Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation

Chris Duran[1,2], Nikki Appleby[1,2], David Edwards*[,1,2] and Jacqueline Batley[1,2,3]

[1]*Australian Centre for Plant Functional Genomics, School of Land, Crop and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia;* [2]*Institute for Molecular Biosciences, University of Queensland, Brisbane, QLD 4072, Australia;* [3]*ARC Centre of Excellence for Integrative Legume Research, University of Queensland, Brisbane, QLD 4072, Australia*

Abstract: Molecular genetic markers represent one of the most powerful tools for the analysis of genomes and enable the association of heritable traits with underlying genomic variation. Molecular marker technology has developed rapidly over the last decade and two forms of sequence based marker, Simple Sequence Repeats (SSRs), also known as microsatellites, and Single Nucleotide Polymorphisms (SNPs) now predominate applications in modern genetic analysis. The reducing cost of DNA sequencing has led to the availability of large sequence data sets derived from whole genome sequencing and large scale Expressed Sequence Tag (EST) discovery that enable the mining of SSRs and SNPs, which may then be applied to diversity analysis, genetic trait mapping, association studies, and marker assisted selection. These markers are inexpensive, require minimal labour to produce and can frequently be associated with annotated genes. Here we review automated methods for the discovery of SSRs and SNPs and provide an overview of the diverse applications of these markers.

Keywords: Simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs), molecular markers, expressed sequence tags (ESTs), genome sequencing.

1. INTRODUCTION

Recent advances in molecular biology provide novel tools for addressing evolutionary, ecological and taxonomic research questions. Variation in DNA sequence can be observed with a level of accuracy and throughput that was previously impossible. The bulk of variation at the nucleotide level is often not visible at the phenotypic level. This DNA variation is frequently exploited in molecular genetic marker systems, and the application of molecular markers to advance research and commercial activities is now well established [1]. DNA based markers have many advantages over phenotypic markers in that they are highly heritable, relatively easy to assay, and are not affected by the environment. In modern genetic analysis, two sequence based marker systems, Single Nucleotide Polymorphisms (SNPs) and Simple Sequence Repeats (SSRs), now predominate.

1.1. Single Nucleotide Polymorphisms (SNPs)

The most abundant source of genetic polymorphism are SNPs, representing a single base change between two individuals at a defined location. There are three different categories of SNPs: transitions (C/T or G/A), transversions (C/G, A/T, C/A, or T/G) and small insertions/deletions (indels). SNPs at any particular site could in principle be bi-, tri- or tetra-allelic, however tri- and tetra-allelic SNPs are rare, and in practice SNPs are generally biallelic [2]. This disadvantage, when compared with multiallelic markers such as SSRs, is compensated by the relative abundance of SNPs, which can provide a high density of markers near a locus of interest. SNPs are evolutionarily stable, not changing significantly from generation to generation. This low mutation rate makes SNPs excellent markers for studying complex genetic traits and as a tool for understanding genome evolution [3]. SNPs are direct markers as the exact nature of the allelic variants is provided by the sequence information. This sequence variation can have a major impact on how the organism develops and responds to the environment.

SNPs are now the dominant marker used in biomedical applications due to the availability of the human genome sequence and knowledge of allelic variation derived from the HapMap project [4]. The ability to screen large numbers of individuals for a range of SNP variants enables the prediction of susceptibility to a wide range of diseases and opens the door to the use of personalised medicine based on the patients genotype. SNPs are becoming increasingly used in animal breeding, with particular success being derived from the bovine HapMap project [5]. It is expected that in crop genetics, SNPs will co-exist with other marker systems for several years [1, 6]. However, with the development of new technologies to increase throughput and reduce the cost of SNP development, along with further genome sequencing, the use of SNPs will become more widespread.

1.2. Simple Sequence Repeats (SSRs)

SSRs, also known as microsatellites, are stretches of DNA sequence consisting of short tandem repeats of mono-, di-, tri-, tetra-, penta- and hexa-nucleotides [7]. SSRs are widely distributed throughout genomes and have been found in all prokaryotic and eukaryotic genomes analysed to date [8, 9]. SSR perfect repeats are without interruptions, imper-

*Address correspondence to this author at the Australian Centre for Plant Functional Genomics, Institute for Molecular Biosciences and School of Land, Crop and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia; Tel: +61 (0)7 3346 2615; Fax: +61 (0)7 3346 2101; E-mail: Dave.Edwards@uq.edu.au

1574-8936/09 $55.00+.00

fect repeats are interrupted by non-repeat nucleotides and compound repeats are cases where two or more SSRs are found adjacent to one another. There may also be combinations of these, for example imperfect compound repeats [10]. SSRs are powerful genetic markers, due to their genetic co-dominance, abundance, dispersal throughout the genome, multi-allelic variation, high reproducibility and high level of polymorphism. This high level of polymorphism is due to mutation affecting the number of repeat units. SSRs provide a number of advantages over other molecular markers, namely that multiple SSR alleles may be detected at a single locus using a simple PCR based screen, very small quantities of DNA are required for screening, and analysis is amenable to automated allele detection and sizing [11]. 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.

Studies of the potential biological function and evolutionary relevance of SSRs is leading to a greater understanding of genomes and genomics [12]. SSRs were initially considered to be evolutionally neutral [13], however more recent evidence suggests they may play an important role in genome evolution [14] and provide hotspots of recombination. Functional roles have been attributed to some SSRs. They are thought to be involved in gene expression, regulation and function [15, 16], and have been found to bind nuclear proteins and function as transcriptional activating elements [17]. There is now evidence to suggest that SSRs in non-coding regions are also of functional significance [18].

2. MARKER APPLICATIONS

During the past two decades, several molecular marker technologies have been developed and applied for genome analysis, predominantly assessing the differences between individuals within a species and for the association of genomic regions with heritable traits. However, due to the relatively high cost associated with the development of molecular markers, these methods have only been applied to a limited number of species, predominantly in developed countries. Even in these situations, the application of molecular markers has tended to focus on a small number of traits or genomic regions. The development of association mapping methods demonstrates the requirement to be able to identify and screen large numbers of markers, rapidly and at low cost. The development of bioinformatics systems that improve marker identification with reducing cost will therefore broaden the uptake of markers to include more diverse species and a greater variety of traits. The SNP and SSR markers which can be rapidly and cheaply identified through bioinformatics have many uses in genetics, such as the detection of alleles associated with disease, genome mapping, association studies, genetic diversity, paternity assessment, forensics and inferences of population history [19, 20].

2.1. Genetic Diversity

Genetic diversity is the sequence variation within species. Information on genetic diversity and relationships among and between individuals, populations, plant varieties, animal breeds and species is of importance to plant and animal breeders for the improvement of crop plants and animal breeds, for conservation biology and for studying the evolutionary ecology of populations. Genetic diversity studies can identify alleles that might affect the ability of the organism to survive in its existing habitat, or might enable it to survive in more diverse habitats. This knowledge is valuable for germplasm conservation, individual, population, variety or breed identification.

SSR molecular markers are frequently used to assess genetic variation within and between populations [21] and there have been many studies describing genetic diversity in a wide range of species. SNPs within specific genes or genomic regions have also been used to infer phylogenetic relationships between species. However, the advent of next generation sequencing technology enables genetic diversity assessment on a genome wide scale. As the cost of genome sequencing continues to decrease the genomes of individuals will be sequenced rather than genotyped [22]. Whole genome studies of genetic diversity allow unprecedented insight into the forces contributing to genetic diversity in a species.

