

Review

Genomic Databases for Crop Improvement

Kaitao Lai ^{1,2}, Michal T. Lorenc ^{1,2} and David Edwards ^{1,2,*}

¹ School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia; E-Mails: k.lai1@uq.edu.au (K.L.); m.lorenc@uq.edu.au (M.T.L.)

² Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, QLD 4072, Australia

* Author to whom correspondence should be addressed; E-Mail: dave.edwards@uq.edu.au; Tel.: +61-7-3346-7084; Fax: +61-7-3365-1176.

Received: 11 January 2012; in revised form: 13 March 2012 / Accepted: 15 March 2012 /

Published: 20 March 2012

Abstract: Genomics is playing an increasing role in plant breeding and this is accelerating with the rapid advances in genome technology. Translating the vast abundance of data being produced by genome technologies requires the development of custom bioinformatics tools and advanced databases. These range from large generic databases which hold specific data types for a broad range of species, to carefully integrated and curated databases which act as a resource for the improvement of specific crops. In this review, we outline some of the features of plant genome databases, identify specific resources for the improvement of individual crops and comment on the potential future direction of crop genome databases.

Keywords: bioinformatics; next-generation sequencing; molecular markers; single-nucleotide polymorphisms

1. Introduction

The majority of DNA sequence and expressed gene sequence data generated today comes from the next- or second-generation sequencing (NGS/2GS) technologies. NGS technologies produce vast quantities of short data rather than Sanger sequencing at a relatively low cost and short time. Genomics is undergoing a revolution, driven by advances in DNA sequencing technology, and this data flood is having a major impact on approaches and strategies for crop improvement. NGS technologies have

been applied for sequenced genomes of a number of cereal crop species including rice, *Sorghum* and maize. A quality sequence of rice that covers 95% of the 389 Mb genome has been produced [1]. The *Sorghum bicolor* (L.) Moench genome has been assembled in size of 730-megabase, placing ~98% of genes in their chromosomal context [2]. The draft nucleotide sequence of the 2.3-gigabase genome of maize has also been improved [3]. One of the challenges encountered by researchers is to translate this abundance of data into improved crops in the field. There remains a gap between genome data production and next-generation crop improvement strategies, but this is being rapidly closed by far sighted companies and individuals with the ability to combine the ability to mine the genomic data with practical crop-improvement skills. Bioinformatics can be defined as the structuring of biological information to enable logical interrogation, and databases are a key part of the bioinformatics toolbox. Numerous databases have been developed for genomic data, on a range of platforms and to suite a variety of different purposes (see Table 1 for examples). These range from generic DNA sequence or molecular marker databases, to those hosting a variety of data for specific species.

Table 1. Examples of genomic databases related to crop improvement.

Database Name	Web Link	References
autoSNPdb	http://autosnpdb.appliedbioinformatics.com.au/	[4,5]
<i>Brachypodium</i> database	http://www.brachypodium.org/	[6]
<i>Brassica</i> genome gateway	http://www.brassicagenome.net	[7]
<i>Brassica rapa</i> genome database	http://brassicadb.org/	[8]
DNA Data Bank of Japan (DDBJ)	http://ddbj.sakura.ne.jp/	[9]
European bioinformatics institute EnSEMBL plants	http://plants.ensembl.org/	[10,11]
European Molecular Biology Laboratory (EMBL) nucleotide sequence database	http://www.ebi.ac.uk/embl/	[12,13]
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	[14–16]
Graingenes	http://wheat.pw.usda.gov/	[17–19]
Gramene	http://www.gramene.org/	[20]
International Crop Information System (ICIS)	http://www.icis.cgiar.org	[21]
International Nucleotide Sequence Database Collaboration (INSDC)	http://www.insdc.org/	[9]
Legume Information System (LIS)	http://www.comparative-legumes.org/	[22,23]
MaizeGDB	http://www.maizegdb.org/	[24–26]
Maize sequence database	http://www.maizesequence.org/	[3]
Oryzabase	http://www.shigen.nig.ac.jp/rice/oryzabase/	[27]
Panzea	http://www.panzea.org/	[28]
Phytozome	http://www.phytozome.net/	[29]
PlantsDB	http://mips.helmholtz-muenchen.de/plant/genomes.jsp	[30]
PlantGDB	http://www.plantgdb.org/	[31,32]
The Plant Ontology	http://www.plantontology.org/	[33]
Plaza	http://bioinformatics.psb.ugent.be/plaza/	[34]
Rice Genome Annotation Project	http://rice.plantbiology.msu.edu/	[35]
SSR Primer	http://flora.acpfg.com.au/ssrprimer2/	[36]

