Structure and organization of the wheat genome – the number of genes in the hexaploid wheat genome

Devos KM[1,2], Costa de Oliveira A[3,5], Xu X[1,2], Estill JC[2], Estep M[2], Jogi A[3], Morales M[4], Pinheiro J[3], San Miguel P[6] and Bennetzen JL[3]

[1]Dept. of Crop and Soil Sciences, University of Georgia, Athens, GA 30602, USA; [2]Dept. of Plant Biology, University of Georgia, Athens, GA 30602, USA; [3]Dept. of Genetics, University of Georgia, Athens, GA 30602, USA; [4]Dept. of Biochemistry & Molecular Biology, University of Georgia, Athens, GA 30602; [5]Plant Genomics and Breeding Center, Federal University of Pelotas, Pelotas, Brazil 96001-970; [6]Purdue Genomics Facility, Purdue University, West Lafayette, IN 47907

ABSTRACT

One hundred and nine BAC clones, representing 0.066% of the hexaploid wheat genome, were sequenced and annotated for gene content. Annotation was done manually or by manual curation of the output of the automated DAWG-PAWS annotation pipeline. Gene numbers were affected by subjective decisions taken by the annotators and also by sample size. A simulation study revealed that extrapolation of gene numbers for the entire wheat genome from less than 1 Mb of DNA led to estimates that could vary by more than 7-fold. The variation decreased with increasing sample size, but remained at 20% of the mean for our sample size of 109 BAC clones, corresponding to 11.1 Mb. A conservative estimate is that the wheat genome contains between 164,000 and 334,000 protein-encoding genes, including pseudogenes.

INTRODUCTION

Milestones in the genetic and genomic analyses of wheat over the past 20 years include the construction of genetic and deletion maps (e.g. Gale et al. 1995; Mickelson-Young et al. 1995; Nelson et al. 1995), the identification of comparative relationships with rice and other grass species (Kurata et al. 1994; Van Deynze et al. 1995; La Rota and Sorrells 2004), the development of some one million expressed sequenced tags (ESTs) (Lazo et al. 2004), the large scale mapping of these ESTs on the deletion maps which led to insights into the structure and evolution of the wheat genome (Qi et al. 2004), the construction of BAC libraries for the diploid (Lijavetzky et al. 1999; Moullet et al. 1999), tetraploid (Cenci et al. 2003) and hexaploid wheat genomes (Allouis et al. 2003) which heralded the start of map-based cloning in wheat, and the targeted sequencing of BACs or small contigs (e.g. Wicker et al. 2001; SanMiguel et al. 2002; Griffiths et al. 2006) which provided the first detailed information on the organization of genes in the wheat genome. The next big targets in wheat genomics are the construction of wheat physical maps, sequencing of the wheat genome, and the development of functional genomics tools. The ultimate goal is to understand the function of all genes in the wheat genome so that targeted improvement of wheat can be achieved in a highly efficient manner.

Gene numbers in different plant species are not expected to vary greatly. In Arabidopsis, 27,235 protein-encoding genes have been annotated (The Arabidopsis Genome Initiative 2000; http://www.arabidopsis.org/). In rice, the latest estimate places the number of genes around 32,000 (The Rice Annotation Project 2007). In maize, an ancient tetraploid, gene numbers are expected to be in the range of 37,000 to 63,000 (Haberer et al. 2005; Liu et al. 2007). However, early estimates of the number of genes present in the hexaploid wheat genome varied from being in line with the expectations to being nearly three times as high. Annotation of BAC end-sequences for genes and extrapolation to the entire genome suggested that there are 108,000 genes in the hexaploid wheat genome (Paux et al. 2006). In stark contrast, sample sequencing of a shot-gun library of hexaploid wheat DNA led to a prediction of around 295,900 genes (Rabinowicz et al. 2005). It is common knowledge that gene numbers, at least in early annotation efforts and particularly in large genomes, tend to be highly inflated because gene fragments and transposable elements are often mis-annotated as genes (Bennetzen et al. 2004). However, reannotation of the Rabinowicz et al. (2005) dataset by Paux et al. (2006) resulted in similar gene numbers. The discrepancy between both studies in the number of genes estimated to be present in the wheat genome thus appears to be the result of inherent differences in the data set rather than in the method of annotation. We will present a new estimate of the total gene number in hexaploid wheat, based on the annotation of 11.1 Mb of DNA sequence from the Chinese Spring wheat genome.