2.2. Genetic Mapping

Molecular markers have revolutionised genome mapping over the last two decades, offering the potential for generating very high density genetic maps that can be used to develop haplotypes for genes or regions of interest, and complete genome mapping is now becoming a reality. Genetic mapping places molecular genetic markers in linkage groups based on their co-segregation in a population. The genetic map predicts the linear arrangement of markers on a chromosome and maps are prepared by analysing populations derived from crosses of genetically diverse parents, and estimating the recombination frequency between genetic loci. Many types of markers can be used for map construction, with population size and marker density being important for map resolution. Genetic maps provide an insight into the genome organisation of an organism and may be used to study synteny between related species and rearrangement across taxa. The use of common molecular genetic markers across related species permits the comparison of linkage maps. This allows the translation of information between model species with sequenced genomes and non-model species [23]. Furthermore, the integration of molecular marker data with genomics, proteomics and phenomics data allows researchers to link sequenced genome data with observed traits, bridging the genome to phenome divide. In recent advances, genome wide sets of SNP markers have been developed in model plant and animal species, such as dog [24], rat [25] and *Arabidopsis thaliana* [26].

2.3. Association Studies

One aim of genetic studies is to associate the genotype with the heritable phenotype [27]. The quantitative pattern of inheritance of complex traits arises from the segregation of the alleles of multiple genes which are often modified by environmental factors. The systematic mapping of genes contributing to a continuously variable trait was not feasible before the use of molecular markers. The production of genetic linkage maps first enabled quantitative trait loci (QTL) to be mapped. Association mapping is a further statistical method to identify genetic loci associated with phenotypic trait variation. Association mapping shares much in common with QTL mapping. QTL mapping generally involves the

use of structured populations and relatively distant markers can segregate with the QTL, providing a wide genetic region within which the gene is located. The use of unstructured populations in association mapping means that they represent many more recombination events and are often many generations from a common ancestor, providing the potential of a greater resolution for a set population size.

2.4. Marker Assisted Selection

While unethical in human populations, genetic selection has contributed to increased productivity of crops and improvements in animal breeds. However, many of these genetic gains have been through traditional breeding methods involving phenotypic selection of traits or from pedigree and heterosis data. With the development of molecular techniques, marker assisted selection (MAS) is now used to enhance traditional breeding programs to improve both crops and animals, and modern plant and animal breeding is dependent on molecular markers for the rapid and precise analysis of germplasm and trait mapping [28]. Molecular markers are complementary tools to traditional selection, used to select parental genotypes in breeding programs, eliminate linkage drag in back-crossing and select for traits that are difficult to measure using phenotypic assays. They can increase our understanding of phenotypic characteristics and their genetic association, which may modify the breeding strategy. MAS allows the breeder to achieve early selection of a trait in a breeding program, and it is particularly useful when the trait is under complex genetic control, or when phenotypic trials are unreliable or expensive. By increasing favourable allele frequency early in the breeding process, a larger number of small populations can be carried forward in the breeding process, each of which has been prescreened to remove or reduce the frequency of unfavourable alleles.

2.5. Diagnostics

The association of molecular markers with human disease has led to the identification of genes and genetic mutations responsible for several heritable diseases such as sickle cell anaemia [29], cystic fibrosis [30], Huntington's disease [31] and phenylketonuria [32]. Research in this field has advanced rapidly since the sequencing of the human genome and the establishment of the human HapMap project which aims to catalogue a range of human genetic diversity [4]. Commercial companies now offer human genotyping services which predict genetic predisposition to disease and provide an insight into genetic ancestry [33]; (https://www.23andme.com/). Marker technology has also advanced the field of forensics with subsequent benefits to society through the crime prevention.

3. COMPUTATIONAL MOLECULAR MARKER DISCOVERY METHODS

As with most molecular markers, the factor limiting the implementation of SNPs and SSRs is the initial cost of their development. Previously, the discovery of SSR loci was limited to the construction of genomic DNA libraries enriched for SSR sequences, followed by DNA sequencing of the clones and analysis of the sequence for the presence of SSRs [34]. This process is both time consuming and expensive due to the large amount of specific sequencing required. SNP discovery involves finding differences between two sequences. Traditionally this has been performed through PCR amplification of genes/genomic regions of interest from multiple individuals selected to represent diversity in the species or population of interest, followed by either direct sequencing of these amplicons, or the more expensive method of cloning and sequencing. Sequences are then aligned and any polymorphisms identified. This approach is frequently prohibitively expensive and time consuming for the identification of the large number of SNPs required for most applications such as genetic mapping and association studies.

In silico methods of SNP and SSR discovery are now being adopted, providing cheap and efficient methods for marker identification. Large quantities of sequence data have been generated internationally through Expressed Sequence Tag (EST) or genome sequencing projects and these provide a valuable resource for the mining of molecular markers. Sequence data generation is undergoing a revolution with the release of 'next generation' technologies (Table **1**). These technologies offer the potential to rapidly re-sequence either whole eukaryotic genomes or representative samples of genomes. While the large volume of next generation sequencing data are generally produced at the expense of sequence quality, the over sampling of genome data enables the differentiation between true SNPs and sequence error. In one of the first examples of this application, a total of 36,000 maize SNPs were identified in data from a single run of the Roche 454 GS20 DNA sequencer [35]. More recently, the complete genome of DNA structure pioneer, James D. Watson was re-sequenced using Roche 454 technology, while an anonymous African male of the Yoruba people of Ibadan, who participated in the international HapMap project was completely sequenced using Applied Biosystems SOLiD sequencing technology. Whole genome sequencing is the most robust method to identify the great variety of genetic diversity in a population and gain a greater understanding of the

Table 1. Comparison of Current DNA Sequencing Technologies

Sequencing Machine	ABI 3730	Roche GSFLX	Illumina Solexa	AB SOLiD	Helicos HeliScope	Pacific Bioscience
Launched	2000	2007	2006	2007	2008	2009
Read length (bp)	800-1100	250-400	35-50	25-35	28	long
Reads per run	96	400 K	60 M	85 M	85 M	?
Throughput per run	0.1 MB	100 MB	3 GB	3 GB	2 GB	?
Cost per GB	>$2500k	$84k	$6k	$5.8k	?	?

relationship between the inherited genome and observed heritable traits. The continued rapid advances in genome sequencing technology will lead to whole genome sequencing becoming the standard method for genetic polymorphism discovery. For a comprehensive list of the genome sequencing initiatives, see http://www.ncbi.nlm.nih.gov/genomes/-static/gpstat.html. To date, there are over 1700 prokaryote genome sequencing projects and over 340 eukaryote genome sequencing projects. These numbers are set to increase rapidly with the expansion of next generation sequencing technology and this data will be used for rapid, inexpensive molecular marker discovery.

