain packaging and/or helper elements from transfection (white boxes), selectively isolating the target population, and/or selectively digesting off-target templates with a methylation-specific restriction endonuclease. Equivalent transfections were performed in parallel using an additional prep containing a fifth unique reporter barcode pair, which would act as an external reference to compare total yields from separate wells. At harvest, experimental cells were spiked with specific volumes of the appropriate reference cells. Biological replicates were initiated at the point of transfection. Each individual column is a read count result from one biological replicate. No measurement replicates were performed. Sample types with a common reference were scaled by an equal factor for ease of comparison with other sample types.

The fold outputs of each barcoded subpopulation (relative to WT α) were strongly consistent across a range of input ratios, leading to approximately linear relationships between input ratios and output ratios in the ranges measured. Least-squares regressions were carried out with the y -intercept constrained to zero, in recognition that an input of zero cannot yield a nonzero output. It was noted that a power relationship between input and output also satisfied this condition and in some instances led to modestly improved R^2 values. The linear patterns, shown in Figure 6.1.5, meant that the RFO could be measured from the gradient of the regressions, with error taken from the standard error of the gradient (S.E.G.). The RFOs of WT β and the non-wild-type variants, according to this method, are shown in Figure 6.1.6a.

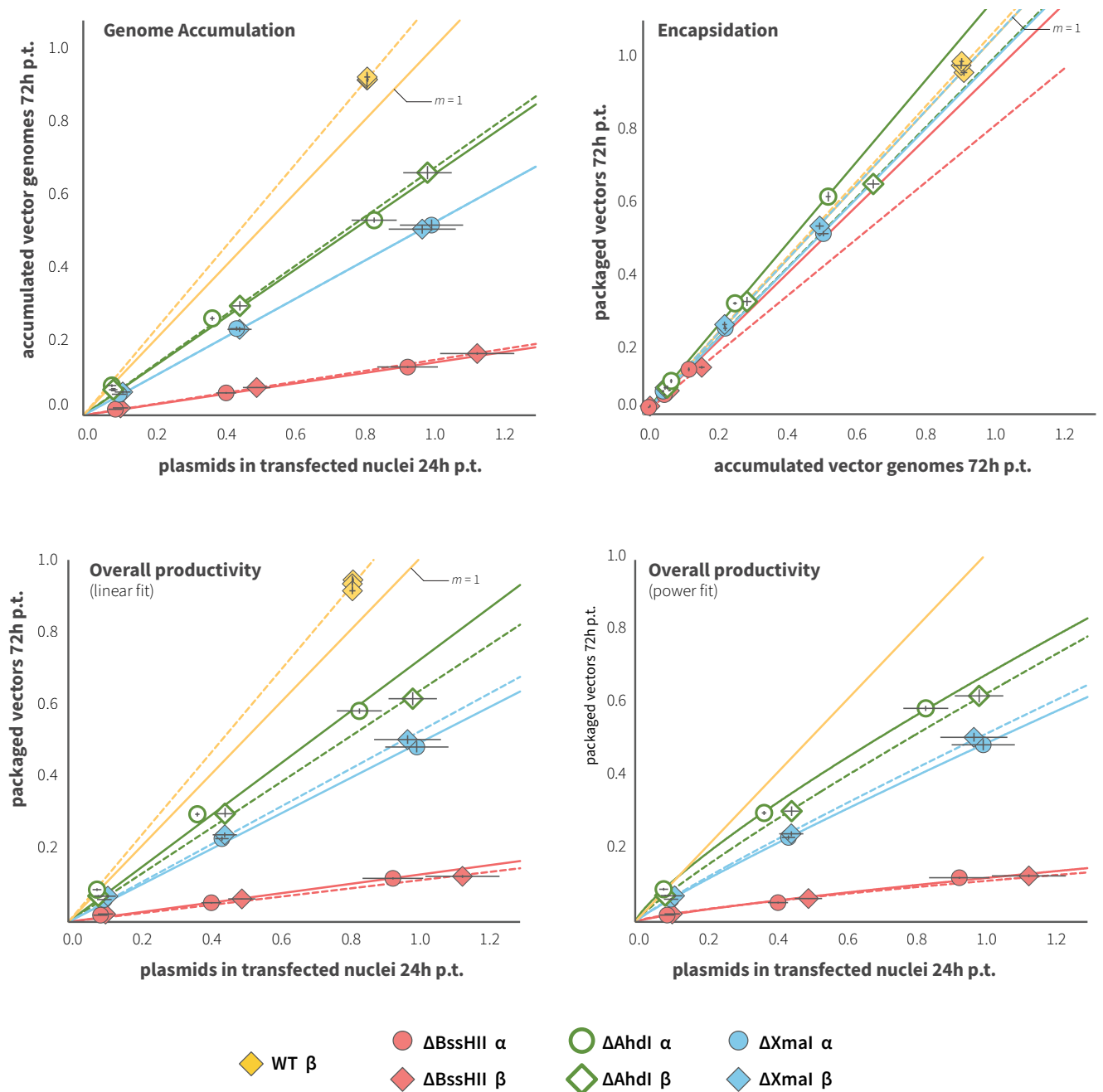


Figure 6.1.5 Linearly driven subpopulation changes allowed RFO to be calculated from gradients | Each barcoded rAAV transfer plasmid variant was co-transfected with pWT in three different proportions and the variant-to-WT α ratios in population x vs population y were compared. Three different populations were analysed in this study: plasmids in transfected cell nuclei, accumulated ssDNA vector genomes and packaged ssDNA vector genomes. Highly consistent fold differences occurred, resulting in a minimum R^2 value of 0.993 (median = 0.9996, max = 0.99998) for the 21 datasets tested. Each x and y value was the average of five biological replicates (or three in the case of nuclear plasmid extracts). The gradient of each regression indicates the change in ratio between two sample types. For any two barcoded subpopulations, a difference in gradient reflects a change in relative abundance from population x to population y; the lower of the pair is decreased and the upper is increased. WT α was used as an arbitrary reference with a constant value of 1.0 relative to itself (solid yellow line). White fill is used to distinguish Δ AhdI data for readers with colour vision deficiencies and has no other significance.

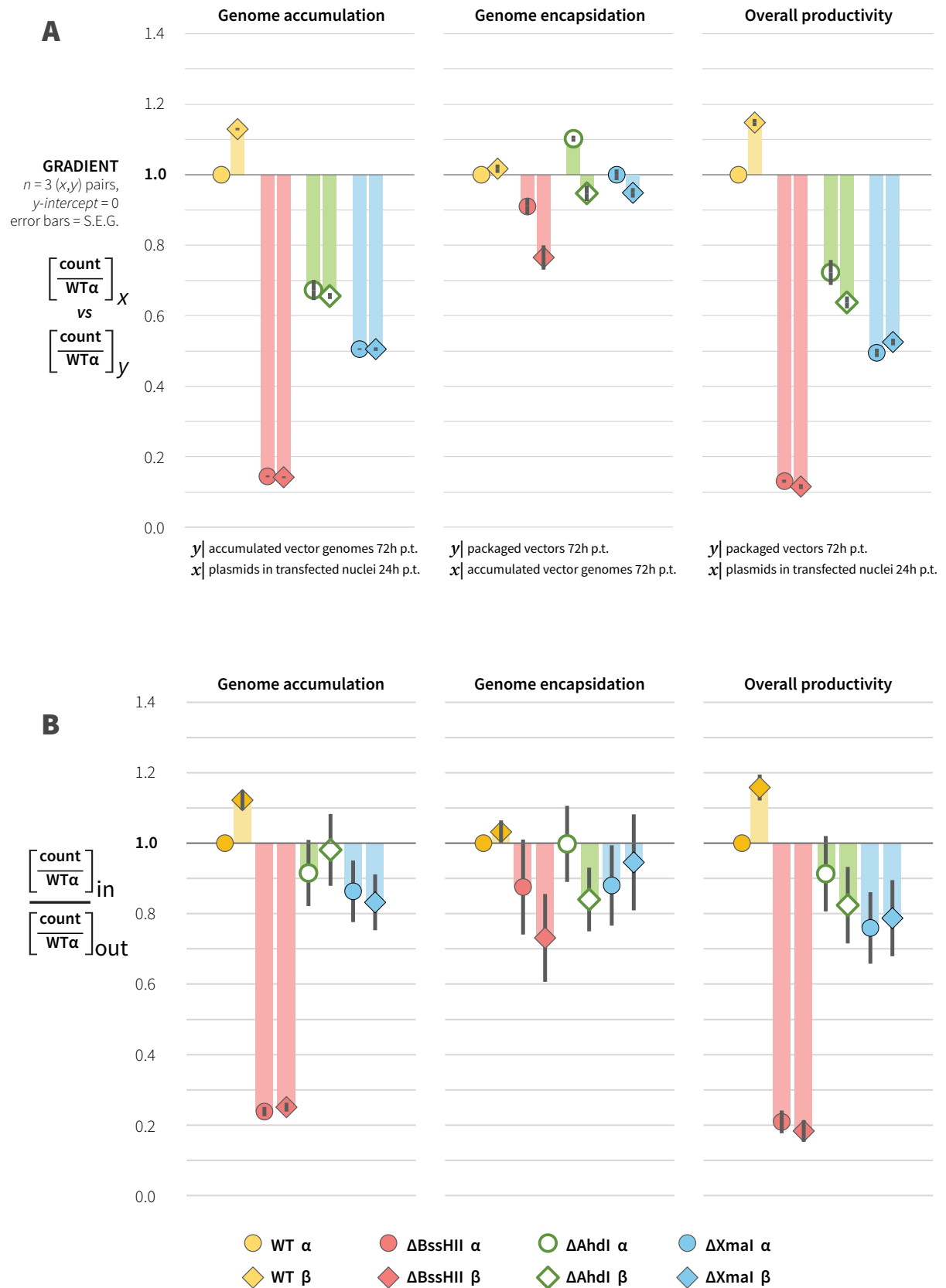


Figure 6.1.6 Productivity of 3'ITR variant plasmids co-transfected with a wild-type-derived plasmid

Figure 6.1.6 Productivity of 3'ITR variant plasmids co-transfected with a wild-type-derived plasmid | (A)

Highly replicable, linear changes in subpopulation composition occurred within the WT-mix transfections. Plasmids with non-wild-type 3'ITRs yielded fewer packaged vectors than the WT variant per mole present in cell nuclei 24h p.t. The fold increase in Δ BssHII-coded reads between nuclear plasmid and packaged vectors was 0.1x that of WT α . 3'ITRs missing the CC' (Δ XmaI) or BB' (Δ AhdI) and hairpin sequences showed ~0.5x and ~0.6-0.7x the fold increase of WT α . Differences in the accumulation of free vector genomes had a strong influence on relative plasmid-to-vector yield, considerably outweighing that of genome packaging. Highly replicable differences in efficiency between α and β subpopulations were observed in some instances despite having been cloned, cultured, and isolated as a single plasmid preparation. **(B)** The four subclonal isolates were separately transfected and pooled at the point of harvest. Ratios of each variant to the WT α reference were measured and inputs and outputs were subject to RFO analysis based on averages. Unlike the variants mixed prior to harvest, the ratio of WT to non-WT variants was not varied. These data did not conflict with the pre-mixed transfection, at least so far as the overall profile. In contrast to the pre-mixed transfections, these transfections demonstrated unambiguous differences in the total yield of transfected wells as a result of less efficiency of non-WT variants. | S.E.G = standard error of the gradient; p.t. = post transfection. White fill is used to distinguish Δ AhdI data for readers with colour vision deficiencies and has no other significance.

The pWT variants resulted in a higher yield of the packaged rAAV-DJ vector than each of the non-wild-type variants, per molar unit of the plasmid measured in producer cell nuclei 48h prior. For the variant containing a 118bp 3'ITR deletion (p Δ BssHII), the measured vector yields were 12.9% $\pm$ 0.4% (α) and 11.4% $\pm$ 0.6% (β) that of the WT α . Deletion of the CC' hairpin sequence (p Δ XmaI) was more impactful than the BB' deletion (p Δ AhdI), though both were more productive than the p Δ BssHII subclone. The BB' deletion resulted in 72.1% $\pm$ 3.5% (α) and 63.6% $\pm$ 1.6% (β), while the CC' resulted in 49.2% $\pm$ 1.2% (α) and 52.4% $\pm$ 0.9% (β) the vector output of WT α . The WT β subpopulation produced approximately 10% more packaged vector genomes than the WT α subpopulation, translating to an approximately 10% lower RFO for the non-WT variants if measured against WT β . There was a lesser difference in population composition between packaged and accumulated vector genomes than between accumulated genomes and their plasmid precursors. When accumulated genomes were modelled as the input to packaged vectors, the RFOs clustered more closely (though not equal to) 1.0. This indicated that the reduced fold output of vectors from plasmids was primarily (though not completely) attributable to differences in the output of packageable vector genomes.

6.1.1.3 Productivity of plasmid 3'ITR variants in individual transfections

Subclonal 3'ITR plasmid variants were transfected both as mixtures, to simulate ITR heterogeneity, and as standalone preps, to compare the behaviour of the variants in isolation. For standalone transfections, the contents of each well were pooled at the point of harvest and then processed as a single sample. The reporter composition of these samples was then analysed using qRBA in the same manner as the variants that had been pooled prior to transfection.

The fold outputs of these standalone transfections relative to the WT α reporter (Figure 6.1.6b) resembled those of the mixed transfections in some respects and contrasted in others. In profile, the overall productivity of these plasmid variants (nuclear plasmids $\rightarrow$ packaged rAAV-DJ) recapitulated those observed in the mixed transfections, although the specific RFOs were not equal. The differences between α and β populations for each variant were well matched between the transfections pooled prior to or at the point of harvest – most notably the higher genome accumulation efficiency of WT β , the lower encapsidation efficiency of Δ BssHII β , and the higher encapsidation efficiency of Δ AhdI α . Minor exceptions include Δ BB' α genome accumulation and Δ XmaI genome encapsidation. Compared to WT-mixed transfections, the non-WT variants in standalone transfections showed a higher RFO in genome accumulation and a lower RFO in genome encapsidation than in the mixed transfections. In approximate terms, the RFO of genome accumulation for standalone transfections was 30% higher for the Δ AhdI variant and 40% higher for the Δ XmaI and Δ BssHII variants than in the mixed transfections. For genome encapsidation, the RFO was 10% lower for the Δ AhdI variant and 5-10% lower for the Δ XmaI and Δ BssHII variants. Importantly, since the reporter composition of standalone samples reflects the contents of whole wells, the non-WT variants displayed consistently lower overall productivity (nuclear plasmid $\rightarrow$ packaged rAAV-DJ) than the WT-only sample in terms of the well's total yield.

6.1.1.4 α - and β -specific behaviours

To ensure prep-to-prep variables were minimised between α - and β -coded plasmid partners, mixed α - and β -coded inserts were ligated into a 3'ITR variant fragment in a single ligation reaction, then transforming and propagating the circularised products as single non-monoclonal batches. Despite this design, α - and β -coded versions of the same reporter variant behaved in some instances as distinct populations according to qRBA results. The β : α ratio of pWT plasmids was anomalously low ($\sim 0.8\times$) in the nuclei of transfected producer cells 24h p.t. compared to other sample types. This measurement was consistent across all nuclear plasmid samples despite variable total quantities and the spread of WT β : α ratios in these samples could not be distinguished from a standard normal distribution by Shapiro-Wilks ($p = 0.96$) or D'Agostino-Pearson ($p = 0.91$) tests, albeit with a sample size of 11 (with one outlier excluded after a Tukey fence was applied with k equal to $2\times$ the interquartile range). The measurement indicates that the pWT α variant was transfected more efficiently than the pWT β , and that WT β plasmids produced higher amounts of eukaryotically synthesised vector genomes. Another

instance was the $\Delta BB'$ variant, whose α version produced packaged vectors from packageable genomes with the highest efficiency, but whose β variant was slightly less efficient than the pWT-derived genomes in this regard.

6.1.1.5 Total vector yield from plasmids with variable 3'ITR integrity

An external reference spiked into samples at the point of harvest allowed the total output population size to be compared, in addition to its reporter composition (Figure 6.1.7a). It is pertinent to emphasize in advance that these data are considered inconclusive due to the risk of inaccuracy evident from Chapter 4 – specifically the risk associated with adding numerous reporter counts to obtain a total. The results are shared for completeness and are subjected to a tentative analysis. Taken at face value, no clear impact on the collective vector yield was observed (relative to the WT-only transfection) when WT-coded input in transfected cell nuclei varied from ~80% to ~25%. Using pipette volumetry and plasmid A_{260} spectrophotometry, it would have been assumed that culture wells were transfected with equal quantities of the rAAV plasmid. qRBA measurements of nuclear plasmid content indicated that this was not the case. Since transfection efficiency was nonequal, formal comparisons of the output vector quantities based on variance or t statistics (n of 5 notwithstanding) would not have been unmeaningful. If taken to be accurate, results were such that a larger scale study would have been required to seek non-obvious trends by other statistical means.

Certain qualitative observations suggesting decreases in vector yield were noted but not consistent trends. For example, WT_40 exhibited a marked reduction in vector fold output compared to WT-only, though the WT_82 transfection was in fact slightly higher and WT_21 showed a lesser reduction despite a lower %WT. $\Delta AhdI$ -only and $\Delta XmaI$ -only transfections were comparable to WT_40 in genome and vector output, which complicates the view that WT% *per se* increases those outputs. Perhaps most conspicuously, vector yield was highest in the WT-only transfection. However, fold output was not reduced consistently with %WT. For the WT-mix transfections, the strongest predictor of output identified in post-hoc analyses was the quantity of plasmid in transfected cell nuclei 48h prior to harvest, irrespective of wild-type 3'ITR content ($R^2 = 0.9996$; $n = 3$ x, y points; $b \neq 0$) (Figure 6.1.7b). The WT-only transfection did not conform to that particular regression, suggesting that other parameters, such as the composition of non-WT variants, could influence vector yield.

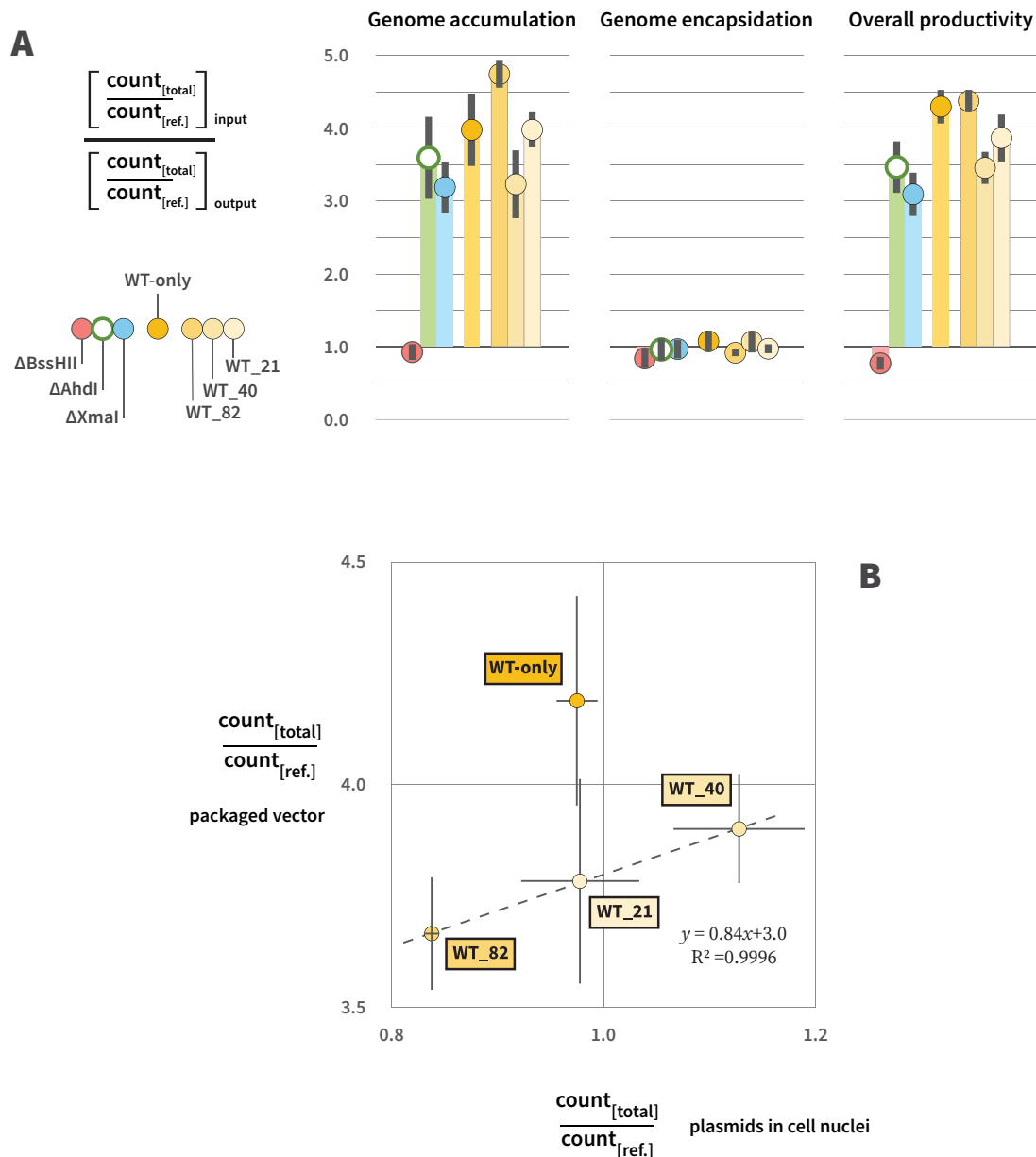


Figure 6.1.7 Products of separate wells measured using an exogenous reference spiked into samples at harvest| (A) The relative fold outputs of whole wells were calculated using a fifth reporter pair as an exogenous reference. This was done with the intention of testing whether ITR integrity affected the total yield of vector products, but regrettably this aspect of the study is treated as inconclusive in light of the results shared in Chapter 5.3, due to the risk of inaccuracy when adding counts from numerous reporter variants. The data are disclosed above nonetheless for completeness, or simply for interest. **(B)** The count data obtained gave the impression that the total nuclear plasmid content was a stronger predictor of total vector yield than %WT plasmid content. Horizontal and vertical error bars are the standard deviations of three replicates (x) and five replicates (y).

6.1.1.6 Quantitative consistency of biological replicates

Transfections in this study were carried out in triplicate (nuclear plasmids) or pentaplicate (accumulated and encapsidated genomes) and each replicate was measured once using qRBA. That is, the study performed biological replicates but not technical replicates. Each reporter barcode count in a sample was fully binarised and the RSD for each possible barcode pair, excluding self-pairs and reciprocals, were calculated. The RSD was calculated from the sample standard deviation (not the population standard deviation) in all cases. Five published reports specifically concerning qPCR or dPCR measurement of vector concentration were identified since 2014^{8,250,252–254} to the time of submission, and the variability of those measurements was collated into a single chart (Fig. 6.1.8) and compared to this study's observations. The purpose of this was to add context to this study's observations (not to perform a 'benchmark' comparison between any particular reports). The RSDs were comparable with technical replicates of purified rAAV samples using qPCR or dPCR.

As in Chapter 4, ζ_s was found to be a central determinant of qRBA measurement precision. Recall that ζ is the standard deviation for a qRBA measurement predicted by a binomial distribution model, and ζ_s is a value for ζ inferred from the ratio and readcount of the data itself (since the underlying parameters are unknown). The variability of biological replicates in this section was compared to ζ_s , which was again found to have a strong predictive relationship. The SDs measured in this section were plotted against the predicted value ζ_s and analysed in Figure 6.1.9. In that figure, ratios involving either the α or β reference barcode are analysed separately from ratios involving only barcodes endogenous to the sample. Also included in this figure are the results of repeated NGS measurements of a single sample, which were generated in Section 2.3 of Chapter 4. The log residuals (that is, $\log[\text{SD}] - \log[\zeta_s]$) were normally (albeit asymmetrically) distributed according to a D'Agostino-Pearson test ($p = 0.19$ and $p = 0.87$ respectively). A Shapiro-Wilks test of the same data detected a lack of normality ($p = 0.084$ and $p = 0.003$), although the effect size (KS-D) was equal to 0.05 in both cases, which was indicated by the online statistical resource used to represent a 'very small' difference between the residual distribution and a normal distribution.

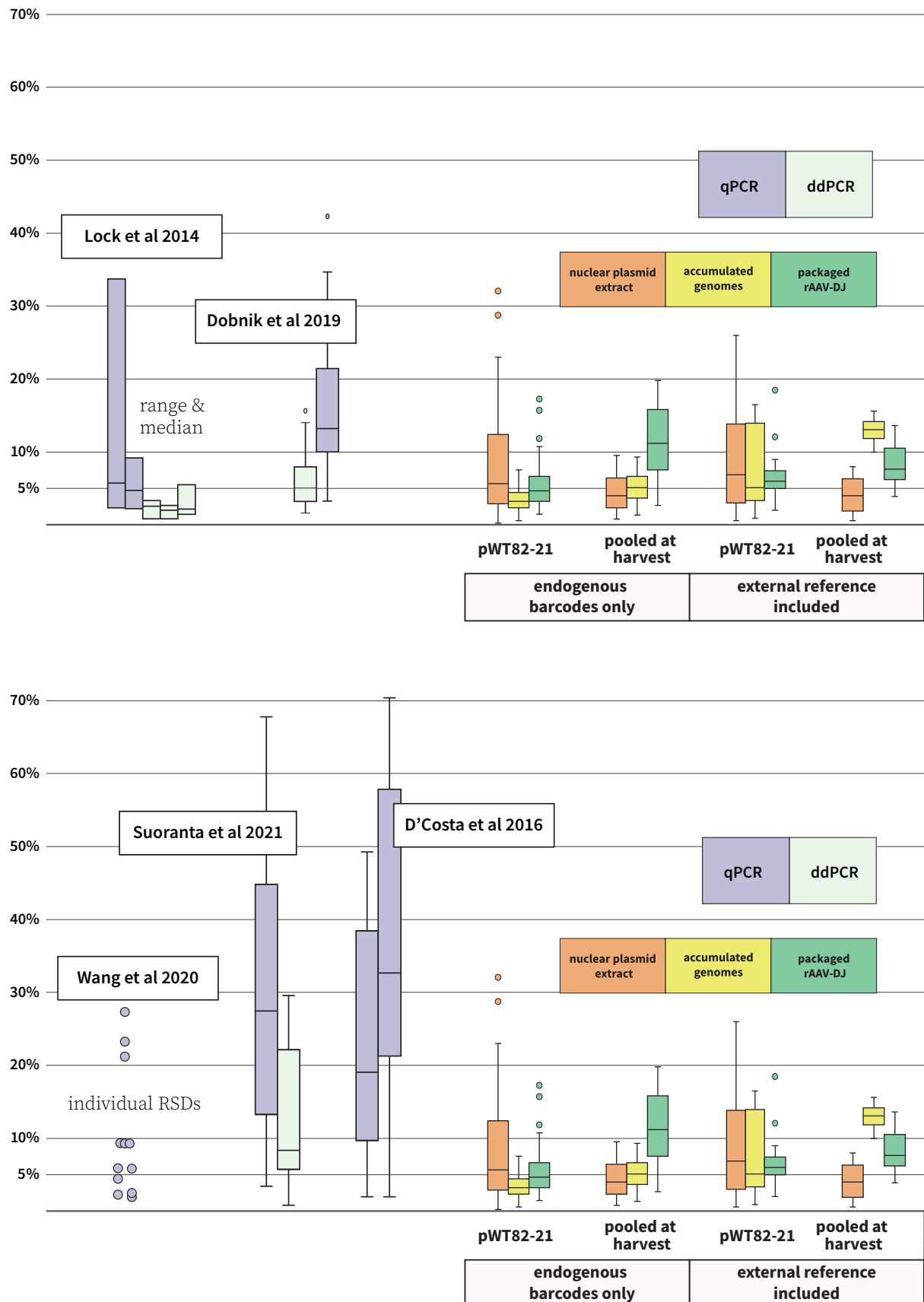


Figure 6.1.8 Consistency of biological qRBA replicates compared with technical replicates reported in the literature

Figure 6.1.8 Consistency of biological qRBA replicates compared with technical replicates reported in the literature

Five scientific reports were identified from 2014 to 2021 concerning qPCR and/or ddPCR quantification of rAAV vectors. Relative standard deviations (RSD) otherwise known as Coefficients of variation (CVs) reported in these publications were collated and compared to those of qRBA biological replicates in the present study. NOTE STRICTLY that the results of those publications represented here are not intended to convey greater or lesser scientific success, and that any such inferences can easily mislead without due regard to their individual contexts. These data were consolidated as a general survey of q/dPCR precision encountered outside of this research project, so that precision within this research project can be viewed in context. The previously published RSD values are the result of replicated technical measurement of single rAAV samples, and represent random error introduced by the measurement process itself. For the present project, this same summary statistic was applied to biological replicates, where replication was initiated from the point of transfection. It therefore represents the random error incurred from measurement, as well as all upstream processes including the biological experiment itself, extraction and purification of the nucleic acid targets, and preparation of those targets for preparative PCR. The RSDs are derived from $n=5$ vector genome samples or $n=3$ nuclear plasmid samples, and calculated from the sample standard deviation (not the population standard deviation). Each possible barcode ratio was measured, excluding reciprocals and self-pairs. Pairs that included one of the reference barcodes were separated from pairs that did not include a reference barcode, so that the efficacy of external references incorporated into qRBA experiments could be considered. For simplicity, RSDs of stored plasmids (shown in Figure 6.1.3), which are technical replicates, were omitted from this figure.

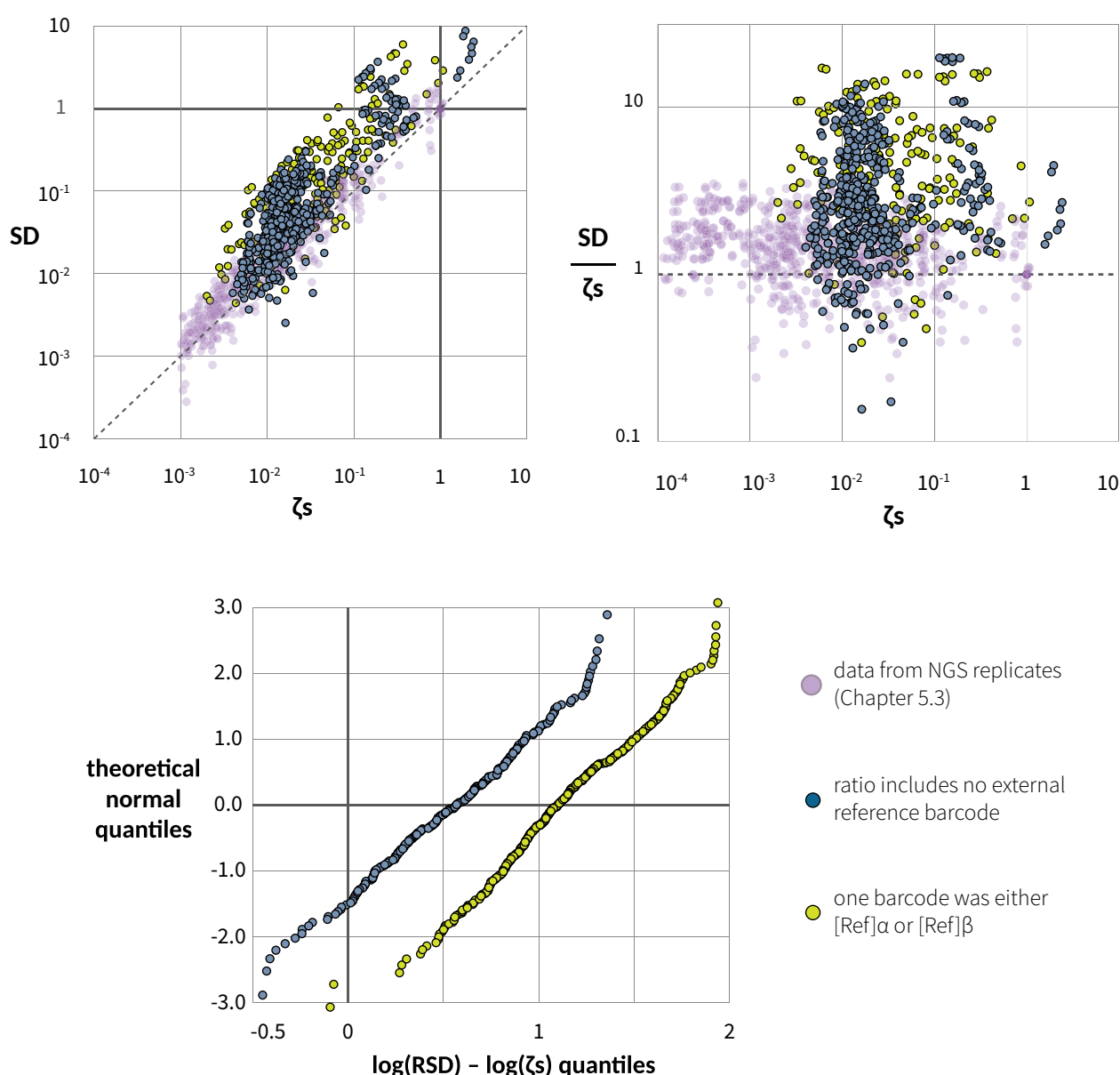


Figure 6.1.9 Measurement variability of biological replicates compared with ζ_s | The precision of qRBA data is related to both the readcount and the reporter ratio being measured. The statistic ζ ('zeta') has been developed in this thesis as a predictor of qRBA precision based on a binomial statistical model, which was reported in detail in Chapter 5. Another statistic, ζ_s ('zeta s') was defined, which estimates the value of ζ value by inferring the underlying parameters from the data (since the parameters in reality are unknown). The biological replicates in this study were compared to two benchmarks: ζ_s , and a previous dataset comprising Illumina sequencing replicates of a single sample. As in Figure 6.1.8, barcode ratios that included an external reference were considered separately from those where both reporters were endogenous to the sample. Differences between ζ_s and the empirical data were normally distributed according to a D'Agostino-Pearson test ($p = 0.19$ and $p = 0.87$ respectively for ratios with and without an external reference respectively). Shapiro-Wilks tests found the data not to be non-normal ($p = 0.084$ and $p = 0.003$) though with a difference characterised as 'very small' ($KD-S = 0.05$ and 0.05) by the online statistical resource used. <<https://www.statskingdom.com/shapiro-wilk-test-calculator.html>>

6.1.2 Discussion

6.1.2.1 Relevance of Chapter 4.4 to the interpretation of these results

The observation that conjugate qRBA assays disagreed on the relative abundance of reporter variants dramatically affected the course of this thesis. Plans to publish the studies in this chapter were withheld until the nature and extent of systematic error could be better understood. However, a thorough exploration of the systematic qRBA bias leads me to conclude that the results of this study are far from unsalvageable. To reiterate, the data in Section 4.4 showed that for any specific pair of reporter variants, changes in their measured ratio can be confidently attributed to changes in their actual ratio. A high degree of confidence can be placed in the direction and magnitude of that change, though with lesser confidence in the magnitude as individual ratios extend beyond approximately 0.1 or 10. This follows from the observation that the power relationships between r_s and r were approximately linear within this range. In fact, the modestly improved fit observed in Figure 6.1.5 when a power regression was used may reflect this subtle nonlinear distortion. Though the ratios between alternative pairs of reporter variants may well be reliable, this cannot be assumed by default. This reasoning discussed here will be relevant to both Sections 6.1 and 6.2, and will be particularly significant as we attempt to make sense of the α - and β -specific behaviours observed in both sections.

