Towards a Wheat Phenome Atlas and a Phenome Atlas Toolbox: What are they? What progress?

DeLacy IH[1,2], Dieters MJ[1], Crossa J[3], Godwin ID[1], Arief V[1], Batley J[1,2] Davenport G[3], Dreisigacker S[3], Duveiller E[3], Edwards D[1,2], Huttner E[5,6], Lambrides CJ[1], Manes Y[3], Payne T[3], Singh RP[3], Warburton M[3], Wenzl P[5,6], Kilian A[5,6], McLaren G[4], Braun H-J[3], Crouch J[3], Ortiz R[3], Basford KE[1,2]

[1]School of Land Crop and Food Sciences, The University of Queensland, Brisbane, [2]Australian Centre for Plant Functional Genomics (ACPFG), [3]International Maize and Wheat Improvement Center (CIMMYT), [4]International Rice Research Institute (IRRI), [5]Diversity Arrays Technology P/L Canberra, [6]Triticarte P/L, Canberra

ABSTRACT

A Phenome Map is a representation of all the regions of a genome that influence heritable phenotypic variation for a trait, and a Phenome Atlas consists of the integration of all available phenome maps with a description of the methodologies that were used to produce the maps. A Phenome Atlas Toolbox is a set of tools and methodologies for producing the Phenome Atlas. The Wheat Phenome Atlas (WPA) will be an integration of phenotypic data (17 million data points for 80 traits from 10,000 international field trials collected during more than 40 years) generated by CIMMYT and partners on approximately 13,000 wheat lines (for which pedigrees are known) with greater than 26 million DArT marker data points obtained by genotyping these lines. To generate this amount of phenotypic data would cost over $500 million today.

WHAT IS A PHENOME ATLAS?

A Phenome Atlas enables a sophisticated integration of pedigree, phenotypic, marker and genomic information into a single knowledge system. A Phenome Map is a representation of all the regions of a genome that influence heritable phenotypic variation for a trait, that is a dense QTL map for the trait integrated with complete gene and marker maps. A Phenome Atlas consists of the integration of all available phenome maps with a description of the methodologies that were used to produce the maps

A Phenome Atlas Toolbox is a compilation of data, algorithms, methodologies and software required to produce a Phenome Atlas. The data for the toolbox is a compilation of pedigree, phenotype, marker and sequence information. Clearly much of this information, such as sequence information, will be made available by linkages to publicly available databases. The software is also a compilation of all available software, purpose built, open source or commercial that is required to produce a Phenome Atlas. Hence a Phenome Atlas Toolbox will be an evolving set of the best tools and methodologies for mining and analysing the vast data resources required for developing a phenome atlas.

WHEAT PHENOME ATLAS

Phenome mapping of a species for a given trait requires extensive phenotypic information for the target trait across all individuals in genealogically connected sets of family trees for which the full pedigree details are known. This is available from CIMMYT for nearly half a century of breeding outputs, covering up to 10 generations for many pedigree lineages. CIMMYT has collated extensive historical phenotypic and genealogical information on approximately 13,000 elite wheat breeding lines and seed has been conserved for all lines. The phenotypic data were derived from more than 40 years of international trials (since 1964). These coordinated trials have been conducted by National Agricultural Research Systems (NARS) partners in over 70 wheat producing countries, targeted to six world-wide agro-ecological zones. There are about 17,000,000 phenotypic data points for over 80 economically important traits across the 13,000 wheat lines evaluated in more than 10,000 field trials. The full pedigrees and selection histories of all entries are known and the data cover yield, agronomic, biotic resistance and quality traits. A conservative estimate puts the value of reproducing these pedigree and phenotypic data at over US$500 million. This unique combination of information-comprehensive germplasm will allow wheat to become the first public sector crop species for which a Phenome Atlas can be created.

Phenome mapping also requires extensive genome-wide marker data on all the genotypes which have been phenotyped. Seed of all lines phenotyped in the International Nurseries is available in the CIMMYT wheat germplasm bank. DNA can be extracted and genotyped. DArT markers; an affordable high throughput marker system capable of delivering a genome wide coverage for wheat, which became available in late 2005. Genotyping the full CIMMYT wheat set with these markers is estimated to cost around US$750,000 which is only approximately 0.15% of the phenotyping cost. This would produce approximately 2,000 data values for each of the wheat lines, giving around 26,000,000 marker data values in total.