3.1. *In Silico* SNP Discovery

The dramatic increase in the number of DNA sequences submitted to databases makes the electronic mining of SNPs possible without the need for sequencing. The identification of sequence polymorphisms in assembled sequence data is relatively simple; the challenge of *in silico* SNP discovery is not SNP identification, but rather the ability to distinguish real polymorphisms from the abundant sequencing errors. Current Sanger sequencing produces errors as frequent as one error every one hundred base pairs, whilst some of the next generation technologies are even less accurate with errors as frequent as one in every 25 bp. Several sources of sequence error need to be addressed during *in silico* SNP identification. The most abundant error in Sanger sequencing is incorrect base calling, due to the requirement to obtain the greatest sequence length. These errors are usually identified by the relatively low quality scores for these nucleotides. Further errors are due to the intrinsically high error rate of the reverse transcription and PCR processes used for the generation of cDNA libraries and these errors are not reflected by poor sequence quality scores. A number of methods used to identify SNPs in aligned sequence data rely on sequence trace file analysis to filter out sequence errors by their dubious trace quality [36-38]. The major drawback to this approach is that the sequence trace files required are rarely available for large sequence datasets collated from a variety of sources. In cases where trace files are unavailable, two complementary approaches have been adopted to differentiate between sequence errors and true polymorphisms: (1) assessing redundancy of the polymorphism in an alignment, and (2) assessing co-segregation of SNPs to define a haplotype. These methods are employed in the following applications for *in silico* SNP identification (Table **2**).

3.1.1. SNP Discovery from Trace Files

PolyBayes

PolyBayes [38] uses a Bayesian-statistical model to find differences within assembled sequences based on the depth of coverage, the base quality values and the expected rate of polymorphic sites in the region. Base quality values can be obtained by running the sequence trace files through the PHRED base-calling program [39, 40], and repeats can be removed from sequences using RepeatMasker [41]. The output can be viewed through the Consed alignment viewer [42]. Recent studies using PolyBayes include SNP discovery for white spruce [43] and bird species [44].

PolyPhred

PolyPhred [45] compares sequence trace files from different individuals to identify heterozygous sites. The sequence trace files are used to identify SNPs and can identify positions in the sequence where double peaks occur that are half the height of the adjacent peaks within a window. The quality of a SNP is assigned based on the spacing between peaks; the relative size of called and uncalled peaks; and the dip between peaks. PolyPhred only analyses nucleotides that have a minimum quality as determined by Phred [39, 40]. PolyPhred is integrated with three other programs: phred, phrap and consed. It runs on Unix, and provides output that

Table 2. Applications for in Silico SNP and SSR Discovery

Tool	URL	Reference
PolyBayes	http://bioinformatics.bc.edu/marthlab/PolyBayes	[38]
PolyPhred	http://droog.mbt.washington.edu/	[45]
SNPDetector	http://lpg.nci.nih.gov/	[48]
NovoSNP	http://www.molgen.ua.ac.be/bioinfo/novosnp/	[50]
AutoSNP	http://acpfg.imb.uq.edu.au	[53]
MISA	http://pgrc.ipk-gatersleben.de/misa/	[63]
SSRIT	http://www.gramene.org/db/searches/ssrtool	[66]
RepeatFinder	http://www.cbcb.umd.edu/software/RepeatFinder/	[67]
SPUTNIK	http://espressosoftware.com/pages/sputnik.jsp http://cbi.labri.fr/outils/Pise/sputnik.html)	Unpublished
TROLL	http://wsmartins.net/webtroll/troll.html	[71]
TRF	http://tandem.bu.edu/trf/trf.html	[72]
SSRPrimer	http://hornbill.cspp.latrobe.edu.au http://acpfg.imb.uq.edu.au	[75, 76]
SSRPoly	http://acpfg.imb.uq.edu.au/ssrpoly.php	Unpublished

can be viewed in Consed [42]. Recent examples of the use of PolyPhred include studies in cattle [46] and in humans that have had liver transplants [47].

SNPDetector

SNPDetector [48] uses Phred [39, 40] to call bases and determine quality scores from trace files, and then aligns reads to a reference sequence using a Smith-Waterman algorithm. SNPs are identified where there is a sequence difference and the flanking sequence is of high quality. SNPDetector has been used to find SNPs in 454 data [35] and has been included within a comprehensive SNP discovery pipeline [49].

NovoSNP

NovoSNP [50] requires both trace files and a reference sequence as input. The trace files are base-called using Phred [39, 40] and quality clipped, then aligned to a reference sequence using BLAST [51]. A SNP confidence score is calculated for each predicted SNP. NovoSNP is written in Tcl with a graphical user interface written in Tk and runs on Linux and Windows. NovoSNP has been used in a study of genotype-phenotype correlation for human disease [52].

3.1.2. SNP Discovery Using Redundancy Approaches

AutoSNP

The autoSNP method [53] assembles sequences using CAP3 [54] with the option of pre-clustering with either d2cluster [55] or TGICL [56]. Redundancy is the principle means of differentiating between sequence errors and real SNPs. While this approach ignores potential SNPs that are poorly represented in the sequence data, it offers the advantage that trace files are not required and sequences may be used directly from GenBank. AutoSNP is therefore applicable to any species for which sequence data is available. A co-segregation score is calculated based on whether multiple SNPs define a haplotype, and this is used as a second, independent measure of confidence. AutoSNP is written in Perl and is run from the Linux command line with a FASTA file of sequences as input. The output is presented as linked HTML with the index page presenting a summary of the results. AutoSNP has been applied to several species including maize [57], peach [58] and cattle [59].

SNPServer

SNPServer [60] is a real time implementation of the autoSNP method, accessed via a web server. A single FASTA sequence is pasted into the interface and similar sequences are retrieved from a nucleotide sequence database using BLAST [51]. The input sequence and matching sequences are assembled using CAP3 and SNPs are discovered using the autoSNP method [53]. The results are presented as HTML. Alternatively, a list of FASTA sequences may be input for assembly or a preassembled ACE format file may be analysed. SNPServer has been used in studies including sea anemone [61] and human [62].

3.2. SSR Discovery

The availability of large quantities of sequence data makes it economical and efficient to use computational tools to mine this for SSRs. Flanking DNA sequence may then be used to design suitable forward and reverse PCR primers to assay the SSR loci. Furthermore, when SSRs are derived from ESTs, they become gene specific and represent functional molecular markers. These features make EST-SSRs highly valuable markers for the construction and comparison of genetic maps and the association of markers with heritable traits. Several computational tools are available for the identification of SSRs in sequence data as well as for the design of PCR amplification primers. Due to redundancy in EST sequence data, and with datasets often being derived from several distinct individuals, it is now also possible to predict the polymorphism of SSRs *in silico*. A selection of SSR discovery tools are described below (Table **2**).

MISA

The MIcroSAtellite (MISA) tool (http://pgrc.ipk-gatersleben.de/misa/) identifies perfect, compound and interrupted SSRs. It requires a set of sequences as FASTA and a parameter file that defines unit size and minimum repeat number of each SSR. The output includes a file containing the table of repeats found, and a summary file. MISA can also design PCR amplification primers on either side of the SSR. The tool is written in Perl and is therefore platform independent, but it requires an installation of Primer3 for the primer search [63]. MISA has been applied for SSR identification in moss [64] and coffee [65].

SSRIT

The tool SSRIT (Simple Sequence Repeat Identification Tool) (http://www.gramene.org/db/searches/ssrtool) uses Perl regular expressions to find perfect SSR repeats within a sequence. It can detect repeats between 2 and 10 bases in length, but eliminates mononucleotide repeats. The output is a file of SSRs in tabular format. A web based version is available that will take a single sequence, and a stand alone version is also available for download. SSRIT has been applied to rice [66].