In lay terms, qRBA measurements are highly capable of ‘staying still’ if a change in the reporter ratio has not occurred. This comports with the $\alpha:\beta$ ratios measured in the stored plasmid mixes, were consistent across each of the three dilutions. The $WT\beta:\alpha$ ratio, though increasing by ~12% in the course of genome rescue and replication, was unchanged in the course of encapsidation (Fig. 6.1.6, upper right panel). The clarity of this trend required highly precise and well-aligned measurements in the $WT\beta:\alpha$ ratio of all accumulated genome samples and all packaged vector samples. We might speculate that this was made likely only by the pair’s ~1:1 ratio in all samples (more specifically ~0.8), and that a different range would be less permissive. This was shown to be untrue by the $\Delta XmaI\beta$ according to the same analysis, which exhibited a packaging fold output exceedingly close to 1.0 relative to $WT\alpha$. This too required a very high degree of consensus between three sets of five *biologically replicated* accumulated genome samples with an equivalent three sets of packaged vector samples, in which the $\Delta XmaI\beta:WT\alpha$ ratios measured were ~0.06, ~0.22 and ~0.46.

Another highly encouraging observation was the level of corroboration between partnered reporters in RFO measurements, particularly in genome accumulation of partnered Δ BssHII and Δ XmaI reporters. Both of these sample sets are instances where a precise consensus regarding fold reduction was achieved, not only between two uniquely barcoded variants but from a range of three possible input ratios. In other words, the linear RFO measured was not dependent on the choice of reporter partner or the input ratio. This adds significant credence to deviations observed between α and β variants in these two plasmids, which in turn creates a strong precedent to suppose this phenomenon is valid in general terms.

Perhaps the most heavily impacted outcome, unfortunately, is the total population sizes measured relative to the reference spike, because it requires the addition of eight unique UMI variants with potentially variable biases. This may or may not have caused the ambiguity of those results on the whole. Because of these ambiguities and the results of Chapter 4, the most sensible decision is to disregard this data and reattempt if improvements to qRBA measurement can be made that allow us to confidently add this many unique populations.

6.1.2.2 Distinct behaviours of partnered 3'ITR plasmid variants

Figure 6.1.1 illustrated the method by which partner reporters were prepared. To minimise the risk of prep-to-prep variation (especially in 3'ITR composition), α - and β - coded pairs were not prepared as individual preps and mixed prior to transfection. Instead, the barcoded region was mixed at the point where the insert was cleaved in preparation for two-piece ligation. A very strong track record for this particular ligation permitted the transformant cells to be propagated directly (since colony screening would defeat the intended purpose) with the assurance that no off-target products would be generated, which was verified with agarose gel analysis of multiple restriction digests and an uncleaved plasmid sample. The inserts were therefore ligated into the same set of 3'ITR variant recipients, propagated within a single culture, extracted and purified as a single volume, stored under the same conditions and transfected from the same PEI transfection mixture. It is therefore reasonable that the distinct and replicable behaviours observed would strain credulity, and fair that scepticism is applied. However, the phenomenon does not comport with the patterns of error discussed at length in Chapter 4, particularly in light of the observations highlighted above.

The phenomenon may make more sense if we suppose that the 'distinct' behaviours are in fact aggregates of heterogenous behaviours within each subpopulation with the appearance of

unity. This would not require us to accept that the α and β variants diverged in some distinct category; it may be that both groups are heterogenous and nonidentical in profile. As mentioned in Chapter 5, recent discoveries in bacterial cloning have cast great doubt on the assumption that transformed bacterial colonies expand homogenously in terms of their plasmid content. In that same Chapter (Figure 5.5.3) an example was shared of plasmid 3'ITR instability whose pattern differed between two clones picked from the same plate of transformants. The presence of the wild-type 3'ITR sequence in the population indicated that the colony as a whole was derived from an intact copy, yet at some stage in the outgrowth the emergence of AhdI-negative or XmaI-negative subclonal variants was apparently fixed. While these points don't offer a direct explanation for the subclone-specific behaviours observed in this (and the following) experiment, it does help to account for how it could be feasible.

6.1.2.3 Quantitative consistency

Figure 6.1.8 compared the variation of replicate qRBA samples with coefficients of variation collated from various articles in the literature concerning the quantification of rAAV vectors. Out of consideration for these authors I would stress yet again that the varied results collated in this figure are not intended to imply superior or inferior research outcomes, and making such inferences without regard to their individual contexts is highly inadvisable. In contrast with these external data (and with Chapter 4), the coefficients of variation (aka RSDs) presented from this study are the results of biological replicates subject to single measurements.

Based on previous feedback it is worth making what is meant by 'biological replicate' excessively clear. To use to accumulated vgs as an example, the following process was replicated from beginning to end five times, generating five samples. A well of adherent HEK293 cells was transfected with a plasmid containing an rAAV transgene sequence with reporter barcodes. Inside the cells, the transgene sequences were converted into the replicationally active ssDNA form and propagated via rolling hairpin replication with varying efficiencies. After 72 hours, the well's contents were harvested (in some cases combined with other wells) and the DNA released from the pelleted cells via minimally mechanical chemolysis. The low-MW ssDNA genomes of interest were separated from the chromosomal DNA following an overnight high-salt incubation via centrifugation and retrieval of the supernatant. The dissolved DNA was separated using phenol-chloroform extraction and purified from the aqueous phase using ethanol precipitation. To selectively measure eukaryotically derived PCR-targets, a sample of DNA was treated with the restriction endonuclease DpnI and then subject to a 10min 80°C

incubation to inactivate the enzyme. PCR was performed on the DpnI-treated low-MW DNA extracts to produce a reporter amplicon and this amplicon was subject to population sequencing (150bp paired end MiSeq) to sort and tally copies according to the plasmid-of-origin. The RSDs plotted in Figure 6.1.8 are each the result of five samples – each measured once – not one sample measured five times.

Replication was initiated from the point of transfection, and thereby faithfully exposes samples to all downstream sources of experimental uncertainty including variation in the biological systems, extraction and preparation of the samples, and *then* the quantification process itself. Further, we can recognise that the qRBA measurand is a unitless ratio between two populations of interest. It therefore conveys directly what would otherwise require two distinct measurements, both of which contribute uncertainty. This principle applies (for example) to comparisons to experimental controls, benchmark groups or endogenous references. Also unrepresented is the volumetric error component inherent to qPCR and ddPCR, whose measurands are concentration. This result unquestionably affirms the arguments put forward in Chapter 3.3, which spoke to the feasibility of this outcome.

To explore the validity of ζ_s in this more complex experimental setting, the data set as box-and-whisker plots in Figure 6.1.8 were compared to ζ_s in Figure 6.1.9. In the context of Figure 6.1.9, we have not hypothesised that ζ_s will be a direct approximant of precision (as was hypothesised in Chapter 4). The numerous sources of variation upstream from the measurement process are fully expected to decrease its predictive power to some extent. What this cursory analysis was able to show was that ζ_s could act as a valid heuristic where qRBA analysis is involved. Interestingly, the empirical data appeared better approximated by ζ_s as ζ_s decreased.

6.1.2.4 The effects of plasmid deletions on rAAV vector production in adherent HEK293T producer cells

Sequence deletions within the 3'ITR had clear impacts on vector generation. This was particularly interesting to analyse in the pre-pooled transfections, where the elimination of inter-well variables and the ability to process the mix directly (rather than combining at harvest) offered a remarkably clear window into the subpopulation dynamics in the intracellular production process. The fold changes from population to population were consistent across a range of input ratios, reflecting a linear (or pseudo-linear) response to the

%WT variable. This phenomenon may represent a path to novel scientific insight regarding the kinetics of intracellular rAAV vector generation processes.

Unsurprisingly the greatest relative fold decreases (~90%) were observed in the Δ BssHII variant, which contained only vestiges of the ITR sequence in the 3' position and included no terminal resolution site. Following this, the second most impacted variant (~50% relative decrease) lacked a CC' hairpin sequence, perhaps contrary to expectation, given that the BB' variant was missing a ribosome binding element (RBE'). The rank and magnitude of these differences are based on a strong corroboration between partnered reporters as to the outcomes of genome accumulation. Loss in packaged vector yield was predominantly due to defects in genome accumulation, as the genome and packaged vector samples reflected a much less dramatic change in subpopulation composition. It is important to raise again that the identity encoded by the endogenous UMI's in this experiment was the plasmid variant of origin and is not assumed to inform us of ITR identity downstream from transfection due to 'repair' of ITR sequences. Notably, changes in reporter composition between replicated and packaged genomes would suggest structural differences in replicated ITRs to account for differences in packaging efficiency, if we recognise that according to the null hypothesis, the abundance of reporter variants within packaged capsids is simply a reflection of their abundance in the packageable population.

The most notable difference between these and the samples pooled at the point of harvest was the heightened efficiency of the two Δ 21bp variants when transfected individually. The Δ AhdI and Δ XmaI variants in the pre-pooled samples showed genome accumulation RFOs of ~0.7 and ~0.5 respectively but were measured at approximately ~0.9 and ~0.8 in individual transfections. This could be indicative of a competitive effect exerted by the wild-type-derived plasmid products.

6.2 Truncations in plasmid 3'ITRs and gene transfer to murine liver and skeletal muscle

This study sought to establish whether deletions in plasmid ITR sequences can affect the gene transfer efficacy of the recombinant vector product. The WT, Δ AhdI, Δ XmaI and Δ BssHII plasmid variants described in the previous section were used in individual productions of

rAAV9 vectors, which were combined into a single mixture and delivered systemically to male juvenile $Fah^{-/-}Rag^{-/-}Il2rg^{-/-}$ (FRG) mice (an immune deficient strain). Transduced tissues were harvested one week ($n = 6$) and eight weeks ($n = 6$) post-administration and qRBA analysis was performed on DNA and mRNA extracts to test whether changes in the reporter barcode composition had occurred (Fig. 6.2.1). The null hypothesis for this experiment is that no change in reporter barcode composition of the exogenous nucleic acids will be detected relative to the vector preparation. This outcome would indicate that deletions in the plasmid 3'ITR sequence did not influence the performance of the vector produced. Each plasmid variant was used to produce the vector in separate batches, which allowed the variants to be added in roughly measured quantities and excluded the possibility of inter-genome recombination (and therefore misattribution of reporter barcodes).

The present context of ITR research factored strongly into the design of this experiment. The impact of ITR heterogeneity on vector performance has received little attention, meaning there is no clear path available for us to begin to understand the problem. To initiate a rigorous line of investigation, we require robust evidence to establish whether plasmid ITR mutations are a variable of concern in vector quality control. By preparing and testing the vector products in a single injection mix, differences in relative fold output can withstand scrutiny created by confounding variables, whose methods for analysis are either underdeveloped or infrequently implemented. These include the proportion of full and empty capsids, process-related contaminants, storage and degradation, vector aggregation, or inaccurate/imprecise dosing. Another aspect was the decisive emphasis on function, rather than the physical and genetic attributes of the vector prep that may be influenced by ITR integrity. An accepted limitation of this approach is that mechanistic interpretations of the results would be highly speculative.

Another important decision was to design an eight-week *in vivo* study as a starting point, rather than the more usual progression from *in vitro* results. For this specific research goal, adherent cultures with limited lifespans and/or dysregulated DDR responses would be a poor environment to compare the effects of ITR integrity in a clinical setting. Results from such experiments would be largely uninformative since they will not recreate the slow processes by which ssAAV transgene cassettes take root in transduced cells and maintain transcriptional activity over time. This, not short-term transcriptional activity, is the functional requirement for recipients of rAAV gene-addition therapies. They may depend, among other things, on physiological DDR activity, tissue-specific DDR pathways and the pharmacokinetics of

transduction. A null result in tissue culture would unduly reduce the motivation to pursue the question *in vivo*.

Figure 6.2.1 An exploration into the effects of plasmid ITR deletion on short- and medium-term *in vivo* gene transfer | **(A)** This experiment sought to determine whether ITR truncations in rAAV plasmids lead to differences in stable gene transfer by the vector product. Four barcoded plasmids derived from an intact or mutant 3'ITR copy were used to produce batches of an rAAV9 vector, which were combined prior to purification and later subjected to qRBA measurement and RFO analysis. The labels on the conical tubes refer to missing 3'ITR restriction sites in the plasmids used to produce the rAAV9 vector. Readers are referred to earlier in the chapter (Fig. 6.1.2) for more complete descriptions of the 3'ITR mutations. The reporter composition of the pooled vector was measured using qRBA using six technical replicates (equivalent to the number of mice analysed). Sample standard deviations are shown immediately to the right of the column so as to make them visualisable. Skeletal muscle and liver tissues were analysed at one and eight weeks following systemic administration of the vector via tail vein injection. **(B)** The results of qRBA measurements for each variant, tissue type and time point are summarised as percentages of the total population. Data shown are averages, though individual variations between mice shown in later figures are noted.

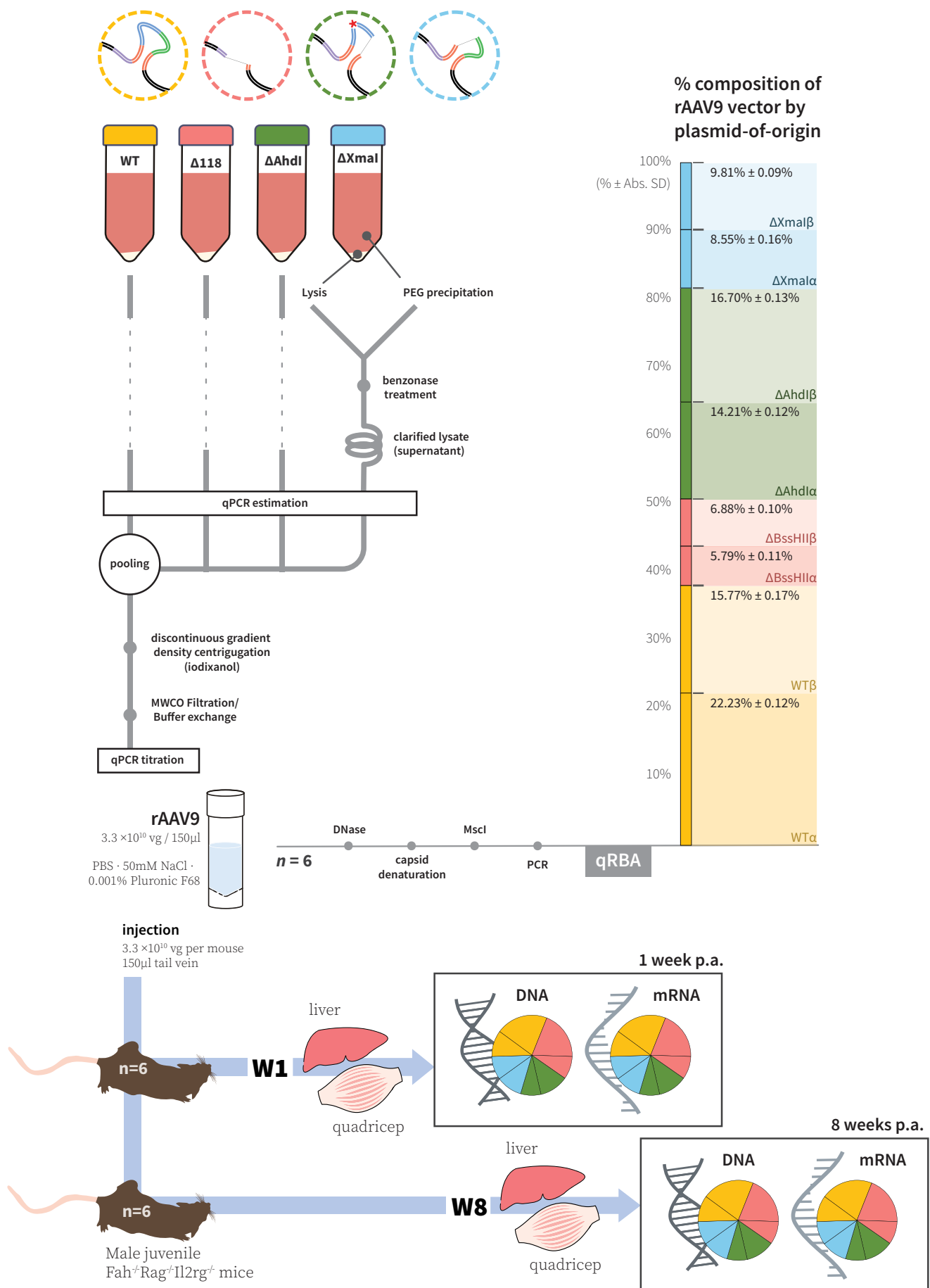


Figure 6.2.1 An exploration into the effects of plasmid ITR deletion on short- and medium-term *in vivo* gene transfer

6.2.2 Results

6.2.2.1 Production of an rAAV vector mix from four unique plasmid 3'ITR variants

Crude vector preparations were made using the 3'ITR plasmid variants described in Section 6.1. This allowed vector quantities to be estimated by qPCR prior to pooling, so that no reporter variant would be strongly underrepresented. After pooling, the vector was purified from the crude extract via discontinuous density gradient ultracentrifugation and molecular weight cut-off filtration. The yield (measured via qPCR) was sufficient for 3.3×10^{10} vg/mouse. The efficiency of production was limited by poor yields of the Δ BssHII variant, which occupied 32 of 76 150mm dishes cultured for these batches. Unfortunately, the central nervous system was effectively excluded from analysis due to this limitation because of the high doses of vector required to achieve sufficient trafficking across the blood-brain barrier. The reporter variant composition of the vector prep was measured using qRBA, with the methodology and results summarised in Figure 6.2.1a. The fold discrepancy between the least (Δ BssHII α) and most (WT α) abundant reporter variants was just under four-fold.

6.2.2.2 Comparison of gene transfer, transgene stability, and transcription, and functional outcomes

The vector was administered systemically via tail-vein injection into twelve male juvenile FRG mice. At one week and eight weeks post administration, six mice were sacrificed and qRBA was performed on tissue DNA and mRNA extracts. Attempts to amplify the vector DNA target from brain tissue were unsuccessful from week 1 and so no attempts were made to extract mRNA from brain tissue at week 8. Reporter compositions in liver and skeletal muscle is summarised multinomially in Figure 6.2.1b. As in Section 6.1, the WT α reporter variant was appointed as the common reference for binarisation of the data and RFO analysis.

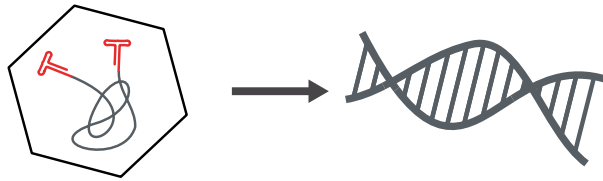
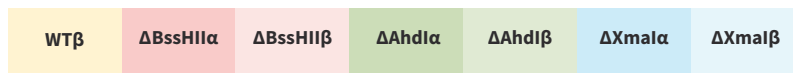
The vector mix was modelled as the input to each individual DNA extract 1 week and 8 weeks post-injection. The measurement uncertainty attributed to DNA measurements from individual mice is ζ_s , expanded by a coverage factor of six (to reduce the risk of underestimating uncertainty). The error attributable to the vector mix was available empirically ($n = 6$). Results are shown in Figure 6.2.2. As was observed in the transfection experiment, the two partnered WT coded variants exhibited markedly distinct tendencies compared to other partnered reporters. Both the variation and magnitude of difference between WT α and WT β were notable

compared to other variants. At one week p.a., the WT β group in three of six mice resulted in the lowest RFOs in both organs. They were also the only pair to exhibit a clear reversal in the direction of the RFO at week 8 compared to week 1. Because the WT behaviour was subclone-specific, the behaviour of other variants could not be described relative to the WT-derived vectors in sweeping terms. Variation between individual animals also made generalisations more difficult, with one or two samples often violating general trends.

The Δ AhdI and Δ XmaI variants both tended to achieve an increase in gene transfer at Week 1 that approached 10% or poorly distinguishable depending on the individual, which represents a consistent increase from roughly 10-20% relative to WT β in Mice 3-6, or as high as 40% compared to WT β in Mice 1 and 2. Generally, the relative fold output of the Δ AhdI and Δ XmaI variants was higher at eight weeks than at one week. The most dramatic contrasts in gene transfer other than WT β were seen in Δ BssHII pair. At week one, the Δ BssHII showed an RFO ranging from 0.9-1.0 relative to WT α , and closer to $\sim$ 1.0 relative to WT β (or $\sim$ 1.3-1.4 relative to Mice 1 and 2). The RFO of the Δ BssHII pair was markedly decreased at eight weeks, ranging from $\sim$ 0.9 to below 0.6.

To assess expression of the vector transgene, the RFO of mRNA from DNA was measured in each individual mouse, with the results shown in Figure 6.2.3. The theoretical error attributed to the RNA and DNA measurements was $6\times\zeta_s$. The range of RFOs for transcription of qRBA targets was more expansive than for delivery by the rAAV9 vector, spanning from as high as 1.75 to just lower than 0.5, as well as showing a stronger tendency toward α - and β - specific behaviours. The Δ XmaI variants both displayed a more consistently higher expression at one week and a reduced and less consistent RFO at week 8. The Δ AhdI α variant behaved similarly to its β counterpart in the liver at week one, and was otherwise more actively transcribed by a factor of $\sim$ 1.1 to 1.25. Notably, the WT β variant was more highly expressed than the WT α by a factor ranging from $\sim$ 1.25 to 1.75, but less highly expressed or closer to par after eight weeks.

The clearest difference between liver and skeletal muscle tissue was expression of the transgene from vectors derived from the p Δ BssHII plasmid subclone at eight weeks. At one week the pair were the only group to exhibit a consistent reduction in RFO. This direction of the RFO underwent a reversal eight weeks in muscle tissue, but not in the liver tissue. This pair displayed α - and β - specific tendencies, with Δ BssHII β expressing more transcripts at both timepoints.



DNA transfer and persistence

Tissue DNA extracts vs injected vector
(**RFO**, relative to WTα)

$$\text{error bars} = \text{RFO} \times \sqrt{(\text{RSD}_{\text{Vec}})^2 + \left[\frac{6(\zeta_S)_{\text{DNA}}}{r_{\text{DNA}}} \right]^2}$$

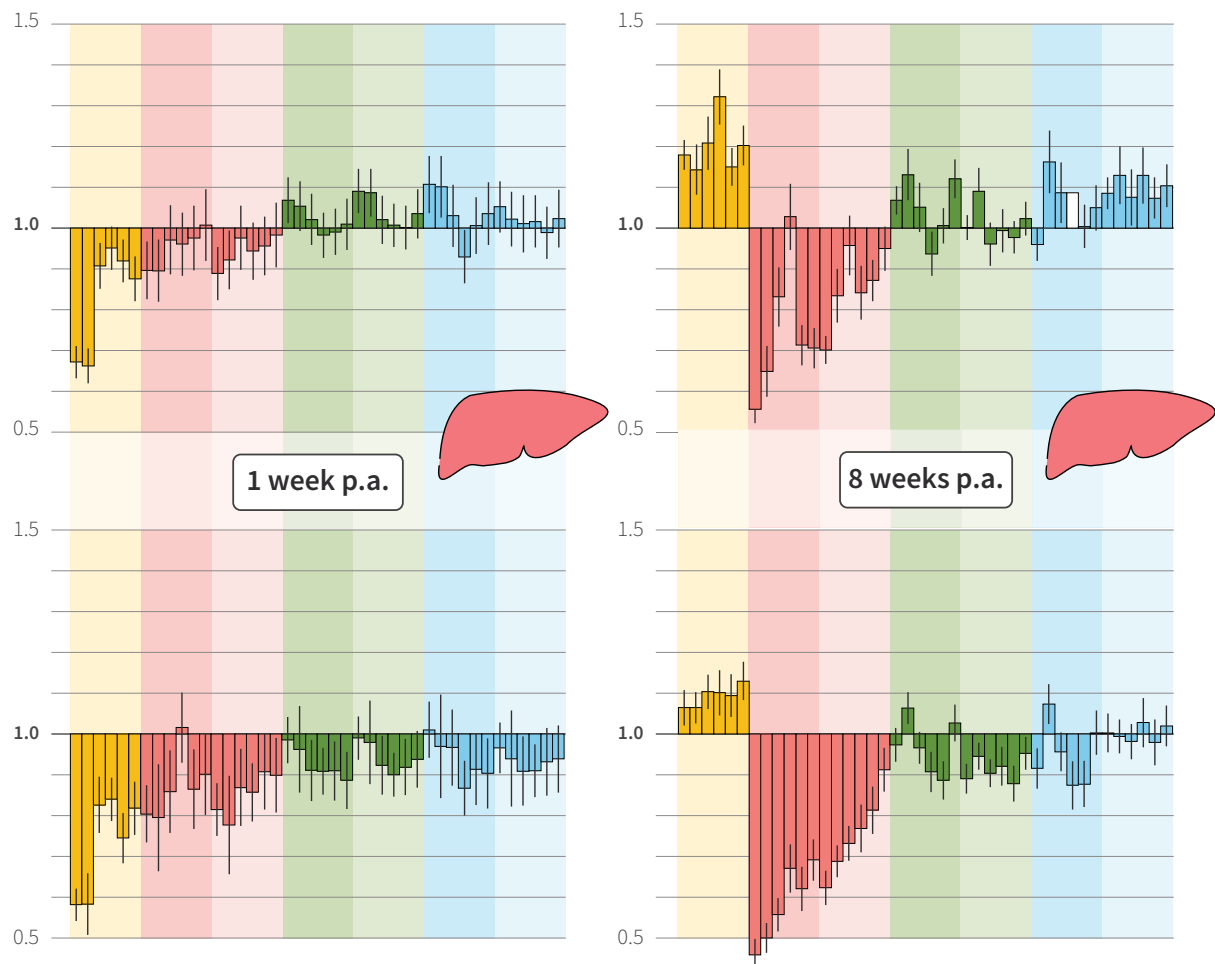
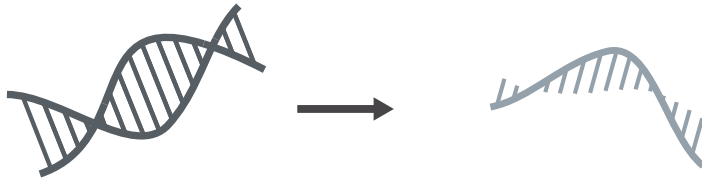
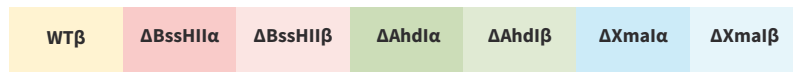


Figure 6.2.2 Short- and medium-term rAAV9-mediated gene transfer in murine liver and skeletal muscle | The RFO of gene transfer was measured, with the injected vector mix treated as input to both time points and DNA qRBA targets extracted from liver and skeletal muscle treated as outputs. Each column represents an individual mouse, and each set of error bars was calculated using the empirical error of the vector mix and a parametric error estimated from the DNA sample ($n = 1$), expanded by a factor of 6. The RFO of WTβ at Week 1 underwent an inversion by Week 8. In skeletal muscle, WTα outperformed all other variants in Week 1 whereas WTβ outperformed all other variants at Week 8. In the liver, the higher-RFO WT variant tended to behave comparably with the two Δ22bp variants. The poor gene transfer performance of the ΔBssHII variant was far more apparent at Week 8 than at one Week 1. The white box in the upper right panel represents an outlier (RFO = 2.6, caused by an abnormally high ΔXmaIα count in the liver of mouse 4), which was assumed to be erroneous and excluded from analysis.



Transgene transcription

RNA tissue extracts to DNA tissue extracts
(RFO, relative to WTα)

$$\text{error bars} = \text{RFO} \times \sqrt{\left[\frac{6(\zeta_S)_{\text{DNA}}}{r_{\text{DNA}}} \right]^2 + \left[\frac{6(\zeta_S)_{\text{RNA}}}{r_{\text{RNA}}} \right]^2}$$

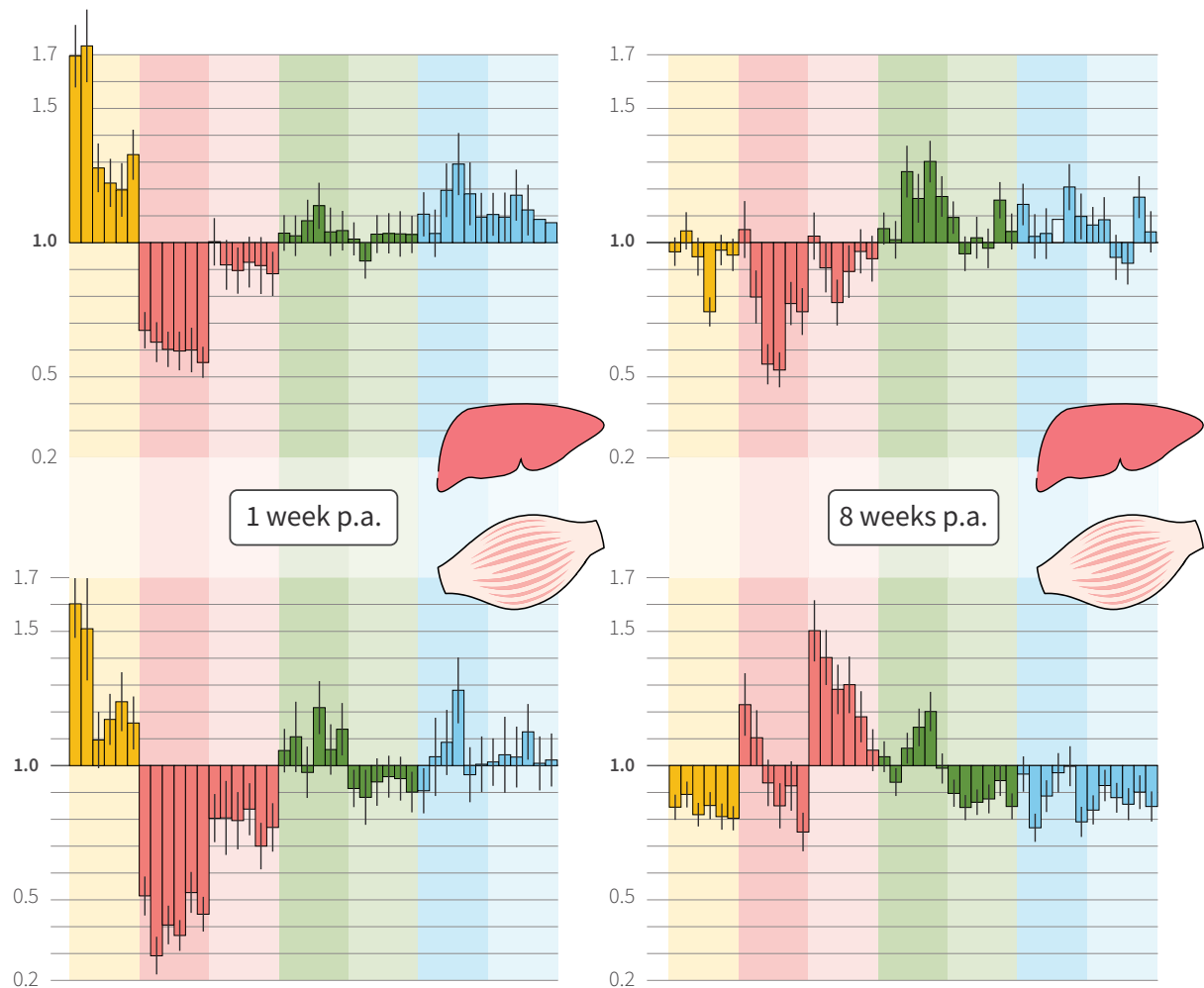


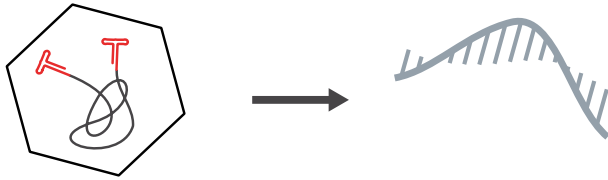
Figure 6.2.3 Transcription of transgene DNA from rAAV9 reporter variants | The RFO of transcription was measured using intracellular vector DNA targets as inputs and RNA targets from the same tissue as outputs. Transcription of qRBA targets resulted in a wider range of relative efficiencies compared to delivery of qRBA targets by the vector. The number of transcripts-per-WTβ-copy was similar to the number of transcripts-per-WTα-copy in the liver at Week 8, but lower by 20-70% at Week 1 (both in liver and muscle) and higher by ~10% at Week 8 in the muscle. Greater α vs β discrepancies were seen for the non-WT variants in this context than for gene transfer, most prominently in the ΔBssHII variants, though less so for the ΔXmaI variants. Opposing tendencies were observed for the WTβ variant in gene transfer compared to gene transcription, which were later formally analysed.

A directly functional perspective was taken by modelling the injected vector mix as input to mRNA transcripts measured one and eight weeks after administration (Fig 6.2.4). Variation between individual mice was notably more moderate in this analysis. Only two groups were consistently disadvantageous in strictly functional terms – $\Delta\text{BssHII}\alpha$ and $\text{WT}\alpha$. The ΔBssHII pair performed poorly overall, with noted exception of $\Delta\text{BssHII}\beta$ in skeletal muscle tissue at eight weeks. The best performance overall (performance taken to mean transcription at eight weeks) was seen in the ΔAhdI group, though this was specific to the α variant. The β variant was comparable with $\text{WT}\alpha$. In skeletal muscle the $\Delta\text{AhdI}\alpha$ variant gave the highest performance consistently ranging between ~ 1.2 to 1.3 times the fold output of $\text{WT}\alpha$. $\Delta\text{AhdI}\alpha$ averaged an RFO of ~ 1.3 in the liver in three individuals (Mice 3, 5 and 6) while the other three were comparable to $\text{WT}\beta$ at ~ 1.1 . The ΔXmaI variants performed comparably with $\text{WT}\beta$ though less well in skeletal muscle compared to the liver and with more notable individual variation seen in the β variant in the liver.

6.2.2.3 vector genome transfer was inversely correlated to vector genome transcription

Where $\text{WT}\beta$ was concerned, the RFO of expression tended to counteract the disadvantage in the number of stably transduced transgenes. To test this observation, transcription RFOs for $\text{WT}\beta$ in both tissues and time points ($n = 24$) were compared to those of gene transfer (Fig. 6.2.5a). An inverse correlation was confirmed, and a least-squares regression returned a highly predictive exponential model ($y = e^{-1.12x-1.27}$) with an R^2 of 0.99 and residuals (that is, $y_{[\text{observed}]} - e^{-1.12x-1.27}$) averaging 2.4%. To test this relationship in variants other than $\text{WT}\beta$ the same comparison was made, revealing most data from ΔAhdI and ΔXmaI to cluster along the same regression. However, separation of data into tissues, timepoints and α/β reporters revealed that specific trends were dependent on these categories, rather than randomly deviating from the $\text{WT}\beta/\text{WT}\alpha$ model (Fig. 6.2.5b). Regression statistics were not obtained, due to the level of dispersion and the low number of replicates. Notably though, in qualitative terms, there were no categories wherein $(_):\text{WT}\alpha$ ratios that were raised through gene transfer were *not* correlated with lowered $(_):\text{WT}\alpha$ ratios through transcription. The ΔBssHII variant conformed well to the $\text{WT}\beta/\text{WT}\alpha$ model, but only the β variant of ΔBssHII and only in the muscle at Week 8 ($n = 6$). These data were shown in Figure 6.2.5a for illustration. Other ΔBssHII data exhibited variable tendencies, as with the ΔAhdI and ΔXmaI variants.