Table 1. *Cont.*

Database Name	Web Link	References
SSR taxonomy tree	http://appliedbioinformatics.com.au/projects/ssrta taxonomy/php/	[36]
SOL Genomics Network (SGN)	http://solgenomics.net/	[37]
SoyBase	http://soybase.org/	[38]
TAGdb	http://flora.acpfg.com.au/tagdb/	[39]
The Crop Expressed Sequence Tag database, CR-EST	http://pgrc.ipk-gatersleben.de/cr-est/	[40]
The Triticeae Repeat Sequence Database (TREP)	http://wheat.pw.usda.gov/ITMI/Repeats/	[41]
Wheat genome information	http://www.wheatgenome.info	[42]

1.1. Generic Databases

The largest of the DNA sequence repositories is the International Nucleotide Sequence Database Collaboration (INSDC), made up of the DNA Data Bank of Japan (DDBJ) at The National Institute of Genetics in Mishima, Japan [9], GenBank at the National Center of Biotechnology Information (NCBI) in Bethesda, USA [15,16], and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, maintained at the European Bioinformatics Institute (EBI) in the UK [13]. Daily data exchange between these groups ensures coordinated international coverage [43].

Since the introduction of advanced next-generation sequencing technology, the storage and interrogation of this data is becoming an expanding challenge [44,45]. The ability to search the vast quantity of this data is made feasible by the development of custom databases such as TAGdb (<http://flora.acpfg.com.au/tagdb/>) [39], but it is increasingly the assembled and annotated genome data which are applied for crop-improvement applications [46].

While it is valuable to maintain all public nucleic acid sequences in one location, the size of this resource limits the ability to visualize this data. Genome viewers, which place genomic data within the context of sequenced or partially sequenced genomes, provide more context-orientated data interrogation. There are two main generic web-based tools to view plant genomes: Ensembl [10] and GBrowse [47,48]. Both are widely used and it is not uncommon to find similar genome information hosted on both systems. A key development in genome databases was the establishment and adoption of a standard file format for genome data [49], and data in the current version, GFF3 can be visualized and searched using a wide range of tools from custom GBrowse databases to stand alone bioinformatics tools such as Biomatters Geneious [50].

There are several resources which collate genome data for multiple plant species. Gramene (<http://www.gramene.org/>) [20] is an Ensembl-based genome viewer and database hosting information on a variety of crop species, but based around the rice, maize and Arabidopsis genomes [18]. A similar resource is hosted by the EBI (<http://plants.ensembl.org/>) [10]. PlantGDB is a resource for comparative plant genomics [31,32] and hosts sequence data for >70,000 plant species with a focus on complete sequencing of reference species, Arabidopsis, rice, maize and Medicago truncatula. Plaza (<http://bioinformatics.psb.ugent.be/plaza/>) [34] hosts pre-computed comparative genomics data sets for a range of species [34]. Phytozome (<http://www.phytozome.net/>) [29] also hosts genome data for

numerous plant species and provides several genomes using the GBrowse format. With 25 complete plant genomes, phytozome is one of the most comprehensive plant genome databases currently available [18]. In addition, PlantsDB is a generic database hosting data for multiple plant species. This database is hosted by MIPS (<http://mips.helmholtz-muenchen.de/plant/genomes.jsp>) [30].

While genome and transcript sequence information makes up the bulk of genome data maintained within public databases, it is often the differences between individuals and varieties which are the most valuable for crop-improvement applications. A major focus of crop genetic research in recent decades has been the development of molecular genetic markers associated with important traits. Genetic markers can be assayed with a variety of techniques [51]. Early molecular genetic markers technologies such as restriction fragment length polymorphisms have been replaced by more high throughput methods, including amplified fragment length polymorphisms (AFLPs), diversity array technologies (DArT) and simple sequence repeats (SSRs) also known as microsatellites. Another important and crop-improvement-oriented database is the maize database Panzea (<http://www.panzea.org/>) [28], which hosts data on genomic diversity in a large germplasm collection including genetic data, trait phenotypes, allele frequencies, phenotyping environments, genetic analysis tools and so on. The Panzea database The Panzea databases comprises the genotypic and phenotypic data and genetic marker information. This database design is based on the Genomic Diversity and Phenotype Data Model (GDPDM) (<http://www.maizogenetics.net/gdpdm/>) [20].