RepeatFinder

RepeatFinder [67] (http://www.cbcb.umd.edu/software/-RepeatFinder/) finds SSRs in four steps: (1) finds all exact repeats using RepeatMatch or REPuter [68]; (2) merges repeats together into repeat classes, for example repeats that overlap; (3) merges all of the other repeats that match those already merged, into the same classes and (4) matches all repeats and classes against each other in a non-exact manner using BLAST. The input is a genome or set of sequences, and the output is a file containing the repeat classes and number of merged repeats found in each class. RepeatFinder finds perfect, imperfect and compound repeats, and was not designed specifically to find SSRs so can find repeats of any length. It runs on Unix or Linux and has been used to identify SSRs in peanut [69].

Sputnik

Sputnik is a commonly used SSR finder as it is fast, efficient and simple to use. It uses a recursive algorithm to search for repeats with length between 2 and 5, and it finds perfect, compound and imperfect repeats. It requires sequences in FASTA format and uses a scoring system to call each SSR. The output is a file of SSRs in tabular format. Unix, Linux and windows versions of sputnik are available from http://espressosoftware.com/pages/sputnik.jsp and

http://cbi.labri.fr/outils/Pise/sputnik.html (PISE enabled version). Sputnik has been applied for SSR identification in many species including Arabidopsis and barley [70]

TROLL

The SSR identification tool Tandem Repeat Occurrence Locator (TROLL) [71] (http://wsmartins.net/webtroll/troll.-html) draws a keyword tree and matches it with a technique adapted from bibliographic searches, based on the *Aho-Corasick* algorithm. It has drawbacks in that it doesn't handle very large sequences and cannot process large batches of sequences as the tree takes up large amounts of memory.

Tandem Repeats Finder (TRF)

Tandem Repeats Finder (TRF) [72] (http://tandem.bu.-edu/trf/trf.html) can find very large SSR repeats, up to a length of 2000 bp. It uses a set of statistical tests for reporting SSRs, which are based on four distributions of the pattern length, the matching probability, the indel probability and the tuple size. TRF finds perfect, imperfect and compound SSRs, and is available for Linux. TRF has been used for SSR identification in Chinese shrimp [73] and cowpea [74].

3.2.1. Compound Methods

The following computational SSR finders combine previously created methods to produce extended output.

SSRPrimer

SSRPrimer [75, 76] combines Sputnik and the PCR primer design software Primer3 to find SSRs and associated amplification primers. The scripts take multiple sequences in FASTA format as input and produce lists of SSRs and associated PCR primers in tabular format. This web-based tool is also available as a stand alone version for very large datasets. SSRPrimer has been applied to a wide range of species including *Brassica* [77-80], citrus [81], mint [82], strawberry [83], *Eragrostis curvula* [84], *Sclerotinia* [85] and shrimp [86].

SSRPoly

SSRPoly (http://acpfg.imb.uq.edu.au/ssrpoly.php) is currently the only tool which is capable of identifying polymorphic SSRs from DNA sequence data. The input is a file of FASTA format sequences. SSRPoly includes a set of Perl scripts and MySQL tables that can be implemented on UNIX, Linux and Windows platforms.

4. DATA STORAGE

Large-scale discovery projects are uncovering vast quantities of marker data. As the data size increases, the storage and logical organisation of the information becomes an important challenge. Marker databases vary between centralised repositories that integrate a variety of data for several species, to small specialised databases designed for very specific purposes. The larger repositories tend to lack detailed analytic tools, while the smaller systems may include further species specific data integration. dbSNP is becoming the default repository for SNP data, and there are a wide variety of additional marker databases specific to particular species. The most commonly used marker databases are detailed below (Table **3**).

dbSNP

The Single Nucleotide Polymorphism database, dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi), was developed by NCBI to provide a public-domain repository for simple genetic polymorphisms [87, 88]. Although dbSNP includes data on markers such as SSRs and insertion/deletion polymorphisms, SNPs are the primary data type, comprising 97.8% of the database [87]. Table (**4**) presents a summary of species represented. Access is provided via a web interface and there are several ways to query the database. Users can search using a known SNP id or use BLAST to compare a known sequence with sequences in the database. Alternatively, dbSNP can be queried using Entrez or Locuslink. dbSNP currently hosts over 52 million refSNP clusters for 44 organisms. Of these clusters, around 16 million (30%) have been validated.

HapMap

The HapMap Consortium collates and catalogues information on human genetic polymorphisms [89]. There are two primary methods to access the data: GBrowse [90] and Bio-Mart [91], and both methods are tailored to specific types of users. GBrowse is a genome browser and is a component of the GMOD project (http://www.gmod.org). Using the GBrowse feature of HapMap, users may browse a region of the genome or search with a specific SNP id (Fig. **1**).

Table 3. Details of Commonly Used Marker Storage Databases

Database	URL	Reference
dbSNP	www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi	[87, 88]
HapMap	www.hapmap.org/	[89]
IBISS	www.livestockgenomics.csiro.au/ibiss/	[90]
MPD SNP Tools	www.jax.org/phenome	[91]
Gramene	www.gramene.org/	[92-94]
GrainGenes	www.graingenes.org/	[95, 96]
TAIR	www.arabidopsis.org/	[97-99]
MaizeGDB	www.maizegdb.org/	[100]
AutoSNPdb	http://acpfg.imb.uq.edu.au/	Unpublished

Table 4. List of Species Represented in dbSNP that have over 500 Validated SNPs

Organism	dbSNP Build	Genome Build	Number of Submissions	Number of RefSNP Clusters (# validated)
Homo sapiens	129	36.3	50529995	18,045,964 (6,587,300)
Mus musculus	128	37.1	18645060	14,380,528 (6,447,366)
Gallus gallus	128	2.1	3641959	3,293,383 (3,280,002)
Oryza sativa	128	4.1	5872081	5,418,373 (22,057)
Canis familiaris	126	2.1	3526996	3,301,322 (217,525)
Bos taurus	128	3.1	2233086	2,223,033 (14,371)
Pan troglodytes	127	0.0	1544900	1,543,217 (112,654)
Danio rerio	128	2.1	700855	662,322 (3,091)
Rattus norvegicus	126	4.1	47711	43,628 (1,605)
Macaca mulatta	128	1.1	789	780 (519)

Clicking the SNP location in the GBrowse viewer opens an information page, providing full details of the SNP locus. The HapMap project maintains over 3.1 million characterised human SNPs which have been genotyped in a geographically diverse selection of 270 individuals [92].

IBISS

The Interactive Bovine *in silico* SNP Database (IBISS) has been created by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) of Australia. It is a collection of 523 448 Bovine SNPs identified from 324,031 Bovine ESTs using a custom analysis pipeline [93]. The da-