WT β	Δ BssHII α	Δ BssHII β	Δ AhdI α	Δ AhdI β	Δ XmaI α	Δ XmaI β
------------	--------------------------	-------------------------	------------------------	-----------------------	------------------------	-----------------------



‘Functional transduction’

RNA tissue extracts vs injected vector)
(RFO, relative to WT α)

$$\text{error bars} = \text{RFO} \times \sqrt{(\text{RSD}_{\text{Vec}})^2 + \left[\frac{6(\zeta_S)_{\text{RNA}}}{r_{\text{RNA}}} \right]^2}$$

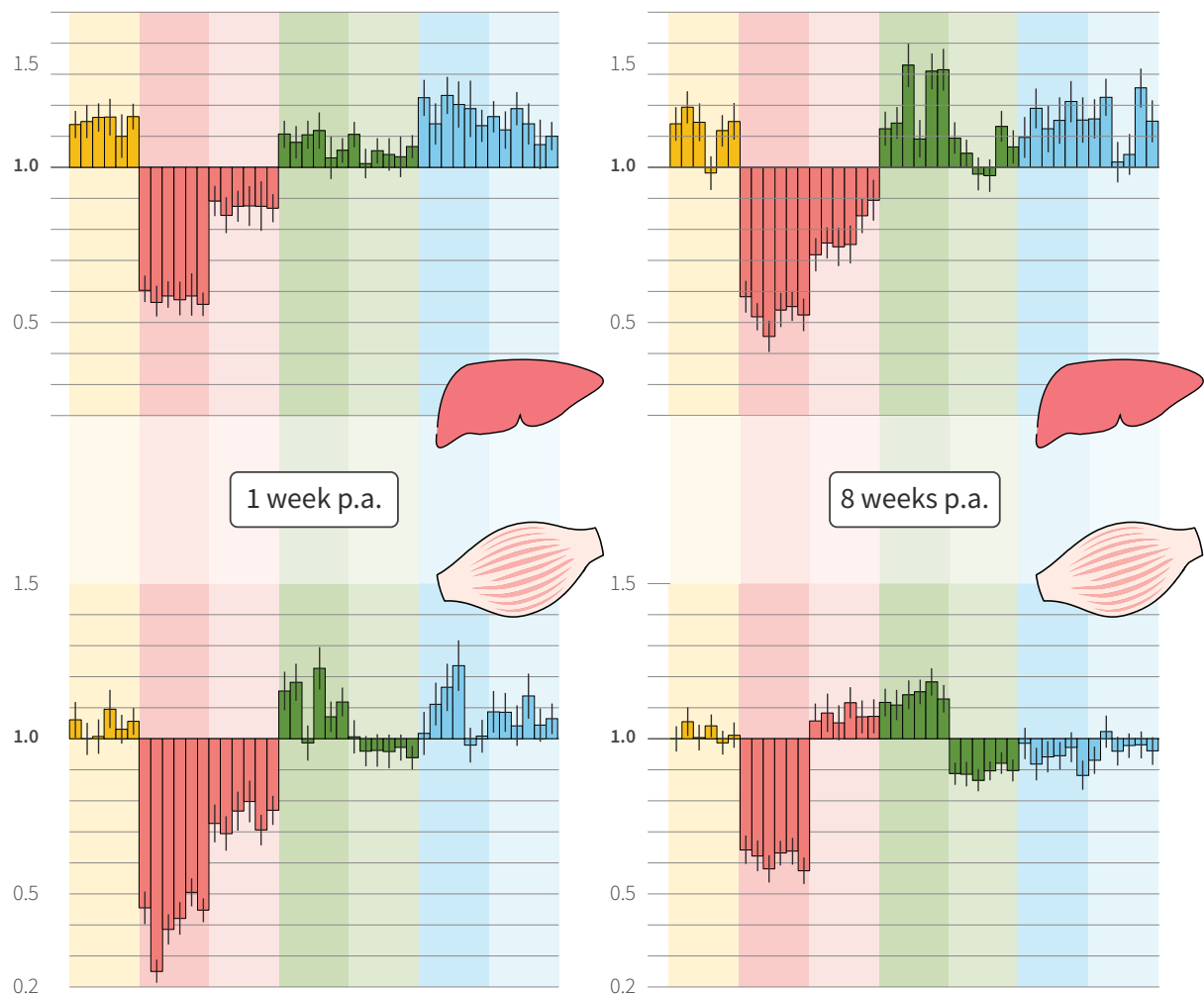


Figure 6.2.4 Production of mRNA transcripts as a functional output of the injected vector | The mRNA generated in target tissues was compared with the injected mixture as functional indicator of performance. A lesser degree of variation between individual mice is noted in these data, in spite of the more distant relationship between input and output. Outcomes at 8 weeks differed between liver and muscle, particularly in the discrepancy between the α and β variants of Δ BssHII and Δ AhdI targets. Relative to WT α , the Δ XmaI variant underperformed slightly in skeletal muscle but produced ~10-40% more mRNA transcripts in the liver. Interestingly, the performance of WT β was ~10-20% higher in the liver at both time points and generally divergent from WT α , but not obviously advantageous or disadvantageous in skeletal muscle at either time point.

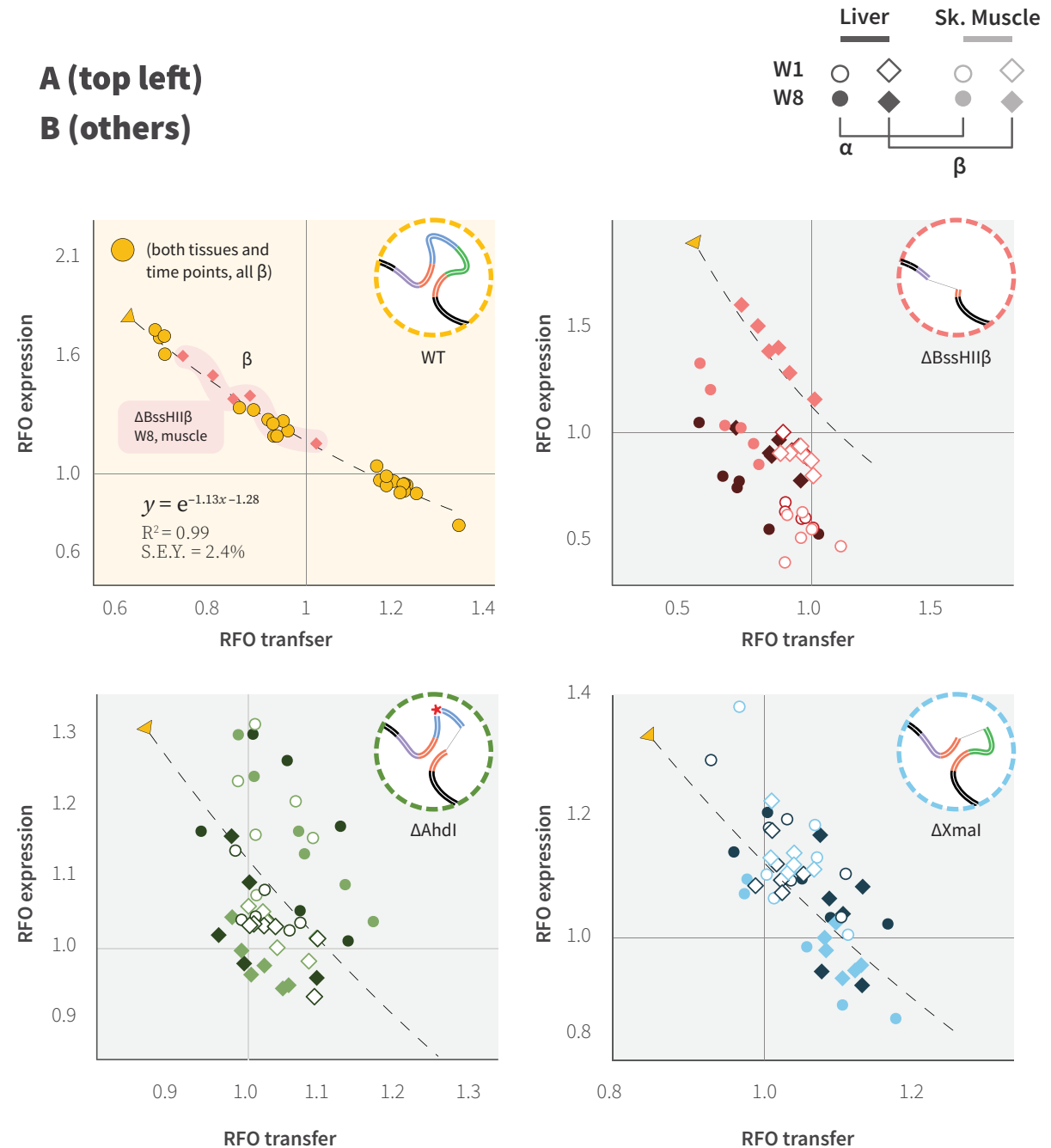


Figure 6.2.5 An increase in ratio through gene transfer correlated to a decrease in ratio through transcription | (A) For WTβ, it was observed that in mice with higher RFOs for gene transfer showed lower RFOs for transcription. To test this correlation formally, RFOs of gene transfer (for both time points and tissue types) were compared to those of gene transcription for each mouse. Gene transfer and gene transcription were found to correlate negatively, and fit well to the exponential model $y = e^{-1.13x-1.28}$, showing an R^2 of 0.99 and a standard residual of 2.4%. In skeletal muscle at week 8, the ΔBssHIIβ target also conformed to this model. **(B)** To explore this relationship for targets other than WTβ, the same analysis was applied to ΔBssHII, ΔAhdI and ΔXmal. Here, the data appeared to conform to this regression for the most part (albeit with notably higher dispersion) but separate analysis of tissues, time points and α/β subclones indicated that individual trends were dependent on these categories. Regression statistics for these trends would potentially be unreliable due to the higher levels of variation and lower number of replicates ($n = 6$), and so was not performed. Shown in all (B) plots is the regression model for WTβ, which is marked by the yellow triangle.

6.2.3 Discussion

In extending this investigation from vector production into vector efficacy, we continued to observe intraclonal (α vs β) variations in behaviour. In fact, α and β reporter partners hosted among the most marked differences between groups observed in the experiment. As with the production study, the WT-derived and Δ BssHII-derived products were the strongest in exhibiting this phenomenon, with the WT α and WT β variants undergoing a reversal in output from the first to the eighth week p.a., and the Δ BssHII β variant displaying a remarkable contrast in yield of mRNA in skeletal muscle (but not in the liver) at eight weeks. This difference was less pronounced when vector DNA was modelled as the input to transcription but was nonetheless clear.

The decision to pursue a longer-term transduction experiment was vindicated. In functional terms, the result with the most significance was the RFO of injected vectors into transgene-derived mRNA transcripts. Here, though the differences were less pronounced than in other analyses, we can observe how restriction to an earlier time point would have affected the reported outcomes for non-WT variants, particularly in skeletal muscle. Other points of contrast were less functionally relevant but enriching to the analysis, in particular the drastic reduction of the WT β subpopulation in stable episomes seven weeks after the initial sample.

The data obtained for this report became clearer and more comprehensible when RFO analysis was performed on individual mice, rather than on the average of multiple mice. The demonstrated precision of qRBA measurement, and the tendency for alternative tissue and nucleic acid types to corroborate, lent confidence that marked variations originated within the sample and not from statistical error or problems in sample preparation. This is perhaps best exhibited in the clarity of the regression formed by the RFOs of transfer and transcription between the WT α and WT β pair. To assist in interpreting this relationship: the more the WT β :WT α ratio changed from the vector mix to the intracellular DNA, the more it changed *in the opposite direction* from intracellular DNA to RNA. While the exponential mode of regression, (as opposed to some other mode) may be an artefact of qRBA error, the presence of a strongly predictive inverse relationship cannot be discounted. The quality of the regression strongly indicates some deterministic relationship between the two processes, where an increased presence of DNA is related inversely to its transcription. Interestingly, in skeletal muscle (by Week 8), the effect was to approximately neutralise discrepancies in behaviour

between the WT α and WT β variants. In the liver, however, the effect amounted to a ~10-20% higher yield of mRNA transcripts (except for Mouse 4).

It is useful to note that the inverse correlation in RFO for the WT α / β pair was highly predictive and consistent between tissue times and time points, whereas other pairings with WT α were not. It informs us, in the least, that such a regression was not guaranteed to occur by some artefact of qRBA/RFO analysis that we have not accounted for. Such a regression, assuming it existed for a given non-WT variant, was dependent on the time point, the tissue, and the variant in question. One suggested factor that would cause DNA and RNA targets to differ in abundance (as with conventional qPCR) is the potential for partial genomes to be packaged, containing the PCR target but not the transcriptional promoter. An experiment might be designed to explore this possibility, where the results of a vector pool such as this were analysed following isopycnic fractionation (which separates partial and total genomes) and compared to an indiscriminate method. Importantly though, this factor would not explain the time or tissue dependence of the transfer-transcription relationship. For an area where so little knowledge is possessed, we should be especially purposeful not to limit explanations to phenomena we are presently aware of.

The use of ζ_s to infer variance of NGS results is (according to a non-expert search of the literature) original to this thesis, and therefore not yet subject to review by statisticians or metrologists. The decision to deploy the metric was informed in large part by the formal validation performed in Chapter 4. Based on the evidence in that chapter the advice would be to find ζ_s and increase it by a factor of ~1.9. To lower the risk of underreporting uncertainty, this was raised to a factor of six. Strictly, this is not intended to act as a confidence interval. At present, the most important factor is to ensure the method for estimating uncertainty is clearly defined and accessible. In any case, the *least* valid option is to disregard uncertainty from unreplicated measurements simply because an empirical measure is unavailable.

6.3 Conclusions

Compared to all possible outcomes the results of this investigation decrease the potential alarm surrounding ITR integrity. The Δ AhdI and Δ XmaI variants did not display a consistent adverse impact on transgene expression achievable from the rAAV9 vector preparation tested. The performance of the Δ BssHII variant in this study was very poor, though this variant is a lesser

concern in light of other observations in the study: the loss of both hairpin arm sequences was a very infrequent occurrence in the colonies screened, and the Δ BssHII variant was challenging to produce in quantities adequate for the study. This encouragingly suggests that the abundance of this variant in wild-type-derived plasmid preps is low and that that abundance would only be reduced further by the production process. These are welcome results.

This should not be interpreted to mean that the urgency to investigate the problem has been reduced. On the contrary, this study has demonstrated unequivocally that rAAV vectors can package genomes with variable therapeutic efficacy, and that plasmid ITR integrity is a very likely contributor to this variation. Though non-threatening, the Δ BssHII variant in particular has demonstrated this phenomenon. Initially, it was posited that the moderate differences in Δ BssHII packaging efficiency (albeit into a different capsid variant), compared to markedly reduced genome accumulation, was a result of repair and enrichment of intact ITR genomes. The mouse study has demonstrated that the Δ BssHII-derived vector genomes, in addition to being competently packaged, led to very poor vector outcomes *in vivo*. This observation, along with the intraclonal variability of wild-type-derived products, should incite concerted efforts to understand the impact of ITR integrity.

One of the most essential responses to this study will be to investigate the validity, causes and implications of intraclonal behavioural differences. The assumption that these differences are artefacts of qRBA analysis is risky and not supported by the patterns of error presented in Chapter 4. If valid, this could represent a longstanding confounding variable with broad implications for rAAV research, both present and historical. Another response should be to survey more broadly the inverse relationship between taking up residency in transduced cells and producing mRNA transcripts. This consistent tendency has strong implications for how the presence of rAAV-delivered DNA in treated tissues is to be interpreted.

Finally, this investigation has vindicated the significant potential for quantitative reporter barcode analysis as a tool not only for technical measurement, but for the study of replication, translocation, expression, and modification of nucleic acids in general. The reason we might view the divergent intraclonal behaviours with incredulity could simply be that we have been previously unequipped to observe the phenomenon so unambiguously. Particularly without some degree of foreknowledge, this would be strenuous to convincingly demonstrate if each target of interest required individual handling and quantification with conventional PCR

techniques. The same applies equally to the linearly driven population shifts in rAAV production and the highly predictive relationship between transfer and expression of the WT α and WT β vector variants. By contrast, the experience of this qRBA study has been one where these numerous and intriguing paths to insight fell into our laps unintended. This speaks to the outstanding potential of qRBA as a complement to existing techniques in bioanalysis and strongly condones the promulgation of this strategy to the wider community of molecular biological research.

7

Discussion & conclusions

7.1 Discussion

This thesis undertook two distinct research pursuits, each explored in dedicated sections, allow their ideas and original evidence to develop without interruption. This discussion will reflect on these two pursuits in chronological terms, which will more naturally convey their shared relevance. As stated in the Preface, these two endeavours were not conceived as two parts of the same unitary project goal. Ascribing such a goal retrospectively would have had inelegant results. However, as was intended, each project directly served the strategic interests of the other and fortified its independent achievements. The success of the rAAV production assays and *in vivo* experiments attested strongly to the inherent potential and practical relevance of qRBA in molecular biology. Metrological studies in Chapter 4 rigorously validated the quantitative nature those experiments and supplied me with the empirical information needed to distil credible observations from the initial body of ITR-related results. As a result, this thesis was able to advance our current body of evidence concerning ITR integrity as a variable in rAAV production and gene delivery definitively. In addition, novel and intriguing observations that arose incidentally in the course of qRBA analysis were vindicated. These outcomes strongly encourage us to develop and expand qRBA in bioanalysis to exploit its fullest potential.

My doctoral research took shape in the field of rAAV vectorology, where the plasmid precursors to research and clinical vector materials are known to contain heterogenous ITRs. Weaknesses in the methodological infrastructure, the recalcitrant reputation of ITRs, and the progress made despite these problems, have persistently delayed efforts to understand the influence of plasmid ITR heterogeneity in rAAV research and translation. This situation is changing rapidly. Where previous techniques have failed to regard ITR integrity as a continuous variable, newer population sequencing technologies offer copy-to-copy ITR sequence data. More pressingly,

rAAVs are rapidly entering the clinical landscape. Here, the pressure from markets and their regulators to comprehend, measure and manage upstream determinants of batch success are incomparable with the translational research environment. One of this thesis's endeavours was to supply the field with reliable evidence as to whether ITR integrity requires more urgent attention.

To produce this evidence, I relied on UMIs endogenous to the target in question (the rAAV vector genome sequence) to reveal differences in process efficiency via their representation in Illumina short-read population sequence data. The expectation was that by avoiding countless sources of inter-sample variability inherent to q/dPCR approaches, a far more granular view of vector generation in producer cells could be attained. To implement this strategy, plasmids clonally derived from copies with or without intact 3'ITRs were fitted with clone-specific UMIs and mixed to simulated a continuous range of ITR integrity. Although the true fraction of intact ITRs was unknown, it would be directly proportional to fraction of wild-type-coded DNA targets. By measuring the UMI composition of these mixtures at different stages of intracellular production (and counting them against external 'reference cells') I could test the null hypothesis that barcode profiles would go unchanged. This investigation then extended to *in vivo* administration, pooling individual preps of an rAAV9 vector before purification and analysing the barcode composition of DNA and RNA extracted from murine liver and skeletal muscle 1 and 8 weeks after injection. This emulated previous studies on capsid variants⁵¹⁻⁵³ and transgene promoter variants⁵⁴, testing the null hypothesis that vectors products from plasmid 3'ITR variants would exhibit equal delivery, stability and expressivity over that period.

A key initial outcome was that measurement consistency across groups of wholly replicated experimental samples were comparable with previously published q/dPCR precision in purified rAAV samples. Strong linear relationships between input and output ratios of barcoded subpopulations could be clearly observed, representing consistent and highly replicable fold-outputs across the three %WT mixes tested. Interestingly, despite clear and consistent reductions in productivity within plasmid mixture (and between separate transfections pooled at harvest), the effect of the %WT variable on absolute yield was inconclusive. In murine liver and skeletal muscle, an unanticipated degree of inter-animal variability somewhat impeded the outcome of primary interest. However, variant-specific tendencies were plainly evident and all samples, in spite of inter-animal variability, exhibited a clear consensus as to negative correlations between persistence of a subpopulation over 1 and 8 weeks and the transcriptional

activity of that subpopulation. Puzzling observations arose from the use of ‘partnered’ α and β barcodes, the count data for which indicated unmistakeable differences in process efficiency between plasmids clonally derived from the same plasmid 3’ITR variant and expanded as a single culture. These observations comported with recent updates to our understanding of clonal mosaicism²³⁹, and were given further credence by consistent pairwise behaviours demonstrated in certain processes as opposed to others. Nonetheless, the validity of the overall dataset could not rest on these speculations.

Each of these outcomes, desired and undesired alike, strongly encouraged promulgation of this method and a deeper formal understanding of UMIs as quantitative reporters. Though not quite narratively adjacent, validating and sharing the concept of qRBA had a clear strategic rationale beyond the defence of my results. This obviously included the growing body of vectorological research utilising reporter barcodes. Less tangibly though, there was a disadvantage to sharing this evidence with an audience as yet unfamiliar with the assay design and the lines of reasoning it represented. Particularly where ITRs were concerned, establishing credibility for the qRBA approach was prudent. Further still, it would serve to engage thinkers from a larger community of researchers to contribute to qRBA learning and problem solving.

Several problems needed to be addressed to carry this out effectively. The technique needed to be conceptualised in terms accessible across a range of biological specialties, whose conception of a UMIs would vary from those in rAAV or lipid nanoparticle research. qRBA needed to be disentangled from a landscape of applications characteristically more stochastic and error prone, wherefrom UMI count analysis comes pre-associated with correlative modes of thinking. The disorganised language associated with UMIs was also an impediment. In response to these issues I undertook concerted efforts establish the concept of a ‘reporter barcode’, to fully articulate the conceptual boundaries of qRBA, and rationalise how these boundaries supported quantitative bioanalysis.

Another hurdle was in evaluating the precision of Illumina-based qRBA. For a pipette or an analytical balance, this would involve performing a sufficient number of valid technical replicates and reporting their variability with respect to their average⁷¹. By contrast, qRBA is a fundamentally stochastic process and its precision is governed by multiple parameters with specific relationships. To evaluate the precision of the Illumina qRBA process, I first needed to hypothesise what these were. Standard models for count analysis were derived from the

bi/multinomial distribution²⁵. This posed a minor barrier, because a Bernoulli process could have different possible ‘precisions’ depending on which barcode’s fraction was taken as the measurand. Though not a problem in a purely mathematical sense, taking the barcode ratio was better suited to the metrological thought process because the RSD was equal irrespective of how two reporter barcodes were categorised. A method of estimating the SD of this ratio was hypothesised and tested with to success. When expanded by a factor of 1.9 in the empirical setting, the predictiveness of this estimate (termed ζ) was indistinguishable between real Illumina qRBA replicates and idealised Bernoulli processes simulated *in silico*.

The accuracy of qRBA was tested by constructing plasmids containing two independent reporter PCR targets, with the rationale that the plasmid ratio was stoichiometrically guaranteed to be equal according to either target. Regrettably, this experiment succeeded in falsifying the null hypothesis – that qRBA of either target would return equal results. A pattern of nonrandom error was revealed, leading to misalignments approaching two-fold in the outer ranges. Further analysis demonstrated that sequencing errors contributed to less than 0.1% of the difference in barcode ratio across the measured range. Analyses also demonstrated that the misalignments did not conform with the binomial probability model, including its inherent asymmetries. In the greater context, these were of course highly desirable possibilities to exclude. Nonetheless, the error was not regarded as idiopathic. Trends in the misaligned measurements were consistent with that of a composite power function, indicating some mechanism whose effect is to exponentiate the true reporter ratio. A hyperbolic transformation of the data revealed that error parameters for a given barcode pair were not common to all samples, appearing instead to correlate with the reporter ratio itself. These patterns were not consistent with those expected from uneven amplification efficiency in the preparative PCR step, indicating that PCR was either a partial explanation or an invalid one.

This experiment found that qRBA measurements, if not accurate, are highly deterministic. The specific patterns of qRBA misalignment implied that qRBA data can be reliably interpreted in useful ways whether or not the error can ultimately be explained or excluded. A key finding was that changes in the ratio between two specific reporters from one population to another are strongly valid in direction and magnitude (though comparisons across more than two reporters are less reliable). This heuristic applies with greatest validity when r is near to 1 and becomes less reliable as r exceeds ~ 0.1 or 10. Stated more generally, changes to a barcode ratio are attributable to differences in the physical reporter variant ratio.

This scientific detour therefore proved extremely valuable to my analysis of ITR integrity in the rAAV vector system. Describing vector/plasmid performance ‘relative to the wild-type’ was not always possible because the α - and β -coded WT group themselves exhibited divergent behaviours. This was particularly true in the *in vivo* setting, where some of the most significant contrasts observed were attributable to the partner variants of the WT group. While prep-to-prep variations in plasmid performance could be attributed to variables such as physical plasmid integrity, plasmid quantification errors or bacterial endotoxins, these partnered reporter variants had co-existed in the same tube from the point of ligation and had been cultured and extracted as a single plasmid preparation. Alterations to the WT α :WT β ratio were not observed merely as variation between technical replicates, but reflected a consensus between three sets of transfection replicates from two unique sample types (comprising 26 samples in total). In the case of the Δ AhdI and Δ BssHII encapsidation efficiency, the range of ():WT input ratios meant that a consensus on α and β divergence could be observed even more robustly via the linear trend agreed upon by these 26 individual samples. More encouragingly still, all of the non-WT plasmid ITR variants exhibited well-matched behaviours in their α - and β -coded subpopulations, demonstrating definitively that the α / β discrepancies were not guaranteed to occur an artefact of qRBA measurement. Divergences between α and β variants of the same plasmid prep are not explained by any established vectorological knowledge. They therefore represent a newly observed phenomenon with similar adverse implications as ITR heterogeneity. That is to say, that some variable of unknown nature or quantity has been demonstrated to influence rAAV vector production and *in vivo* activity.

The establishment of ζ as an estimate of qRBA measurement variability also had direct significance to the experiment – and vice versa. The same analysis performed on NGS replicates in Chapter 4 was applied to the experimental replicates produced in Chapter 6.1. Although more widely dispersed than measurement replicates (as expected), the residuals between ζ_s and SD were normally distributed (albeit somewhat non-uniform over the ζ_s range). This supported the validity of ζ as a guiding principle for anticipating uncertainty in the experimental design process. This is analogous to the range of suitable target concentrations useful for Poisson statistics in dPCR applications⁶⁴. Validation of ζ significantly influenced the *in vivo* vector

studies. Despite the tendency for tissues or nucleic acid types to corroborate, the inter-animal variation observed in this setting non-trivial and cause for initial concern. Using ζ to estimate uncertainty attributable to the Illumina qRBA process (expanded to $6\cdot\zeta$ for conservativeness) substantiated the decision to perform RFO analysis on an animal-by-animal basis, rather than on averages of all six mice. Using the former approach, a negative correlation between gene delivery efficacy and transcriptional activity could readily be observed¹⁸⁹. The WT α and WT β pair were unique among reporter variants in that both tissues and time points conformed to a unified exponential regression ($n = 24$; S.E.Y = 2.4%). For non-WT-derived vectors, trends displayed a dependence on each of these factors.

7.2 Conclusions

Each of these investigational areas provides its own powerful impetus to validate qRBA more expansively so that it may be exploited to its fullest potential molecular biology. The nonrandom error observed in this thesis should not be regarded as idiopathic. The broader implications for NGS count analysis demand we investigate what leads conjugate qRBA measurements to misalign, and whether these causes are intrinsic to Illumina sequencing or the methodological concept in general. The success of the modified binomial model in this thesis encourages us to refine the model under the guidance of subject matter experts. In spite of the nonrandom error, this thesis has argued and empirically demonstrated that qRBA should not be thought of as less stochastic or error prone variant of more complex reporter screens or transcriptomic analysis – it is a fundamentally quantitative technique that can be understood according to binomial probability theory and exploited to meet bioanalytical challenges underserved by PCR-based methods. It offers not just variety to nucleic acid quantification, but outstanding precision.

The scope of research into ITR integrity was narrow – to determine whether ITR integrity could act as an upstream determinant of vector yield or performance. Unambiguous evidence was generated to show that intracellular genome processivity in adherent HEK293 cells was impacted by variation in the 3'ITR sequence of the rAAV transfer plasmid. Further to this, unidentified variables affecting subpopulations of the same plasmid prep were newly observed, which may or may not relate to its intact ITR content. These unidentified plasmid attributes were shown to affect processivity in producer cells, and go on to affect the stability of rAAV

transgenes and their transcriptional activity. Intra-prep variability itself was so pronounced that the effect of ITR variation was not always easy to demarcate, making it difficult to determine how variants performed against the wild-type in certain processes. Whether or not these sources of variation ultimately prove burdensome is secondary to our present need to identify and understand them. To this end, this thesis has showcase the capabilities of qRBA strategies and established important groundwork toward its greater adoption.

7.3 References

1. Rosner, B. *Fundamentals of biostatistics*. (Cengage Learning, 2015).
2. Wypij, D. in *Wiley StatsRef Stat. Ref. Online* 1–7 (John Wiley & Sons, Ltd, 2014).
3. Westhaus, A., Cabanes Creus, M., Rybicki, A., Baltazar, G., Navarro, R. G., Zhu, E., Drouyer, M., Knight, M., Albu, R. F., Ng, B. H., Kalajdzic, P., Kwiatek, M., Hsu, K., Santilli, G., Gold, W., Kramer, B., Gonzalez-Cordero, A., Thrasher, A. J., Alexander, I. E. & Lisowski, L. High throughput in vitro, ex vivo and in vivo screen of AAV vectors based on physical and functional transduction. *Hum. Gene Ther.* **31**, hum.2019.264 (2020).
4. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
5. Galvis, A. E., Fisher, H. & Camerini, D. NP-40 Fractionation and Nucleic Acid Extraction in Mammalian Cells. *Bio-protocol* **7**, (2017).
6. Bardelli, M., Zarate-Perez, F., Agundez, L., Jolinon, N., Michael Linden, R., Escalante, C. R. & Henckaerts, E. Analysis of replicative intermediates of adeno-associated virus through Hirt extraction and Southern blotting. *Bio-protocol* **7**, (2017).
7. Arad, U. Modified Hirt procedure for rapid purification of extrachromosomal DNA from mammalian cells. *Biotechniques* **24**, 760–762 (1998).
8. Dobnik, D., Kogovšek, P., Jakomin, T., Košir, N., Žnidarič, M. T., Leskovec, M., Kaminsky, S. M., Mostrom, J., Lee, H. & Ravnikar, M. Accurate quantification and characterization of adeno-associated viral vectors. *Front. Microbiol.* **10**, (2019).

9. Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.* **122**, e59 (2018).
10. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **82**, 801–811 (2021).
11. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
12. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).
13. Manley, L. J., Ma, D. & Levine, S. S. Monitoring error rates in Illumina sequencing. *J. Biomol. Tech.* **27**, 125–128 (2016).
14. Franklin, J. in *Neo-Aristotelian Perspect. Metaphys.* (eds. Novotný, D. & Novák, L.) 221–244 (Routledge, 2013).
15. Mari, L., Wilson, M. & Maul, A. in *Meas. Across Sci. Dev. a Shar. Concept Syst. Meas.* (eds. Mari, L., Wilson, M. & Maul, A.) 81–121 (Springer International Publishing, 2023).
16. Tal, E. Old and new problems in philosophy of measurement. *Philos. Compass* **8**, 1159–1173 (2013).
17. Lewis, J., Feltus, A., Ensor, M., Ramanathan, S. & Daunert, S. Peer reviewed: Applications of reporter genes. *Anal. Chem.* **70**, 579A–585A (2011).
18. Hajibabaei, M. The golden age of DNA metasystematics. *Trends Genet.* **28**, 535–537 (2012).
19. Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J. & van Opijnen, T. A decade of advances in transposon-insertion sequencing. *Nat. Rev. Genet.* **21**, 526–540 (2020).
20. Johnson, M. S., Martsul, A., Kryazhimskiy, S. & Desai, M. M. Higher-fitness yeast genotypes are less robust to deleterious mutations. *Science (80-.)*. **366**, 490–493 (2019).
21. Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P. & Deutschbauer, A. Rapid quantification of

- mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, e00306-15 (2015).
22. Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S. & Sherlock, G. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
 23. Bram, V. den B., Toon, S., Maarten, F. & Jan, M. Experimental design, population dynamics, and diversity in microbial experimental evolution. *Microbiol. Mol. Biol. Rev.* **82**, e00008-18 (2018).
 24. Davidsson, M., Wang, G., Aldrin-Kirk, P., Cardoso, T., Nolbrant, S., Hartnor, M., Mudannayake, J., Parmar, M. & Björklund, T. A systematic capsid evolution approach performed in vivo for the design of AAV vectors with tailored properties and tropism. *Proc. Natl. Acad. Sci.* **116**, 27053–27062 (2019).
 25. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nat. Preced.* **1**, 1–10 (2010).
 26. Sena, J. A., Galotto, G., Devitt, N. P., Connick, M. C., Jacobi, J. L., Umale, P. E., Vidali, L. & Bell, C. J. Unique molecular identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci. Rep.* **8**, 13121 (2018).
 27. Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D. & Weng, Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 531 (2018).
 28. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. & Linnarsson, S. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
 29. Vially, P.-J., Sater, V., Viennot, M., Bohers, E., Vergne, N., Berard, C., Dauchel, H., Lecroq, T., Celebi, A., Ruminy, P., Marchand, V., Lanic, M.-D., Dubois, S., Penther, D., Tilly, H., Mareschal, S. & Jardin, F. Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers. *BMC Bioinformatics* **22**, 120 (2021).
 30. Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. & Taipale,

- J. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
31. MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M. S., Ducar, M. D., Meyerson, M. & Thorner, A. R. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 30 (2018).
 32. McCloskey, M. L., Stöger, R., Hansen, R. S. & Laird, C. D. Encoding PCR products with batch-stamps and barcodes. *Biochem. Genet.* **45**, 761–767 (2007).
 33. Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St.Onge, R. P., Tyers, M., Koller, D., Altman, R. B., Davis, R. W., Nislow, C. & Giaever, G. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* (80-.). **320**, 362–365 (2008).
 34. Smith, A. M., Heisler, L. E., Mellor, J., Kaper, F., Thompson, M. J., Chee, M., Roth, F. P., Giaever, G. & Nislow, C. Quantitative phenotyping via deep barcode sequencing. *Genome Res.* **19**, 1836–1842 (2009).
 35. Adachi, K., Enoki, T., Kawano, Y., Veraz, M. & Nakai, H. Drawing a high-resolution functional map of adeno-associated virus capsid by massively parallel sequencing. *Nat. Commun.* **5**, 1–14 (2014).
 36. Dahlman, J. E., Kauffman, K. J., Xing, Y., Shaw, T. E., Mir, F. F., Dlott, C. C., Langer, R., Anderson, D. G. & Wang, E. T. Barcoded nanoparticles for high throughput in vivo discovery of targeted therapeutics. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2060–2065 (2017).
 37. Lokugamage, M. P., Sago, C. D. & Dahlman, J. E. Testing thousands of nanoparticles in vivo using DNA barcodes. *Curr. Opin. Biomed. Eng.* **7**, 1–8 (2018).
 38. Sago, C. D., Lokugamage, M. P., Paunovska, K., Vanover, D. A., Monaco, C. M., Shah, N. N., Castro, M. G., Anderson, S. E., Rudoltz, T. G., Lando, G. N., Tiwari, P. M., Kirschman, J. L., Willett, N., Jang, Y. C., Santangelo, P. J., Bryksin, A. V. & Dahlman, J. E. High-throughput in vivo screen of functional mRNA delivery identifies nanoparticles for endothelial cell gene editing. *Proc. Natl. Acad. Sci.* **115**, E9944–E9952 (2018).