An expressed sequence tag (EST) represents a short sub-sequence of a cDNA sequence. EMBL or GenBank have sub-sections for EST sequences. The crop expressed sequence tag database, CR-EST (<http://pgrc.ipk-gatersleben.de/cr-est/>) [40], provides access to more than 200,000 sequences derived from 41 cDNA libraries of four species: barley, wheat, pea and potato [40].

SSRs are short stretches of DNA sequence occurring as tandem repeats of mono-, di-, tri-, tetra-, penta- and hexa-nucleotides. They are highly polymorphic due to mutation affecting the number of repeat units. The value of SSRs is due to their genetic co-dominance, abundance, dispersal throughout the genome, multi-allelic variation and high reproducibility. The hypervariability of SSRs among related organisms makes them excellent markers for genotype identification, analysis of genetic diversity, phenotype mapping and marker assisted selection [52,53]. SSRs demonstrate a high degree of transferability between species, as PCR primers designed to an SSR within one species frequently amplify a corresponding locus in related species, enabling comparative genetic and genomic analysis.

With the continued advances in DNA sequencing technologies, single-nucleotide polymorphisms (SNPs) have come to dominate high throughput molecular marker analysis. SNPs are the ultimate form of molecular genetic marker, as a nucleotide base is the smallest unit of inheritance, and a SNP represents a single-nucleotide difference between two individuals at a defined location. SNPs are direct markers as the sequence information provides the exact nature of the allelic variants. Furthermore, this sequence variation can have a major impact on how the organism develops and responds to the environment. SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers near a locus of interest. SNPs at any particular site could in principle involve four different nucleotide variants, but in practice they are generally biallelic. This disadvantage, when compared with multiallelic markers such as SSRs, is compensated by the relative abundance of SNPs. The high density of SNPs makes them valuable for genome mapping, and in particular they allow the generation of ultra-high density genetic maps and haplotyping systems for

genes or regions of interest, and map-based positional cloning. SNPs are used routinely in crop breeding programs, for genetic diversity analysis, cultivar identification, phylogenetic analysis, characterization of genetic resources and association with agronomic traits [54,55].

SSR Primer (<http://flora.acpfg.com.au/ssrprimer2/>) [36] is a web-based tool that enables the real time discovery of SSRs within submitted DNA sequences, with the concomitant design of PCR primers for SSR amplification [56]. Alternatively, users may browse an SSR Taxonomy Tree (<http://appliedbioinformatics.com.au/projects/ssrtaxonomy/php/>) [36] to identify pre-determined SSR amplification primers for species represented within the GenBank database [36].

The SNP discovery software autoSNP [57,58] identifies SNPs and insertion/deletion (indel) polymorphisms from bulk sequence data using two measures of confidence; redundancy, defined as the number of times a polymorphism occurs at a locus in a sequence alignment; and co-segregation of SNPs to define a haplotype. AutoSNP software has recently been extended to database format, autoSNPdb, which permits complex queries and provides detailed genomic and functional information [4,5]. Where the sequence trace files are available, the SNP discovery tool PolyPhred [59,60] can make use of the base pair quality scores to further differentiate between true SNP polymorphisms and random sequence error. The recent developments in next-generation sequence data have led to the identification of large numbers of SNPs in a range of plant genomes and these approaches are likely to dominate SNP discovery in the coming years [61].

The increased throughput for the discovery and application of molecular genetic markers has led to the requirement for databases hosting the results of molecular marker analysis. These may be integrated within other database systems such as Gramene [20], the Legume Information System (LIS) [22,23], or Graingenes [17,18].

One of the principal uses of molecular genetic markers is the production of genetic maps and the mapping of heritable traits. While mapping data may be described as lists, graphical representations are more readily understood. The genetic map viewer CMap, developed by the GMOD consortium [62] is valuable for the validation of traits that map to the same position in different populations and also for the linkage between crop genetic maps and sequenced model genomes, enabling the identification of candidate genes for genetically mapped traits. A recent addition, CMap3D [63], enables the comparison of a larger number of maps in 3D space.

The linking of genomic data with agronomic traits remains one of the greatest challenges in the application of genome data for crop improvement [64,65]. Several databases have been developed to assist in this endeavor. The International Crop Information System (ICIS) [21] is a database system that hosts integrated management information for crop improvement, including details on diverse germplasm and traits. One challenge in developing trait databases is the establishment of functional ontologies. The Plant Ontology (<http://www.plantontology.org/>) [33] is a controlled vocabulary (ontology) that describes plant anatomy and morphology and stages of growth and development for all plants [33] and this database of ontologies is becoming the standard for comparative physiology and for linking genes with potential function.

1.2. Species Focused Databases

It would be impossible to detail all available plant genetic and genomic databases, however some of the main ones are listed below along with a brief description of their content.