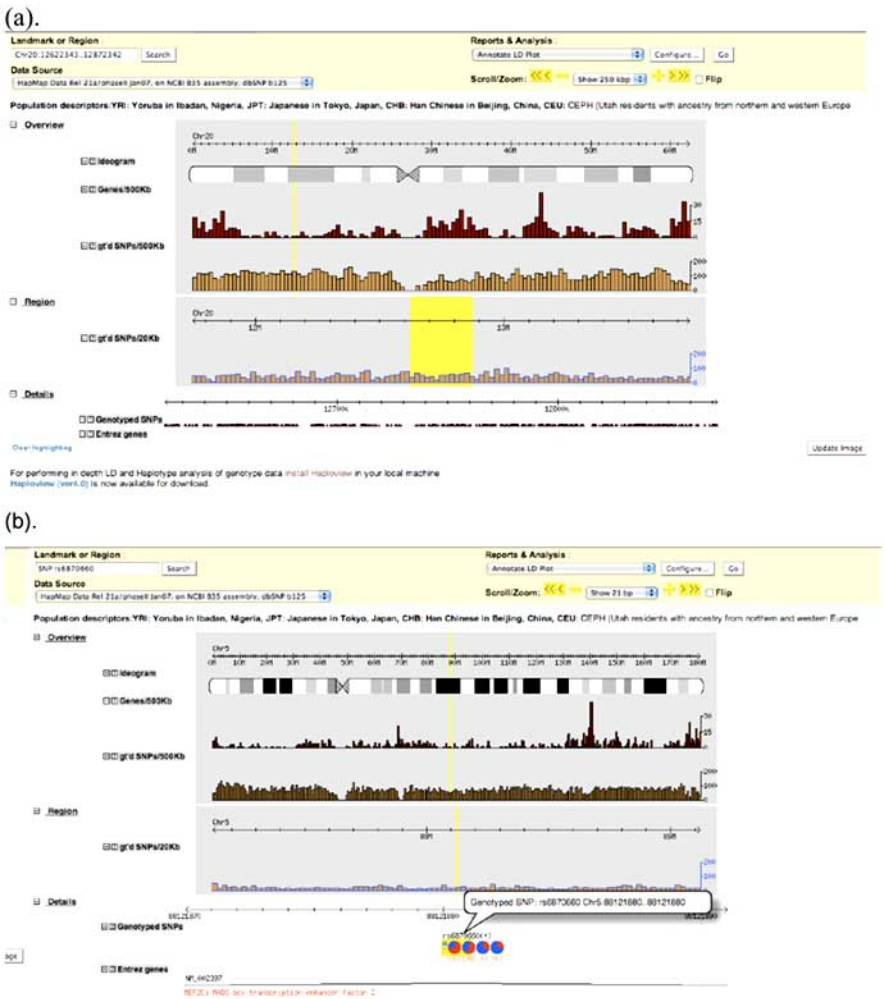

Fig. (1). Examples of HapMap searches: (**a**). Chromosomal region search, showing the population of genotyped SNPs as a custom GBrowse track (bottom). (**b**). SNP id search, showing the specific genome location for the desired SNP.

tabase can be searched by keyword, accession id or by BLAST comparison with an entry sequence. Users can also browse for markers using a linked genome browser.

MPD SNP Tools

The Jackson Laboratory's Mouse Phenome Database (MPD) (www.jax.org/phenome) aims to facilitate the research of human health issues through mouse models. As well as a wealth of trait information on mice, MPD also hosts a collection of over 10 million mouse SNPs (http://-www.jax.org/phenome/snp.html).

Gramene

Gramene is an online comparative mapping database for rice and related grass species [94-96]. Gramene contains information on cereal genomic and EST sequences, genetic maps, relationships between maps, details of rice mutants, and molecular genetic markers. The database uses the sequenced rice genome as its reference and annotates this genome with various data types. As well as the genome browser, Gramene also incorporates a version of the comparative map viewer, CMap. This allows users to view genetic maps and comparative genetic mapping information and provides a link between markers on genetic maps and the sequenced genome information.

GrainGenes

GrainGenes integrates genetic data for Triticeae and Avena [97, 98]. The database includes genetic markers, map locations, alleles and key references for barley, wheat, rye, oat and related wild species. Graingenes also provides access to genetic data using CMap.

TAIR

The Arabidopsis Information Resource (TAIR) (http://-www.arabidopsis.org/) provides an extensive web-based resource for the model plant *Arabidopsis thaliana* [99-101]. Data includes gene, marker, genetic mapping, protein sequence, gene expression and community data within a relational database.

MaizeGDB

MaizeGDB [102] combines information from the original MaizeDB and ZmDB [103, 104] repositories with sequence data from PlantGDB [105-107]. The system maintains information on maize genomic and gene sequences, genetic markers, literature references, as well as contact information for the maize research community.

AutoSNPdb

AutoSNPdb implements the autoSNP pipeline within a relational database to enable the efficient mining of the identified SNP and indel polymorphisms and the detailed interrogation of the data. A web-based application enables searching and visualisation of the data, including the display of sequence alignments and SNPs (Fig. **2**). All sequences are annotated by comparison with GenBank and UniRef90, as well as through comparison with reference genome sequences. The system allows researchers to query the results of SNP analysis to identify SNPs between specific groups of individuals or within genes of predicted function. AutoSNPdb is currently available for barley, rice and *Brassica* species and is available at: http://acpfg.imb.uq.edu.au/.

5. DATA VISUALISATION

The effective visualisation of large amounts of data is as critical an issue as its storage. Increasing volumes of data permit researchers to draw, with increasing confidence, comparative links across the genome to phenome divide. Visualisation tools, combined with the ability to dynamically categorise data, allow the identification trends and relationships at varying tiers of resolution. Current visualisation techniques for markers broadly fall into two categories: graphical map viewers and genome browsers. Map viewers display markers as a representation of a genetic linkage map. Genome browsers generally host a greater quantity of annotation data and may be linked to related genetic map viewers.

5.1. Graphical Map Viewers

The NCBI map viewer (http://www.ncbi.nih.gov/map-view) uses sets of graphically-aligned maps to visualise molecular genetic markers, genome assemblies and other anno-

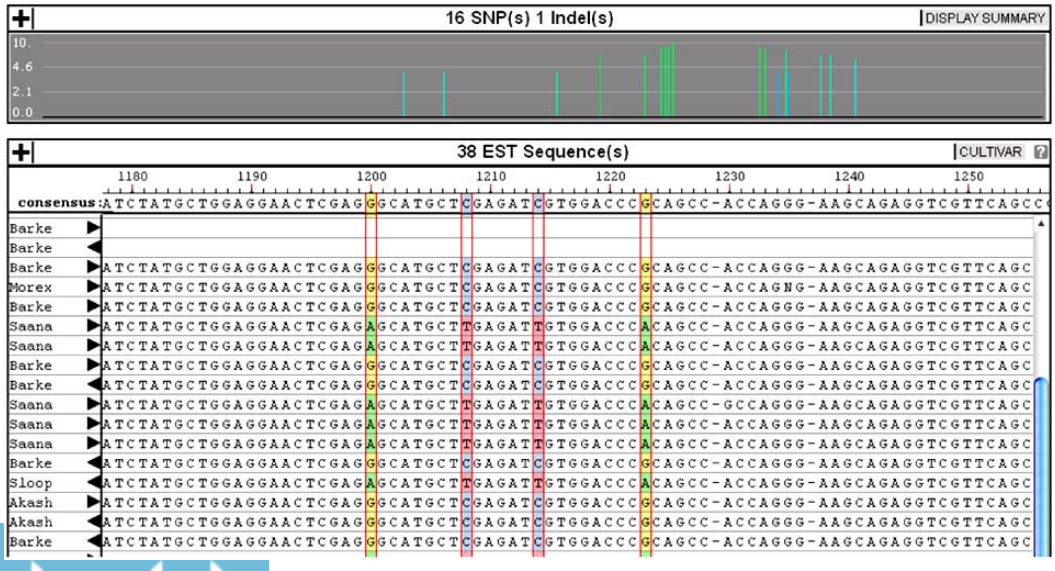

Fig. (2). AutoSNPdb showing the overview of the SNPs in this assembly and the aligned sequences with the SNPs highlighted.