MATERIALS AND METHODS

BAC selection and sequencing

One hundred and nine BACs were selected randomly from among the 1,200,000 clones in the Chinese Spring BAC library (Allouis et al. 2003). Preparation of shotgun libraries, sequencing of 576 to 768 subclones for each BAC, base calling and sequence assembly were performed as previously described (Devos et al. 2005).

BAC annotation for genes

Two different methods were used for annotation. Seventy BAC clones were annotated manually and 67 clones, 29 of which had been manually annotated, were run through the DAWG-PAWS (*D*istributed *A*nnotation *W*orking *G*roup – *P*ipeline to *A*nnotate *W*heat *S*equences) wheat annotation pipeline (http://dawgpaws.sourceforge.net), followed by manual curation.

The first step in the manual annotation was to use the gene prediction program FGENESH with the monocot training set (www.softberry.com) to identify putative genes. Gene structures were then subjected to BLASTN searches against the Triticeae Repeat database (TREP) (http://wheat.pw.usda.gov/ITMI/Repeats/) and the TIGR Gramineae repeat database (http://www.tigr.org//tdb/e2k1/plant.repeats/). Predicted genes that did not match transposable elements were subjected to BLASTX searches against the 'nr Peptide Sequence Section' of GenBank. Only sequences that showed homology at an E-value of <E^{-10} to genes or ESTs derived from species belonging to genera other than *Triticum* and *Aegilops* were considered true genes.

The DAWG-PAWS pipeline conducts *de novo* gene prediction and annotation of transposable elements, and homology-based searches at both the DNA and protein level against a variety of repeat, EST, gene indices and protein databases. The DAWG-PAWS output of the 67 clones was visualized in the Apollo Genome Annotation Curation Tool (Lewis et al. 2002; http://apollo.berkeleybop.org/current/index.html) and used to manually identify genes.

Effect of sample size on gene estimations

To establish the amount of DNA that needs to be annotated to reliably extrapolate gene estimates for the entire wheat genome, the 109 BAC clones annotated were resampled with replacement, and gene numbers from random selections of 5, 10, 15, *etc.* BAC clones were extrapolated to the 16,800 Mb wheat genome. Where the number of genes annotated differed between two annotation methods, gene numbers were averaged over the two methods. To establish 95% confidence intervals of the means, the sampling was repeated 1000 times.

RESULTS

Manual gene annotation

The gene prediction software FGENESH identified 1,378 gene structures in ~7.3 Mb of genomic DNA, corresponding to 70 randomly selected BAC clones. BLAST analyses demonstrated that 71.8% of the predicted genes had homology to transposable elements in the TREP and/or *Gramineae* repeat database. The remaining 28.2% of the sequences did not have significant homology to known repeats. Some 50% of these displayed homology with wheat ESTs. A BLASTX search against GenBank proteins, however, identified only 84 sequences (6.1% of the predicted genes) that had homology at the E^{-10} value with functional or hypothetical proteins in species other than wheat or its close relatives. These sequences were considered true genes.

Annotation using the DAWG-PAWS pipeline

Annotation of 67 BAC clones, totalling ~6.6 Mb of DNA, led to the identification of 103 genes. Criteria for what constitutes a gene comprised identification by a gene prediction program, lack of homology to known repeats, and homology to Arabidopsis, rice or other grass proteins. Twenty-eight of the DAWG-PAWS annotated BAC clones had also been annotated manually. A total of 24 and 33 genes were annotated manually and using the DAWG-PAWS pipeline, respectively, on these 28 BAC clones.

"Gene-free" BACs

Of these 109 BACs, 31 (28%) were annotated as lacking any genes. Not surprisingly, the percentage of these gene-free BACs was higher (31%) among those with inserts smaller than 100 kb than it was (25%) among those with inserts of 100 kb or more.