ENABLING TECHNOLOGIES

There have been three important technologies developed over the last 20 years which have matured sufficiently to enable the construction of a WPA. These are (1) ASREM: a software program which enables the analysis of the large set of phenotypic data available using modern general mixed model statistical approaches, (2) ICIS: a general databasing platform which enables the collation and storage of all data required and (3) A high throughput marker system: suitable for genotyping the large number of genotypes.

ASREML is a fully developed software platform which can be used for performing a general mixed model analysis on the large amount of phenotypic data available from the CIMMYT international nurseries and the high through put genotyping with DArT markers.

ICIS, the International Crop Information System (Fox and Skovmand 1996) has become the open-source standard for information management systems for plant breeding data. It offers greatly enhanced functionality for integrating phenotype and pedigree data with genotype data and information on genes, markers, QTLs and maps derived from them.

DArT markers are an affordable high-density marker system that generates genome-wide fingerprints (Jaccoud *et al.* 2001). Although this technology does not require DNA sequence information, it generates sequence ready markers (clones). DArT has already been used for mapping and association studies in hexaploid wheat (Wenzl *et al.* 2004; Akbari *et al.* 2006).

METHODOLOGY

A sample of seed of every line that has been tested in the International Nurseries is deposited in the CIMMYT wheat germplasm bank and is publicly available. Because the pedigrees of these interconnected and related families are known and cover up to 10 generations they are the equivalent of a Mega-Recombinant Inbred Line population (MRIL). We believe that this MRIL with the associated pedigree and phenotypic information is unique in terms of its scale and depth. With the advent of an affordable high density marker system such as that provided by DArT it is now possible to fully genotype (with more than 2,000 markers per genotype) all these lines. We envisage that other genotyping platforms (such as SNP) will become routinely available in the future and thus will be added to the information pool for improving the phenome atlas.

Given this information, a full functional description of the genome can be made for any trait by whole genome applications or by linking each segment of a chromosome passed down a family tree from different ancestors (or founders) to differences in the phenotypes of their descendents. Given that the pedigrees are known, each segment of the chromosome can be identified by high density markers tracked through the family trees. These segments can be linked to function as the ancestors (founders) and descendents have been phenotyped. The identification of segments of chromosomes passed from founders to descendents is crucial to this analysis and is dependent on discrimination of 'identity by descent (ibd)' from 'identity by state (ibs)' or 'identity by function (ibf)'. Being able to distinguish ibd from ibs is dependent on knowing pedigrees and to distinguish ibf from ibs and ibd requires knowing the phenotype in addition to the genotype.

Statistical methodologies for identifying the linkage between phenotype and genotype in this type of dataset are being developed in a number of research organisations across the world. These need to be collected, collated and compared in order to evaluate the underlying procedures and associated software. The methods and software need to (1) be packaged into a phenome map toolbox, (2) be scaled to handle the massive amount of data available, and (3) develop methods and software to update the atlas as new information becomes available.

PROGRESS THUS FAR

A study (Crossa *et al.* 2007) used DArT markers to find associations with resistance to stem rust, leaf rust, yellow rust, and powdery mildew, plus grain yield in five Elite Spring Wheat Yield Trials (ESWYT). Two linear mixed models using data from up to 60 environments for each ESWYT which modelled genotype by environment interactions were used to assess marker-trait associations (MTA). An integrated map containing 813 DArT markers and 831 other markers was constructed. Several linkage disequilibrium clusters bearing multiple hosts plant resistant genes were found. Most of the associated markers were found in genomic regions where previous reports had found genes or quantitative trait loci (QTL) influencing the same traits, providing an independent validation of this approach. Many new chromosome regions for disease resistance and grain yield were identified in the wheat genome.