39. Dobrowolski, C., Paunovska, K., Schrader Echeverri, E., Loughrey, D., Da Silva Sanchez, A. J., Ni, H., Hatit, M. Z. C., Lokugamage, M. P., Kuzminich, Y., Peck, H. E., Santangelo, P. J. & Dahlman, J. E. Nanoparticle single-cell multiomic readouts reveal that cell heterogeneity influences lipid nanoparticle-mediated messenger RNA delivery. *Nat. Nanotechnol.* **17**, 871–879 (2022).
40. Ni, H., Hatit, M. Z. C., Zhao, K., Loughrey, D., Lokugamage, M. P., Peck, H. E., Cid, A. Del, Muralidharan, A., Kim, Y., Santangelo, P. J. & Dahlman, J. E. Piperazine-derived lipid nanoparticles deliver mRNA to immune cells in vivo. *Nat. Commun.* **13**, 4766 (2022).
41. Huayamares, S. G., Lokugamage, M. P., Rab, R., Da Silva Sanchez, A. J., Kim, H., Radmand, A., Loughrey, D., Lian, L., Hou, Y., Achyut, B. R., Ehrhardt, A., Hong, J. S., Sago, C. D., Paunovska, K., Echeverri, E. S., Vanover, D., Santangelo, P. J., Sorscher, E. J. & Dahlman, J. E. High-throughput screens identify a lipid nanoparticle that preferentially delivers mRNA to human tumors in vivo. *J. Control. Release* **357**, 394–403 (2023).
42. Schoenmaker, L., Witzigmann, D., Kulkarni, J. A., Verbeke, R., Kersten, G., Jiskoot, W. & Crommelin, D. J. A. mRNA-lipid nanoparticle COVID-19 vaccines: Structure and stability. *Int. J. Pharm.* **601**, 120586 (2021).
43. Marsic, D., Govindasamy, L., Currin, S., Markusic, D. M., Tseng, Y. S., Herzog, R. W., Agbandje-McKenna, M. & Zolotukhin, S. Vector design Tour de Force: integrating combinatorial and rational approaches to derive novel adeno-associated virus variants. *Mol. Ther.* **22**, 1900–1909 (2014).
44. Marsic, D., Méndez-Gómez, H. R. & Zolotukhin, S. High-accuracy biodistribution analysis of adeno-associated virus variants by double barcode sequencing. *Mol. Ther. - Methods Clin. Dev.* **2**, 15041 (2015).
45. Weinmann, J. & Grimm, D. Next-generation AAV vectors for clinical use: an ever-accelerating race. *Virus Genes* **53**, 707–713 (2017).
46. Herrmann, A.-K. & Grimm, D. High-throughput dissection of AAV–host interactions: the fast and the curious. *J. Mol. Biol.* **430**, 2626–2640 (2018).

47. Szumska, J. & Grimm, D. Boosters for adeno-associated virus (AAV) vector (r)evolution. *Cytotherapy* **25**, 254–260 (2023).
48. Maheshri, N., Koerber, J. T., Kaspar, B. K. & Schaffer, D. V. Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nat. Biotechnol.* **24**, 198–204 (2006).
49. Meumann, N., Cabanes-Creus, M., Ertelt, M., Navarro, R. G., Lucifora, J., Yuan, Q., Nien-Huber, K., Abdelrahman, A., Vu, X.-K., Zhang, L., Franke, A.-C., Schmithals, C., Piiper, A., Vogt, A., Gonzalez-Carmona, M., Frueh, J. T., Ullrich, E., Meuleman, P., Talbot, S. R., Odenthal, M., Ott, M., Seifried, E., Schoeder, C. T., Schwäble, J., Lisowski, L. & Büning, H. Novel adeno-associated virus (AAV) serotype 2 capsid variants for improved liver-directed gene therapy. *Hepatology* (2022).
50. Cabanes-Creus, M., Hallwirth, C. V, Westhaus, A., Ng, B. H., Liao, S. H. Y., Zhu, E., Navarro, R. G., Baltazar, G., Drouyer, M., Scott, S., Logan, G. J., Santilli, G., Bennett, A., Ginn, S. L., McCaughan, G., Thrasher, A. J., Agbandje-McKenna, M., Alexander, I. E. & Lisowski, L. Restoring the natural tropism of AAV2 vectors for human liver. *Sci. Transl. Med.* **12**, eaba3312 (2020).
51. Cabanes-Creus, M., Navarro, R. G., Zhu, E., Baltazar, G., Liao, S. H. Y., Drouyer, M., Amaya, A. K., Scott, S., Nguyen, L. H., Westhaus, A., Hebben, M., Wilson, L. O. W., Thrasher, A. J., Alexander, I. E. & Lisowski, L. Novel human liver-tropic AAV variants define transferable domains that markedly enhance the human tropism of AAV7 and AAV8. *Mol. Ther. - Methods Clin. Dev.* **24**, 88–101 (2022).
52. Cabanes-Creus, M., Westhaus, A., Navarro, R. G., Baltazar, G., Zhu, E., Amaya, A. K., Liao, S. H. Y., Scott, S., Sallard, E., Dilworth, K. L., Rybicki, A., Drouyer, M., Hallwirth, C. V, Bennett, A., Santilli, G., Thrasher, A. J., Agbandje-McKenna, M., Alexander, I. E. & Lisowski, L. Attenuation of heparan sulfate proteoglycan binding enhances in vivo transduction of human primary hepatocytes with AAV2. *Mol. Ther. - Methods Clin. Dev.* **17**, 1139–1154 (2020).
53. Cabanes-Creus, M., Navarro, R. G., Liao, S. H. Y., Baltazar, G., Drouyer, M., Zhu, E., Scott, S., Luong, C., Wilson, L. O. W., Alexander, I. E. & Lisowski, L. Single amino acid insertion allows functional transduction of murine hepatocytes with human liver tropic

- AAV capsids. *Mol. Ther. - Methods Clin. Dev.* **21**, 607–620 (2021).
54. Westhaus, A., Cabanes-Creus, M., Jonker, T., Sallard, E., Navarro, R. G., Zhu, E., Baltazar Torres, G., Lee, S., Wilmott, P., Gonzalez-Cordero, A., Santilli, G., Thrasher, A. J., Alexander, I. E. & Lisowski, L. AAV-p40 bioengineering platform for variant selection based on transgene expression. *Hum. Gene Ther.* **33**, 664–682 (2022).
 55. Kularatne, R. N., Crist, R. M. & Stern, S. T. The future of tissue-targeted lipid nanoparticle-mediated nucleic acid delivery. *Pharmaceuticals* **15**, at <https://doi.org/10.3390/ph15070897> (2022)
 56. Yang, L., Gong, L., Wang, P., Zhao, X., Zhao, F., Zhang, Z., Li, Y. & Huang, W. Recent advances in lipid nanoparticles for delivery of mRNA. *Pharmaceutics* **14**, at <https://doi.org/10.3390/pharmaceutics14122682> (2022)
 57. Rogers-Broadway, K. & Karteris, E. Amplification efficiency and thermal stability of qPCR instrumentation: Current landscape and future perspectives. *Exp. Ther. Med.* **10**, 1261–1264 (2015).
 58. Svec, D., Tichopad, A., Novosadova, V., Pfaffl, M. W. & Kubista, M. How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments. *Biomol. Detect. Quantif.* **3**, 9–16 (2015).
 59. Mamedov, T. G., Pienaar, E., Whitney, S. E., TerMaat, J. R., Carvill, G., Goliath, R., Subramanian, A. & Viljoen, H. J. A fundamental study of the PCR amplification of GC-rich DNA templates. *Comput. Biol. Chem.* **32**, 452–457 (2008).
 60. Fan, H., Wang, J., Komiyama, M. & Liang, X. Effects of secondary structures of DNA templates on the quantification of qPCR. *J. Biomol. Struct. Dyn.* **37**, 2867–2874 (2019).
 61. Ruiz-Villalba, A., van Pelt-Verkuil, E., Gunst, Q. D., Ruijter, J. M. & van den Hoff, M. J. B. Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). *Biomol. Detect. Quantif.* **14**, 7–18 (2017).
 62. Ma, S., Huang, Y. & van Huystee, R. B. Improved plant RNA stability in storage. *Anal. Biochem.* **326**, 122–124 (2004).
 63. Fabre, A.-L., Colotte, M., Luis, A., Tuffet, S. & Bonnet, J. An efficient method for long-

- term room temperature storage of RNA. *Eur. J. Hum. Genet.* **22**, 379–385 (2014).
64. Basu, A. S. Digital Assays Part I: Partitioning Statistics and Digital PCR. *SLAS Technol.* **22**, 369–386 (2017).
65. Ding, Y., Howes, P. D. & deMello, A. J. Recent Advances in Droplet Microfluidics. *Anal. Chem.* **92**, 132–149 (2020).
66. Tellinghuisen, J. & Spiess, A.-N. Statistical uncertainty and its propagation in the analysis of quantitative polymerase chain reaction data: Comparison of methods. *Anal. Biochem.* **464**, 94–102 (2014).
67. Bilgrau, A. E., Falgreen, S., Petersen, A., Kjeldsen, M. K., Bødker, J. S., Johnsen, H. E., Dybkær, K. & Bøgsted, M. Unaccounted uncertainty from qPCR efficiency estimates entails uncontrolled false positive rates. *BMC Bioinformatics* **17**, 159 (2016).
68. Gustavo González, A. & Ángeles Herrador, M. A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles. *TrAC Trends Anal. Chem.* **26**, 227–238 (2007).
69. Ferrero, A. & Salicone, S. Measurement uncertainty. *IEEE Instrum. Meas. Mag.* **9**, 44–51 (2006).
70. van der Veen, A. M. H. & Cox, M. G. Error analysis in the evaluation of measurement uncertainty. *Metrologia* **40**, 42 (2003).
71. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP & OIML. *Guide to the Expression of Uncertainty in Measurement*. (International Organisation for Standardisation, 1995).
72. Mazumdar, M., Banerjee, S. & Van Epps, H. L. in *Stat. Methods Mol. Biol.* (eds. Bang, H., Zhou, X. K., van Epps, H. L. & Mazumdar, M.) 563–598 (Humana Press, 2010).
73. Košir, A. B., Divieto, C., Pavšič, J., Pavarelli, S., Dobnik, D., Dreo, T., Bellotti, R., Sassi, M. P. & Žel, J. Droplet volume variability as a critical factor for accuracy of absolute quantification using droplet digital PCR. *Anal. Bioanal. Chem.* **409**, 6689–6697 (2017).
74. Pinheiro, L. B., Coleman, V. A., Hindson, C. M., Herrmann, J., Hindson, B. J., Bhat, S. & Emslie, K. R. Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Anal. Chem.* **84**, 1003–1011 (2012).

75. Borror, C. M. in *Wiley StatsRef Stat. Ref. Online* (2014).
76. Smith, E. P. in *Wiley StatsRef Stat. Ref. Online* (2014).
77. Cameron, J. M. in *Wiley StatsRef Stat. Ref. Online* (2014).
78. Possolo, A. in *Wiley StatsRef Stat. Ref. Online* 1–8 (2018).
79. Meyer, V. R. Measurement uncertainty. *J. Chromatogr. A* **1158**, 15–24 (2007).
80. Yanai, R. D., Mann, T. A., Hong, S. D., Pu, G. & Zukswert, J. M. The current state of uncertainty reporting in ecosystem studies: a systematic evaluation of peer-reviewed literature. *Ecosphere* **12**, e03535 (2021).
81. Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014).
82. Bustin, S. & Nolan, T. Talking the talk, but not walking the walk: RT-qPCR as a paradigm for the lack of reproducibility in molecular research. *Eur. J. Clin. Invest.* **47**, 756–774 (2017).
83. Na, J., Yang, H., Bae, S. & Lim, K. Analysis of statistical methods currently used in toxicology Journals. *Toxicol. Res.* **30**, 185–191 (2014).
84. Briner, W. & Kirwan, J. Experimental toxicology: Issues of statistics, experimental design, and replication. *Neurotoxicology* **58**, 137–142 (2017).
85. Ko, W.-R., Hung, W.-T., Chang, H.-C. & Lin, L.-Y. Inappropriate use of standard error of the mean when reporting variability of study samples: A critical evaluation of four selected journals of obstetrics and gynecology. *Taiwan. J. Obstet. Gynecol.* **53**, 26–29 (2014).
86. Auer, P. L. & Doerge, R. W. Statistical design and analysis of RNA sequencing data. *Genetics* **185**, 405–416 (2010).
87. Weinmann, J., Weis, S., Sippel, J., Tulalamba, W., Remes, A., Andari, J. El, Herrmann, A.-K., Pham, Q. H., Borowski, C., Hille, S., Schönberger, T., Frey, N., Lenter, M., VandenDriessche, T., Müller, O. J., Chuah, M. K., Lamla, T. & Grimm, D. Identification of a myotropic AAV by massively parallel in vivo evaluation of barcoded capsid variants.

Nat. Commun. 2020 111 **11**, 1–12 (2020).

88. Pratt, J. W. A normal approximation for binomial, F, Beta, and other common, related tail probabilities, II. *J. Am. Stat. Assoc.* **63**, 1457–1483 (1968).
89. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **16**, 101–117 (2001).
90. Jones, G. M., Busby, E., Garson, J. A., Grant, P. R., Nastouli, E., Devonshire, A. S. & Whale, A. S. Digital PCR dynamic range is approaching that of real-time quantitative PCR. *Biomol. Detect. Quantif.* **10**, 31–33 (2016).
91. Mirasol, F. Industry pushes forward with cell and gene therapies. *Pharm. Technol.* **44**, (2020).
92. Koonin, E. V. & Starokadomskyy, P. Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 125–134 (2016).
93. Flint, S. J., Racaniello, V. R., Rall, G. F., Skalka, A. M. & Enquist, L. W. *Principles of Virology*. (ASM Press Washington, DC, 2015).
94. Lundstrom, K. Viral vectors in gene therapy. *Diseases* **6**, at <https://doi.org/10.3390/diseases6020042> (2018)
95. Wright, J. F. Transient transfection methods for clinical adeno-associated viral vector production. *Hum. Gene Ther.* **20**, 698–706 (2009).
96. Nathwani, A. C., Reiss, U., Tuddenham, E., Chowdary, P., McIntosh, J., Riddell, A., Pie, J., Mahlangu, J. N., Recht, M., Shen, Y.-M., Halka, K. G., Meagher, M. M., Nienhuis, A. W., Davidoff, A. M., Mangles, S., Morton, C. L., Junfang, Z. & Radulescu, V. C. Adeno-associated mediated gene transfer for hemophilia B: 8 year follow up and impact of removing “empty viral particles” on safety and efficacy of gene transfer. *Blood* **132**, 491 (2018).
97. Sehara, Y., Fujimoto, K., Ikeguchi, K., Katakai, Y., Ono, F., Takino, N., Ito, M., Ozawa, K. & Muramatsu, S. Persistent expression of dopamine-synthesizing enzymes 15 years after gene transfer in a primate model of Parkinson’s disease. *Hum. Gene Ther. Clin. Dev.* **28**,

- 74–79 (2017).
98. Berns, K. I. & Giraud, C. in *Adeno-Associated Virus Vectors Gene Ther.* (eds. Berns, K. I. & Giraud, C.) 1–23 (Springer Berlin Heidelberg, 1996).
 99. Sonntag, F., Schmidt, K. & Kleinschmidt, J. A. A viral assembly factor promotes AAV2 capsid formation in the nucleolus. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10220–5 (2010).
 100. Cao, M., Chiriva-Internati, M. & Hermonat, P. L. AAV2 X increases AAV6 rep/cap-driven rAAV production. *Virology* **482**, 84–88 (2015).
 101. Lisowski, L., Tay, S. S. & Alexander, I. E. Adeno-associated virus serotypes for gene therapeutics. *Curr. Opin. Pharmacol.* **24**, 59–67 (2015).
 102. Zincarelli, C., Soltys, S., Rengo, G. & Rabinowitz, J. E. Analysis of AAV serotypes 1–9 mediated gene expression and tropism in mice after systemic injection. *Mol. Ther.* **16**, 1073–1080 (2008).
 103. Grimm, D., Kay, M. A. & Kleinschmidt, J. A. Helper virus-free, optically controllable, and two-plasmid-based production of adeno-associated virus vectors of serotypes 1 to 6. *Mol. Ther.* **7**, 839–850 (2003).
 104. Grimm, D. & Zolotukhin, S. E Pluribus Unum: 50 years of research, millions of viruses, and one goal—tailored acceleration of AAV evolution. *Mol. Ther.* **23**, 1819–1831 (2015).
 105. Wu, Z., Asokan, A. & Samulski, R. J. Adeno-associated virus serotypes: vector toolkit for human gene therapy. *Mol. Ther.* **14**, 316–327 (2006).
 106. Balakrishnan, B. & Jayandharan, G. R. Basic biology of adeno-associated virus (AAV) vectors used in gene therapy. *Curr. Gene Ther.* **14**, 86–100 (2014).
 107. Ding, W., Zhang, L., Yan, Z. & Engelhardt, J. F. Intracellular trafficking of adeno-associated viral vectors. *Gene Ther.* **12**, 873–880 at <https://doi.org/10.1038/sj.gt.3302527> (2005)
 108. Vihinen-Ranta, M., Suikkanen, S. & Parrish, C. R. Pathways of cell infection by parvoviruses and adeno-associated viruses. *J. Virol.* **78**, 6709–6714 (2004).
 109. Nonnenmacher, M. & Weber, T. Intracellular transport of recombinant adeno-

- associated virus vectors. *Gene Ther.* **19**, 649–658 (2012).
110. Ward, P., Elias, P. & Linden, R. M. Rescue of the adeno-associated virus genome from a plasmid vector: evidence for rescue by replication. *J. Virol.* **77**, 11480–11490 (2003).
 111. Wang, X. S., Ponnazhagan, S. & Srivastava, A. Rescue and replication of adeno-associated virus type 2 as well as vector DNA sequences from recombinant plasmids containing deletions in the viral inverted terminal repeats: selective encapsidation of viral genomes in progeny virions. *J. Virol.* **70**, 1668–77 (1996).
 112. Brister, J. R. & Muzyczka, N. Mechanism of Rep-mediated adeno-associated virus origin nicking. *J. Virol.* **74**, 7762–71 (2000).
 113. Ward, P. in *Parvoviruses* (eds. Kerr, J. R., Cotmore, S. F., Bloom, M. E., Linden, R. M. & Parrish, C. R.) 190–211 (Hodder Arnold, 2005).
 114. Duan, D., Yan, Z., Yue, Y. & Engelhardt, J. F. Structural analysis of adeno-associated virus transduction circular intermediates. *Virology* **8**, 8–14 (1999).
 115. Zhou, X., Zeng, X., Fan, Z., Li, C., McCown, T., Samulski, R. J. & Xiao, X. Adeno-associated virus of a single-polarity DNA genome is capable of transduction in vivo. *Mol. Ther.* **16**, 494–499 (2008).
 116. Timpe, J., Bevington, J., Casper, J., Dignam, J. & Trempe, J. Mechanisms of adeno-associated virus genome encapsidation. *Curr. Gene Ther.* **5**, 273–284 (2005).
 117. Wang Y, L. C., Wang L, L. Y. & Aslanidi GV, J. G. The adeno-associated virus genome packaging puzzle. *J. Mol. Genet. Med.* **09**, (2015).
 118. Wilmott, P., Lisowski, L., Alexander, I. E. & Logan, G. J. A user's guide to the inverted terminal repeats of adeno-associated virus. *Hum. Gene Ther. Methods* **30**, 206–213 (2019).
 119. Maurer, A. C. & Weitzman, M. D. Adeno-associated virus genome interactions important for vector production and transduction. *Hum. Gene Ther.* **31**, 499–511 (2020).
 120. Laugel, M., Lecomte, E., Ayuso, E., Adjali, O., Mével, M. & Penaud-Budloo, M. in *Recent Adv. Canine Med.* (ed. Fonseca-Alves, D. C. E.) (IntechOpen, 2022).

121. Spear, I. S., Fife, K. H., Hauswirth, W. W., Jones, C. J. & Berns, K. I. Evidence for two nucleotide sequence orientations within the terminal repetition of adeno-associated virus DNA. *J. Virol.* **24**, 627–34 (1977).
122. Lusby, E., Fife, K. H. & Berns, K. I. Nucleotide sequence of the inverted terminal repetition in adeno-associated virus DNA. *J. Virol.* **34**, 402–9 (1980).
123. *Regulation (EC) No 1394/2007 of the European Parliament and of the council of 13 November 2007 on advanced therapy medicinal products and amending Directive 2001/83/EC and Regulation (EC) No 726/2004.* L 324/121 (2007).
124. Gottlieb, S. Director of the Center for Biologics Evaluation and Research on new policies to advance development of safe and effective cell and gene therapies. *U.S. Food Drug Adm. Cent. Biol. Eval. Res.* (2019).
125. Eisenman, D. The United States' regulatory environment is evolving to accommodate a coming boom in gene therapy research. *Appl. Biosaf.* **24**, 147–152 (2019).
126. Rittié, L., Athanasopoulos, T., Calero-Garcia, M., Davies, M. L., Dow, D. J., Howe, S. J., Morrison, A., Ricciardelli, I., Saudemont, A., Jespers, L. & Clay, T. M. The landscape of early clinical gene therapies outside Oncology. *Mol. Ther.* **27**, 1706–1717 (2019).
127. Kuzmin, D. A., Shutova, M. V, Johnston, N. R., Smith, O. P., Fedorin, V. V, Kukushkin, Y. S. & van der Loo, J. C. M. The clinical landscape for AAV gene therapies. *Nat. Rev. Drug Discov.* **20**, 173–175 (2021).
128. Dobrowsky, T., Gianni, D., Pieracci, J. & Suh, J. AAV manufacturing for clinical use: Insights on current challenges from the upstream process perspective. *Curr. Opin. Biomed. Eng.* **20**, 100353 (2021).
129. Iglesias-Lopez, C., Agustí, A., Obach, M. & Vallano, A. Regulatory framework for advanced therapy medicinal Products and rAAV DNA in Europe and United States. *Front. Pharmacol.* **10**, at (2019)
130. Burdett, T. & Nuseibeh, S. Changing trends in the development of AAV-based gene therapies: a meta-analysis of past and present therapies. *Gene Ther.* 1–13 (2022).
131. Bulcha, J. T., Wang, Y., Ma, H., Tai, P. W. L. & Gao, G. Viral vector platforms within the

- gene therapy landscape. *Signal Transduct. Target. Ther.* **6**, 53 (2021).
132. Lee, C. S., Bishop, E. S., Zhang, R., Yu, X., Farina, E. M., Yan, S., Zhao, C., Zeng, Z., Shu, Y., Wu, X., Lei, J., Li, Y., Zhang, W., Yang, C., Wu, K., Wu, Y., Ho, S., Athiviraham, A., Lee, M. J., Wolf, J. M., Reid, R. R. & He, T.-C. Adenovirus-mediated gene delivery: Potential applications for gene and cell-based therapies in the new era of personalized medicine. *Genes Dis.* **4**, 43–63 (2017).
 133. Keeler, A., Elmallah, M. & Flotte, T. Gene therapy 2017: Progress and future directions. *Clin Transl Sci* **10**, 242–248 (2017).
 134. Ylä-Herttuala, S. Endgame: Glybera finally recommended for approval as the first gene therapy drug in the European Union. *Mol. Ther.* **20**, 1831–1832 (2012).
 135. Miraldi Utz, V., Coussa, R. G., Antaki, F. & Traboulsi, E. I. Gene therapy for RPE65-related retinal disease. *Ophthalmic Genet.* **39**, 671–677 (2018).
 136. Hoy, S. M. Onasemnogene abeparvovec: first global approval. *Drugs* **79**, 1255–1262 (2019).
 137. Larkin, H. First FDA-approved gene therapy for hemophilia. *J. Am. Med. Assoc.* **329**, 14 (2023).
 138. Ozelo, M. C., Mahlangu, J., Pasi, K. J., Giermasz, A., Leavitt, A. D., Laffan, M., Symington, E., Quon, D. V., Wang, J.-D., Peerlinck, K., Pipe, S. W., Madan, B., Key, N. S., Pierce, G. F., O'Mahony, B., Kaczmarek, R., Henshaw, J., Lawal, A., Jayaram, K., Huang, M., Yang, X., Wong, W. Y. & Kim, B. Valoctocogene roxaparvovec gene therapy for Hemophilia A. *N. Engl. J. Med.* **386**, 1013–1025 (2022).
 139. Keam, S. J. Eladocogene exuparvovec: first approval. *Drugs* **82**, 1427–1432 (2022).
 140. Senior, M. After Glybera's withdrawal, what's next for gene therapy? *Nat. Biotechnol.* **35**, 491–492 (2017).
 141. Darrow, J. J. Luxturna: FDA documents reveal the value of a costly gene therapy. *Drug Discov. Today* **24**, 949–954 (2019).
 142. Jayasundera, K., Abuzaitoun, R., Lacy, G., Abalem, M., Saltzman, G., Ciulla, T. & Johnson, M. Challenges of cost-effectiveness analyses of novel therapeutics for inherited

- retinal diseases. *Am. J. Ophthalmol.* **235**, 90–97 (2022).
143. Sabatini, M. T., Xia, T. & Chalmers, M. Pricing and market access challenges in the era of one-time administration cell and gene therapies. *Pharmaceut. Med.* **36**, 265–274 (2022).
 144. Shukla, V., Seoane-Vazquez, E., Fawaz, S., Brown, L. & Rodriguez-Monguio, R. The landscape of cellular and gene therapy products: authorization, discontinuations, and cost. *Hum. Gene Ther. Clin. Dev.* **30**, 102–113 (2019).
 145. Meunier, A., Longworth, L., Kowal, S., Ramagopalan, S., Love-Koh, J. & Griffin, S. Distributional cost-effectiveness analysis of health technologies: data requirements and challenges. *Value Heal.* **26**, 60–63 (2023).
 146. Lloyd-Williams, H. & Hughes, D. A. A systematic review of economic evaluations of advanced therapy medicinal products. *Br. J. Clin. Pharmacol.* **87**, 2428–2443 (2021).
 147. Escandell, J. M., Pais, D. A. M., Carvalho, S. B., Vincent, K., Gomes-Alves, P. & Alves, P. M. Leveraging rAAV bioprocess understanding and next generation bioanalytics development. *Curr. Opin. Biotechnol.* **74**, 271–277 (2022).
 148. Yang, O., Tao, Y., Qadan, M. & Ierapetritou, M. Process design and comparison for batch and continuous manufacturing of recombinant adeno-associated virus. *J. Pharm. Innov.* (2022).
 149. Rumachik, N. G., Malaker, S. A., Poweleit, N., Maynard, L. H., Adams, C. M., Leib, R. D., Cirolia, G., Thomas, D., Stamnes, S., Holt, K., Sinn, P., May, A. P. & Paulk, N. K. Methods matter: Standard production platforms for recombinant AAV produce chemically and functionally distinct vectors. *Mol. Ther. - Methods Clin. Dev.* **18**, 98–118 (2020).
 150. McGillicuddy, N., Floris, P., Albrecht, S. & Bones, J. Examining the sources of variability in cell culture media used for biopharmaceutical production. *Biotechnol. Lett.* **40**, 5–21 (2018).
 151. Whitford, W. G. in *Process Control. Intensification, Digit. Contin. Biomanufacturing* 93–136 (2022).
 152. Masri, F., Cheeseman, E. & Ansorge, S. Viral vector manufacturing: how to address

- current and future demands? *Cell Gene Ther. Insights* **5**, 949–970 (2019).
153. Cotmore, S. F. & Tattersall, P. Parvoviruses: small does not mean simple. *Annu. Rev. Virol.* **1**, 517–537 (2014).
 154. Cotmore, S. F., Agbandje-McKenna, M., Chiorini, J. A., Mukha, D. V., Pintel, D. J., Qiu, J., Soderlund-Venermo, M., Tattersall, P., Tijssen, P., Gatherer, D. & Davison, A. J. The family Parvoviridae. *Arch. Virol.* **159**, 1239–1247 (2014).
 155. Cotmore, S. F. & Tattersall, P. Parvovirus diversity and DNA damage responses. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).
 156. Berns, K. I. & Linden, R. M. The cryptic life style of adeno-associated virus. *BioEssays* **17**, 237–245 (1995).
 157. Weitzman, M. D. & Linden, R. M. in *Methods Mol. Biol.* **807**, 1–23 (2012).
 158. Geoffroy, M.-C. & Salvetti, A. Helper functions required for wild type and recombinant adeno-associated virus growth. *Curr. Gene Ther.* **5**, 265–271 (2005).
 159. Thomson, B. J., Weindler, F. W., Gray, D., Schwaab, V. & Heilbronn, R. Human herpesvirus 6 (HHV-6) is a helper virus for adeno-associated virus type 2 (AAV-2) and the AAV-2 rep gene homologue in HHV-6 can mediate AAV-2 DNA replication and regulate gene expression. *Virology* **204**, 304–311 (1994).
 160. Moore, A. R., Dong, B., Chen, L. & Xiao, W. Vaccinia virus as a subhelper for AAV replication and packaging. *Mol. Ther. - Methods Clin. Dev.* **2**, 15044 (2015).
 161. Ogston, P., Raj, K. & Beard, P. Productive replication of adeno-associated virus can occur in human papillomavirus type 16 (HPV-16) episome-containing keratinocytes and is augmented by the HPV-16 E2 protein. *J. Virol.* **74**, 3494 LP – 3504 (2000).
 162. Yalklnoglu, A. Ö., Heilbronn, R., Bürkle, A., Schlehofer, J. R. & Hausen, H. Zur. Dna amplification of adeno-associated virus as a response to cellular genotoxic stress. *Cancer Res.* **48**, 3123–3129 (1988).
 163. Berns, K. I. Parvovirus replication. *Microbiol. Rev.* **54**, 316–329 at <https://doi.org/10.1128/mmbr.54.3.316-329.1990> (1990)

164. Xiao, X., Xiao, W., Li, J. & Samulski, R. J. A novel 165-base-pair terminal repeat sequence is the sole cis requirement for the adeno-associated virus life cycle. *J. Virol.* **71**, 941–948 (1997).
165. Yakobson, B., Hrynko, T. A., Peak, M. J. & Winocour, E. Replication of adeno-associated virus in cells irradiated with UV light at 254 nm. *J. Virol.* **63**, 1023–30 (1989).
166. Yakobson, B., Koch, T. & Winocour, E. Replication of adeno-associated virus in synchronized cells without the addition of a helper virus. *J. Virol.* **61**, 972–81 (1987).
167. Linden, R. M., Ward, P., Giraud, C., Winocour, E. & Berns, K. I. Site-specific integration by adeno-associated virus. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 11288–11294 (1996).
168. Kotin, R. M., Siniscalco, M., Samulski, R. J., Zhu, X., Hunter, L., Laughlin, C. A., McLaughlin, S., Muzyczka, N., Rocchi, M. & Berns, K. I. Site-specific integration by adeno-associated virus. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2211–2215 (1990).
169. Schnepf, B. C., Chulay, J. D., Ye, G. J., Flotte, T. R., Trapnell, B. C. & Johnson, P. R. Recombinant adeno-associated virus vector genomes take the form of long-lived, transcriptionally competent episomes in human muscle. *Hum. Gene Ther.* **27**, 32–42 (2016).
170. Schnepf, B. C., Clark, K. R., Klemanski, D. L., Pacak, C. A., Johnson, P. R. & Flotte, T. R. Genetic fate of recombinant adeno-associated virus vector genomes in muscle. *J. Virol.* **77**, 3495–504 (2003).
171. Penaud-Budloo, M., Le Guiner, C., Nowrouzi, A., Toromanoff, A., Chérel, Y., Chenuaud, P., Schmidt, M., von Kalle, C., Rolling, F., Moullier, P. & Snyder, R. O. Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. *J. Virol.* **82**, 7875–85 (2008).
172. Duan, D., Sharma, P., Yang, J., Yue, Y., Dudus, L., Zhang, Y., Fisher, K. J. & Engelhardt, J. F. Circular intermediates of recombinant adeno-associated virus have defined structural characteristics responsible for long-term episomal persistence in muscle tissue. *J. Virol.* **72**, 8568–77 (1998).
173. Hernandez, Y. J., Wang, J., Kearns, W. G., Loiler, S., Poirier, A. & Flotte, T. R. Latent adeno-associated virus infection elicits humoral but not cell-mediated immune

- responses in a nonhuman primate model. *J. Virol.* **73**, 8549–8558 (1999).
174. Schnepf, B. C., Jensen, R. L., Chen, C.-L., Johnson, P. R. & Clark, K. R. Characterization of adeno-associated virus genomes isolated from human tissues. *J. Virol.* **79**, 14793–14803 (2005).
 175. La Bella, T., Imbeaud, S., Peneau, C., Mami, I., Datta, S., Bayard, Q., Caruso, S., Hirsch, T. Z., Calderaro, J., Morcrette, G., Guettier, C., Paradis, V., Amaddeo, G., Laurent, A., Possenti, L., Chiche, L., Bioulac-Sage, P., Blanc, J. F., Letouze, E., Nault, J. C. & Zucman-Rossi, J. Adeno-associated virus in the liver: Natural history and consequences in tumour development. *Gut* **69**, 737–747 (2020).
 176. Marcus-Sekura, C. J. & Carter, B. J. Chromatin-like structure of adeno-associated virus DNA in infected cells. *J. Virol.* **48**, 79–87 (1983).
 177. Brown, M. & Weber, J. Adenoassociated virus has a unique chromatin structure. *Can. J. Biochem.* **60**, 1001–1005 (1982).
 178. Okada, T., Uchibori, R., Iwata-Okada, M., Takahashi, M., Nomoto, T., Nonaka-Sarukawa, M., Ito, T., Liu, Y., Mizukami, H., Kume, A., Kobayashi, E. & Ozawa, K. A histone deacetylase inhibitor enhances recombinant adeno-associated virus-mediated gene expression in tumor cells. *Mol. Ther.* **13**, 738–746 (2006).
 179. Smith-Moore, S., Neil, S. J. D., Fraefel, C., Michael Linden, R., Bollen, M., Rowe, H. M. & Henckaerts, E. Adeno-associated virus Rep proteins antagonize phosphatase PP1 to counteract KAP1 repression of the latent viral genome. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3529–E3538 (2018).
 180. Samulski, R. J., Srivastava, A., Berns, K. I. & Muzyczka, N. Rescue of adeno-associated virus from recombinant plasmids: gene correction within the terminal repeats of AAV. *Cell* **33**, 135–143 (1983).
 181. Ward, P., Urcelay, E., Kotin, R., Safer, B. & Berns, K. I. Adeno-associated virus DNA replication in vitro: activation by a maltose binding protein/Rep 68 fusion protein. *J. Virol.* **68**, 6029–6037 (1994).
 182. Ward, P., Clément, N. & Linden, R. M. cis effects in adeno-associated virus type 2 replication. *J. Virol.* **81**, 9976–89 (2007).