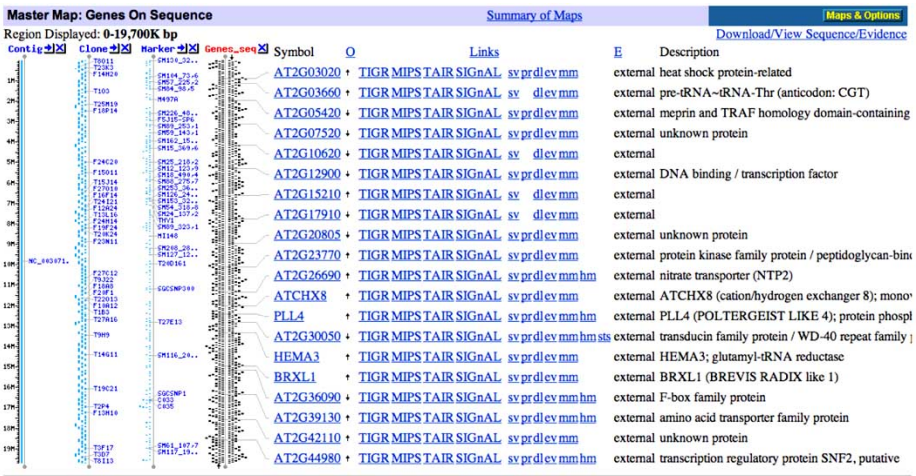

Fig. (3). The NCBI map viewer displaying the overall view for *Arabidopsis thaliana* chromosome 2.

tations [108]. It allows users to show multiple levels of annotation in tandem for a given chromosomal segment (Fig. **3**). As well as allowing users to view the map graphically, NCBI also provides a function to download raw mapping data in a tabular format.

CMap is a tool for viewing and comparing genetic and physical maps and has been applied successfully for the comparison of maps within and between related species [94]. CMap was originally developed for the Gramene project (http://www.gramene.org/CMap/) and has since been applied for the comparison of genetic maps of *Brassica* [109], sheep, cattle, pig, wallaby [110], honeybee, grasses [94, 111], peanut [112], Rosaceae [113] and legumes [114]. As an extension to CMap; CMap3D allows researchers to compare multiple genetic maps in three-dimensional space (Fig. **4**). CMap3D accesses data from CMap databases, with specifications defined by the Generic Model Organism Database (GMOD) (http://www.gmod.org/CMap).

5.2. Genome Browsers

Several software packages have been developed for the visualisation of genome information. EnsEMBL was developed by the European Bioinformatics Institute (EBI) and the Sanger Centre to visualise data from the Human Genome Project [115, 116]. It has since been extended to a variety of eukaryotic organisms, including plants [94, 117]. In contrast, GBrowse was designed to be a generic genome browser [90] which has been applied to the genomes of a wide variety of species including *C. Elegans*, Drosophila, Honeybee, Cattle and Human. The UCSC Genome browser is another popular browser which has been developed principally for vertebrate genomes [118].

EnsEMBL, GBrowse and the UCSC Genome Browser display annotations as customisable 'tracks' along selected regions of genome sequence. One advantage of displaying data in this format is that relationships between markers, predicted gene structures, trait annotations and other forms

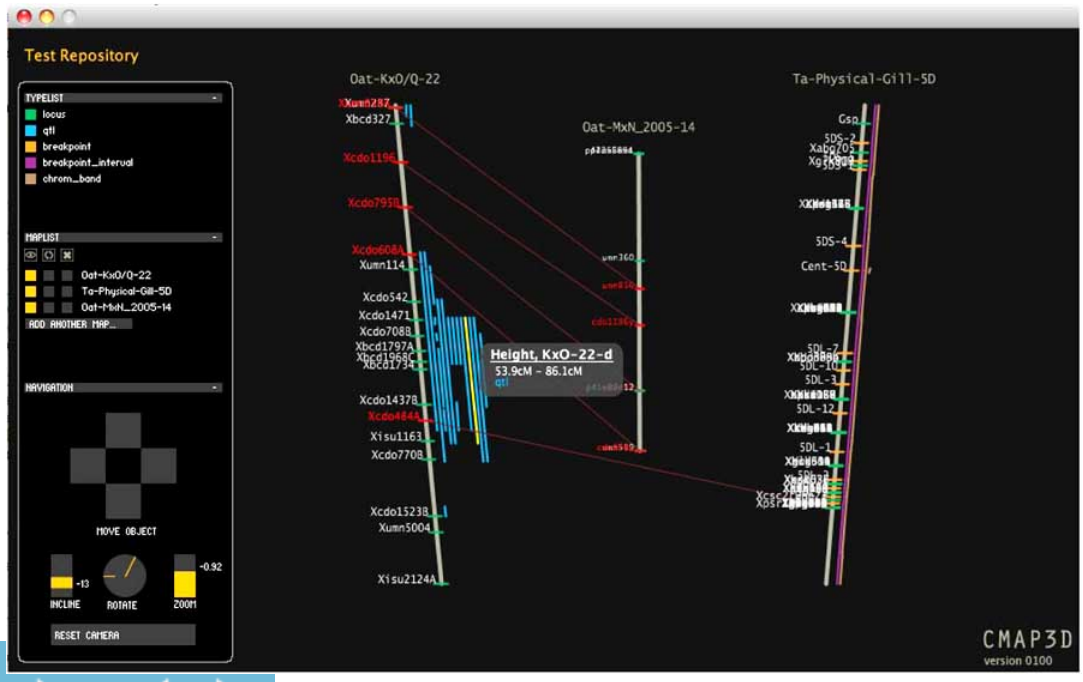

Fig. (4). The CMap3D map viewer displaying associations between three linkage groups and mapped traits.

of biological annotation can be viewed in a clear and intuitive manner.

CONCLUDING REMARKS

Genetic markers have played a major role in our understanding of heritable traits. In the current genomics era, molecular genetic markers are bridging the divide between these traits and increasingly available genome sequence information. Conversely, the increasing quantity of genome sequence information is a valuable source of new genetic markers. Bioinformatics tools have been developed to mine sequence data for markers and present these in a biologist friendly manner. With the expansion of next generation sequencing technologies, there will be a rapid growth in associated marker information and the use of these markers for diverse applications from crop breeding to predicting human disease risks, impacting both food production and human health for future generations.

REFERENCES

[1] Gupta PK, Roy JK, Prasad M. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr Sci* **2001**; 80: 524-535.

[2] Doveri S, Lee D, Maheswaran M, Powell W (2008). Molecular markers: History, features and applications. In Principles and Practices of Plant Genomics, Volume 1, C.K.a.A.G. Abbott, ed. (Enfield, USA: Science Publishers), pp. 23-68.

[3] Syvanen AC. Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2001**; 2: 930-942.

[4] Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. *Nature* **2005**; 437: 1299-1320.

[5] Khatkar MS, Zenger KR, Hobbs M, *et al.* A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle. *Genetics* **2007**; 176: 763-772.

[6] Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* **2002**; 5: 94-100.

[7] Tautz D, Schlotterer C. Concerted Evolution, Molecular Drive and Natural-Selection - Reply. *Curr Bio* **1994**; 4: 1166-1166.

[8] Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evolut* **2001**; 18: 1161-1167.

[9] Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res* **2000**; 10: 967-981.

[10] Weber JL. Informativeness of Human (Dc-Da)N.(Dg-Dt)N Polymorphisms. *Genomics* **1990**; 7: 524-530.