GrainGenes (<http://wheat.pw.usda.gov/>) [18] is a genetic database for Triticeae, oats, and sugarcane. GrainGenes (Matthews *et al.*, 2003; Carollo *et al.*, 2005) [18,19]. Comprehensive information includes genetic markers, map locations, alleles, key references and disease symptoms. The Triticeae Repeat Sequence Database (TREP) (<http://wheat.pw.usda.gov/ITMI/Repeats/>) [41] contains a collection of repetitive DNA sequences from different *Triticeae* species which can be used for the development of molecular markers.

While *Brachypodium distachyon* is not grown as a crop, this species has many qualities that make it a model for studies in temperate grasses and cereals, including a small genome (~ 300 Mbp), small physical stature, self-fertility, a short lifecycle, and simple growth requirements. The *B. distachyon* genome was sequenced in 2010 [66] and the *Brachypodium* database which includes a GBrowse based genome viewer is available at <http://www.brachypodium.org/> [6].

The maize genome was sequenced in 2009 [3] and there are several databases hosting information on this important crop. These include MaizeGDB [25,26] (<http://www.maizegdb.org/>) [24] based on GBrowse, and maizesequence.org (<http://www.maizesequence.org/>) based on Ensembl [3].

Rice was one of the first crop genomes to be sequenced and there are now numerous resources available to mine this genomic information. Oryzabase is an integrated rice science database established in 2000 (<http://www.shigen.nig.ac.jp/rice/oryzabase/>) [27]. The database hosts information on genetic resources, chromosome maps, genes and rice mutants. This is complemented by a rice genome annotation project [35] which presents data using GBrowse (<http://rice.plantbiology.msu.edu/>) [35].

Although wheat is an extremely important crop, advances in genomics have been limited by its large and highly complex genome. Assemblies of the gene rich regions for the group 7 chromosomes have been completed [67,68], and annotated sequences, including a large number of SNP polymorphisms are available at <http://www.wheatgenome.info> [42].

A central portal for *Brassica* data is maintained at Brassica.info, with links to genetic marker, map and a range of diverse Brassica related information. The recently sequenced *Brassica rapa* genome [7] is hosted at <http://brassicadb.org/> [8] in a database named BRAD [8], with a second database which contains *Brassica* repeat information at <http://www.BrassicaGenome.net> [7]. Both of these databases use GBrowse.

The Legume Information System (<http://www.comparative-legumes.org/>) [22,23] supports basic research in the legumes by relating data from multiple crop and model species, and by helping researchers traverse among various data types [22,23]. It currently hosts data for seventeen species and includes GBrowse databases for Glycine max (soybean), Lotus japonicus (birdsfoot trefoil), Medicago truncatula (barrel medic) and Cajanus cajan (pigeonpea). It is complemented by detailed soybean data hosted at SoyBase (<http://soybase.org/>) [38].

The SOL Genomics Network (SGN) (<http://solgenomics.net/>) [37] is a clade oriented database containing genomic, genetic, phenotypic and taxonomic information for plant genomes, with a focus on the Euasterid clade, which includes Solanaceae (e.g., tomato, potato, eggplant, pepper and petunia)

and Rubiaceae (coffee) [37]. As well as being a resource for basic crop research, SGN maintains databases with a specific focus on giving breeders direct links to breeder-relevant tools and data.

2. Conclusions and Future Direction

There are currently a range of databases dedicated to generic genome data or focusing on specific crops or clades. Both the type and volumes of data have increased greatly over the last few years and this trend looks to continue. Some of the early database formats are either no longer used or have limited applications [69–71], however several newer web tools are now becoming predominant. These include the GBrowse genome viewer [47,48] and associated open source bioinformatics developments as well as the Ensembl system [10]. As genome technology continues to advance and an increasing number of crop genomes become available, an expanding number of these databases will be developed. One of the main challenges facing crop bioinformatics researchers is to make the ever increasing volume and types of data available in a suitable format for analysis [72]. This includes new high-throughput plant phenotype data as well as the increasing volumes of genotypic diversity data. It will be the association of this diversity data with heritable phenotypes which will likely drive genome database development over the coming years [73,74]. These databases therefore will require the implementation of appropriate statistical tools for association of high-density genotype and high-throughput phenotype data.

Acknowledgments

The authors would like to acknowledge funding support from the Grains Research and Development Corporation (Project DAN00117) and the Australian Research Council (Projects LP0882095, LP0883462 and LP110100200). Support from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC) is gratefully acknowledged.