183. Nony, P., Tessier, J., Chadeuf, G., Ward, P., Giraud, A., Dugast, M., Linden, R. M., Moullier, P. & Salvetti, A. Novel cis-acting replication element in the adeno-associated virus type 2 genome is involved in amplification of integrated rep-cap sequences. *J. Virol.* **75**, 9991–9994 (2001).
184. Pereira, D. J., McCarty, D. M. & Muzyczka, N. The adeno-associated virus (AAV) Rep protein acts as both a repressor and an activator to regulate AAV transcription during a productive infection. *J. Virol.* **71**, 1079–88 (1997).
185. Hüser, D., Khalid, D., Lutter, T., Hammer, E.-M., Weger, S., Heßler, M., Kalus, U., Tauchmann, Y., Hensel-Wiegel, K., Lassner, D. & Heilbronn, R. High prevalence of infectious adeno-associated virus (AAV) in human peripheral blood mononuclear cells indicative of T lymphocytes as sites of AAV persistence. *J. Virol.* **91**, e02137-16 (2017).
186. Kyöstiö, S. R., Wonderling, R. S. & Owens, R. A. Negative regulation of the adeno-associated virus (AAV) P5 promoter involves both the P5 rep binding site and the consensus ATP-binding motif of the AAV Rep68 protein. *J. Virol.* **69**, 6787–96 (1995).
187. Jayandharan, G. R., Zhong, L., Li, B., Kachniarz, B. & Srivastava, A. Strategies for improving the transduction efficiency of single-stranded adeno-associated virus vectors in vitro and in vivo. *Gene Ther.* **15**, 1287–1293 (2008).
188. Zhao, W., Wu, J., Zhong, L. & Srivastava, A. Adeno-associated virus 2-mediated gene transfer: Role of a cellular serine/threonine protein phosphatase in augmenting transduction efficiency. *Gene Ther.* **14**, 545–550 (2007).
189. Lentz, T. B. & Samulski, R. J. Insight into the mechanism of inhibition of adeno-associated virus by the Mre11/Rad50/Nbs1 complex. *J. Virol.* **89**, 181–194 (2015).
190. Choi, V. W., McCarty, D. M. & Samulski, R. J. Host cell DNA repair pathways in adeno-associated viral genome processing. *J. Virol.* **80**, 10346–56 (2006).
191. Cervelli, T., Palacios, J. A., Zentilin, L., Mano, M., Schwartz, R. A., Weitzman, M. D. & Giacca, M. Processing of recombinant AAV genomes occurs in specific nuclear structures that overlap with foci of DNA-damage-response proteins. *J. Cell Sci.* **121**, 349–57 (2008).
192. Song, S., Lu, Y., Choi, Y. K., Han, Y., Tang, Q., Zhao, G., Berns, K. I. & Flotte, T. R. DNA-

- dependent PK inhibits adeno-associated virus DNA integration. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2112–2116 (2004).
193. Song, S., Laipis, P. J., Berns, K. I. & Flotte, T. R. Effect of DNA-dependent protein kinase on the molecular fate of the rAAV2 genome in skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4084–4088 (2001).
 194. Schwartz, R. A., Carson, C. T., Schuberth, C. & Weitzman, M. D. Adeno-associated virus replication induces a DNA damage response coordinated by DNA-dependent protein kinase. *J. Virol.* **83**, 6269–6278 (2009).
 195. Inagaki, K., Ma, C., Storm, T. A., Kay, M. A. & Nakai, H. The role of DNA-PKcs and artemis in opening viral DNA hairpin termini in various tissues in mice. *J. Virol.* **81**, 11304–11321 (2007).
 196. Choi, Y. K., Nash, K., Byrne, B. J., Muzyczka, N. & Song, S. The effect of DNA-dependent protein kinase on adeno-associated virus replication. *PLoS One* **5**, (2010).
 197. Nash, K., Chen, W. & Muzyczka, N. Complete in vitro reconstitution of adeno-associated virus DNA replication requires the minichromosome maintenance complex proteins. *J. Virol.* **82**, 1458–64 (2008).
 198. Millet, R., Jolinon, N., Nguyen, X.-N., Berger, G., Cimorelli, A., Greco, A., Bertrand, P., Odenthal, M., Büning, H. & Salvetti, A. Impact of the MRN complex on adeno-associated virus integration and replication during coinfection with herpes simplex virus 1. *J. Virol.* **89**, 6824–6834 (2015).
 199. Ryan, J. H., Zolotukhin, S. & Muzyczka, N. Sequence requirements for binding of Rep68 to the adeno-associated virus terminal repeats. *J. Virol.* **70**, 1542–53 (1996).
 200. Schmidt, M., Katano, H., Bossis, I. & Chiorini, J. A. Cloning and Characterization of a Bovine Adeno-Associated Virus. *J. Virol.* **78**, 6509–6516 (2004).
 201. Tellez, J., Van Vliet, K., Tseng, Y.-S., Finn, J. D., Tschernia, N., Almeida-Porada, G., Arruda, V. R., Agbandje-McKenna, M. & Porada, C. D. Characterization of naturally-occurring humoral immunity to AAV in sheep. *PLoS One* **8**, e75142 (2013).
 202. Péntzes, J. J., Pham, H. T., Benkö, M. & Tijssen, P. Novel parvoviruses in reptiles and

- genome sequence of a lizard parvovirus shed light on Dependoparvovirus genus evolution. *J. Gen. Virol.* **96**, 2569–2579 (2015).
203. Farkas, S. L., Zádori, Z., Benko, M., Essbauer, S., Harrach, B. & Tijssen, P. A parvovirus isolated from royal python (*Phyton regius*) is a member of the genus Dependovirus. *J. Gen. Virol.* **85**, 555–561 (2004).
 204. Grimm, D., Pandey, K., Nakai, H., Storm, T. A. & Kay, M. A. Liver transduction with recombinant adeno-associated virus is primarily restricted by capsid serotype not vector genotype. *J. Virol.* **80**, 426–439 (2006).
 205. Yoon-Robarts, M., Blouin, A. G., Bleker, S., Kleinschmidt, J. A., Aggarwal, A. K., Escalante, C. R. & Linden, R. M. Residues within the B' Motif Are Critical for DNA Binding by the Superfamily 3 Helicase Rep40 of Adeno-associated Virus Type 2. *J. Biol. Chem.* **279**, 50472–50481 (2004).
 206. Zhong, L., Zhou, X., Li, Y., Qing, K., Xiao, X., Samulski, R. J. & Srivastava, A. Single-polarity recombinant adeno-associated virus 2 vector-mediated transgene expression In vitro and In vivo: Mechanism of transduction. *Mol. Ther.* **16**, 290–295 (2008).
 207. Chen, Z. Y., Yant, S. R., He, C. Y., Meuse, L., Shen, S. & Kay, M. A. Linear DNAs concatemerize in vivo and result in sustained transgene expression in mouse liver. *Mol. Ther.* **3**, 403–410 (2001).
 208. Salganik, M., Aydemir, F., Nam, H.-J., McKenna, R., Agbandje-McKenna, M. & Muzyczka, N. Adeno-associated virus capsid proteins may play a role in transcription and second-strand synthesis of recombinant genomes. *J. Virol.* **88**, 1071–1079 (2014).
 209. Thomas, C. E., Storm, T. A., Huang, Z. & Kay, M. A. Rapid uncoating of vector genomes is the key to efficient liver transduction with pseudotyped adeno-associated virus vectors. *J. Virol.* **78**, 3110–3122 (2004).
 210. Wang, J., Xie, J., Lu, H., Chen, L., Hauck, B., Samulski, R. J. & Xiao, W. Existence of transient functional double-stranded DNA intermediates during recombinant AAV transduction. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13104–13109 (2007).
 211. Hauck, B., Zhao, W., High, K. & Xiao, W. Intracellular viral processing, not single-stranded DNA accumulation, is crucial for recombinant adeno-associated virus

- transduction. *J. Virol.* **78**, 13678–13686 (2004).
212. Miao, C. H., Nakai, H., Thompson, A. R., Storm, T. A., Chiu, W., Snyder, R. O. & Kay, M. A. Nonrandom transduction of recombinant adeno-associated virus vectors in mouse hepatocytes in vivo: cell cycling does not influence hepatocyte transduction. *J. Virol.* **74**, 3793–3803 (2000).
 213. Yang, J., Zhou, W., Zhang, Y., Zidon, T., Ritchie, T. & Engelhardt, J. F. Concatamerization of adeno-associated virus circular genomes occurs through intermolecular recombination. *J. Virol.* **73**, 9468–77 (1999).
 214. Choi, V. W., Samulski, R. J. & McCarty, D. M. Effects of adeno-associated virus DNA hairpin structure on recombination. *J. Virol.* **79**, 6801–6807 (2005).
 215. Fisher, K. J., Gao, G. P., Weitzman, M. D., DeMatteo, R., Burda, J. F. & Wilson, J. M. Transduction with recombinant adeno-associated virus for gene therapy is limited by leading-strand synthesis. *J. Virol.* **70**, 520–532 (1996).
 216. Nakai, H., Storm, T. A. & Kay, M. A. Recruitment of single-stranded recombinant adeno-associated virus vector genomes and intermolecular recombination are responsible for stable transduction of liver in vivo. *J. Virol.* **74**, 9451–9463 (2000).
 217. Domenger, C. & Grimm, D. Next-generation AAV vectors-do not judge a virus (only) by its cover. *Hum. Mol. Genet.* **28**, R3–R14 at <https://doi.org/10.1093/hmg/ddz148> (2019)
 218. Cervelli, T., Palacios, J. A., Zentilin, L., Mano, M., Schwartz, R. A., Weitzman, M. D. & Giacca, M. Processing of recombinant AAV genomes occurs in specific nuclear structures that overlap with foci of DNA-damage-response proteins. *J. Cell Sci.* **121**, 349–357 (2008).
 219. Cataldi, M. P. & McCarty, D. M. Hairpin-end conformation of adeno-associated virus genome determines interactions with DNA-repair pathways. *Gene Ther.* **20**, 686–93 (2013).
 220. Julien, L., Chassagne, J., Peccate, C., Lorain, S., Piétri-Rouxel, F., Danos, O. & Benkhelifa-Ziyyat, S. RFX1 and RFX3 transcription factors interact with the D sequence of adeno-associated virus inverted terminal repeat and regulate AAV transduction. *Sci. Rep.* **8**, 210 (2018).

221. Ling, C., Wang, Y., Lu, Y., Wang, L., Jayandharan, G. R., Aslanidi, G. V, Li, B., Cheng, B., Ma, W., Lentz, T., Ling, C., Xiao, X., Samulski, R. J., Muzyczka, N. & Srivastava, A. Enhanced transgene expression from recombinant single-stranded D-sequence-substituted adeno-associated virus vectors in human cell lines in vitro and in murine hepatocytes in vivo. *J. Virol.* **89**, 952–961 (2015).
222. Qing, K., Hansen, J., Weigel-Kelley, K. A., Tan, M., Zhou, S. & Srivastava, A. Adeno-associated virus type 2-mediated gene transfer: role of cellular FKBP52 protein in transgene expression. *J. Virol.* **75**, 8968–8976 (2001).
223. Qing, K., Li, W., Zhong, L., Tan, M., Hansen, J., Weigel-Kelley, K. A., Chen, L., Yoder, M. C. & Srivastava, A. Adeno-associated virus type 2-mediated gene transfer: role of cellular T-cell protein tyrosine phosphatase in transgene expression in established cell lines in vitro and transgenic mice in vivo. *J. Virol.* **77**, 2741–2746 (2003).
224. Ling, C., Yin, Z., Li, J., Zhang, D., Aslanidi, G. & Srivastava, A. Strategies to generate high-titer, high-potency recombinant AAV3 serotype vectors. *Mol. Ther. - Methods Clin. Dev.* **3**, 16029 (2016).
225. Earley, L. F., Conatser, L. M., Lue, V. M., Dobbins, A. L., Li, C., Hirsch, M. L. & Samulski, R. J. Adeno-associated virus serotype-specific inverted terminal repeat sequence role in vector transgene expression. *Hum. Gene Ther.* **31**, 151–162 (2020).
226. Logan, G. J., Dane, A. P., Hallwirth, C. V, Smyth, C. M., Wilkie, E. E., Amaya, A. K., Zhu, E., Khandekar, N., Ginn, S. L., Liao, S. H. Y., Cunningham, S. C., Sasaki, N., Cabanes-Creus, M., Tam, P. P. L., Russell, D. W., Lisowski, L. & Alexander, I. E. Identification of liver-specific enhancer–promoter activity in the 3' untranslated region of the wild-type AAV2 genome. *Nat. Genet.* **49**, 1267–1273 (2017).
227. Yoon-Robarts, M., Blouin, A. G., Bleker, S., Kleinschmidt, J. A., Aggarwal, A. K., Escalante, C. R. & Linden, R. M. Residues within the B' motif are critical for DNA binding by the superfamily 3 helicase Rep40 of adeno-associated virus type 2. *J. Biol. Chem.* **279**, 50472–50481 (2004).
228. Yan, Z., Zak, R., Zhang, Y. & Engelhardt, J. F. Inverted terminal repeat sequences are important for intermolecular recombination and circularization of adeno-associated virus genomes. *J. Virol.* **79**, 364–379 (2005).

229. Zhou, Q., Tian, W., Liu, C., Lian, Z., Dong, X. & Wu, X. Deletion of the B-B' and C-C' regions of inverted terminal repeats reduces rAAV productivity but increases transgene expression. *Sci. Rep.* **7**, 1–13 (2017).
230. Savy, A., Dickx, Y., Nauwynck, L., Bonnin, D., Merten, O.-W. & Galibert, L. Impact of inverted terminal repeat integrity on rAAV8 production using the baculovirus/Sf9 cells system. *Hum. Gene Ther. Methods* **28**, 277–289 (2017).
231. Samulski, R. J., Chang, L. S. & Shenk, T. A recombinant plasmid from which an infectious adeno-associated virus genome can be excised in vitro and its use to study viral replication. *J. Virol.* **61**, 3096–101 (1987).
232. Saveliev, A., Liu, J., Li, M., Hirata, L., Latshaw, C., Zhang, J. & Wilson, J. M. Accurate and rapid sequence analysis of adeno-associated virus plasmids by Illumina next-generation sequencing. *Hum. Gene Ther. Methods* **29**, 201–211 (2018).
233. Leach, D. R. F. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays* **16**, 893–900 (1994).
234. Cromie, G. A., Millar, C. B., Schmidt, K. H. & Leach, D. R. F. Palindromes as substrates for multiple pathways of recombination in *Escherichia coli*. *Genetics* **154**, 513–522 (2000).
235. Sinden, R. R., Zheng, G. X., Brankamp, R. G. & Allen, K. N. On the deletion of inverted repeated DNA in *Escherichia coli*: effects of length, thermal stability, and cruciform formation in vivo. *Genetics* **129**, 991–1005 (1991).
236. Bzymek, M. & Lovett, S. T. Evidence for two mechanisms of palindrome-stimulated deletion in *Escherichia coli*: single-strand annealing and replication slipped mispairing. *Plasmid* **158**, 527–40 (2001).
237. Bowater, R. P., Bohálová, N. & Brázda, V. Interaction of proteins with inverted repeats and cruciform structures in nucleic acids. *Int. J. Mol. Sci.* **23**, (2022).
238. West, S. C. & Körner, A. Cleavage of cruciform DNA structures by an activity from *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* **82**, 6445–6449 (1985).
239. Tomoiaga, D., Bubnell, J., Herndon, L. & Feinstein, P. High rates of plasmid

- cotransformation in *E. coli* overturn the clonality myth and reveal colony development. *Sci. Rep.* **12**, 11515 (2022).
240. Tai, P. W. L., Xie, J., Fong, K., Seetin, M., Heiner, C., Su, Q., Weiand, M., Wilmot, D., Zapp, M. L. & Gao, G. Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras. *Mol. Ther. Methods Clin. Dev.* **9**, 130–141 (2018).
 241. Tran, N. T., Heiner, C., Weber, K., Weiand, M., Wilmot, D., Xie, J., Wang, D., Brown, A., Manokaran, S., Su, Q., Zapp, M. L., Gao, G. & Tai, P. W. L. AAV-genome population sequencing of vectors packaging CRISPR components reveals design-influenced heterogeneity. *Mol. Ther. - Methods Clin. Dev.* (2020).
 242. Namkung, S., Tran, N. T., Manokaran, S., He, R., Su, Q., Xie, J., Gao, G. & Tai, P. W. L. Direct ITR-to-ITR nanopore sequencing of AAV vector genomes. *Hum. Gene Ther.* **33**, 1187–1196 (2022).
 243. Radukic, M. T., Brandt, D., Haak, M., Müller, K. M. & Kalinowski, J. Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. *NAR Genomics Bioinforma.* **2**, lqaa074 (2020).
 244. Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V, Wang, C. & Carter, A. B. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).
 245. Petri, K., Fronza, R., Gabriel, R., Käppel, C., Nowrouzi, A., Linden, R. M., Henckaerts, E. & Schmidt, M. Comparative next-generation sequencing of adeno-associated virus inverted terminal repeats. *Biotechniques* **56**, 269–273 (2014).
 246. Lecomte, E., Leger, A., Penaud-Budloo, M. & Ayuso, E. in *Adeno-Associated Virus Vectors. Methods Mol. Biol.* (ed. Castle, M. J.) **1950**, 93–114 (Humana Press, 2019).
 247. Penaud-Budloo, M., Lecomte, E., Guy-Duché, A., Saleun, S., Roulet, A., Lopez-Roques, C., Tournaire, B., Cogné, B., Léger, A., Blouin, V., Lindenbaum, P., Moullier, P. & Ayuso, E. Accurate identification and quantification of DNA species by next-generation

- sequencing in adeno-associated viral vectors produced in insect cells. *Hum. Gene Ther. Methods* **28**, 148–162 (2017).
248. Tran, N. T., Lecomte, E., Saleun, S., Namkung, S., Robin, C., Weber, K., Devine, E., Blouin, V., Adjali, O., Ayuso, E., Gao, G., Penaud-Budloo, M. & Tai, P. W. L. Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity. *Hum. Gene Ther.* **33**, 371–388 (2022).
 249. Shmidt, A. A. & Egorova, T. V. PCR-based analytical methods for quantification and quality control of recombinant adeno-associated viral vector preparations. *Pharmaceuticals* **15**, 23 (2021).
 250. Suoranta, T., Laham-Karam, N. & Ylä-Herttuala, S. Optimized protocol for accurate titration of adeno-associated virus vectors. *Hum. Gene Ther.* (2021).
 251. Aurnhammer, C., Haase, M., Muether, N., Hausl, M., Rauschhuber, C., Huber, I., Nitschko, H., Busch, U., Sing, A., Ehrhardt, A. & Baiker, A. Universal real-time PCR for the detection and quantification of adeno-associated virus serotype 2-derived inverted terminal repeat sequences. *Hum. Gene Ther. Methods* **23**, 18–28 (2012).
 252. Lock, M., Alvira, M. R., Chen, S. J. & Wilson, J. M. Absolute determination of single-stranded and self-complementary adeno-associated viral vector genome titers by droplet digital PCR. *Hum. Gene Ther. Methods* **25**, 115–125 (2014).
 253. Wang, Y., Menon, N., Shen, S., Feschenko, M. & Bergelson, S. A qPCR method for AAV genome titer with ddPCR-level of accuracy and precision. *Mol. Ther. - Methods Clin. Dev.* **19**, 341–346 (2020).
 254. D’Costa, S., Blouin, V., Broucque, F., Penaud-Budloo, M., Fçranois, A., Perez, I. C., Le Bec, C., Moullier, P., Snyder, R. O. & Ayuso, E. Practical utilization of recombinant AAV vector reference standards: focus on vector genomes titration by free ITR qPCR. *Mol. Ther. - Methods Clin. Dev.* **3**, 16019 (2016).
 255. Gimpel, A. L., Katsikis, G., Sha, S., Maloney, A. J., Hong, M. S., Nguyen, T. N. T., Wolfrum, J., Springs, S. L., Sinskey, A. J., Manalis, S. R., Barone, P. W. & Braatz, R. D. Analytical methods for process and product characterization of recombinant adeno-associated virus-based gene therapies. *Mol. Ther. - Methods Clin. Dev.* **20**, 740–754

- (2021).
256. Fu, X., Chen, W.-C., Argento, C., Clarner, P., Bhatt, V., Dickerson, R., Bou-Assaf, G., Bakhshayeshi, M., Lu, X., Bergelson, S. & Pieracci, J. Analytical strategies for quantification of adeno-associated virus empty capsids to support process development. *Hum. Gene Ther. Methods* **30**, 144–152 (2019).
 257. Wright, J. *Product-related impurities in clinical-grade recombinant AAV vectors: Characterization and risk assessment. Biomedicines* **2**, (2014).
 258. van der Loo, J. C. M. & Wright, J. F. Progress and challenges in viral vector manufacturing. *Hum. Mol. Genet.* **25**, R42–R52 (2016).
 259. Senapathy, P., Tratschin, J. D. & Carter, B. J. Replication of adeno-associated virus DNA. Complementation of naturally occurring rep- mutants by a wild-type genome or an ori-mutant and correction of terminal palindrome deletions. *J. Mol. Biol.* **179**, 1–20 (1984).
 260. Bohenzky, R. A., Lefebvre, R. B. & Berns, K. I. Sequence and symmetry requirements within the internal palindromic sequences of the adeno-associated virus terminal repeat. *Virology* **166**, 316–327 (1988).

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

Mathematics in this thesis: scope and accessibility

Work not included in this thesis

Image attributions

2.1 REAGENTS AND EQUIPMENT

_Toc143892462

_Toc143892463

_Toc143892464

2.2 STANDARD PROTOCOLS

2.2.1 DNA Ligation

2.2.2 Restriction digestion

2.2.3 Agarose gel electrophoresis

2.2.4 Establishing monoclonal plasmid preparations

2.2.5 Tissue DNA extraction

2.2.6 Tissue RNA extraction

2.2.7 Construction of cDNA template from RNA samples

2.2.8 Tissue culture

2.2.9 Triple-transfection of HEK293 cells

2.2.10 Measurement of vector concentration using qPCR

2.2.11 Data analysis and visualisation

2.3 METHODS RELATING TO CHAPTER 4

2.3.1 Construction of plasmids containing conjugate reporter barcodes

2.3.2 qRBA measurement of reporter ratios in plasmid samples

2.4 METHODS RELATING TO CHAPTER 6.1 (RAAV PRODUCTION)

2.4.1 Subclonal isolation of 3'ITR plasmid variants

2.4.2 Preparation of variant mixes simulating plasmid ITR heterogeneity

2.4.3 Transfection of HEK293T cells in 6-well plates

2.4.4 Cell harvest and spiking

2.4.5 Extraction of target nucleic acid species from harvested cells (and a novel cellular lysis technique)

2.4.6 Pre-PCR treatments and qRBA measurement

2.5 METHODS RELATING TO CHAPTER 6.2 (RAAV GENE DELIVERY)

2.5.1 Production and crude extraction of rAAV9 vectors from 3'ITR plasmid variants

2.5.2 Animal handling

3.1 INTRODUCTION

3.1.1 The Illumina mechanism of (relative) molecular counting

3.2 WHAT ARE REPORTER BARCODES, AND WHAT IS QRBA?

3.3 THE SUBSTANTIAL ANALYTICAL BENEFITS OF QRBA MEASUREMENT

3.4 RELATIVE FOLD OUTPUT AND COMPLETE BINARISATION

3.5 A BINOMIAL DISTRIBUTION MODEL FOR QRBA COUNT ANALYSIS

3.5.1 Preface

3.5.2 Statement and basis of the model

3.5.2 A primer on the binomial distribution

3.5.3 Complete binarisation, use of the ratio, and the SD of the ratio

3.6 CONCLUSIONS

_Toc143892501

INTRODUCTION & OUTLINE

4.1 DESIGN AND CONSTRUCTION OF EXPERIMENTAL *CONJUGATE* REPORTER PLASMIDS

4.2 TESTING Z USING *IN SILICO* SIMULATED BERNOULLI PROCESSES

4.2.1 Results

4.2.2 Discussion of ζ and r_s as expressions of the binomial distribution

4.3 EMPIRICALLY TESTING THE BINOMIAL MODEL OF ILLUMINA-DERIVED QRBA MEASUREMENT

4.3.1 Breakdown of count data

4.3.2 Analysis of SDr_s and its relationship with ζ_s

4.3.3 Discussion of qRBA precision

4.4 SYSTEMATIC ERROR IN QRBA MEASUREMENT

4.4.1 Introduction

4.4.2 Result of conjugate qRBA measurements and *ad hoc* analyses

4.4.3 *Ad hoc* analyses involving composite functions

_Toc143892515

4.4.4 Discussion – sources of qRBA measurement error

4.4.5 Discussion – how should existing count data be interpreted?

4.5 CONCLUSIONS

5.1 INTRODUCTION

5.2 THE AAV VIRUS AND ITS VECTOR DERIVATIVE – A SNAPSHOT

5.3 THE BIOTECHNOLOGICAL CONTEXT OF PLASMID ITR INSTABILITY

5.4 A DEEPER DIVE INTO AAV GENOME BIOLOGY

5.5 THE ITR INSTABILITY PROBLEM

5.5.1 What is the ITR instability problem?

5.5.2 Dull, elusive, and non-fatal

5.6 CONCLUSIONS

INTRODUCTION

6.1 PLASMID ITR HETEROGENEITY AND RAAV PRODUCTION IN ADHERENT HEK293T CELLS

_Toc143892529

6.1.1 Results

6.1.1.1 Subclonal isolation of 3'ITR plasmid variants and preparation of rAAV plasmid mixtures

_Toc143892532

6.1.1.2 Productivity of plasmid 3'ITR variants co-transfected with pWT

6.1.1.3 Productivity of plasmid 3'ITR variants in individual transfections

6.1.1.4 α - and β -specific behaviours

6.1.1.5 Total vector yield from plasmids with variable 3'ITR integrity

_Toc143892537

6.1.1.6 Quantitative consistency of biological replicates

6.1.2 Discussion

6.1.2.1 Relevance of Chapter 4.4 to the interpretation of these results

6.1.2.2 Distinct behaviours of partnered 3'ITR plasmid variants

6.1.2.3 Quantitative consistency

6.1.2.4 The effects of plasmid deletions on rAAV vector production in adherent HEK293T producer cells

6.2 TRUNCATIONS IN PLASMID 3'ITRS AND GENE TRANSFER TO MURINE LIVER AND SKELETAL MUSCLE

6.2.2 Results

- 6.2.2.1 Production of an rAAV vector mix from four unique plasmid 3'ITR variants
- 6.2.2.2 Comparison of gene transfer, transgene stability, and transcription, and functional outcomes
- 6.2.2.3 vector genome transfer was inversely correlated to vector genome transcription

_Toc143892549

6.2.3 Discussion

6.3 CONCLUSIONS

7.1 DISCUSSION

7.2 CONCLUSIONS

7.3 REFERENCES

Appendix A. A synopsis of efforts to develop methods for quantifying intact ITRs

This doctorate, which commenced in 2019, was initially focused on efforts to measure the fraction of rAAV plasmid copies with wild-type ITR sequences. The ability to measure this variable was viewed as a prerequisite to studying its effects. This endeavour stood to benefit enormously from the involvement of career biometrologists. The plan was not only to perform the first validated quantification of intact plasmid ITRs, but to establish a collection of methods suited to a range of context-specific demands on stringency and routine implementability. These would include everyday spot checks as well as validated assays traceable to the International System of units (SI) appropriate for GMP (good manufacturing practice) environments.

One of the more significant challenges for this project was a technical prerequisite to the main research goal. Namely, preparing a sufficient quantity of DNA materials that were homogenous with respect to the ITR sequence, which would act as reference samples to test various methodologies. Three iterations of this plan gained momentum but ultimately none reached a point of success. All involved fractionation of a homogenous PCR amplicon or restriction product via ion exchange high-performance chromatography (IE-HPLC). The first iteration was a restriction product, which was cleaved at the closest available restriction sites to the ITR. In for a full-length ITR this was 154 base pairs. After obtaining homogenous fractions of the ITR, solutions of the DNA fragments would be reconstituted in gravimetrically determined quantities and analysed with orthogonal electrophoretic, chromatographic, and mass spectrometric techniques.

Fractionation of the 154bp fragment required two stages, one to separate it from the bulk of plasmid restriction products and another to successfully separate the wild-type ITR from mutant variants. The fractionation process was highly labour intensive. Multi-milligram preparative restriction digests were performed for a wild-type-derived AAV plasmid preparation and another from a $\Delta BB'$ ($\Delta AhdI$) variant. The 154bp ITR-containing fragment represented just 6% of the plasmid mass, making the loading capacity of the column for bulk restriction product separation a rate limiting factor. Efforts to complete these first rounds of

separation were quickly interrupted by the COVID-19 pandemic, and collaborations were delayed further still while the Bioanalysis team undertook urgent research in aid to public health efforts.

On return to this collaboration, I had developed two alterations to the strategy, which together represented the second iteration of the plan to generate ITR analytical materials. Having demonstrated success in PCR amplification of full-length ITR sequences, one modification was to adopt a preparative PCR strategy in place of a preparative restriction digest strategy. This had two distinct advantages. Firstly, the PCR product would constitute most if not all of the DNA mass loadable onto the chromatography column, making this stage of the process much more expedient. Secondly, restriction variants of the AAV2 ITR could be prepared without the need to subclonally isolate them, since the PCR template could instead first be treated with enzymes such as AhdI and XmaI (which target the BB' and CC' hairpin sequences respectively).

The other modification was to include an adjacent ~400bp to the ITR, which would contain elements utilisable in the context of assay development. It would contain (1) a ddPCR target region, allowing the ITR analytical material to be quantified in absolute terms, and (2) restriction sites that would three fragments (including the ITR fragment), where the ITR had the intermediate length, which would act as endogenous standards. Rather than attempting to find the ratio of the wild-type ITR sequence to the various mutants in a plasmid prep, it would be measured relative to the endogenous standards cleaved from the same parent molecule.

The obstruction to success here was failure of the target to amplify with practicable efficiency – a problem that was shown to be specific to the primer pairs utilised. After protracted failed attempts, a systematic troubleshooting process was carried out that proved further failure was ensured. Success seen with alternative primer sets and the very same plasmid template proved to be a useful positive control in this context, as was a mutant clone of the plasmid that lacked any secondary-structure-forming elements. I was able to 'prove' that the DNA polymerases (Taq DNA polymerase with betaine supplementation and Q5 DNA polymerase) were capable of replicating AhdI⁺XmaI⁺ ITR sequences from both loci of the plasmid, and that the unsuccessful primers were capable of amplifying the target in the absence of hairpin-forming elements.

This result led to the third iteration of potential ITR analytical materials – a decisive reversion back to the preparative restriction digest strategy, while retaining the concept of the

endogenous standard and ddPCR target. After subclonally isolating an additional 3'ITR mutants I prepared EcoRI+PstI restriction products collectively amounting to >11mg for IE-HPLC fractionation. Here difficulties arose again, this time in attempting to develop a chromatographic protocol with a predictable retention pattern, either for the first round of bulk separation or the second round of finer ITR-variant separation. Unrelated technical problems with the instruments available prevented access for an indeterminate period, which catalysed a discussion around whether to press ahead or cut our losses and choose a different research question altogether. Certainly, I have learned this is one of the least comfortable experiences in research – needing to compare a known switching cost with an unknown ‘staying’ cost. By this time, I had produced results from a pilot version of the qRBA study of rAAV production reported in Chapter 6.1, which had not demonstrated a clear effect on vector yield when producer cells were transfected with variable fractions of a wild-type-derived rAAV plasmid prep. Long read sequencing technologies were also gaining traction in the rAAV area and unlike when the project had commenced, had the potential to serve research into ITR integrity in the absence of SI-traceable assays. With these factors at play, it was decided to pursue the project described in the following appendix.