[11] Schlotterer C. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **2000**; 109: 365-371.

[12] Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **2003**; 4: R13.

[13] Awadalla P, Ritland K. Microsatellite variation and evolution in the Mimulus guttatus species complex with contrasting mating systems. *Mol Biol and Evolut* **1997**; 14: 1023-1034.

[14] Moxon ER, Wills C. DNA microsatellites: Agents of evolution? *Sci Am* **1999**; 280: 94-99.

[15] Gupta M, Chyi YS, Romeroseverson J, Owen JL. Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. *Theor Appl Genet* **1994**; 89: 998-1006.

[16] Kashi Y, King D, Soller M. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* **1997**; 13: 74-78.

[17] Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* **2002**; 11: 2453-2465.

[18] Mortimer J, Batley J, Love C, Logan E, Edwards D. Simple Sequence Repeat (SSR) and GC distribution in the Arabidopsis thaliana genome. *J of Plant Biotechnol* **2005**; 7: 17-25.

[19] Brumfield RT, Beerli P, Nickerson DA, Edwards SV. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecol Evol* **2003**; 18: 249-256.

[20] Collins A, Lau W, De la Vega FM. Mapping genes for common diseases: The case for genetic (LD) maps. *Hum Heredity* **2004**; 58: 2-9.

[21] Vigouroux Y, Mitchell S, Matsuoka Y, *et al.* An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* **2005**; 169: 1617-1630.

[22] Bennett ST, Barnes C, Cox A, Davies L, Brown C. Toward the $1000 human genome. *Pharmacogenomics* **2005**; 6: 373-382.

[23] Moore G, Devos KM, Wang Z, Gale MD. Cereal Genome Evolution - Grasses, Line up and Form a Circle. *Curr Biol* **1995**; 5: 737-739.

[24] Lindblad-Toh K, Wade CM, Mikkelsen TS, *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**; 438: 803-819.

[25] Nijman IJ, Kuipers S, Verheul M, Guryev V, Cuppen E. A genome-wide SNP panel for mapping and association studies in the rat. *BMC Genomics* **2008**; 9.

[26] Schmid KJ, Sorensen TR, Stracke R, *et al.* Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana. Genome Res* **2003**; 13: 1250-1257.

[27] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **2003**; 33: 228-237.

[28] Koebner R, Summers R. The impact of molecular markers on the wheat breeding paradigm. *Cell Mol Biol Lett* **2002**; 7: 695-702.

[29] Wu DY, Ugozzoli L, Pal BK, Wallace RB. Allele-Specific Enzymatic Amplification of Beta-Globin Genomic DNA for Diagnosis of Sickle-Cell Anemia. *Proc Natl Acad Sci USA* **1989**; 86: 2757-2760.

[30] Rommens JM, Iannuzzi MC, Kerem BS, *et al.* Identification of the Cystic-Fibrosis Gene - Chromosome Walking and Jumping. *Science* **1989**; 245: 1059-1065.

[31] Andrew SE, Goldberg YP, Kremer B, *et al*. The Relationship between Trinucleotide (CAG) Repeat Length and Clinical-Features of Huntingtons-Disease. *Nat Genet* **1993**; 4: 398-403.

[32] Konecki DS, Lichterkonecki U. The Phenylketonuria Locus - Current Knowledge About Alleles and Mutations of the Phenylalanine-Hydroxylase Gene in Various Populations. *Hum Genet* **1991**; 87: 377-388.

[33] Amos J, Patnaik M. Commercial molecular diagnostics in the US: The human genome project to the clinical laboratory. *Hum Mut* **2002**; 19: 324-333.

[34] Edwards KJ, Barker JHA, Daly A, Jones C, Karp A. Microsatellite libraries enriched for several microsatellite sequences in plants. *Biotechniques* **1996**; 20: 758-&.

[35] Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. *Plant J* **2007**; 51: 910-918.

[36] Garg K, Green P, Nickerson DA. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res* **1999**; 9: 1087-1092.

[37] Kwok PY, Carlson C, Yager TD, Ankener W, Nickerson DA. Comparative-Analysis of Human DNA Variations by Fluorescence-Based Sequencing of PCR Products. *Genomics* **1994**; 23: 138-144.

[38] Marth GT, Korf I, Yandell MD, *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat Genet* **1999**; 23: 452-456.

[39] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **1998**; 8: 175-185.

[40] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **1998**; 8: 186-194.

[41] Mallon AM, Strivens M. DNA sequence analysis and comparative sequencing. *Methods* **1998**; 14: 160-178.

[42] Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. *Genome Res* **1998**; 8: 195-202.

[43] Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J. Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. *BMC Genomics* **2006**; 7: (06 July 2006).

[44] Sironi L, Lazzari B, Ramelli P, Gorni C, Mariani P. Single nucleotide polymorphism discovery in the avian Tapasin gene. *Poult Sci* **2006**; 85: 606-612.

[45] Nickerson DA, Tobe VO, Taylor SL. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* **1997**; 25: 2745-2751.

[46] Lee SH, Park EW, Cho YM, *et al*. Confirming single nucleotide polymorphisms from expressed sequence tag datasets derived from three cattle cDNA libraries. *J Biochem Mol Biol* **2006**; 39: 183-188.

[47] Wang WL, Zhang GL, Wu LH, *et al.* Efficacy of hepatitis B immunoglobulin in relation to the gene polymorphisms of human leukocyte Fc gamma receptor III (CD16) in Chinese liver transplant patients. *Chinese Med J* **2007**; 120: 1606-1610.

[48] Zhang JH, Wheeler DA, Yakub I, *et al.* SNPdetector: A software tool for sensitive and accurate SNP detection. *Plos Comput Biol* **2005**; 1: 395-404.

[49] Matukumalli LK, Grefenstette JJ, Hyten DL, Choi I-Y, Cregan PB, Van Tassell CP. SNP-PHAGE - High throughput SNP discovery pipeline. *BMC Bioinform* **2006**; 7: Article No.: 468.

[50] Weckx S, Del-Favero J, Rademakers R, *et al.* novoSNP, a novel computational tool for sequence variation discovery. *Genome Res* **2005**; 15: 436-442.

[51] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J Mol Biol* **1990**; 215: 403-410.

[52] Dierick I, Baets J, Irobi J, *et al.* Relative contribution of mutations in genes for autosomal dominant distal hereditary motor neuropathies: a genotype-phenotype correlation study. *Brain* **2008**; 131: 1217-1227.

[53] Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* **2003**; 19: 421-422.

[54] Huang XQ, Madan A. CAP3: A DNA sequence assembly program. *Genome Res* **1999**; 9: 868-877.

[55] Burke J, Davison D, Hide W. d2_cluster: A validated method for clustering EST and full-length cDNA sequences. *Genome Res* **1999**; 9: 1135-1142.

[56] Pertea G, Huang XQ, Liang F, *et al*. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **2003**; 19: 651-652.

[57] Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol* **2003**; 132: 84-91.

[58] Lazzari B, Caprera A, Vecchietti A, Stella A, Milanesi L, Pozzi C. ESTree db: a tool for peach functional genomics. *BMC Bioinform* **2005**; 6.

[59] Corva P, Soria L, Schor A, *et al.* Association of CAPN1 and CAST gene polymorphisms with meat tendemess in Bos taurus beef cattle from Argentina. *Genet Mol Biol* **2007**; 30: 1064-1069.