References and Notes

1. The map-based sequence of the rice genome. The map-based sequence of the rice genome. *Nature* **2005**, *436*, 793–800.
2. Paterson, A.H.; Bowers, J.E.; Bruggmann, R.; Dubchak, I.; Grimwood, J.; Gundlach, H.; Haberer, G.; Hellsten, U.; Mitros, T.; Poliakov, A.; *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **2009**, *457*, 7229, 551–556.
3. Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.S.; Pasternak, S.; Liang, C.Z.; Zhang, J.W.; Fulton, L.; Graves, T.A.; *et al.* The B73 maize genome: Complexity, diversity, and dynamics. *Science* **2009**, *326*, 1112–1115.
4. Duran, C.; Appleby, N.; Clark, T.; Wood, D.; Imelfort, M.; Batley, J.; Edwards, D. AutoSNPdb: An annotated single nucleotide polymorphism database for crop plants. *Nucl. Acid. Res.* **2009**, *37*, D951–D953.
5. Duran, C.; Appleby, N.; Vardy, M.; Imelfort, M.; Edwards, D.; Batley, J. Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnol. J.* **2009**, *7*, 326–333.

6. Larré, C.; Penninck, S.; Bouchet, B.; Lollier, V.; Tranquet, O.; Denery-Papini, S.; Guillon F.; Rogniaux, H. *Brachypodium distachyon* grain: Identification and subcellular localization of storage proteins. *J. Exp. Bot.* **2010**, *61*, 1771–1783.
7. Wang, X.; Wang, H.; Wang, J.; Sun, R.; Wu, J.; Liu, S.; Bai, Y.; Mun, J.-H.; Bancroft, I.; Cheng, F.; *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **2011**, *43*, 1035–1039.
8. Cheng, F.; Liu, S.; Wu, J.; Fang, L.; Sun, S.; Liu, B.; Li, P.; Hua, W.; Wang, X.; Cheng, F.; *et al.* BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biology* **2011**, *11*, doi:10.1186/1471-2229-11-136.
9. Sugawara, H.; Ogasawara, O.; Okubo, K.; Gojobori, T.; Tatenno Y. DDBJ with new system and face. *Nucl. Acids Res.* **2008**, *36*, D22–D24.
10. Flicek, P.; Amode, M.R.; Barrell, D.; Beal, K.; Brent, S.; Chen, Y.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; *et al.* Ensembl 2011. *Nucl. Acid. Res.* **2011**, *39*, D800–D806.
11. Kersey, P.; Lawson, D.; Birney, E.; Derwent, P. S.; Haimel, M.; Herrero, J.; Keenan, S.; *et al.* Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucl. Acids Res.* **2010**, *38*, D563–D569.
12. Kulikova, T.; Akhtar, R.; Aldebert, P.; Althorpe, N.; Andersson, M.; Baldwin, A.; *et al.* EMBL Nucleotide Sequence Database in 2006. *Nucl. Acids Res.* **2007**, *35*, D16–D20.
13. Sterk, P.; Kulikova, T.; Kersey, P.; Apweiler, R. The EMBL nucleotide sequence and genome reviews databases. In *Methods in Molecular Biology*; Edwards, D., Ed.; Humana Press: Totowa, NJ, USA, 2007; Volume 406, pp. 1–21.
14. Karsch-Mizrachi, I.; Nakamura, Y.; Cochrane, G.; The international nucleotide sequence database collaboration. *Nucl. Acids Res.* **2012**, *40*, D33–D37.
15. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucl. Acid. Res.* **2009**, *37*, 26–31.
16. Wheeler, D.L.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Edgar, R.; Federhen, S.; *et al.* Database resources of the national center for biotechnology information. *Nucl. Acid. Res.* **2008**, *36*, D13–D21.
17. O’Sullivan, H. GrainGenes—A genomic database for Triticeae and Avena. In *Methods in Molecular Biology*; Edwards, D., Ed.; Humana Press: Totowa, NJ, USA, 2007; Volume 406, pp. 301–314.
18. Carollo, V.; Matthews, D.E.; Lazo, G.R.; Blake, T.K.; Hummel, D.D.; Lui, N.; Hane, D.L.; Anderson, O.D. GrainGenes 2.0: An improved resource for the small-grains community. *Plant Physiol.* **2005**, *139*, 643–651.
19. Matthews, D.; Carollo, V.L.; Lazo, G.R.; Anderson, O.D. GrainGenes, the genome database for small-grain crops. *Nucl. Acids Res.* **2003**, *31*, 183–186.
20. Youens-Clark, K.; Buckler, E.; Casstevens, T.; Chen, C.; DeClerck, G.; Derwent, P.; Dharmawardhana, P.; Jaiswal, P.; Kersey, P.; Karthikeyan, A.S.; *et al.* Gramene database in 2010: Updates and extensions. *Nucl. Acid. Res.* **2011**, *39*, D1085–D1094.
21. Fox, P.N.; Skovman, B. The International Crop Information System (ICIS)—connects genebank to breeder to farmer’s field. Plant adaptation and crop improvement. CAB International: Wallingford, Oxon, UK, 1996; pp. 317–326.