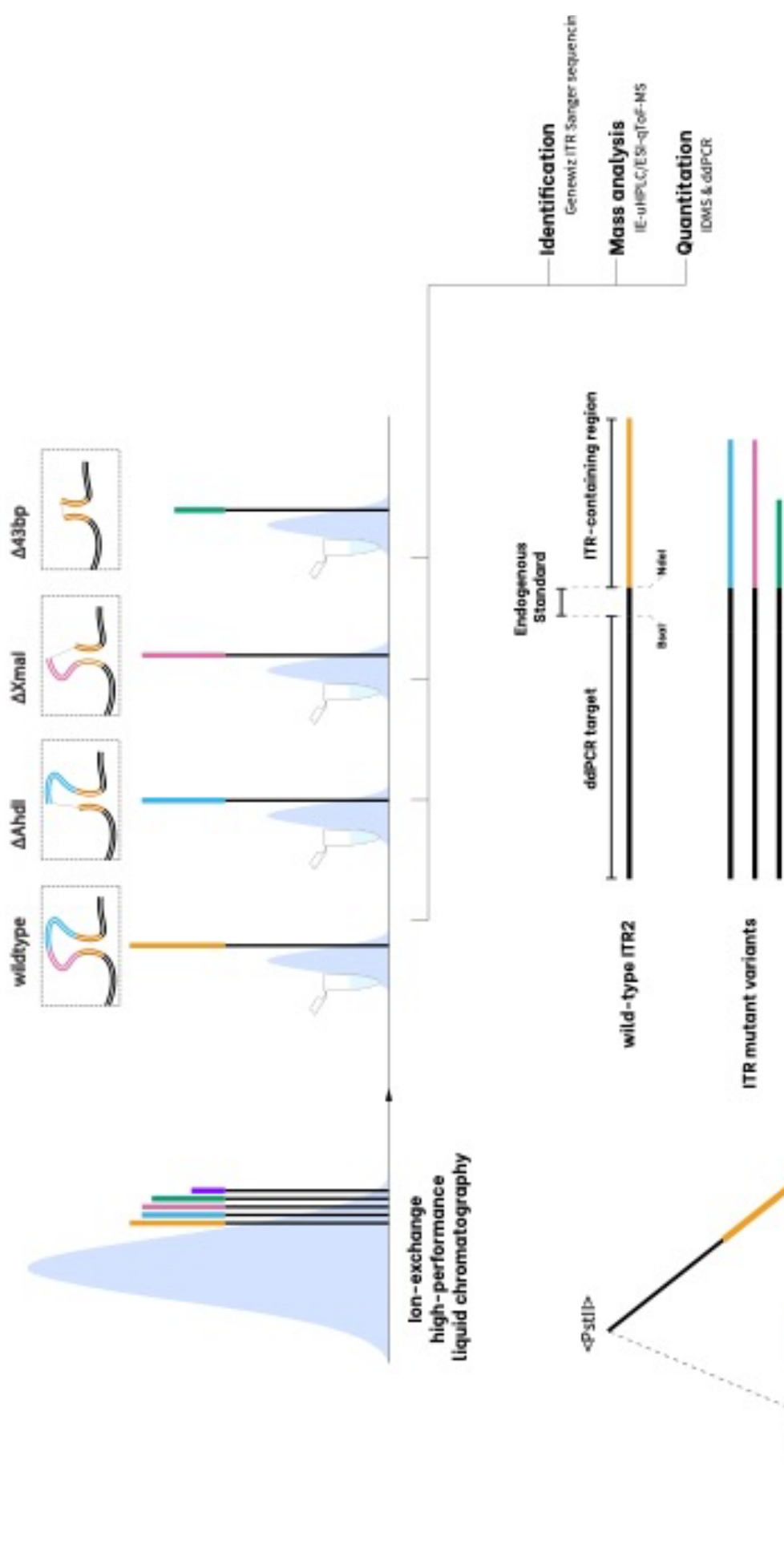
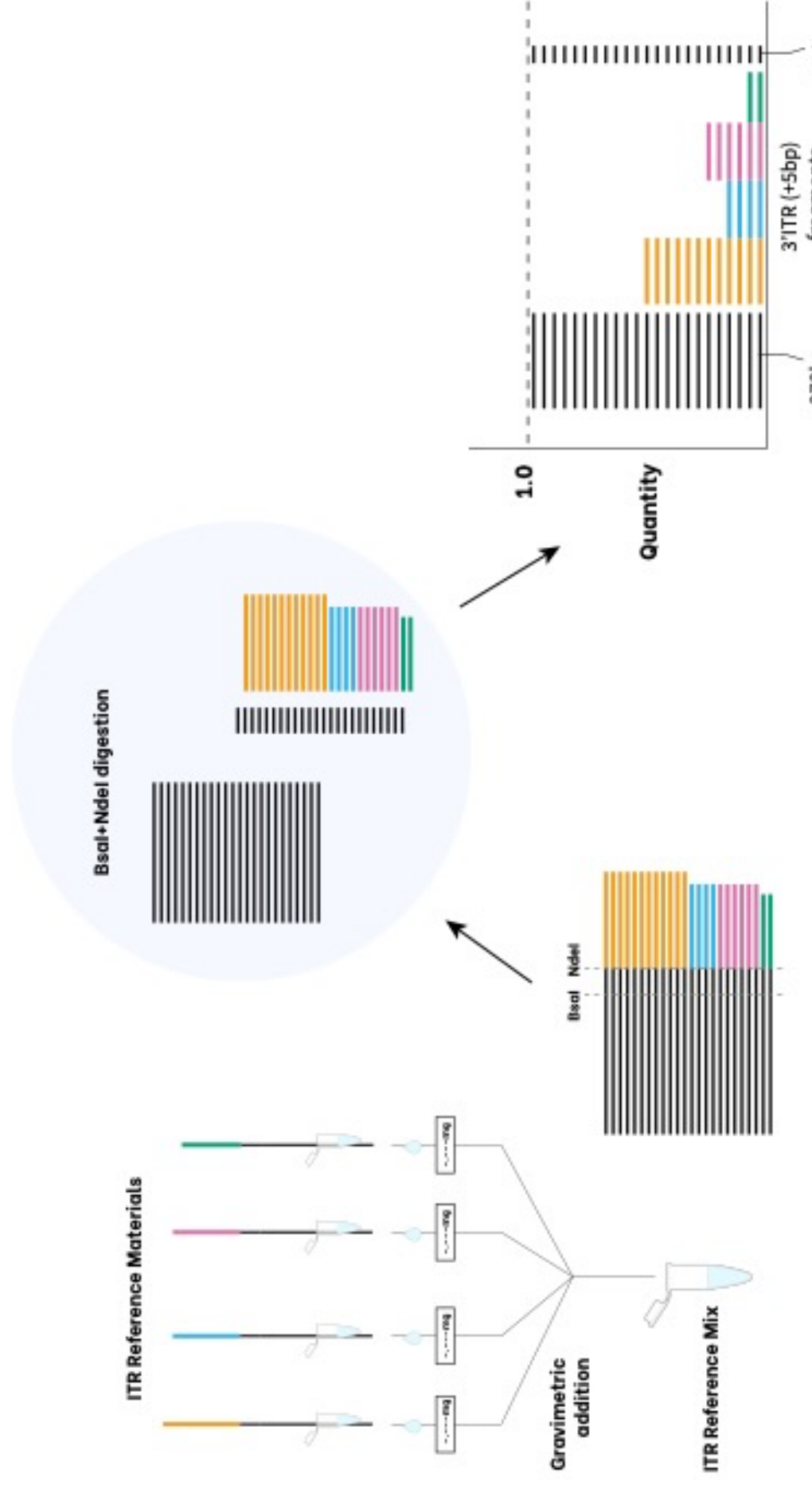


Figure A. Research was commenced in pursuit of validated techniques for the measurement of full-length ITR sequences in plasmid DNA preparations. The design ultimately incorporated a native segment of adjacent to the ITR and put it to use as a target for orthogonal analysis and an endogenous stan

Figure A. (Cont.)



Appendix B. A synopsis of efforts to develop a mass-spectrometric qRBA method

In Chapter 3 I raised an analogy to help explain qRBA, likening mixtures of barcoded DNA variants to grains of sand with specific colours in a mixed pile. This analogy is intended to highlight the resilience of the measurand – the ratio between reporter variants – against processes that affect their absolute quantities. While hopefully illuminating in the context of massively parallel sequencing, the analogy may come across as quaint to those with experience in legacy techniques in analytical biochemistry, relying routinely on spiked or endogenous species mixed with an analyte, allowing it to be quantified in relative terms. In that Chapter, I also emphasised that qRBA should not strictly be viewed in association with massively parallel sequencing. As an alternative to my original endeavour, I was enthusiastic to continue work in a metrology-focused environment by validating the concept of qRBA with mass spectrometry as the ‘counting’ mechanism.

To achieve this, the ‘UMI’ would consist not of a nucleobase sequence variation *per se*, but rather a restriction site in a variable location, such that the reporter amplicon would be digested to yield products with unique mass-to-charge (m/z) ratios. The two most obviously attractive qualities of this technique are (1) that the c parameter of the binomial model (analogous to NGS readcount) would be raised by orders of magnitude, and (2) the mechanistic simplicity of reporter fragments colliding with an ion detector (as opposed to the complex mechanisms of NGS sequencing). Less readily appreciated, but highly significant, is the powerful metrological control implicit to the technique; each reporter is represented by not two, but four equimolar, reporter-specific m/z variants. The number of restriction products is two (though this number could be raised through additional designs) and the number of molecules present in each product is two (the two complementary DNA strands).

To assess the feasibility of this concept, we utilised a PCR amplicon without any variable sequence, but which hosted alternative restriction recognition sites that could be used as surrogates for m/z reporter barcodes. In addition to testing the assay, this gave us the opportunity to identify the practical elements that affected feasibility. Biochemically, the MS-qRBA process needed to be designed with minimal involvement of nonvolatile salts and/or large biomolecules, which adversely affect the tandem HPLC-MS pipeline. Restriction endonucleases are of course obligatory, but these enzymes would be required to be active in

BSA-free reaction buffers. Thankfully BSA is less often than not unnecessary for restriction endonuclease activity, serving rather to maximise buffer compatibility. Another practical hurdle is purification of the restriction products. Here, a balance needed to be struck between minimising fragment size (in the interest of m/z resolution) and the decreased recovery yields characteristic to DNA molecule as they reduce in size.

Using capillary gel analysis, I was able to show that magnetic bead separation was superior to MWCO filtration in terms of yield and recovery bias, however HPLC-MS test samples exhibited an undesirable degree of noise after one round of purification. The other challenge that arose in the early stages was deconvolution – as an initial strategy we had relied on a minimum mass difference of 1%, but for the fragments in question this still resulted in charge states that differed in m/z by less than the value of a proton. Neither of these results precluded further progress, but it was decided that in conjunction, these hurdles meant further efforts would more appropriately suited to post-doctoral mass spectrometric research, rather than the final twelve months of a doctoral candidacy.

$$cpq = \text{variance} = \text{SD}^2$$

$$\text{SD} = \sqrt{cpq}$$

$$\text{RSD}_{cp_s} = \frac{\sqrt{cpq}}{cp} \quad \text{RSD}_{cq_s} = \frac{\sqrt{cpq}}{cq}$$

$$\begin{aligned} \text{RSD}_{\left[\frac{p_s}{q_s}\right]} &= \frac{\sqrt{cpq}}{cp} + \frac{\sqrt{cpq}}{cq} \\ &= \frac{q\sqrt{cpq}}{cpq} + \frac{p\sqrt{cpq}}{cpq} \\ &= \frac{\sqrt{cpq}(p+q)}{cpq} = \frac{\sqrt{cpq} \cdot 1}{cpq} \\ &= \frac{(cpq)^{0.5}}{(cpq)^1} \\ &= \frac{1}{(cpq)^{0.5}} \end{aligned}$$

$$\text{RSD}_{\left[\frac{p_s}{q_s}\right]} = (cpq)^{-0.5} = \frac{1}{\sqrt{cpq}}$$

Appendix D $r_{[luc2]}$ and $r_{[egfp]}$ can be modelled as composite power functions

In this appendix readers will note many vestigial references to areas in Chapter 5, its former home (now Chapter 4 of this revised manuscript). The ability to characterise the input-output relationship between ‘known’ r and measured r would contribute greatly to our understanding of the error observed and allow us to hypothesise its sources with greater acuity. Unfortunately, the input-output relationship for these measurements is not observable because r here is unknown. We can easily plot the two outputs $r_{[luc2]}$ and $r_{[egfp]}$ as (x, y) coordinates respectively and use (log-log) linear regression to describe y in terms of x . This tells us how $r_{[luc2]}$ relates to $r_{[egfp]}$ but not how either relate to r . If the two resulted successfully in a $y = x$ regression model, the reason, we would fairly assume, is that both measure r without systematic error. Any other outcome requires further contemplation to interpret why the data exist the way they do, and how alternative parameters would have shaped the regression. Available to us are the two outputs $r_{[luc2]}$ and $r_{[egfp]}$, and the knowledge that they share the same input plasmid ratio r .

This section will describe how the data introduced in Section 5.4.1 can be modelled as the result of two outputs that relate to r by way of a power relationship. It will show how the $r_{[luc2]}$ and $r_{[egfp]}$ can be expressed in composite with respect to the more generic variable x . This model will refine our understanding of the bias observed and allow us to constrain the underlying variables to a limited range of solutions. It will allow the error to be described by two pairs of coordinates on the x - y plane, which will be valid regardless of which reporter barcode is classified as p or which reporter assay is plotted on the y -axis. It will also help us to make sense of a hyperbolised representation of the data (introduced in a following section) that will show patterns of bias to be Mix-specific rather than (predominantly) barcode-specific. The scope of this model will not include conclusions as to its mechanism. It will also treat the functions described as deterministic, though in reality their outputs are governed by a binomial distribution. The variable r is treated as the standalone value rather than the quotient of two operands p and q .

To model $r_{[egfp]}$ and $r_{[luc2]}$ as functions of the variable r , we will refer to them as $f(r)$ and $g(r)$ respectively.

$$r_{[egfp]} = f(r)$$

$$r_{[luc2]} = g(r)$$

Power functions take the form $f(x) = ax^b$. They can be expressed by a single term (e.g. ax^b) and include no constant term, in contrast (for example) to a quadratic function such as $f(x) = ax^2 + c$. Taking the plasmid ratio r as our independent variable and our conjugate qRBA measurements as dependent variables:

$$r_{[egfp]} = f(r) = ar^b \quad (1)$$

$$r_{[luc2]} = g(r) = cr^d \quad (2)$$

So far as we can be confident in the linearity of the log-log plot in Figure 5.4.2a, we can be confident that the line formed by $g(r)$ and $f(r)$ is not polynomial or exponential; a straight line can only result from a log-log transformation if the input is raised to single exponent, which forms the gradient of the log-log plot. This excludes the possibility of a constant term, which would be a coefficient of r^0 . The log-log relationship must take the form $\log y = m \log x + b$, which can be satisfied by a composite of two power functions. Linear functions (or a composite of two linear functions) also satisfy this requirement, given that they share a single exponent of 1. This possibility can be excluded though, so far as we can be confident that the gradient of the log-log plot is not 1. (Though to be completely accurate, one of either $g(r)$ or $f(r)$ could be linear and can be modelled as ar^1 or cr^1 if this were the case). For some plasmid pairs, $g(r)$ vs $f(r)$ may be fairly approximated by $y = x$, but the model described in this section is still affects how this is interpreted – it acknowledges the presence of some error source, be it large or small, whose effect is to exponentiate r .

With $f(r)$ plotted on the x -axis and $g(r)$ on the y -axis, we can use the terms ar^b and cr^d to define the y -coordinate with respect to x rather than to $f(r)$. To do so, we will employ an *inverse function*. An inverse function, denoted with a superscripted -1 , applies a function ‘in reverse’, such that:

$$f^{-1}[f(r)] = r$$

Note carefully that in this context, -1 in superscript signifies the inverse function, *not* the reciprocal of the function. This section will later address reciprocals of qRBA measurements (that is, $q \div p$ rather than $p \div q$), so it is very important that the two are distinguished. To minimise the risk of confusion, $[f(x)]^{-1}$ is the only way that $1 \div f(x)$ will ever be expressed in this thesis.

Since we have plotted f on the x -axis, we can use x as the input to f^{-1} .

$$ar^b = x \iff \left(\frac{x}{a}\right)^{\frac{1}{b}} = r \quad (a, b \neq 0)$$

Now, we have the function denoted by f and its inverse denoted by f^{-1} .

$$f(r) = x = ar^b \iff f^{-1}(x) = r = \left(\frac{x}{a}\right)^{\frac{1}{b}}$$

With r expressed in terms of x , we can express $g(r)$ in terms of x . From (2):

$$\begin{aligned} g(r) &= g\left[\left(\frac{x}{a}\right)^{\frac{1}{b}}\right] \\ &= c\left[\left(\frac{x}{a}\right)^{\frac{1}{b}}\right]^d \\ g(r) &= c\left(\frac{x}{a}\right)^{\frac{d}{b}} \end{aligned}$$

With $f(r)$ and $g(r)$ plotted as (x, y) coordinates respectively, y becomes a composite function of x whose components are f^{-1} and g , where $(x/a)^{1/b}$ is the output of f^{-1} and the input of g . This concept has been illustrated in Figure 5.4.8a.

$$y = g[f^{-1}(x)] = c\left[\left(\frac{x}{a}\right)^{\frac{1}{b}}\right]^d = c\left(\frac{x}{a}\right)^{\frac{d}{b}}$$

Stated with standard notation: when $y = g(r)$ and $x = f(r)$,

$$y = g \circ f^{-1}(x) = c\left(\frac{x}{a}\right)^{\frac{d}{b}}$$

Likewise, when $y = f(r)$ and $x = g(r)$:

$$y = f \circ g^{-1}(x) = a\left(\frac{x}{c}\right)^{\frac{b}{d}}$$

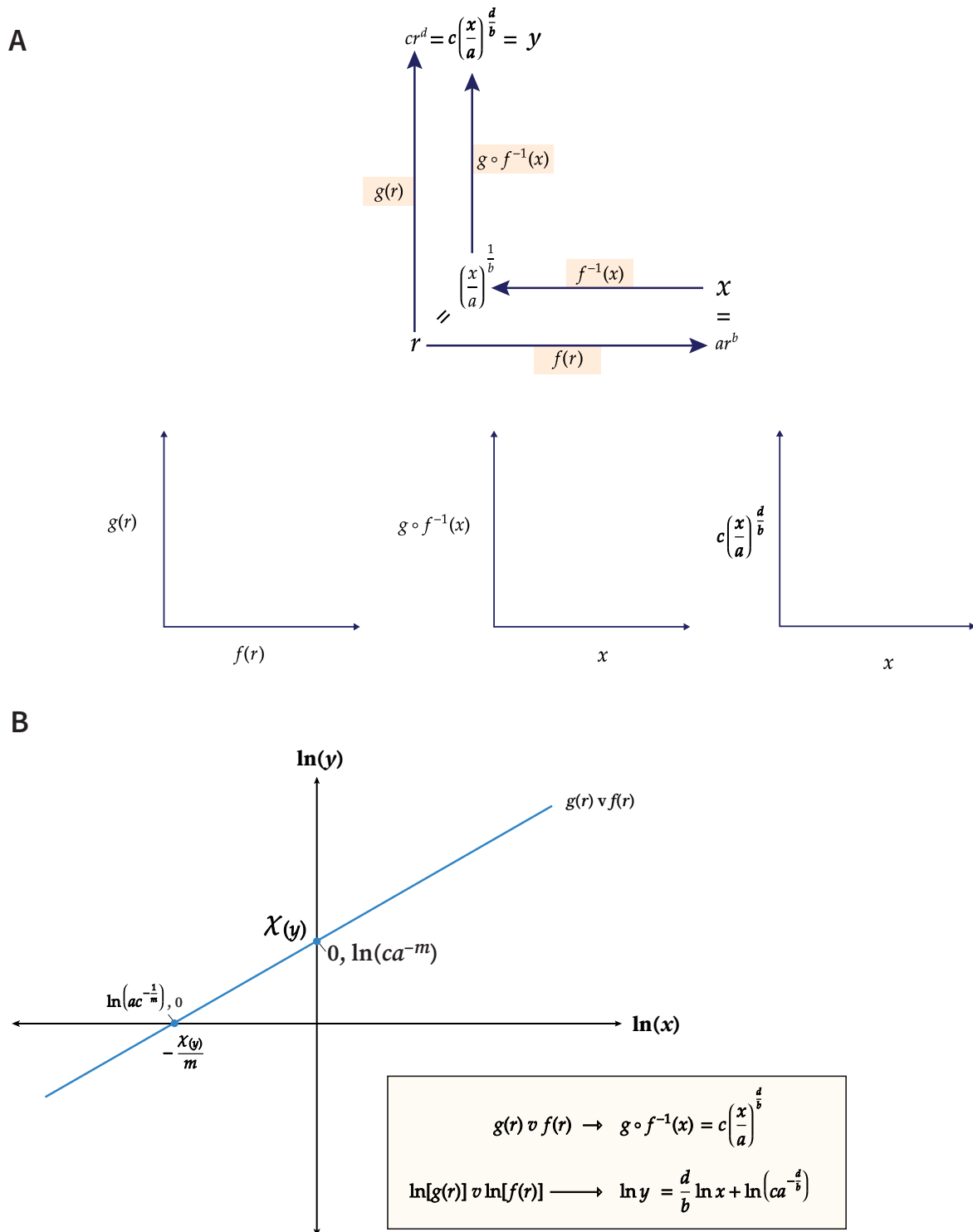


Figure 5.4.8 A composite function can be used to describe the relationship between our two independent variables $r_{[\text{egfp}]}$ and $r_{[\text{luc2}]}$ are both dependent variables relating to the independent variable r . The plots shown thus far in this section do not directly convey y as a function of x . This can however be achieved by expressing y as a composite function of $r_{[\text{egfp}]}$ and $r_{[\text{luc2}]}$. **(A)** This approach is depicted visually. Each of the three smaller plots in the lower panel is a valid way of labelling the x and y axes. The composite function specified is based on the argument in the main text, that each conjugate measurement may have a power relationship to the physical reporter ratio. **(B)** The log-transformed qRBA data forms a straight line, whose gradient and y -intercept relates to the parameters of the component functions. Throughout this section, the term $\chi[y]$ refers to the y -intercept when $\ln(r_{[\text{luc2}]})$ is plotted against $\ln(r_{[\text{egfp}]})$.

We can measure the fold difference between the exponents d and b via a log transformation such that $(x, y) \rightarrow (\ln x, \ln y)$. **The curve $g \circ f^{-1}(x)$, will be converted to a straight line of the form $y = mx + \chi_{[y]}$, where the gradient m will be equal to $d:b$ and the y -intercept $\chi_{[y]}$ ('chi-y') will be $\ln(ca^{-m})$** (Figure 5.4.8b). Generally speaking:

$$y = ax^b \iff \ln y = b \ln x + \ln a$$

Applying this to our composite function $g \circ f^{-1}(x)$:

$$\begin{aligned} \ln[g \circ f^{-1}(x)] &= \ln c \left(\frac{x}{a} \right)^{\frac{d}{b}} \\ &= \frac{d}{b} (\ln x - \ln a) + \ln c \\ &= \frac{d}{b} \ln x + \left(\ln c - \frac{d}{b} \log a \right) \\ \ln[g \circ f^{-1}(x)] &= \frac{d}{b} \ln x + \ln \left(ca^{-\frac{d}{b}} \right) \end{aligned}$$

Simplifying $d:b$ to m (the gradient), we have:

$$\ln[g \circ f^{-1}(x)] = m \ln x + \ln(ca^{-m})$$

This expression also identifies the y -intercept $\chi_{[y]}$ as $\ln(ca^{-m})$. Given the constants m and $\chi_{[y]}$, we can draw an exponential curve $c = e^{\chi_{[y]}} a^m$, along which lie all possible values of (c, a) as (x, y) coordinates. This curve approaches $y = x$ as the exponent ratio (m) approaches 1 and as ca^{-m} approaches 1 (that is, as $\chi_{[y]}$ approaches 0).

$$\begin{aligned} \chi_{[y]} &= \ln(ca^{-m}) \\ e^{\chi_{[y]}} &= ca^{-m} \\ e^{\chi_{[y]}} &= \frac{c}{a^m} \\ c &= e^{\chi_{[y]}} a^m \end{aligned}$$

When $y = f(x)$ and $x = g(y)$, the log-log plot can be defined through simple 45° mirror symmetry. Its gradient will be m^{-1} and its y -intercept will be the x -intercept of $\ln g(r)$ vs $\ln f(r)$. Since the gradient m is rise-over-run, that x -intercept ($\chi_{[x]}$) is simple to derive:

$$\chi_{[x]} = -\frac{1}{m} \chi_{[y]}$$

$$\chi_{[x]} = -\frac{1}{m} \ln(ca^{-m})$$

Another iteration of this regression depends on ‘which way up’ $p:q$ is measured. This can easily be a source of confusion. While it is facile to divide [Q] reads by [P] reads (rather than [P] by [Q]) the result will not be $f(r^{-1})$. If $f(r) = r$, then $[f(r)]^{-1} = f(r^{-1}) = r^{-1}$. However, this is not the case where $f(r) \neq r$. In our dataset, neither $r_{[luc2]}$ nor $r_{[egfp]}$ can be assumed to equal to pq^{-1} , meaning that neither $r_{[luc2]}^{-1}$ nor $r_{[egfp]}^{-1}$ can be assumed equal to qp^{-1} . That is to say:

$$\checkmark \quad \frac{p}{q} \longrightarrow r_{[luc2]}$$

$$\otimes \quad \frac{q}{p} \longrightarrow \frac{1}{r_{[luc2]}}$$

A different function is required to relate x and y values to the reciprocal of r . Fortunately, the translational symmetry of these functions lets us refrain from defining more functions in this section. By taking the reciprocal of [P]/[Q], the composite function already defined is simply translated such that the new (log-log) x and y intercepts are the originals multiplied by -1 . We can show this for $g \circ f^{-1}(x)$ and assume the same for $f \circ g^{-1}(x)$ due to 45° mirror symmetry.

$$\frac{1}{f(r)} = \frac{1}{ar^b} \quad \frac{1}{g(r)} = \frac{1}{cr^d}$$

$$x = \frac{1}{ar^b} \iff r = \left(\frac{1}{xa}\right)^{\frac{1}{b}}$$

$$y = \frac{1}{g(r)}$$

$$= \frac{1}{c \left[\left(\frac{1}{xa} \right)^{\frac{1}{b}} \right]^d}$$

$$= \frac{1}{c \left(\frac{1}{xa} \right)^{\frac{d}{b}}}$$

$$y = \frac{(xa)^{\frac{d}{b}}}{c}$$

Applying the log-log transformation gives us the same gradient as $g \circ f^{-1}(x)$, a y -intercept equal to $-\chi_{[y]}$ (that is, equal to $-\ln(ca^{-m})$) and an x -intercept equal to $-\chi_{[x]}$.

$$\begin{aligned}
 \ln y &= \ln \left[\frac{(xa)^m}{c} \right] \\
 &= m (\ln x + \ln a) - \ln c \\
 &= m \ln x + m \ln a - \ln c \\
 &= m \ln x - (\ln c - m \ln a) \\
 \ln y &= m \ln x - \ln(ca^{-m})
 \end{aligned}$$

For the x -intercept:

$$\begin{aligned}
 m \ln x - \ln(ca^{-m}) &= 0 \\
 \ln x &= \frac{1}{m} \ln(ca^{-m}) \\
 \ln x &= -\chi_{[x]}
 \end{aligned}$$

The symmetrical relationship between any of the four possible iterations of this regression is illustrated can be seen in Figure 5.4.9. Four different regressions are possible depending on which output is plotted on the x -axis, and which reporter barcode is classified as P. The four intersects that occur when each of these iterations are plotted are the vertices of a rhombus. At each of these vertices, the x coordinate and y coordinate are equal (that is, $x, y = x, x$). Because of symmetry, only two different coordinates are necessary to identify all eight that locate the four vertices. These two coordinates can be calculated from any of the four possible regressions. This simplifies the analysis, but does not directly contribute to the overall argument of the Chapter and so will be shown in Appendix D. The two coordinates are:

$$\frac{m \ln(ca)}{m+1} \quad \text{and} \quad \frac{m \ln(ca)}{m-1}$$

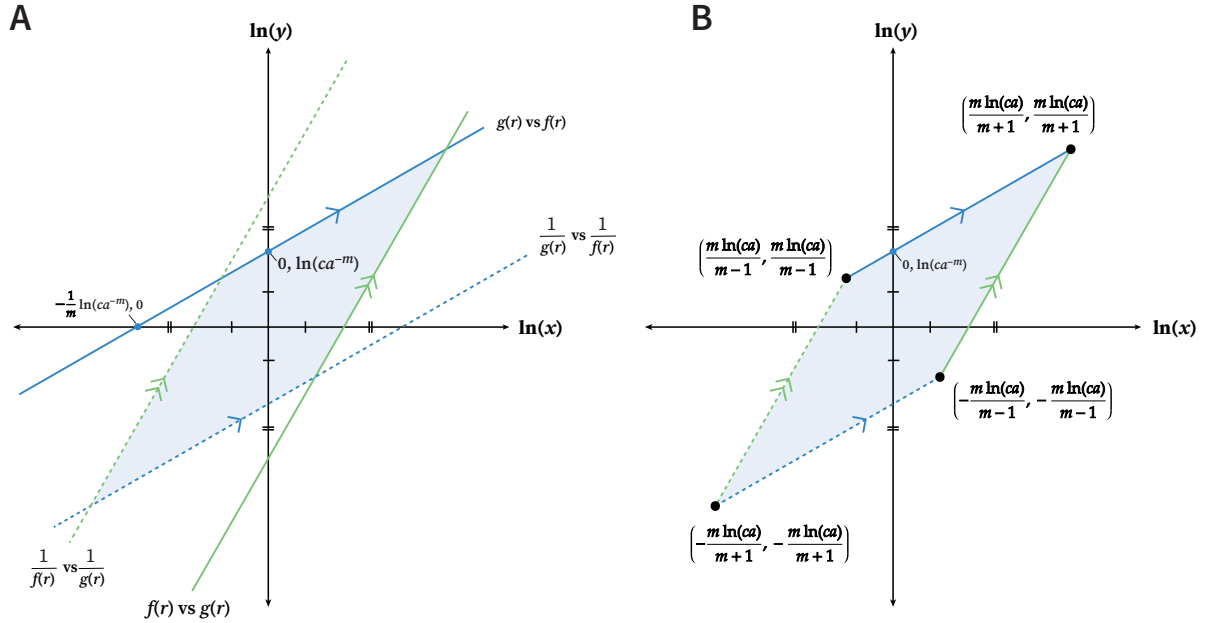


Figure 5.4.9 Symmetry between different arrangements of data allows the error pattern to be generalised | (A) In this experiment, the choice was available to plot either $r_{[luc2]}$ or $r_{[egfp]}$ on the y axis, and to measure the quotient one way up or the other. (That is, to assign either barcode in a pair as 'p' and the other as 'q'). At first glance this may seem like a dilemma, but in reality, any one of these four possible regressions can describe the relationship in terms that apply to all four. Essentially, when we try to describe the error pattern quantitatively, what we are describing is the rhombus shaded in blue. **(B)** As shown in Appendix E, the vertices of this rhombus can be located from any of the four possible regressions. While this does apply to any set of four straight lines that share this relationship, the coordinates specific to the composite power model suggested in this Chapter are defined on the right.

Appendix E A hyperbolic transformation affirms Mix-specific patterns of error

Unsurprisingly, $(r_{[egfp]}, r_{[luc2]})$ coordinates derived from the same plasmid mixture tended to cluster within a single location, which was not conducive to linear regression analysis. It was therefore not an appropriate tool to distinguish whether patterns of error were Mix-specific. To explore whether Mix-specific trends existed, data were transformed such that y was a hyperbolic function of x , whose vertical and horizontal asymptotes were $\chi_{[y]}$ and m respectively. For the dataset in question, these two parameters tended closely towards 0 and 1 respectively. This placed most transformed data in closer vicinity, within a zone of relatively rapid vertical-horizontal transition (as will be explained). It may be that the asymptotes of these hyperbolae are unable to be estimated with much greater certainty than linear regression, however, results that had clustered amorously in the log-log plot regressed well to a hyperbolic model in this transitional zone. This rather abstract relationship between y and x therefore brought Mix-specific trends into relief that were otherwise poorly observable.

A formal analysis of these plots is counted among those unanticipated tasks that, while potentially fruitful, required specialised know-how and needed to be deferred decisively in the interest of a timely submission. In this short section, the method of hyperbolic transformation is specified and the transformed data are presented so that the Mix-specific trends can be observed.

The hyperbola takes the form:

$$y = \frac{dx + \ln\left(\frac{a^d}{c^b}\right)}{bx + \ln\left(\frac{a^d}{c^b}\right)}$$

Two further functions are defined, s and u , and the above hyperbola is obtained by plotting $u(r)$ on the y -axis and $s(r)$ on the x -axis, giving us the composite function $u \circ s^{-1}(x)$. To reduce crowding, $d - b$ can be expressed below as Δ .

$$x = s(r) = \ln \frac{g(r)}{f(r)} \quad y = u(r) = \frac{\ln g(r)}{\ln f(r)}$$

$$x = \ln \left(\frac{cr^d}{ar^b} \right)$$

$$e^x = \left(\frac{c}{a}\right) r^{d-b}$$

$$r = \left(\frac{ae^x}{c}\right)^{\frac{1}{d-b}}$$

$$x = s(r) = \ln \left(\frac{cr^d}{ar^b}\right) \Leftrightarrow s^{-1}(x) = r = \left(\frac{ae^x}{c}\right)^{\frac{1}{d-b}}$$

$$y = u \circ s^{-1}(x) = \frac{\ln g(r)}{\ln f(r)}$$

$$u \circ s^{-1}(x) = \frac{\ln \left[c \left(\frac{ae^x}{c} \right)^{\frac{d}{\Delta}} \right]}{\ln \left[a \left(\frac{ae^x}{c} \right)^{\frac{b}{\Delta}} \right]}$$

$$= \frac{\ln c + \ln \left[\left(\frac{ae^x}{c} \right)^{\frac{d}{\Delta}} \right]}{\ln a + \ln \left[\left(\frac{ae^x}{c} \right)^{\frac{b}{\Delta}} \right]}$$

$$= \frac{\ln c + \frac{d}{\Delta} [\ln(ae^x) - \ln c]}{\ln a + \frac{b}{\Delta} [\ln(ae^x) - \ln c]}$$

$$= \frac{\ln c + \frac{d}{\Delta} (\ln a + x - \ln c)}{\ln a + \frac{b}{\Delta} (\ln a + x - \ln c)}$$

$$= \frac{\ln c + \frac{1}{\Delta} (\ln a^d + dx - \ln c^d)}{\ln a + \frac{1}{\Delta} (\ln a^b + bx - \ln c^b)}$$

$$= \frac{\Delta^{-1} [\Delta \ln c + (\ln a^d + dx - \ln c^d)]}{\Delta^{-1} [\Delta \ln a + (\ln a^b + bx - \ln c^b)]}$$

$$= \frac{\Delta \ln c + (\ln a^d + dx - \ln c^d)}{\Delta \ln a + (\ln a^b + bx - \ln c^b)}$$

Substituting $(d - b)$ for Δ and expanding:

$$y = \frac{(\ln c^d - \ln c^b) + (\ln a^d + dx - \ln c^d)}{(\ln a^d - \ln a^b) + (\ln a^b + dx - \ln c^b)}$$

Rearranging terms:

$$y = \frac{dx + \ln a^d - \ln c^b + (\ln c^d - \ln c^d)}{bx + \ln a^d - \ln c^b + (\ln a^b - \ln a^b)}$$

$$y = \frac{dx + \ln \left(\frac{a^d}{c^b} \right)}{bx + \ln \left(\frac{a^d}{c^b} \right)}$$

Hyperbolae with this form have a horizontal asymptote $y = d : b$, giving us the asymptote $y = m$. The vertical asymptote is located where the denominator is equal to zero, in this case where $x = \chi_{[y]}$.

$0 = bx + \ln \left(\frac{a^d}{c^b} \right)$ $bx = -\ln \left(\frac{a^d}{c^b} \right)$ $x = -\frac{1}{b} \ln \left(\frac{a^d}{c^b} \right)$	$x = -\ln \left(\frac{a^d}{c^b} \right)^{\frac{1}{b}}$ $= -\ln \left(\frac{a^d}{c^b} \right)$ $= -(\ln a^m - \ln c)$	$= \ln c - \ln a^m$ $x = \ln(ca^{-m})$ $x = \chi_{[y]}$
--	--	---

The results of this transformation are shown in Figure 5.4.10, which spans five pages. All hyperbolae described by the above equation will intersect at (0, 1) (note that if m were 1, the function would not be a hyperbola). Since m and $\ln(ca^{-m})$ were often close to 1 and 0 respectively in the present dataset, this common intersection will tend to be in close proximity to one of either hyperbolic vertices, making this a favourable region for resolving differences in these parameters. For ease of visualisation, the axes in Figure 5.4.10b are drawn along the lines $x = 0$ and $y = 1$. Presented alongside are the average $r_{[egfp]}$ ratios taken from each NGS replicate and PCR sample derived from a given plasmid mixture ($n = 15$). At least one tentative observation is worth highlighting: the position of the asymptotes for a given barcode pair tend to correspond, at least in ordinal terms, with the rank in $r_{[egfp]}$ ratio specific to a Mix. This would imply that a composite power function cannot be used to describe the error of a given barcode pair generally, since it does not allow variation in slope (let alone x -dependent variation in slope) in its log-log plot. Another observation is that in all cases, the position of vertical asymptote (signifying $\chi_{[y]}$) was positive where $r_{[egfp]} > 1$ and negative where $r_{[egfp]} < 1$. In Mix1, the $r_{[egfp]}$ result for the pair [C][D] was measured to be $1.00 \pm 1.3\%$, and the hyperbola had a vertical asymptote near to zero.

Figure 5.4.10 (1 of 5) Regression to a hyperbolic model resolves Mix-specific trends | Six plasmid template mixtures were prepared for this section, each of which was amplified in three uniquely indexed PCR reactions. Each individual plasmid ratio was therefore measured fifteen times - with three PCR replicates and five NGS replicates. These fifteen replicates tended, predictably, to cluster in a single location on the x-y plane. Subtle differences in gradient were therefore poorly detectable using linear regression, and differences in the y-intercept wholly unachievable. **(A)** A hyperbolic regression was achieved using the x,y metrics specified in the upper panel, which produced the regression model shown in blue. Hyperbolae with this form will have a horizontal asymptote at m and a vertical asymptote at $\chi[y]$ since both of these values tended toward 1.0, much of our empirical data will fall within a close proximity in the hyperbolic model, which also be a region of relatively high curvature. These factors make it possible to resolve qualitative differences in the error pattern of datasets that were difficult to distinguish via linear regression. **(B)** For each unique reporter pair, all six Mixes were analysed under a hyperbolic regression model. Since all functions pass through (0,1), these coordinates are where the axis lines in these plots will cross. To the side are shown the average p:q result according to the egfp assay. These data are stacked by rank to help illustrate the tendency for asymptotes in these regressions to correlate to the p:q ratio. White fill is used for distinction of Mix6 data by readers with colour vision deficiencies and has no other significance.