Another study (Arief *et. al.* these proceedings) identified MTA by the joint analysis of phenotype data from a comprehensive set of field trials and dense DArT genome scans (1,447 polymorphic markers). MTA were identified for 21 traits (three rusts, grain yield, five agronomic characters, two quality traits, and 10 other foliar diseases) using data collected from the first 25 years of CIMMYT's ESWYT. MTA were identified for each trait using a t-test with a p-value of 0.001 to declare significance level. This approach identified numerous associations for each trait. The DArT genome scans were consistent across duplicated lines and enabled the identification of introgressed segments based on haplotypes.

Another study (Dreisigacker *et. al.* these proceedings) from a Linkage Disequilibrium (LD) analysis of the 25 ESWYT and a set of 160 synthetic hexaploid wheats (SHW) showed a variable pattern of LD among chromosomes. Different patterns of LD decay were observed for advanced wheat lines and SHW. LD dropped faster in SHW and a lower percentage of loci pairs in LD were detected demonstrating the usefulness of SHW for high-resolution association analyses.

IMMEDIATE PLANS

The International Spring Wheat Yield Nurseries (ISWYN) which included a wide range of world spring wheat lines was distributed from 1964 till 1994: 31 years in all. The ESWYT was first distributed in 1979, is still being circulated and contains mostly elite CIMMYT bred germplasm. Both are yield nurseries and between them cover a large proportion of the international wheat germplasm over the past half century: for example, in the early ISWYNs tall and stem rust susceptible lines were still being entered. Bread wheat is a global crop. To aide in directing decisions in wheat breeding programs, CIMMYT scientists have classified the world wheat environments into 12 agro-ecological zones, six of them for spring bread wheat, called mega-environments (MEs). For example, ME1 is low latitude, high input (irrigated and fertilised) deserts. ME2 is also low latitude but the wheat is rain grown. Since the mid 1980s CIMMYT has targeted these MEs with specialist nurseries which have tailored suites of adapted germplasm. One nursery, the Semi-Arid Wheat Yield Nursery (SAWYT) is targeted at ME4 which are water stressed environments. ME4, like ME2 are wheat growing regions in areas less than 40^{o} latitude but averages less than 500mm rainfall. Another nursery, the High Temperature Wheat Yield Nursery (HTWYT) targeted at ME5 which are heat stress environments.

In the next phase of the WPA development we plan to add the data from the SAWYT and HTWYT nurseries to the association analysis already conducted for the ESWYT. This will target different environments and germplasm than that targeted by the ESWYT analysis. The number of unique lines tested in the ESWYT (25 years), SAWYT (14 years) and HTWYT (12 years) are 686, 520 and 370 respectively, a total of 1480 unique lines of which 599 lines have been DArT genotyped,

As stated earlier, to distinguish ibd from ibs chromosome segments full pedigrees need to be genotyped, that is the genotypes of the test (descendents) lines and their parents (founders) are required. Associations with segments of the chromosome passed through generations of germplasm will discriminate between associations caused by linkage and those caused by other mechanisms. The number of parents for the lines tested in the ESWYT, SAWYT and HTWYT are 517, 434 and 255 respectively, giving 937 unique parents.

It is planned to map the remaining DArT markers and integrate this with other publicly available marker maps. Triticarte will also sequence all DArT markers. In addition, work is progressing on the visualization of the WPA and developing procedures to enable online WEB access.

CONCLUSIONS

The work reported so far shows that the pedigree information, phenotypic data from the CIMMYT international nurseries and the affordable high through put dominant DArT markers can be used to develop a Phenome Atlas.

REFERENCES

Akbari M et. al. (2006) Diversity Arrays Technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theoretical and Applied Genetics* 113: 1409-1420

Crossa J et al. (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 177, 1889-1913.

Fox PN, Skovmand B (1996) The International Crop Information System (ICIS) - Connects Genebank to Breeder to Farmer's Field. In 'Plant Adaptation and Crop Improvement'. (Eds M Cooper and GL Hammer) pp. 317-328. (CAB International: Wallingford, UK)

Gilmour, A. R., et. al. 2006 *ASReml User Guide Release 2.0*. VSN International Ltd., Hemel Hempstead, HP1 1ES, UK.

Jaccoud D et. al. (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research* 29: e25

Wenzl P et. al. (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. *Proceedings of the National Academy of Sciences of the USA* 101: 9915-9920.