22. Gonzales, M.D.; Gajendran, K.; Farmer, A.D.; Archuleta, E.; Beavis, W.D. Leveraging model legume information to find candidate genes for soybean sudden death syndrome using the legume information system. In *Methods in Molecular Biology*; Edwards, D., Ed.; Humana Press: Totowa, NJ, USA, 2007; Volume 406, pp. 245–259.
23. Gonzales, M.D.; Archuleta, E.; Farmer, A.; Gajendran, K.; Grant, D.; Shoemaker, R.; Beavis, W.D.; Waugh, M.E. The legume information system (LIS): An integrated information resource for comparative legume biology. *Nucl. Acid. Res.* **2005**, *33*, D660–D665.
24. Schaeffer, M.L.; Harper, L.C.; Gardiner, J.M.; Andorf, C.M.; Campbell, D.A.; Cannon, E.K.; Sen, T.Z.; Lawrence, C.J. MaizeGDB: curation and outreach go hand-in-hand. *Database.* **2011**, doi: 10.1093/database/bar022
25. Lawrence, C.J. MaizeGDB—The maize genetics and genomics database. In *Methods in Molecular Biology*; Edwards, D., Ed.; Humana Press: Totowa, NJ, USA, 2007; Volume 406, pp. 331–345.
26. Lawrence, C.J.; Schaeffer, M.L.; Seigfried, T.E.; Campbell, D.A.; Harper, L.C. MaizeGDB's new data types, resources and activities. *Nucl. Acid. Res.* **2007**, *35*, D895–D900.
27. Yamazaki, Y.; Sakaniwa, S.; Tsuchiya, R.; Nonomura, K.I.; Kurata, N. Oryzabase: An integrated information resource for rice science. *Breed. Sci.* **2010**, *60*, 544–548.
28. Canaran, P.; Buckler, E.S.; Glaubitz, J.C.; Stein, L.; Sun, Q.; Zhao, W.; Ware, D. Panzea: An update on new content and features. *Nucl. Acids Res.* **2008**, *36*, D1041–D1043.
29. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; *et al.* Phytozome: A comparative platform for green plant genomics. *Nucl. Acid. Res.* **2012**, *40*, D1178–D1186.
30. Mewes, H.W.; Dietmann, S.; Frishman, D.; Gregory, R.; Mannhapt, G.; Mayer, K.F.X.; Münsterkötter, M.; Ruepp, A.; Spannagl, M.; Stümpflen, V.; Rattei, T. MIPS: analysis and annotation of genome information in 2007. *Nucl. Acids Res.* **2008**, *36*, D196–D201.
31. Brendel, V. Gene structure annotation at PlantGDB. In *Methods in Molecular Biology*; Edwards, D., Ed.; Humana Press: Totowa, NJ, USA, 2007; Volume 406, pp. 521–533.
32. Duvick, J.; Fu, A.; Muppirala, U.; Sabharwal, M.; Wilkerson, M.D.; Lawrence, C.J.; Lushbough, C.; Brendel, V. PlantGDB: A resource for comparative plant genomics. *Nucl. Acid. Res.* **2008**, *36*, D959–D965.
33. Avraham, S.; Tung, C.-W.; Ilic, K.; Jaiswal, P.; Kellogg, E.A.; McCouch, S.; Pujar, A.; Reiser, L.; Rhee, S.Y.; Sachs, M.M.; *et al.* The plant ontology database: A community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucl. Acid. Res.* **2008**, *36*, D449–D454.
34. Proost, S.; Van Bel, M.; Sterck, L.; Billiau, K.; Van Parys, T.; Van de Peer, Y.; Vandepoele, K. PLAZA: A comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* **2009**, *21*, 3718–3731.
35. Ouyang, S.; Zhu, W.; Hamilton, J.; Lin, H.; Campbell, M.; Childs, K.; Thibaud-Nissen, F.; Malek, R.L.; Lee, Y.; Zheng, L.; *et al.* The TIGR rice genome annotation resource: Improvements and new features. *Nucl. Acid. Res.* **2007**, *35*, D883–D887.