Reliability of gene number estimates

Simulations of the effect of sample size on total gene number estimates were done for BAC samples representing from 0.003% to 0.066% of the hexaploid wheat genome. The average gene number estimated for each sample size, the 95% confidence intervals on the means and 5% error on the means are shown in Figure 1.

DISCUSSION

Estimating gene numbers in complex plant genomes

The most significant problem in assessing gene content in plant genomes is the common mis-annotation of transposable element (TE) encoded proteins as true plant genes (Bennetzen et al. 2004). Because most or all TEs are expressed in at least some tissue under some treatments, presence of homology in an EST library is not proof that a sequence is a true gene (Bennetzen et al. 2004). Since only a limited amount of wheat repeats have been annotated, a sequence that shows good homology to a wheat EST and no homology to a known repeat may still represent a TE. In addition, many TEs exist in low copy numbers in any given genome, so a lack of a high copy number is also not sufficient proof that a particular open reading frame is a real gene. However, TEs tend to evolve more rapidly than the standard genes in all studied genomes, so good conservation of a DNA sequence across distantly related taxa is a useful criterion to distinguish between candidate genes with true selected function and TE-encoded genes. This approach is not without its own limitations, though, because it would miss those rare genes that are truly novel to a particular lineage and would also still annotate highly conserved (e.g., possibly recent horizontally transferred) TEs as genes.

Annotation of genes within individual sequence reads is particularly fraught with potential errors. In reconstructions with ABI-3730 reads inside known Arabidopsis genes, only about two thirds were convincingly identified (Liu and Bennetzen, 2008). Even more problematic is the greater likelihood of mis-identifying TE-encoded genes as true plant genes because the homology criteria are dropped to the lower level needed to find real genes in a single strand sequence that is generally less than 1.5 kb in length. Depending on the thresholds set by the investigators, and there are no agreed-upon criteria in this field, gene number prediction from single shotgun sequence reads of unassembled data is expected to be widely variable and wildly inaccurate. Hence, gene number predictions should be most accurate from assembled sequence like those from BACs or BAC contigs.

Gene distribution

From earlier research in which markers were localized to cytogenetically-characterized deletion stocks, it was predicted that about 71% of the hexaploid wheat genome is largely or completely lacking in genes (Erayman et al. 2004). Assuming that the inserts in the Chinese Spring BAC library contain a fairly random representation of this hexaploid wheat genome, our annotation techniques indicate that only about 28% of BACs are "gene-free". Taking into account that there will be errors associated with the size of the deletions, and that annotation criteria might differ in the two studies, the fact that only 28% of BAC clones annotated within our study contain no genes and a further 28% of BAC clones contain only a single gene, indicates that gene densities in the wheat genome are much lower than previously predicted.

Predicted numbers of protein-coding genes in the hexaploid wheat genome

Manual annotation of ~7.3 Mb of wheat DNA suggests that the entire bread wheat genome contains ~193,000 genes. Automated annotation of ~6.6 Mb of DNA sequence, followed by manual curation of the DAWG-PAWS pipeline output, led to an estimate of ~262,000 genes. To assess whether the discrepancy in gene number is related to the annotation method, 28 BAC clones were annotated using both methods. Although the same criteria for identifying true genes were used by both the manual annotator and the curator of the DAWG-PAWS output, the same number of genes was identified on only 54% of the BAC clones. Manual curation of the annotation pipeline output led to higher and lower gene numbers in 32% and 14% of the BAC clones, respectively, compared to entirely manual annotation. Some of these differences may be due to the fact that the manual annotation was done before a comprehensive assembly of the DNA sequence had been carried out, while automated annotation was done on largely Phase II sequence. Also, manual and automated blast searches were in some cases done against different database releases and the automated annotation was, of course, more comprehensive than the manual annotation. Possibly more common than differences in the results of the sequence analyses, however, are differences in the way these results are interpreted by the annotator with respect to what constitutes a true gene.