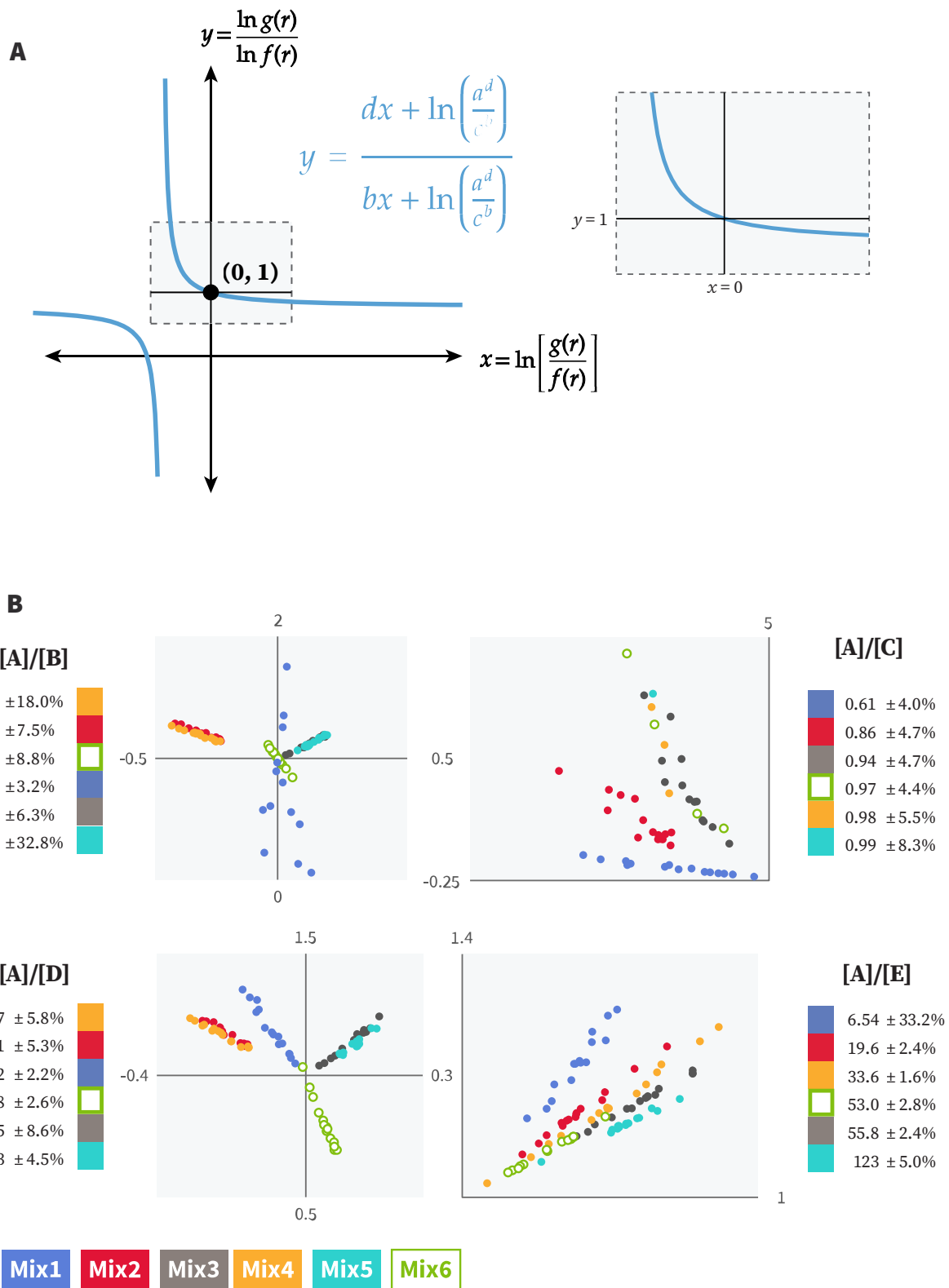


Figure 5.4.10 (1 of 5) Regression to a hyperbolic model resolves Mix-specific trends

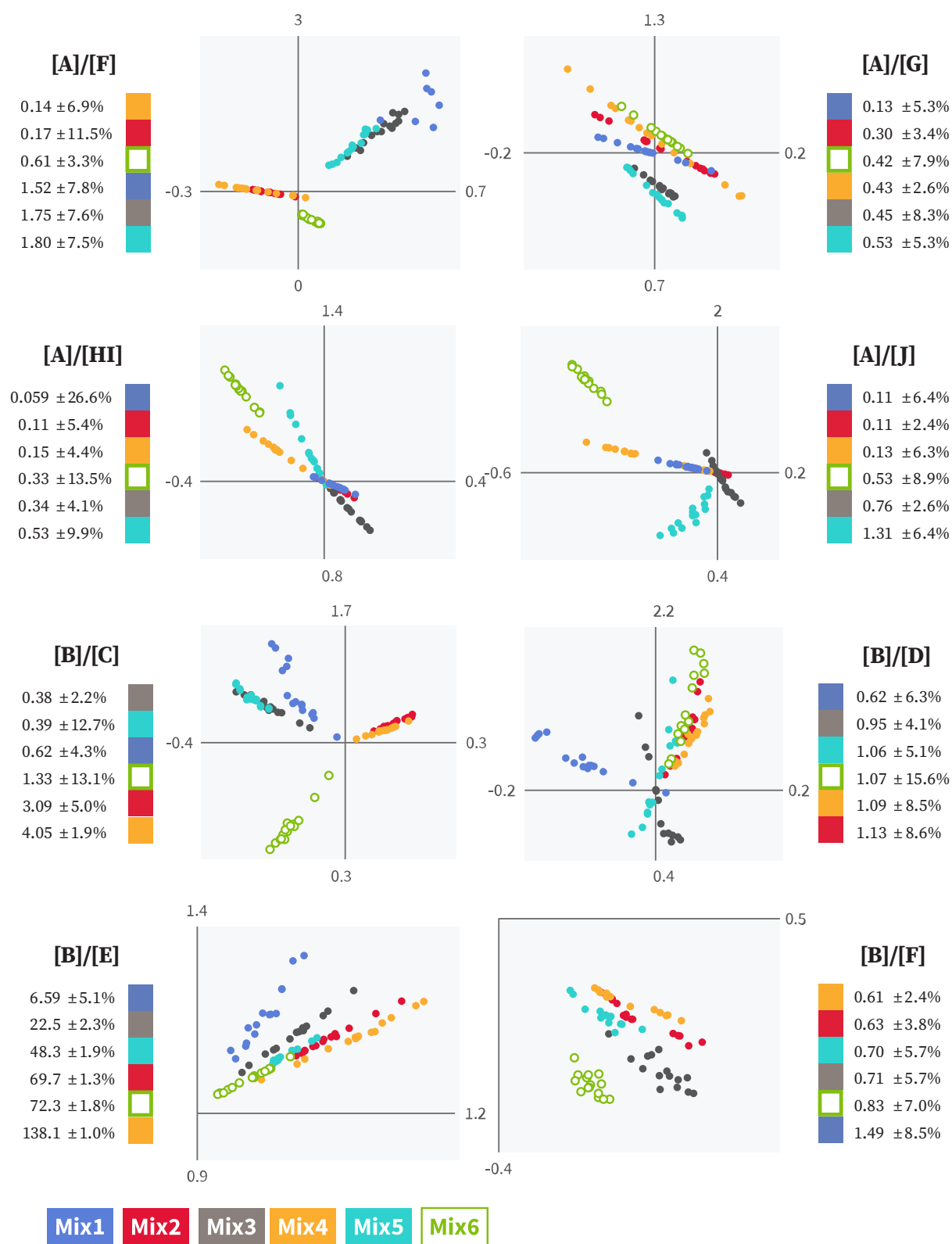


Figure 5.4.10 (2 of 5) Regression to a hyperbolic model resolves Mix-specific trends

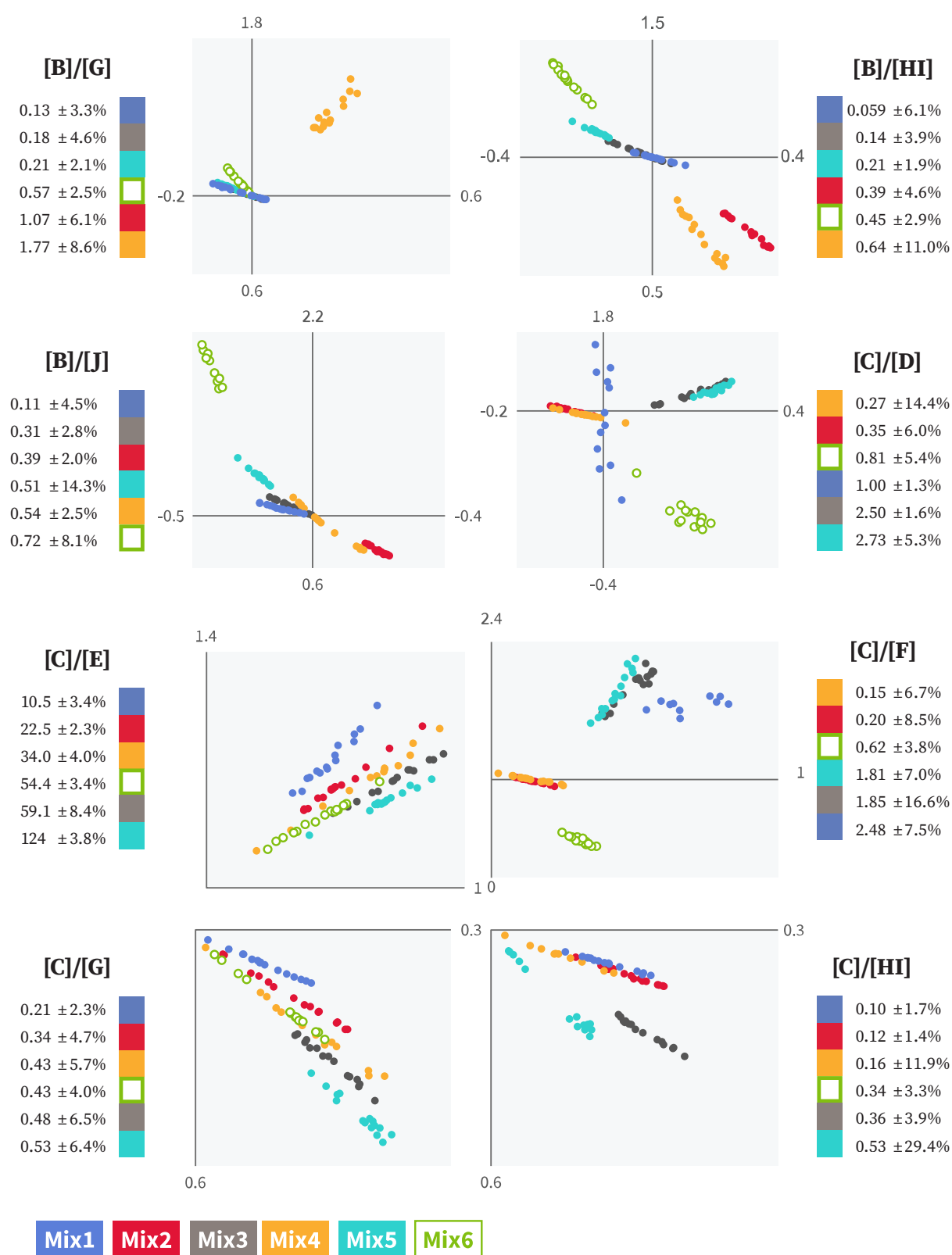


Figure 5.4.10 (3 of 5) Regression to a hyperbolic model resolves Mix-specific trends

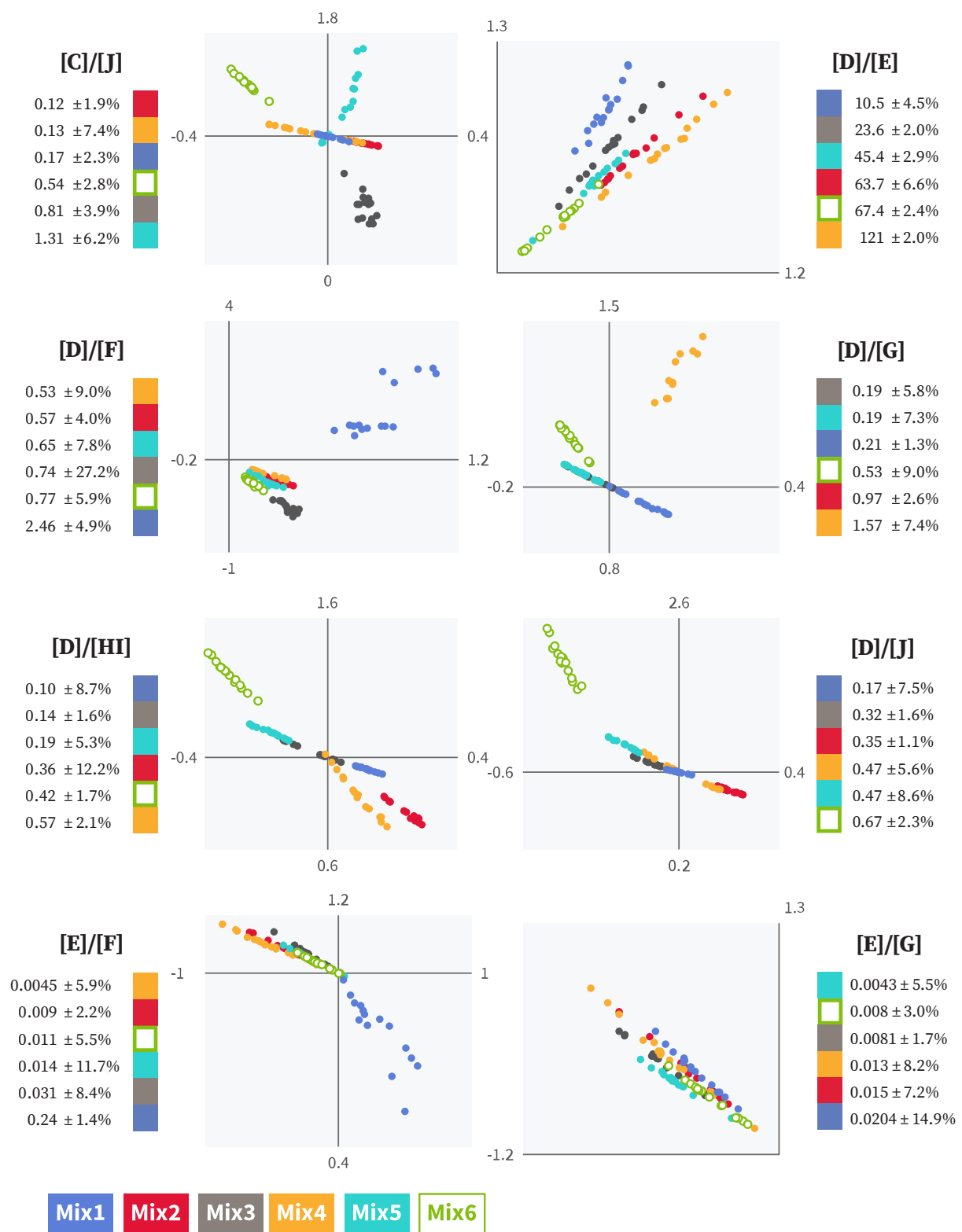


Figure 5.4.10 (4 of 5) Regression to a hyperbolic model resolves Mix-specific trends

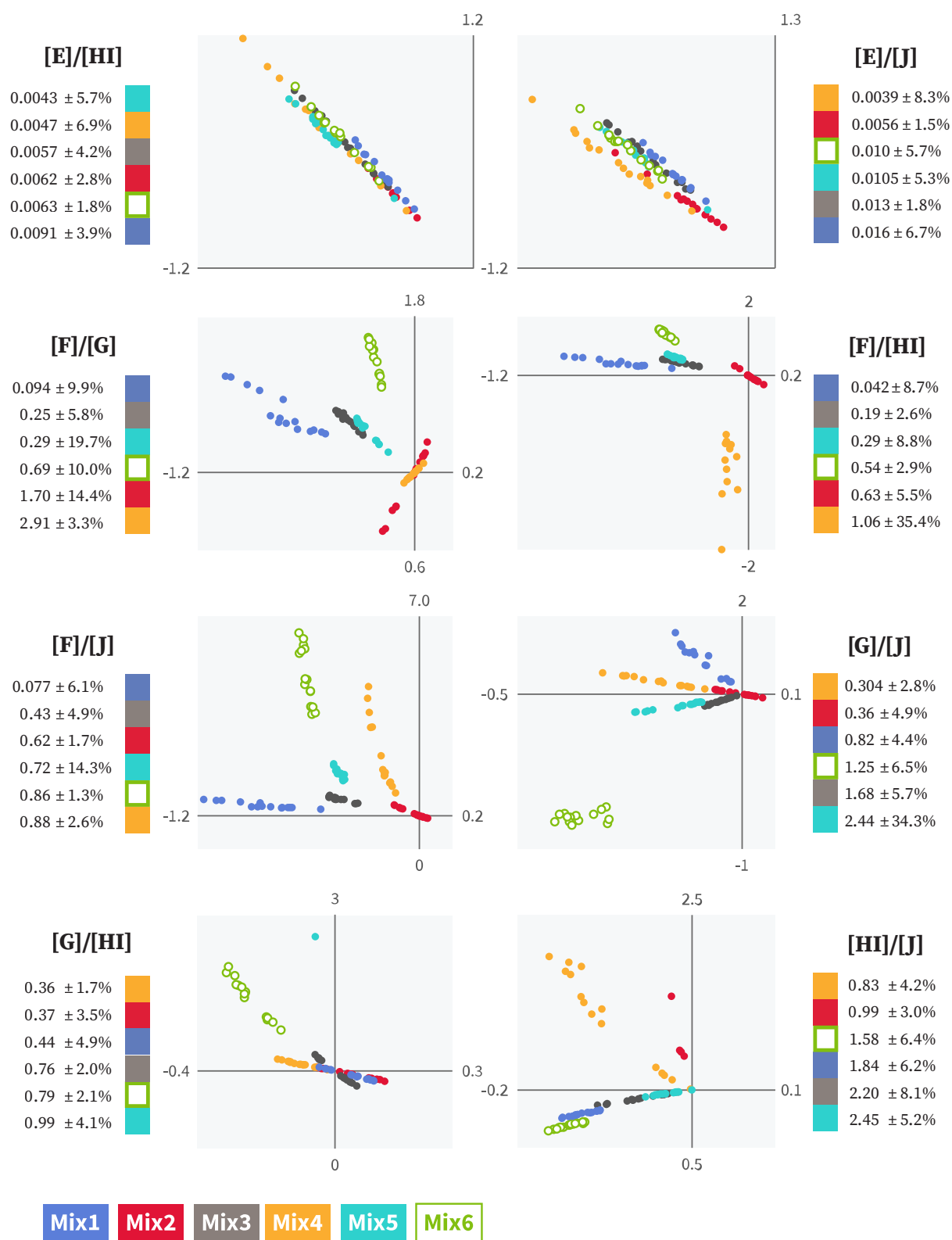


Figure 5.4.10 (5 of 5) Regression to a hyperbolic model resolves Mix-specific trends

Appendix F the error does not directly resemble PCR amplification bias

In Section 5.4.5, measurements of a given reporter barcode pair were shown to regress well to a hyperbolic model when they belonged to a particular plasmid mixture, which was measured in five NGS replicates of three indexed PCR samples. The approximate position of the asymptotes often correlated, at least in rank, to the average $r_{[egfp]}$ measurement (based on an informal analysis). This raises the possibility that the plasmid ratio itself was a parameter. One hypothesis is that the PCR process leads to an r -dependent divergence between the barcode ratio in the template material and that in the indexed PCR products. This hypothesis is compelling since it would explain Mix-specific biases without appealing to skew or NGS sequencing quality, which have not been supported. However, the hypothesis becomes somewhat tenuous when we consult the canonical model of exponential growth. As will be shown, it predicts that the y -intercepts, but not the gradients of log-log plots are affected. This is not compatible with the underlying problem of Mix-specific log-log gradients.

This final brief section will show how these inferences are made from the exponential function. While not exhaustive, this approach is a sensible first step in response to the PCR hypothesis. Firstly, the exponential function will be introduced and a few points of clarity will be made. At any given time (t), the size of an exponentially expanding population (A) is the product of two terms – its initial size (A_0) and a *growth factor* (b) carrying time in the exponent. The growth factor is the variable that defines the doubling time of the population. *Doubling time* is a useful way to refer to the ‘proliferativeness’ of exponentially expanding populations, much like the *half-life* of populations undergoing exponential decay. Note that ‘growth rate’ is incorrect because the growth rate changes continually with time.

$$A = A_0 \cdot (b)^t$$

In the natural sciences, it is conventional to express b as Euler's number (e) raised to the power of a number k . This greatly simplifies many frequently applied operations involving logarithms and calculus. It also harmonises the language of exponential growth and decay with that of intimately related concepts in physical/chemical process kinetics. Stated conventionally:

$$A = A_0 \cdot (e^k)^t$$

By this convention, k , the *growth constant*, becomes the variable that defines doubling time.

$$\text{growth factor} = b = e^k \quad \text{growth constant} = k$$

Consider two populations with sizes P and Q and growth constants k_p and k_q . As the two populations expand, the change in their ratio r can also be modelled with an exponential function – one with a growth constant of $(k_p - k_q)$.

$$\frac{P}{Q} = \frac{P_0}{Q_0} \cdot \frac{(e^{k_p})^t}{(e^{k_q})^t}$$

Since our measurand (r) is the ratio of P_0 and Q_0 :

$$\begin{aligned} f(r) &= r \cdot \frac{(e^{k_p})^t}{(e^{k_q})^t} \\ &= r \cdot e^{(k_p - k_q)t} \end{aligned}$$

Here we can note that the PCR hypothesis requires a difference between k_p and k_q for *egfp* and/or *luc2* reporter assays, since $re^0 = r$. This reflects the fact that exponential growth, in and of itself, does not cause differences in subpopulation composition.

Let us take the *egfp*- and *luc2*-targeted PCR reactions as separate processes with different discrepancies between k_p and k_q .

$$(k_p - k_q)_{[egfp]} = k_f \quad (k_p - k_q)_{[luc2]} = k_g$$

This experiment implemented 25 PCR cycles for both the *egfp*-targeted and *luc2*-targeted qRBA assays. However, though t was a common parameter, labelling t distinctly for $r_{[luc2]}$ and $r_{[egfp]}$ will make this analysis applicable whether or not the number of cycles was equal. So, we will describe the ratio of P to Q in the reporter amplicons as follows:

$$f(r) = re^{t_f k_f} \quad g(r) = re^{t_g k_g}$$

Since r is unknown we are again limited to plotting $g(r)$ against $f(r)$, which we can rephrase in terms of x with the composite method used previously. We find that nonequal amplification efficiency and/or discrepancies in the number of PCR cycles does not result in the power relationship between $r_{[egfp]}$ and $r_{[luc2]}$ that was empirically observed.

$$f(r) = x = re^{t_f k_f} \iff f^{-1}(x) = r = xe^{-t_f k_f}$$

$$\begin{aligned} y &= g \circ f^{-1}(x) \\ &= xe^{-t_f k_f} \cdot e^{t_g k_g} \\ y &= xe^{(t_g k_g - t_f k_f)} \end{aligned}$$

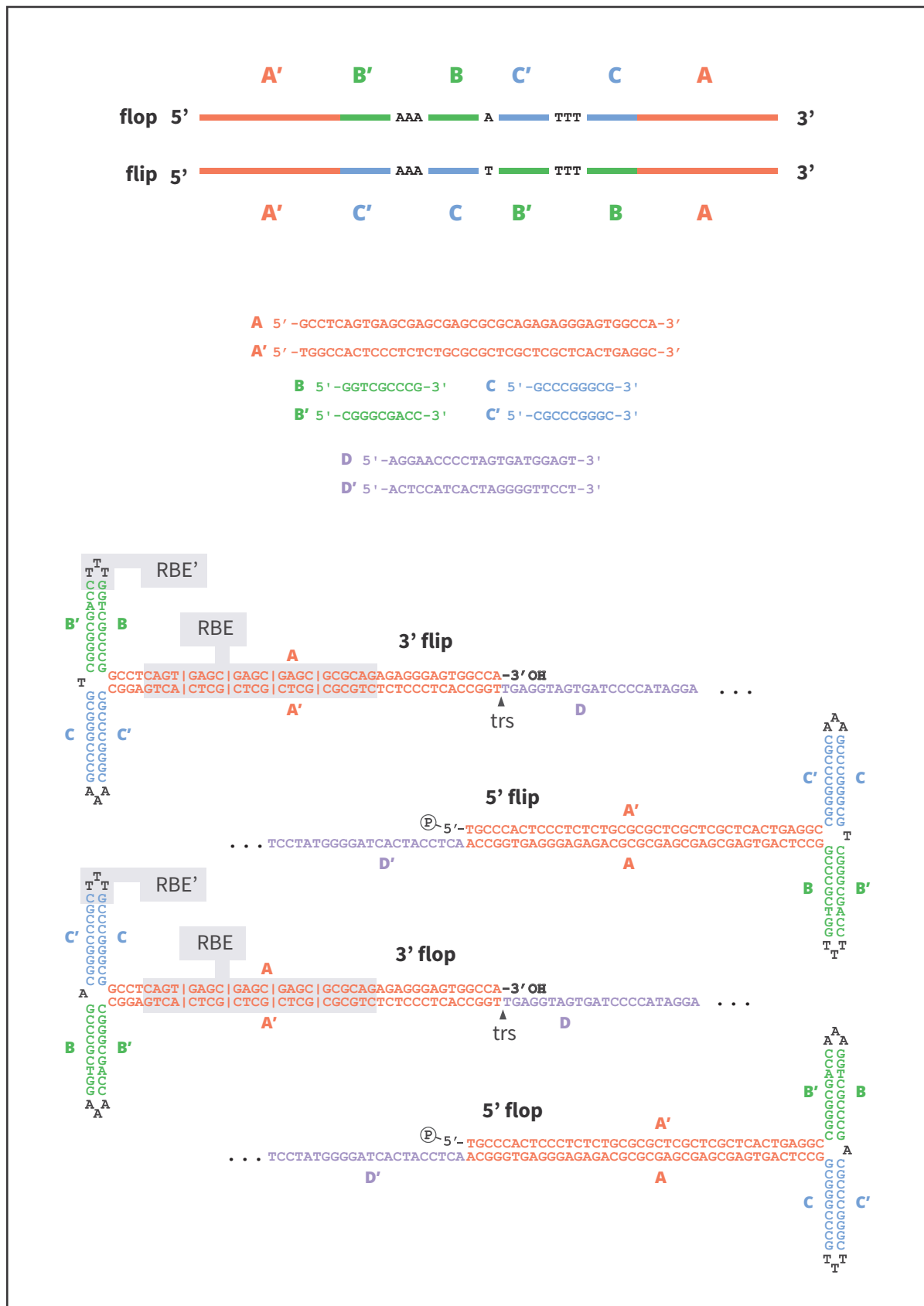
$$\ln[g \circ f^{-1}(x)] = \ln x + (t_g k_g - t_f k_f)$$

Instead, a simple linear function is predicted with a gradient equal to $e^{(t_g k_g - t_f k_f)}$. The log-log plot would have a gradient of 1 and a y -intercept of $(t_g k_g - t_f k_f)$. In such a regression, $\chi_{[y]}$ would be dependent on t_g , k_g , t_f , and k_f , but not on r . The PCR hypothesis (at least superficially) requires us to assume that these terms were variable but consistent between the three indexed PCR samples derived from a given Mix (since their hyperbolic asymptotes were consistent), and that they correlated to r by coincidence.

Appendix G Explainer: The sequence(s) of the AAV2 ITR and their representation

Understanding the primary sequence of AAV's ITRs can be a demanding spatial reasoning task given their nested layout and variable internal configurations. Further, the packaged genome as a whole exists in either sense, is represented in a counterintuitive 3'-to-5' direction, and exists in a multitude of single- and double-stranded structures outside the virion stage of the lifecycle. Worse than creating confusion, this can often be a barrier to entry for a sustained interest in these fascinating pieces of virology. Words such as “symmetrical” and “palindrome”, while conceptually appropriate, become sources of confusion if they lead to any specific assumptions. Those finding themselves challenged can take comfort in the fact that even peer-reviewed literature is prone to errors and inconsistencies where ITR sequences are concerned. For example (as will be explained), the accepted length of AAV2 ITR has been incompatible with the accepted location of its terminal resolution site for more than thirty years. Even the term “ITR”, as the terminal 145(ish) nucleotides of an ssAAV genome have come to be known, is an accepted misnomer, since it is used as a singular noun despite each viral terminus containing three-and-a-half inverted repeats. Why is the initialism trs written in lowercase? This author cannot say. Given the age of this convention, typewriter malfunctions are a plausible explanation.

Opportunities for confusion in this thesis are many, and they are difficult for its author to predict. I have prepared this explainer, if not as required reading then as a reference, to explain the sequence(s) of the AAV2 ITR and how they are annotated. Figure 5.2.2 (duplicated from Chapter 5) will be our point of reference.



5.2.2 Arrangement of inverted repeat sequences at both possible termini of the ssAAV genome | Shown above are reference sequences for 'flip' or 'flop' ITRs at both ends of the genome, including sequence labels that appear in this thesis. RBE(') = replicase binding element; trs = terminal resolution site

The seven inverted terminal repeats of adeno-associated virus

When a base sequence and its reverse complement are found on the same nucleic acid strand, we call them an inverted repeat. There appears to be no official minimum length or maximum separation between the two components for them to qualify as an inverted repeat, though the likelihood of seeing the term used increases with the potential for stable intra-strand base pairing. Self-annealed inverted repeats created a 'hairpin' structure within the ssDNA strand, with any unpaired bases between the repeat creating a 'hairpin loop'. If both strands of a double-stranded inverted repeat are self-annealed, the two opposing hairpins create a 'cruciform' structure with a four-way junction at the centre.

Four unique inverted repeat sequences exist in a packaged AAV genome, three of which appear twice (once per ITR). The BB' and CC' repeats are adjacent to one another and flanked by the AA' repeat. When self-annealed, the BB' and CC' repeats create a three-way junction in place of a simple hairpin turn. The genome contains only one copy of the fourth repeat – the DD' repeat. Its upstream and downstream components are directly inset from the hammerhead-forming repeats and separated by the rest of the viral genome. Each (r)AAV genome terminus can be best described therefore as comprising three-and-a-half inverted repeats. Seven unpaired bases also exist within each terminal hammerhead. A single adenine or thymine separates BB' sequence from the CC' sequence, and both of those inverted repeats contain either a triadenine or trithymine base at their centres, which exist as short hairpin loops at the end of each 'arm' of the T structure. Note that although the 'T' shape is more iconic, a 'Y' shape is predicted *in silico* and more likely accurate¹.

In this thesis, the 'hammerhead' will refer specifically to the terminal 124nt (or bp) of the AAV2 genome, A-to-A' inclusive. If not clear in context, 'hammerhead sequence' will refer specifically to their primary ssDNA or dsDNA sequence while 'hammerhead' refers to the self-annealed structure.

2.4.1.2 | Your left or my left?

By convention, a DNA sequence is depicted with the 5' terminus on the left and the 3' terminus on the right. If the DNA is double-stranded, the uppermost strand will have this orientation.

The leftmost element depicted in an ssAAV genome must therefore be the 3'ITR. This positions the 'uppermost' strand of the genome's dsDNA regions in the 5'-to-3' orientation and allows second-strand synthesis to be depicted in the conventional left-to-right direction. The 5'-to-3' convention and the ssAAV genome are so far, perhaps begrudgingly, in harmony. Admittedly disharmonious however are the apostrophes used to distinguish the upstream and downstream components of an inverted repeat. Outside AAV research, the apostrophe denotes the downstream (3'-most) reverse complementary sequence. For example, an inverted repeat consisting of the sequence x and its reverse complement x' has the implied structure 5'- x ... x' -3'. At some formative time in AAV genetics, an apostrophe system was implemented flowing from left-to-right rather than upstream-to-downstream, meaning that AA' in fact denotes the sequence 5'- A' ... A -3'. Though not officiated, this tendency is adhered to for the sake of consistency in the literature² and in this thesis. Importantly, an outward-to-inward designation (which does tend to occur) should be strictly avoided, since it results in the same annotation being used to denote two different 5'-to-3' base sequences. It is far preferable that any sequence labelled A' (for example) has the same 5'-to-3' base sequence, whatever its position or orientation on the page.

The (r)AAV genome also exists in numerous possible double-stranded forms including chromosomal integrations and recombinant cloning plasmids, which have no objective 5' or 3' end. In dsAAV genomes, the sense of an important gene (viral or recombinant) usually decides the depicted orientation of the dsDNA. The ITR sequence upstream from the start codon is labelled as the 5'ITR, and the one downstream from the stop codon is labelled the 3'ITR. Regrettably, this results in 5'-terminal ssITRs downstream that may or may not map to the dsDNA locus labelled "5'ITR-" (and so too for 3'-terminal ITRs). Thankfully this source of confusion does not arise in this thesis.

2.4.1.3 | What the flop?

The ITR hammerhead sequences, unlike the single-stranded portion of the genome, possess no sense. This is not to say they don't exhibit distinct forward and reverse complementarity. Nor is it because they contain no translated codons. More intriguingly, it is because the forward or reverse complementarity of either hammerhead does not correlate with the strand it is attached to³. Interestingly, the rolling hairpin replication mechanism seen among homotelemic parvoviruses⁴ causes the single-stranded genome and its terminal hammerheads to alternate

complementarity out of phase with one another. Those curious can note in Figure 5.4.1 that the parent and daughter strands receive a mix of template and *de novo* synthesised DNA, the boundary to which is the parental trs. Both molecules depart with one *de novo* synthesised 3'ITR and one pre-existing 5'ITR. The 3'ITRs both switch complementarity, while the 5'-to-3' nucleobase sequence of both 5'ITRs goes unchanged. Over three rounds of unit-length genome duplication (dimers notwithstanding), all eight ($= 2^3$) iterations of the (r)AAV genome sequence will occur.

Observed in 1977⁵ and confirmed in 1980⁶, the two apparent orientations of the ITR hammerhead were whimsically termed 'flip' and 'flop'. Intriguingly, the BB' and CC' sequences appear to have switched positions, though recombination is not necessary to make sense of this. It is a rare but simple consequence of the nested inverted repeat structure; the A and A' sequences have also 'switched positions' from template strand to *de novo* strand, but the result is indistinguishable. The inversion of unpaired As and Ts quietly signifies that the BB'-CC' 'switch' is the expected result of second-strand synthesis.