36. Jewell, E.; Robinson, A.; Savage, D.; Erwin, T.; Love, C.G.; Lim, G.A.C.; Li, X.; Batley, J.; Spangenberg, G.C.; Edwards, D. SSRPrimer and SSR taxonomy tree: Biome SSR discovery. *Nucl. Acid. Res.* **2006**, *34*, W656–W659.
37. Bombarely, A.; Menda, N.; Tecle, I.Y.; Buels, R.M.; Strickler, S.; Fischer-York, T.; Pujar, A.; Leto, J.; Gosselin, J.; Mueller, L.A. The sol genomics network (solgenomics.net): Growing tomatoes using Perl. *Nucl. Acid. Res.* **2011**, *39*, D1149–D1155.
38. Grant, D.; Nelson, R.T.; Cannon, S.B.; Shoemaker, R.C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucl. Acids Res.* **2010**, *38*, D843–D846.
39. Marshall, D.; Hayward, A.; Eales, D.; Imelfort, M.; Stiller, J.; Berkman, P.; Clark, T.; McKenzie, M.; Lai, K.; Duran, C.; *et al.* Targeted identification of genomic regions using TAGdb. *Plant Methods* **2010**, *6*, 19; doi:10.1186/1746-4811-6-19.
40. Künne, C.; Lange, M.; Funke, T.; Miehe, H.; Thiel, T.; Grosse, I.; Scholz, U. CR-EST: A resource for crop ESTs. *Nucl. Acids Res.* **2005**, *33*, D619–D621.
41. Wicker, T.; Buell, C.R. Gene and repetitive sequence annotation in the Triticeae. *Plant Genet. GenomicsCrop. Model.* **2009**, *7*, 407–425.
42. Lai, K.; Berkman, P.J.; Lorenc, M.T.; Duran, C.; Smits, L.; Manoli, S.; Stiller, J.; Edwards, D. WheatGenome.info: An integrated database and portal for wheat genome information. *Plant Cell Physiol.* **2011**, doi: 10.1093/pcp/pcr141.
43. Edwards, D.; Hansen, D.; Stajich, J. DNA sequence databases. In *Applied Bioinformatics*; Edwards, D.; Stajich, J.; Hansen, D.; Eds.; Springer: New York, NY, USA, 2009; pp. 1–11.
44. Batley, J.; Edwards, D. Genome sequence data: Management, storage, and visualization. *Biotechniques* **2009**, *46*, 333–336.
45. Lee, H.; Lai, K.; Lorenc, M.T.; Imelfort, M.; Duran, C.; Edwards, D. Bioinformatics tools and databases for analysis of next generation sequence data. *Brief. Funct. Genomics* **2012**, *11*, 12–24.
46. Edwards, D.; Batley, J. Plant genome sequencing: Applications for crop improvement. *Plant Biotechnol. J.* **2010**, *7*, 1–8.
47. Arnaoudova, E.G.; Bowens, P.J.; Chui, R.G.; Dinkins, R.D.; Hesse, U.; Jaromczyk, J.W.; Martin, M.; Maynard, P.; Moore, N.; Schardl, C.L. Visualizing and sharing results in bioinformatics projects: GBrowse and GenBank exports. *BMC Bioinformatics* **2009**, *10*, A4; doi:10.1186/1471-2105-10-S7-A4.
48. Donlin, M. Using the generic genome browser (GBrowse). *Curr. Protoc. Bioinformatics* **2007**, doi:10.1002/0471250953.bi0909s28.
49. Reese, M.G.; Moore, B.; Batchelor, C.; Salas, F.; Cunningham, F.; Marth, G.T.; Stein, L.; Flicek, P.; Yandell, M.; Eilbeck, K. A standard variation file format for human genome sequences. *Genome Biol.* **2010**, *11*, R88; doi: 10.1186/gb-2010-11-8-r88
50. Drummond, A.J.; Ashton, B.; Buxton, S.; Cheung, M.; Cooper, A.; Duran, C.; Field, M.; Heled, J.; Kearse, M.; Markowitz, S.; *et al.* *Geneious*, Version 5.4; Biomatters Ltd.: Auckland, New Zealand. Available online: <http://www.geneious.com> (accessed on 17 March 2012)
51. Duran, C.; Edwards, D.; Batley, J. Molecular marker discovery and genetic map visualisation. In *Applied Bioinformatics*; Edwards, D., Hanson, D., Stajich, J., Eds.; Springer: New York, NY, USA, 2009.