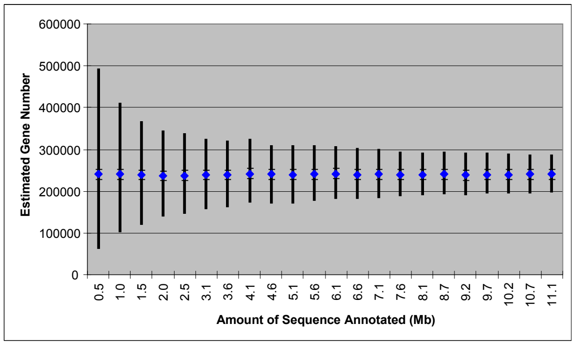

Fig. 1. Simulation of the effect of sample size on gene number estimates. Diamonds show average gene numbers, based on 1000 random samplings. The vertical bars that extend above and below the diamonds indicate 95% confidence intervals. Small horizontal lines above and below the means indicate 5% variation from the mean.

Extrapolations from the number of genes identified in the 28 BAC clones that were both manually and pipeline annotated indicated that the wheat genome comprised 144,000 and 198,000 genes, respectively. These estimates were considerably lower than the 193,000 and 262,000 genes obtained after annotation of 70 and 67 BAC clones, respectively, using these same methods. This raised the question of whether there is minimum amount of sequence that needs to be analyzed to obtain reliable total gene number estimates. A simulation of the effect of sample size showed that estimated gene numbers could vary more than 7-fold, depending on the sample, if only 0.5 Mb of sequence data, corresponding to 0.003% of the wheat genome, were annotated (Fig.1). The sampling effect decreased, at first quickly and then gradually, with increasing sample size. However, even with our sample size of 109 BAC clones, estimates (95% confidence intervals) ranged from 195,000 to 288,000 genes and differed from the mean by as much as 20% (Fig.1). This suggested that 11.1 Mb of DNA sequence is too small of a data set to obtain a reliable gene estimate for the entire wheat genome.

Previous estimates for the gene number in wheat varied from 108,000 (Paux et al. 2006) to 296,000 (Rabinowicz et al. 2005). Considering the large effect of sample size, it is important to infer the error rate associated with these estimates and to establish whether these estimates are significantly different from our projected gene numbers. The whole-genome shotgun sequence data set from Rabinowicz and colleagues consisted of only 0.81 Mb, corresponding to 0.005 % of the wheat genome. This is the range in which gene estimates vary greatly depending on the sample analyzed (Fig.1). In our analysis, gene numbers ranged from less than 100,000 to more than 400,000 when extrapolations were done from less than 1 Mb of sequence. The gene estimate obtained by Rabinowicz et al. (2005) may therefore not be significantly different from that obtained by Paux et al. (2006). The sample analyzed by Paux and colleagues consisted of 11 Mb of BAC-end sequence. While 11 Mb is also insufficient to give a precise tally of the number of genes present in the entire wheat genome, the sampling error associated with this sample size is expected to be around 20% of the mean (Fig.1). If we assume that 108,000 is at the low end of the range, gene numbers based on the annotation of 11 Mb of DNA sequence might be as high as 150,000. This figure is significantly lower than the minimum number of 195,000 estimated in our study. In neither the Paux et al. (2006) study nor in our study was a differentiation was made between functional genes and pseudogenes. The use of different annotation criteria might be one possible explanation for the discrepancy in predicted gene numbers in the two studies. As described above, it is also possible that to the difficulty in identifying genes in 700 bp BAC-end sequence reads, compared to the longer contiguous stretches present in BAC sequence assemblies, might be the reason for the low gene number estimate published in Paux et al. (2006).

In conclusion, the jury is still out on the total number of genes that are present in the hexaploid wheat genome. Annotation criteria, interpretation of these criteria by the annotator, sample type, and sample size are all likely to affect gene number estimations. A conservative estimate based on the annotation of 11.1 Mb of sequence from randomly selected BAC clones, and taking into account the fact that gene numbers varied by as much as 32% when the same sequence was annotated by different people, is that the wheat genome contains between 164,000 and 334,000 protein-encoding genes. Because we have not yet applied criteria to determine what portion of these candidate genes might be pseudogenes, it is currently not clear whether wheat actually contains the exceptionally high gene number (55,000 to 111,000) per diploid genome suggested by these studies.