Importantly, all (full-length) flip hammerheads have identical 5'-to-3' sequences regardless of locus, genome sense, or where in space they are facing. The same is true for flop. By convention, the flip AAV2 ITR has its BB' sequence in the 3'-most position, and a flop AAV2 ITR has its CC' sequence in the 3'-most position. This convention was established in 1980 by authors Lusby, Fife & Berns⁶, who first characterised the dual ITR configurations. Given the painstaking technique and adroit logic demonstrated by these and other authors⁵, we today might find it all the more reassuring that two of these authors mislabelled their 5' flop ITR as 'flip' in the first published AAV2 genome sequence only three years later⁷. A simpler rule is also effective: the single unpaired base at the very centre of the hammerhead is T in a flip ITR and A in a flop ITR. Importantly, the unpaired nucleotides allow us to identify the ITR orientation even after a BB' or CC' deletion has occurred (c.f. Tran et al. 2022⁸).

How long is an AAV2 ITR?

The ITR2 sequence was first published in 1980⁶ three years in advance of the complete AAV2 genome sequence. For more than a decade thereafter, terminal ssAAV restriction fragments with heterogenous lengths led researchers to theorise that the terminal resolution site (where a nick at the A-D' junction separates parent and daughter strands during viral replication) was indefinite by nature⁹. In 1990, Im & Muzyczka¹⁰ observed an exact site of terminal resolution

where it is now indicated in all sources consulted, across one base and to the left (on the page) from the presumed terminal adenine of the 3' ITR (Fig. 5.4.3). The trs defines the 5' terminus of the daughter genome, and by extension, the *de novo* 3' terminus of the parent strand. If the currently accepted trs location is accurate, then Im & Muzyczka unwittingly reduced the putative length of an AAV2 ITR by one nucleotide, from 145nt to 144nt. To my knowledge, the adventitious adenine has yet receive comment – including in a 2019 review styled as a “user’s guide” to the AAV ITRs², which I myself co-authored. This said, publications presenting the more logical terminal sequence would imply that it hasn’t gone unnoticed¹¹. The rAAV research community should perhaps consider a renewed effort to confirm the location of trs, since its currently accepted location would imply that the documented positions of all bases in any (r)AAV reference sequence containing wild-type AAV2 ITRs⁷, worldwide, are incorrect by one base.

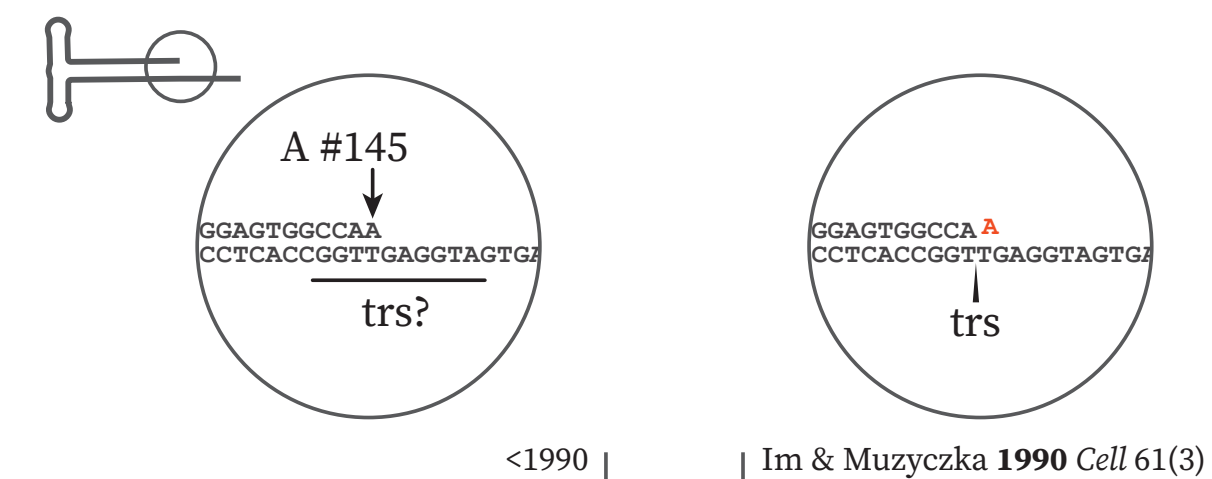


Figure 2.4.3 The currently accepted trs site indicates an ITR length of 144nt | The terminal resolution site (trs) is the location of a single-stranded nick that separates parent and daughter strands in rolling hairpin replication. It defines the boundaries of the A' and D sequences, and 3' terminal nucleotide of the AAV genome. At the the ITR sequence was originally obtained, the existence of a specific location of terminal resolution was questionable. A specific location for the trs was observed in 1990 and soon accepted in AAV and ITR reference sequences. This site was located between the TpT dinucleotide complementary to the terminal ApA originally reported; this directly indicated that the terminal bases of the ITR were CpCpA-OH, and not CpCpApA-OH. While the trs site was accepted, the incompatibility of the terminal ApA went unnoticed has persisted in most sources since. The definite location observed by Im & Muzyczka and the indefinite locations observed in prior experiments have not been reconciled.

Appendix H Explainer: DNA topology and the structural dynamics of double-stranded ITR sequences

Inverted repeats potentiate intra-strand base pairing. In ssDNA, with no duplex structure to compete with, intra-strand base pairing is probable. For a dsDNA molecule, the tendency for nonduplex structures to arise is strongly determined by its topology – more directly how ‘wound’ it is. The biophysical principles behind this have been uncontentious since the 1980s, though surprisingly it is far from common knowledge in the rAAV vector field. A grasp of this theory is greatly prescient to AAV researchers, given the diverse array of structures that ITR-containing sequences will exhibit both in the viral lifecycle and the recombinant vector system. Because of its global relevance to ITR research, this section will explain the concept of topological constraint and its immediate relevance to ITR secondary structure. I have identified only one passing mention of cruciform structures in supercoiled rAAV plasmids by Samulski and colleagues from 1982¹², a time in which the existence of cruciform DNA in living systems was still in question, and then-recent observations of cruciform DNA in supercoiled plasmids¹³ lent credence to certain models proposed by the authors. I am aware of no other sources that acknowledge this connection, and none appear to exist that contemplate the deterministic relationship between topology and the secondary structure of AAV ITRs, including other publications where it has proven thematically relevant^{14,15}.

2.4.2.1 | The topological model of dsDNA

The physical aspects of this model constitute ‘common knowledge’ and have been summarised in consultation with peer-reviewed primary and secondary sources authored by Clauvelin *et al.*¹⁶, Mirkin *et al.*¹⁷ and Vologodskii¹⁸, with additional sources cited in context. Though avoiding a mathematical representation where possible, term I have found unavoidable is ‘*linking number*’, which refers to the number of times from end to end that one strand of the dsDNA helix encircles the other. This is distinct (though not always numerically different) from the number of helical periods. To define the linking number formally, Mirkin suggests that we “assume that one DNA strand is the edge of an imaginary surface and count the number of times that the other DNA strand crosses this surface”. If this feels too abstract, I also offer an analogy: if you were to take a ride on a rollercoaster whose rails were the strands of a DNA double helix, the linking number is the number of times your feet would swing over your head.

Though, topologists may be more amenable to ‘the number of times your head and feet switch positions’ since the ground is as good a reference point as any.

In classical B-DNA, the double-helix tends to turn 100 times per 1050 base pairs but this number is not static. The periodicity of the helix is susceptible to a variety of dynamic lengthening and shortening forces that affect the stacking angle between adjacent base pairs^{19,20}. In nicked or linear dsDNA, the helix is free to cooperate with these forces by rotating at least one open ssDNA terminus around the opposite strand. The linking number is free to vary dynamically and the molecule is perpetually ‘relaxed’. In intact circular dsDNA, the termini are covalently sealed and the linking number cannot be changed. This condition is called *topological constraint*. Unless the period of the helix is exactly equal to that in the relaxed state, shortening and lengthening forces will impart torsional stress onto the molecule that cannot be dispersed. If dsDNA is not relaxed, it is said to be ‘supercoiled’. These torsional stresses possess directionality relative to that of the B-DNA double helix. If the helix has more right-handed turns than the relaxed molecule would, it is positively supercoiled (or ‘overwound’). If it has fewer right-handed turns than the relaxed molecule would, it is negatively supercoiled (or ‘underwound’). Note that left-handed turns are not a feature of B-DNA.

Topologically, each ‘link’ contributing to the linking number manifests as either a twist in the double helix or a 360° bend in the dsDNA polymer referred to as writhe²¹. Returning to the rollercoaster analogy, a twist would lead your feet to swing over your head while your body hurdles in a constant direction. Writhe, by contrast, completes the head-foot inversion via a curved detour carrying you in the reverse direction before bending back onto the original trajectory (Figure 5.4.4). (Vologodskii refers to this as a ‘higher order helix’¹⁸). The equivalence between twist and writhe is so fundamental that writhe in the negative direction increases twist in the positive direction and vice versa, such that the linking number remains the same. (This is what is referred to by the rule *linking number = twist + writhe*). Of course, topologically unconstrained DNA relaxes into a torsionally neutral double helix, rather than a flat ribbon, stabilised by non-topological forces. The linking number of topologically unconstrained dsDNA is defined simply by the periodicity of the relaxed helix. However, when topologically constrained, the number of twists may be too few (or too many) compared to the relaxed state.

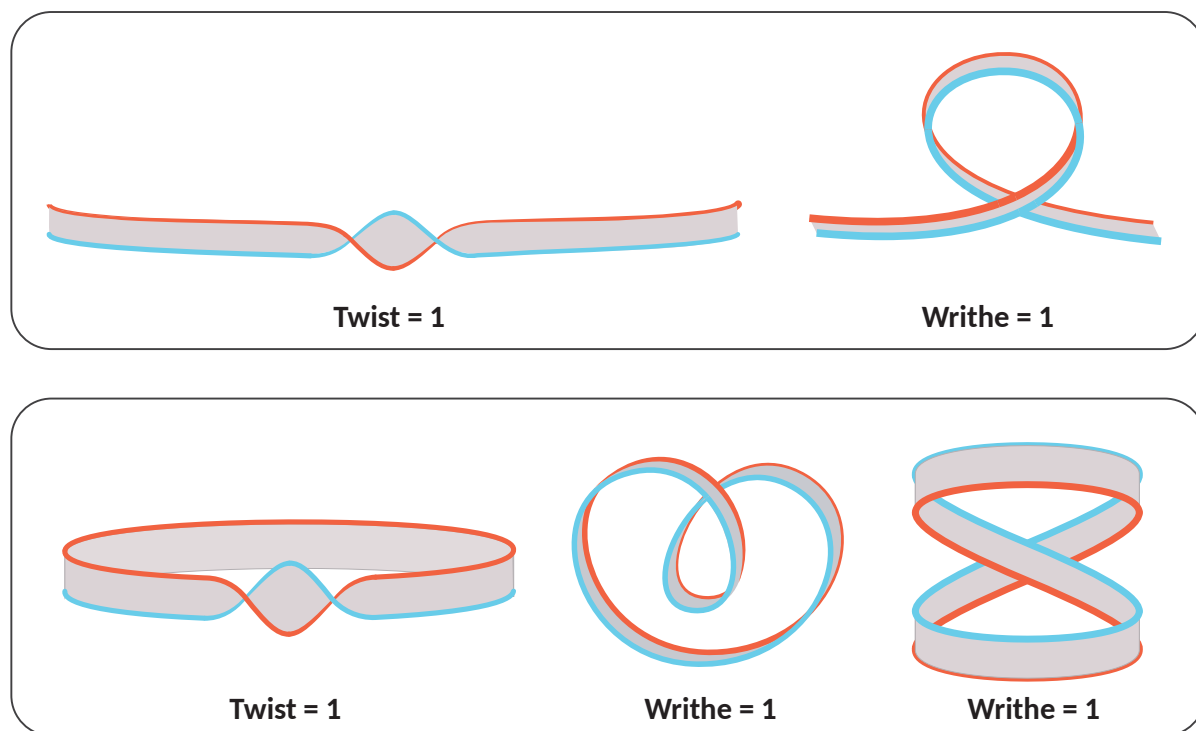


Figure 2.4.4 The equivalence of twist and writhe to topological linking number Each instance that one dsDNA strand passes around the other manifests either as 'twist' or 'writhe'. In the main text, it has been highlighted how a rollercoaster could invert your feet and head and back in two ways, which this figure helps to illustrate, using a flat ribbon model with a linking number of one. The way analogous to 'twist' is an rotational distortion that would allow your body as a whole to travel in a constant direction. The way analogous to writhe is for the track to bend backward and continue bending until you your body is travelling back in the original direction. Note that in each illustration of twist or writhe, the blue strand encircles the red strand once. In the lower panel, the 'molecule' is circular. The rightmost conformation depicts a writhen 'loop' that is distributed throughout the molecule, such that it takes the whole molecule to complete.

The resulting torsional stresses drive the polymer, if unrestricted, to bend by -360° per 'missing' twist, which increases the number of twists to that of relaxed DNA without altering the linking number. Of course, without a specific reason (such as wrapping around a protein) writhe is diffused throughout DNA polymer, not confined to a local 360° loop. In circular DNA, the bend effectively takes the whole DNA ring to complete, giving rise to the familiar nodular structure of intact supercoiled plasmid DNA.

Relaxed B-DNA, while torsionally neutral, is homeostatically non-neutral. In all observed cellular life, DNA is maintained in a negatively supercoiled state by topoisomerase activity²². Negative superhelicity profoundly influences DNA biology through its effects on such things as local unwinding kinetics, protein-groove interactions, histone interactions, and the bending elasticity of the dsDNA polymer^{23–25}. Although eukaryotic chromosomes are not covalently

closed, protein binding, membrane tethering and intra-macromolecular friction all exert topological constraint and prevent negative superhelical torsion from dispersing²⁶.

2.4.2.2 | The cruciform conformer of inverted repeats is unstable in topologically unconstrained dsDNA

Negative superhelical torsion is required for four-way junctions to arise non-transiently from dsDNA inverted repeats²⁷⁻³¹. Cruciform extrusion, like negative writhe, is an outlet for negative superhelical tension in topologically constrained B-DNA, and the torque in the molecule provides the free energy required to nucleate the transition. The tendency of inverted repeats toward a linear or cruciform conformation is not as even-handed as a Watson-Crick pairing model alone would suggest; the unpaired bases are unfavourable but more so is the exposure of water to the nucleotides' hydrophobic surfaces around the hairpins loops and four-way junction. This makes the linear duplex the preferred state by default. In the negatively supercoiled state, the linear and cruciform conformers exist in a dynamic equilibrium, driven

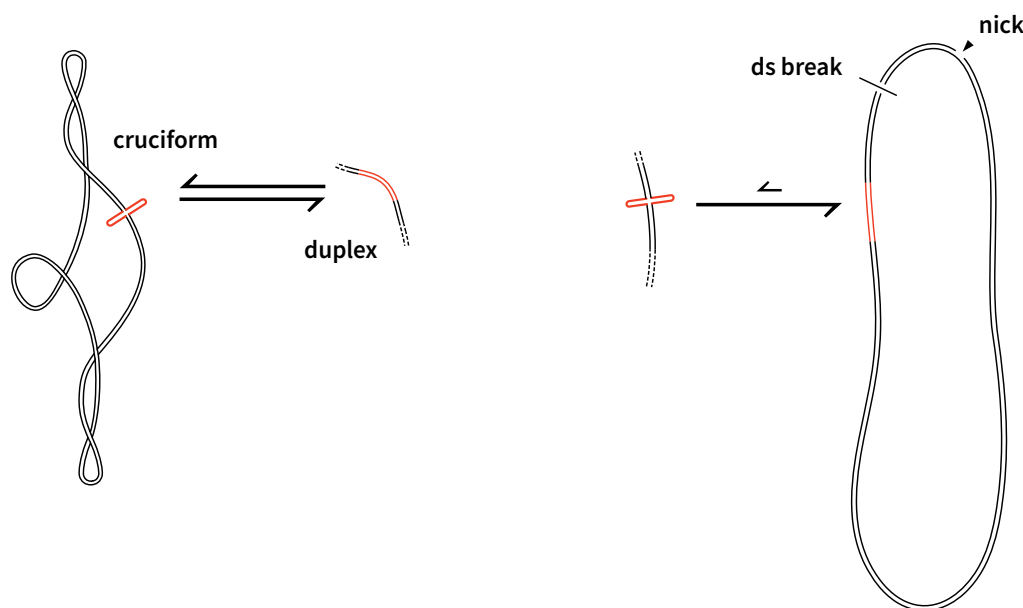


Figure 2.4.5 Topological constraint is a requirement for cruciform conformers to occur in dsDNA | 'Topological constraint' refers to the inability of a dsDNA molecule to change the number of times one strand passes around the other - formally termed its 'linking number'. In the case of closed circular DNA, this is because the linking number is covalently fixed. The drive to change the linking number is a disagreement between the relaxed stacking angle of bases in the dsDNA ladder and the number of turns in the helix, which arise from fluctuations in its physiochemical environmental factors or via the activity of DNA topoisomerases. In topologically constrained DNA, forces that would otherwise drive the linking number to change are indispensible and give rise to torsional stress. Cruciform extrusion is an outlet for negative superhelical torsion, and the torsion also provides the free energy needed to nucleate the process. The duplex conformation is strongly favoured in the absence of this torsion due to hydrophobic interactions otherwise invited around the hairpin loops and fourway junction. The strength of these factors is such that dynamic equilibrium between the duplex and cruciform conformers is observed in spite of the relaxing effect of cruciform extrusion.

to cruciform extrusion by negative superhelical tension and to the linear duplex by non-topological factors (reviewed in some depth by Bickard and colleagues³²) (Fig. 2.4.5).

A small demonstration original to this thesis (Fig. 5.4.6) shows that plasmid ITRs concur with this established model. An rAAV plasmid bearing two ITR2 sequences was treated with T7 endonuclease I (T7) for one hour. Then, the restriction endonuclease BsrGI (NEB) was added to the reaction and allowed to incubate for an additional hour before analysing the products via agarose gel electrophoresis. BsrGI generates distinct restriction products depending on whether the 3'ITR, 5'ITR, both ITRs, or neither ITR was cleaved by T7. Only uncleaved or once-cleaved plasmids resulted in obvious BsrGI products. This illustrates that once the

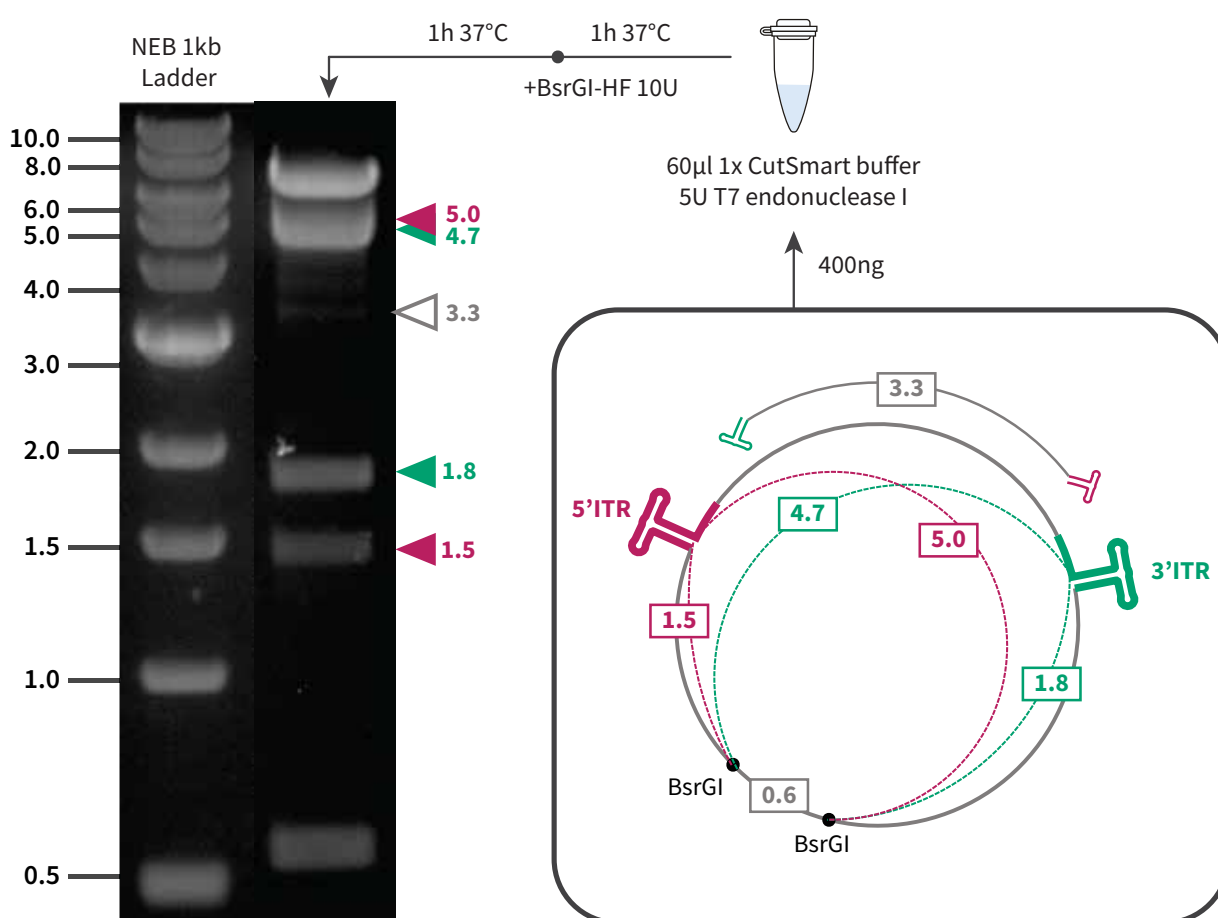


Figure 2.4.6 Effects of topology on the secondary structure of ITR sequences in double-stranded DNA |

T7 endonuclease I is a structure-specific nuclease that cleaves double-stranded DNA diagonally across four-way junctions. This enzyme was used to test for the presence of cruciform structures in the ITR loci. Cleavage in both these loci, followed by digestion with the restriction enzyme BsrGI would result in a 3.3kbp product, which was poorly visualisable in the EtBR-stained agarose gel on the left. The dominant band pattern consistent of BsrGI products that had previously been cleaved in one ITR, neither ITR, but not both ITRs. This result exemplifies the understanding that cruciform structures, such as those formed by dsITRs, are not stably extruded or maintained in the absence of negative superhelical forces, which are rapidly dispersed as soon as the molecule is nicked or cleaved.

plasmid has been topologically deconstrained by a single ITR cleavage event, the uncleaved ITR sequence is poorly susceptible to T7 activity. Cruciform extrusion and relaxation are complex, multi-stage processes whose kinetics are greatly affected by the primary structure and its environment^{29,33,34}, meaning we cannot assume the exact distribution of restriction products in Figure 5.4.6 would be recapitulated in an intracellular environment. This said, context-independent biophysical realities are strongly reflected: torsional forces lead topologically constrained dsDNA to become supercoiled, cruciform extrusion is driven by negative superhelicity, and non-topological factors oppose spontaneous extrusion and maintenance of cruciform structures in relaxed dsDNA.

This simple demonstration therefore compels us to recognise the relationship between ITR structure and genome topology in our models of (r)AAV biology, and I will close these two ‘explainer’ subsections by drawing attention to some examples. It implies, for example, that secondary-structure-specific events involving a dsITR (such as latent genome rescue) are unlikely to follow an initial cleavage or nicking event without additional mechanisms to stabilise or recover the cruciform (c.f. Ward, Elias & Linden 2003³⁵). It excludes the possibility of spontaneous hammerhead reformation from models of genome rescue or replication. The model also predicts that non-ITR hairpin sequences could act as alternative targets for cruciform resolvase activity in plasmid rAAV rescue^{36–38}. In dsDNA the terminal resolution site (located directly at the A'D junction) is potentially inaccessible as a Holliday structure, making it possible that access is topologically regulated. Due to the implications for rAAV genome rescue, the proportion of supercoiled and open circular copies in a plasmid preparation (a routine quality attribute) may have special significance where plasmids containing ITR sequences are intended for use in rAAV production.

The principles of DNA topology also have broad implications for the (r)AAV genome in transduced cells. Most obviously it implies that for the exogenous genome to become negatively supercoiled, it must first assume a topologically constrained dsDNA structure. Generically speaking, the significance of negative supercoiling in eukaryotic genetics is difficult to understate. Prominently, it is a central determinant of histone association and thereby nucleosome development and chromatin structure^{23,39}. There is also a multitude of notable cruciform-specific interactions that have been reviewed in detail by Brázda and colleagues²⁵. Also notable are the dynamic and closely interrelated shifts in superhelicity, nucleosome structure and secondary structure that occur when histone-associated DNA is replicated or

transcribed^{40–42}. This creates an intriguing constellation between transcription of stable rAAV episomes, their helicity, their nucleosome structure, and the conformation of their ITRs. Finally, I note that DNA topology highlights a feature specific to the ITR hammerhead structure: the ‘intra-hairpin’ structures (i.e. the BB’ and CC’ hairpins) can retain their nonduplex conformation after the four-way junction is cleaved. While this does not solve the question of why the hammerhead structure should exist at all, it is one of few suggestions that could not also be satisfied by a simple U-shaped hairpin.

References to Appendices G &H

1. Laugel, M., Lecomte, E., Ayuso, E., Adjali, O., Mével, M. & Penaud-Budloo, M. in *Recent Adv. Canine Med.* (ed. Fonseca-Alves, D. C. E.) (IntechOpen, 2022).
2. Wilmott, P., Lisowski, L., Alexander, I. E. & Logan, G. J. A user’s guide to the Inverted Terminal Repeats of Adeno-Associated Virus. *Hum. Gene Ther. Methods* **30**, (2019).
3. Lusby, E., Bohenzky, R. & Berns, K. I. Inverted terminal repetition in adeno-associated virus DNA: independence of the orientation at either end of the genome. *J. Virol.* **37**, 1083–6 (1981).
4. Ward, P. in *Parvoviruses* (eds. Kerr, J. R., Cotmore, S. F., Bloom, M. E., Linden, R. M. & Parrish, C. R.) 190–211 (Hodder Arnold, 2005).
5. Spear, I. S., Fife, K. H., Hauswirth, W. W., Jones, C. J. & Berns, K. I. Evidence for two nucleotide sequence orientations within the terminal repetition of adeno-associated virus DNA. *J. Virol.* **24**, 627–34 (1977).
6. Lusby, E., Fife, K. H. & Berns, K. I. Nucleotide sequence of the inverted terminal repetition in adeno-associated virus DNA. *J. Virol.* **34**, 402–9 (1980).
7. Srivastava, A., Lusby, E. W. & Berns, K. I. Nucleotide sequence and organization of the adeno-associated virus 2 genome. *J. Virol.* **45**, 555–64 (1983).
8. Tran, N. T., Lecomte, E., Saleun, S., Namkung, S., Robin, C., Weber, K., Devine, E., Blouin, V., Adjali, O., Ayuso, E., Gao, G., Penaud-Budloo, M. & Tai, P. W. L. Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity. *Hum. Gene Ther.* **33**, 371–388 (2022).

9. Fife, K. H., Berns, K. I. & Murray, K. Structure and nucleotide sequence of the terminal regions of adeno-associated virus DNA. *Virology* **78**, 475–487 (1977).
10. Im, D. S. & Muzyczka, N. The AAV origin binding protein Rep68 is an ATP-dependent site-specific endonuclease with DNA helicase activity. *Cell* **61**, 447–57 (1990).
11. Maurer, A. C. & Weitzman, M. D. Adeno-associated virus genome interactions important for vector production and transduction. *Hum. Gene Ther.* **31**, 499–511 (2020).
12. Samulski, R. J., Berns, K. I., Tan, M. & Muzyczka, N. Cloning of adeno-associated virus into pBR322: Rescue of intact virus from the recombinant plasmid in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 2077–2081 (1982).
13. Lilley, D. M. J. In vivo consequences of plasmid topology. *Nature* **292**, 380–382 (1981).
14. Lamartina, S., Ciliberto, G. & Toniatti, C. Selective cleavage of AAVS1 substrates by the adeno-associated virus type 2 Rep68 protein is dependent on topological and sequence constraints. *J. Virol.* **74**, 8831–8842 (2000).
15. Ashktorab, H. & Srivastava, A. Identification of nuclear proteins that specifically interact with adeno-associated virus type 2 inverted terminal repeat hairpin DNA. *J. Virol.* **63**, 3034–9 (1989).
16. Clauvelin, N., Olson, W. K. & Tobias, I. Characterization of the geometry and topology of DNA pictured as a discrete collection of atoms. *J. Chem. Theory Comput.* **8**, 1092–1107 (2012).
17. Mirkin, S. M. in *eLS* 1–11 (John Wiley & Sons Ltd, 2001).
18. Vologodskii, A. V. *Topology and Physics of Circular DNA*. (CRC Press, 1992).
19. Kriegel, F., Matek, C., Dršata, T., Kulenkampff, K., Tschirpke, S., Zacharias, M., Lankaš, F. & Lipfert, J. The temperature dependence of the helical twist of DNA. *Nucleic Acids Res.* **46**, 7998–8009 (2018).
20. Lipfert, J., Doniach, S., Das, R. & Herschlag, D. Understanding nucleic acid–ion interactions. *Annu. Rev. Biochem.* **83**, 813–841 (2014).
21. Vinograd, J. & Lebowitz, J. Physical and topological properties of circular DNA. *J. Gen.*

- Physiol.* **49**, 103–125 (1966).
22. Duprey, A. & Groisman, E. A. The regulation of DNA supercoiling across evolution. *Protein Sci.* **30**, 2042–2056 (2021).
 23. Corless, S. & Gilbert, N. Effects of DNA supercoiling on chromatin architecture. *Biophys. Rev.* **8**, 245–258 (2016).
 24. Cruz-León, S., Vanderlinden, W., Müller, P., Forster, T., Staudt, G., Lin, Y.-Y., Lipfert, J. & Schwierz, N. Twisting DNA by salt. *Nucleic Acids Res.* **50**, 5726–5738 (2022).
 25. Brázda, V., Laister, R. C., Jagelská, E. B. & Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **12**, 33 (2011).
 26. Corless, S. & Gilbert, N. Investigating DNA supercoiling in eukaryotic genomes. *Brief. Funct. Genomics* **16**, 379–389 (2017).
 27. Benham, C. J. Stable cruciform formation at inverted repeat sequences in supercoiled DNA. *Biopolymers* **21**, 679–696 (1982).
 28. Vologodskii, A. V. & Frank-Kamenetskii, M. D. The relaxation time for a cruciform structure in superhelical DNA. *FEBS Lett.* **160**, 173–176 (1983).
 29. Zhabinskaya, D. & Benham, C. J. Competitive superhelical transitions involving cruciform extrusion. *Nucleic Acids Res.* **41**, 9610–9621 (2013).
 30. Benham, C. J., Savitt, A. G. & Bauer, W. R. Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: Complete determination of energetics using a statistical mechanical model. *J. Mol. Biol.* **316**, 563–581 (2002).
 31. Sinden, R. R. & Pettijohn, D. E. Cruciform transitions in DNA. *J. Biol. Chem.* **259**, 6593–6600 (1984).
 32. David, B., Céline, L., Zeynep, B. & Didier, M. FIXME!! Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol. Mol. Biol. Rev.* **74**, 570–588 (2010).
 33. Shlyakhtenko, L. S., Potaman, V. N., Sinden, R. R. & Lyubchenko, Y. L. Structure and

- dynamics of supercoil-stabilized DNA cruciforms. *J. Mol. Biol.* **280**, 61–72 (1998).
34. Eichman, B. F., Vargason, J. M., Mooers, B. H. & Ho, P. S. The Holliday junction in an inverted repeat DNA sequence: sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3971–6 (2000).
 35. Ward, P., Elias, P. & Linden, R. M. Rescue of the adeno-associated virus genome from a plasmid vector: evidence for rescue by replication. *J. Virol.* **77**, 11480–11490 (2003).
 36. Xie, J., Mao, Q., Tai, P. W. L., He, R., Ai, J., Su, Q., Zhu, Y., Ma, H., Li, J., Gong, S., Wang, D., Gao, Z., Li, M., Zhong, L., Zhou, H. & Gao, G. Short DNA hairpins compromise recombinant adeno-associated virus genome homogeneity. *Mol. Ther.* **25**, 1363–1374 (2017).
 37. Tran, N. T., Heiner, C., Weber, K., Weiland, M., Wilmot, D., Xie, J., Wang, D., Brown, A., Manokaran, S., Su, Q., Zapp, M. L., Gao, G. & Tai, P. W. L. AAV-Genome Population Sequencing of Vectors Packaging CRISPR Components Reveals Design-Influenced Heterogeneity. *Mol. Ther. - Methods Clin. Dev.* **18**, 639–651 (2020).
 38. Xu, M., Li, J., Xie, J., He, R., Su, Q., Gao, G. & Tai, P. W. L. High-throughput quantification of in vivo adeno-associated virus transduction with barcoded non-coding RNAs. *Hum. Gene Ther.* **30**, 946–956 (2019).
 39. Gilbert, N. & Allan, J. Supercoiling in DNA and chromatin. *Curr. Opin. Genet. Dev.* **25**, 15–21 (2014).
 40. Mani, P., Yadav, V. K., Das, S. K. & Chowdhury, S. Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS One* **4**, e4399 (2009).
 41. Dayn, A., Malkhosyan, S. & Mirkin, S. M. Transcriptionally driven cruciform formation *in vivo*. *Nucleic Acids Res.* **20**, 5991–5997 (1992).
 42. Luchnik, A. N. DNA conformational transitions induced by supercoiling control transcription in chromatin. *Gene Regul. Syst. Bio.* **8**, GRSB.S13756 (2014).

Appendix I ITR Sanger sequencing traces

The clonally dominant ITR sequences of plasmid isolates used in Chapter 6 were analysed using a specialised Sanger sequencing service provided by GENEWIZ LLC. (NJ). The 3'ITR deletions shown in cartoon form in Chapter 6 are informed by the alignments displayed on the following page. Sequencing of the 5'ITR was less of a technical success, possibly due to being primed too close to the target sequence (~200bp distance as opposed to ~400bp for the 3'ITR), which initially raised concerns about insertions of single adenine bases based on the alignment. While this would not compromise the experiment, the analysis would be less 'neat' as a result. However, two factors led me to conclude that this was most likely erroneous, aside from the overall poorer quality of the trace. The green peak indicating the adenine is does not overlap with alternative peaks, meaning that at face value, it is not characteristic of a heterogeneous sequence. However, the Δ XmaI clone that had been derived from this plasmid contained the wild-type 5'ITR sequence, which all but guarantees that it existed in the parent plasmid. Secondly, the trace overall showed a decline in quality surrounding the hairpin junctions, detracting from the assumption that the adenine insertion was valid. In the 3'ITR of the WT prep, heterogeneous ITR lengths led to dual peaks downstream from the ITR (from the direction of the primer binding site). The 5'ITR, in contrast, was 'clean' downstream from the ITR, indicating a lack of heterogeneity.

A' **C** **C** **B'** **B** **A**

Flu STIR2

S' D Sequence

Figure 2: Schematic representation of the 2D V-sequence. The diagram shows a 2D grid of cells with a color gradient from yellow to red. The top row is labeled "2 V-sequence" and the bottom row is labeled "160° 211°C". The grid is divided into four quadrants by a vertical line. The left quadrant is labeled "5log A" and the right quadrant is labeled "5log A". The top row is labeled "5log A" and the bottom row is labeled "5log A". The grid is composed of cells with different colors: yellow, orange, green, and red. The cells are arranged in a grid that is 160° wide and 211° high. The cells are labeled with "A" and "B" in the top row and "C" and "C" in the bottom row. The cells are arranged in a grid that is 160° wide and 211° high. The cells are labeled with "A" and "B" in the top row and "C" and "C" in the bottom row.