52. Powell, W.; Machray, G.C.; Provan, J. Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* **1996**, *1*, 215–222.
53. Tautz, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucl. Acid. Res.* **1989**, *17*, 6463–6471.
54. Rafalski, A. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **2002**, *5*, 94–100.
55. Batley, J.; Edwards, D. SNP applications in plants. In *Association Mapping in Plants*; Oraguzie, N., Rikkerink, E., Gardiner, S., De Silva, H., Eds.; Springer: New York, NY, USA, 2007; pp. 95–102.
56. Robinson, A.J.; Love, C.G.; Batley, J.; Barker, G.; Edwards, D. Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* **2004**, *20*, 1475–1476.
57. Barker, G.; Batley, J.; O’Sullivan, H.; Edwards, K.J.; Edwards, D. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* **2003**, *19*, 421–422.
58. Batley, J.; Barker, G.; O’Sullivan, H.; Edwards, K.J.; Edwards, D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* **2003**, *132*, 84–91.
59. Bhangale, T.R.; Stephens, M.; Nickerson, D.A. Automating resequencing-based detection of insertion-deletion polymorphisms. *Nat. Genet.* **2006**, *38*, 1457–1462.
60. Stephens, M.; Sloan, J.S.; Robertson, P.D.; Scheet, P.; Nickerson, D.A. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **2006**, *38*, 375–381.
61. Imelfort, M.; Duran, C.; Batley, J.; Edwards, D. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol. J.* **2009**, *7*, 312–317.
62. Youens-Clark, K.; Faga, B.; Yap, I.V.; Stein, L.; Ware, D. CMap 1.01: A comparative mapping application for the Internet. *Bioinformatics* **2009**, *25*, 3040–3042.
63. Duran, C.; Boskovic, Z.; Imelfort, M.; Batley, J.; Hamilton, N.A.; Edwards, D. CMap3D: A 3D visualisation tool for comparative genetic maps. *Bioinformatics* **2010**, *26*, 273–274.
64. Edwards, D.; Batley, J. Bioinformatics: Fundamentals and applications in plant genetics, mapping and breeding. In *Principles and Practices of Plant Genomics*; Kole, C., Abbott, A.G., Eds.; Science Publishers, Inc.: Enfield, NH, USA, 2008; pp. 269–302.
65. Edwards, D. Bioinformatics and plant genomics for staple crops improvement. In *Breeding Major Food Staples*; Kang, M.S., Priyadarshan, P.M., Eds.; Blackwell: Oxford, UK, 2007; pp. 93–106.
66. The international *Brachypodium* initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **2010**, *463*, 763–768.
67. Berkman, P.J.; Skarszewski, A.; Manoli, S.; Lorenc, M.T.; Stiller, J.; Smits, L.; Lai, K.; Campbell, E.; Kubalakova, M.; *et al.* Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **2012**, *124*, 423–432.

68. Berkman, B.J.; Skarszewski, A.; Lorenc, M.T.; Lai, K.; Duran, C.; Ling, E.Y.S.; Stiller, J.; Smits, L.; Imelfort, M.; Manoli, S.; *et al.* Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.* **2011**, *9*, 768–775.
69. Erwin, T.A.; Jewell, E.G.; Love, C.G.; Lim, G.A.C.; Li, X.; Chapman, R.; Batley, J.; Stajich, J.E.; Mongin, E.; Stupka, E.; *et al.* BASC: An integrated bioinformatics system for *Brassica* research. *Nucl. Acid. Res.* **2007**, *35*, D870–D873.
70. Love, C.G.; Robinson, A.J.; Lim, G.A.C.; Hopkins, C.J.; Batley, J.; Barker, G.; Spangenberg, G.C.; Edwards, D. Brassica ASTRA: An integrated database for *Brassica* genomic research. *Nucl. Acid. Res.* **2005**, *33*, D656–D659.
71. Stein, L.D.; Thierry-Mieg, J. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res.* **1998**, *8*, 1308–1315.
72. Berkman, P.J.; Lai, K.; Lorenc, M.T.; Edwards, D. Next generation sequencing applications for wheat crop improvement. *Amer. J. Bot.* **2012**, *99*, 365–371.
73. Edwards, D.; Batley, J., Plant Bioinformatics: From genome to phenome. *Trends Biotech.* **2004**, *22*, 232–237.
74. Duran, C.; Eales, D.; Marshall, D.; Imelfort, M.; Stiller, J.; Berkman, P.J.; Clark, T.; McKenzie, M.; Appleby, N.; Batley, J.; *et al.* Future tools for association mapping in crop plants. *Genome* **2010**, *53*, 1017–1023.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).