REFERENCES

Allouis S, Moore G, Bellec A, Sharp R, Faivre-Rampant P, Mortimer K, Pateyron S, Foote TN, Griffiths S, Caboche M, Chalhoub B (2003) Construction and characterisation of a hexaploid wheat (*Triticum aestivum* L.) BAC library from the reference germplasm 'Chinese Spring'. Cereal Res Commun 31:331-338

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7:732-736

Cenci A, Chantret N, Kong X, Gu Y, Anderson OD, Fahima T, Distelfeld A, Dubcovsky J (2003) Construction and characterization of a half million clone BAC library of durum wheat (*Triticum turgidum* ssp. *durum*). Theor Appl Genet 107:931-939

Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) Analysis and mapping of randomly chosen BAC clones from hexaploid bread wheat. Proc Natl Acad Sci USA 102:19243-19248

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS and Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucl Acids Res 32:3546-3565

Gale MD, Atkinson MD, Chinoy CN, Harcourt RL, Jia J, Li QY, Devos KM (1995) Genetic maps of hexaploid wheat. In: Li ZS, Xin ZY (eds) Proc 8th Int Wheat Genet Symp. China Agricultural Scientech Press, Beijing, pp 29-40

Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G (2006) Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. Nature 439:749-752

Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Birren B, Nusbaum C, Mayer KFX, Messing J (2005) Structure and architecture of the maize genome. Plant Physiol 139:1612-1624

Kurata N, Moore G, Nagamura Y, Foote TN, Yano M, Minobe Y, Gale MD (1994) Conservation of genome structure between rice and wheat. Bio/Technology 12:276-278

La Rota M, Sorrells ME (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. Funct Integr Genomics 4:34-46

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NLV, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XT, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson DO (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. Genetics 168:585-593

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Ricter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews B, Prochnik SE, Smith CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME (2002) Apollo: a sequence annotation editor. Genome Biology 3:Research 0082.1-0082.14

Lijavetzky D, Muzzi G, Wicker T, Keller B, Wing R, Dubcovsky J (1999) Construction and characterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. Genome 42:1176-1182

Liu RY, Vitte C, Ma JX, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL (2007) A GeneTrek analysis of the maize genome. Proc Natl Acad Sci USA 104:11844-11849

Liu RY, Bennetzen JL (2008) ENCHILADA REDUX: How complete is your genome sequence? New Phytol (in press)

Mickelson-Young L, Endo TR, Gill BS (1995) A cytogenetic ladder-map of the wheat homoeologous group-4 chromosomes. Theor Appl Genet 90:1007-1011

Moullet O, Zhang HB, Lagudah ES (1999) Construction and characterisation of a large DNA insert library from the D genome of wheat. Theor Appl Genet 99:305-313

Nelson JC, Sorrells ME, Van Deynze AE, Lu Y-H, Atkinson MD, Bernard M, Leroy P, Faris JD, Anderson JA (1995) Molecular mapping of wheat: Genes and rearrangements in homoeologous groups 4, 5 and 7. Genetics 141:721-731

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463-474

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma XF, Miftahudin, Gustafson JP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NLV, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168:701-712

Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA (2005) Differential methylation of genes and repeats in land plants. Genome Res 15:1431-1440

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A^{m}. Funct Integr Genomics 2:70-80

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815

The Rice Annotation Project (2007) Curated genome annotation of *Oryza sativa* ssp *japonica* and comparative genome analysis with *Arabidopsis thaliana*. Genome Res 17:175-183

Van Deynze AE, Nelson JC, Yglesias ES, Harrington SE, Braga DP, McCouch SR, Sorrells ME (1995) Comparative mapping in grasses: Wheat relationships. Mol Gen Genet 248:744-754

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307-316

ACKNOWLEDGEMENTS

This research was accomplished with funds from the National Science Foundation – Plant Genome Program award No. DBI-0501814.