[60] Savage D, Batley J, Erwin T, *et al*. SNPServer: a real-time SNP discovery tool. *Nucleic Acids Res* **2005**; 33: W493-W495.

[61] Sullivan JC, Reitzel AM, Finnerty JR. Upgrades to StellaBase facilitate medical and genetic studies on the starlet sea anemone, Nematostella vectensis. *Nucleic Acids Res* **2008**; 36: D607-D611.

[62] Pumpernik D, Oblak B, Borstnik B. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Mol Genet Genomics* **2008**; 279: 53-61.

[63] Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet* **2003**; 106: 411-422.

[64] von Stackelberg M, Rensing SA, Reski R. Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biol* **2006**; 6: 9.

[65] Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor Appl Genet* **2007**; 114: 359-372.

[66] Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* **2001**; 11: 1441-1452.

[67] Volfovsky N, Haas BJ, Salzberg SL. A clustering method for repeat analysis in DNA sequences. *Genome Biol* **2001**; 2.

[68] Kurtz S, Schleiermacher C. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **1999**; 15: 426-427.

[69] Jayashree B, Ferguson M, Ilut D, Doyle J, Crouch JH. Analysis of genomic sequences from peanut (Arachis hypogaea). *Electron J Biotechnol* **2005**; 8.

[70] Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* **2000**; 156: 847-854.

[71] Castelo AT, Martins W, Gao GR. TROLL-Tandem Repeat Occurrence Locator. *Bioinformatics* **2002**; 18: 634-636.

[72] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **1999**; 27: 573-580.

[73] Gao H, Kong J. The microsatellites and minisatellites in the genome of Fenneropenaeus chinensis. *DNA Seq* **2005**; 16: 426-436.

[74] Chen XF, Laudeman TW, Rushton PJ, Spraggins TA, Timko MP. CGKB: an annotation knowledge base for cowpea (Vigna unguiculata L.) methylation filtered genomic genespace sequences. *BMC Bioinformatics* **2007**; 8.

[75] Jewell E, Robinson A, Savage D, *et al*. SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery. *Nucleic Acids Res* **2006**; 34: W656-W659.

[76] Robinson AJ, Love CG, Batley J, Barker G, Edwards D. Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* **2004**; 20: 1475-1476.

[77] Batley J, Hopkins CJ, Cogan NOI, *et al*. Identification and characterization of simple sequence repeat markers from *Brassica napus* expressed sequences. *Mol Ecol Notes* **2007**; 7: 886-889.

[78] Burgess B, Mountford H, Hopkins CJ, *et al.* Identification and characterization of simple sequence repeat (SSR) markers derived in silico from *Brassica oleracea* genome shotgun sequences. *Mol Ecol Notes* **2006**; 6: 1191-1194.

[79] Hopkins CJ, Cogan NOI, Hand M, *et al*. Sixteen new simple sequence repeat markers from *Brassica juncea* expressed sequences and their cross-species amplification. *Mol Ecol Notes* **2007**; 7: 697-700.

[80] Ling AE, Kaur J, Burgess B, *et al*. Characterization of simple sequence repeat markers derived *in silico* from *Brassica rapa* bacterial artificial chromosome sequences and their application in Brassica napus. *Mol Ecol Notes* **2007**; 7: 273-277.

[81] Chen CX, Zhou P, Choi YA, Huang S, Gmitter FG. Mining and characterizing microsatellites from citrus ESTs. *Theor Appl Genet* **2006**; 112: 1248-1257.

[82] Lindqvist C, Scheen AC, Yoo MJ, *et al*. An expressed sequence tag (EST) library from developing fruits of an Hawaiian endemic mint (Stenogyne rugosa, Lamiaceae): characterization and microsatellite markers. *BMC Plant Biol* **2006**; 6: 16.

[83] Keniry A, Hopkins CJ, Jewell E, *et al*. Identification and characterization of simple sequence repeat (SSR) markers from *Fragaria x ananassa* expressed sequences. *Mol Ecol Notes* **2006**; 6: 319-322.

[84] Cervigni GDL, Paniego N, Diaz M, *et al.* Expressed sequence tag analysis and development of gene associated markers in a near-isogenic plant system of Eragrostis curvula. *Plant Mol Biol* **2008**; 67: 1-10.

[85] Winton LM, Krohn AL, Leiner RH. Microsatellite markers for *Sclerotinia subarctica* nom. prov., a new vegetable pathogen of the High North. *Mol Ecol Notes* **2007**; 7: 1077-1079.

[86] Perez F, Ortiz J, Zhinaula M, Gonzabay C, Calderon J, Volckaert F. Development of EST-SSR markers by data mining in three species of shrimp: *Litopenaeus vannamei*, *Litopenaeus stylirostris*, and *Trachypenaeus birdy*. *Marine Biotechnol* **2005**; 7: 554-569.

[87] Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **2000**; 28: 352-355.

[88] Sherry ST, Ward MH, Sirotkin K. dbSNP - Database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **1999**; 9: 677-679.

[89] Gibbs RA, Belmont JW, Hardenbol P, *et al.* The international HapMap project. *Nature* **2003**; 426: 789-796.

[90] Stein LD, Mungall C, Shu SQ, *et al*. The Generic Genome Browser: A building block for a model organism system database. *Genome Res* **2002**; 12: 1599-1610.

[91] Kasprzyk A, Keefe D, Smedley D, *et al.* EnsMart: A generic system for fast and flexible access to biological data. *Genome Res* **2004**; 14: 160-169.

[92] Frazer KA, Ballinger DG, Cox DR, *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **2007**; 449: 851-U853.

[93] Hawken RJ, Barris WC, McWilliam SM, Dalrymple BP. An interactive bovine in silico SNP database (IBISS). *Mammalian Genome* **2004**; 15: 819-827.

[94] Jaiswal P, Ni JJ, Yap I, *et al.* Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res* **2006**; 34: D717-D723.

[95] Ware D, Jaiswal P, Ni JJ, *et al.* Gramene: a resource for comparative grass genomics. *Nucleic Acids Res* **2002**; 30: 103-105.

[96] Ware DH, Jaiswal PJ, Ni JJ, *et al.* Gramene, a tool for grass Genomics. *Plant Physiol* **2002**; 130: 1606-1613.

[97] Carollo V, Matthews DE, Lazo GR, *et al.* GrainGenes 2.0. An improved resource for the small-grains community. *Plant Physiol* **2005**; 139: 643-651.

[98] Matthews DE, Carollo VL, Lazo GR, Anderson OD. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res* **2003**; 31: 183-186.

[99] Weems D, Miller N, Garcia-Hernandez M, Huala E, Rhee SY. Design, implementation and maintenance of a model organism database for Arabidopsis thaliana. *Comparative Funct Genom* **2004**; 5: 362-369.

[100] Rhee SY, Beavis W, Berardini TZ, *et al.* The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* **2003**; 31: 224-228.

[101] Huala E, Dickerman AW, Garcia-Hernandez M, *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* **2001**; 29: 102-105.

[102] Lawrence CJ, Dong OF, Polacco ML, Seigfried TE, Brendel V. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res* **2004**; 32: D393-D397.

[103] Dong QF, Roy L, Freeling M, Walbot V, Brendel V. ZmDB, an integrated database for maize genome research. *Nucleic Acids Res* **2003**; 31: 244-247.

[104] Gai XW, Lal S, Xing LQ, Brendel V, Walbot V. Gene discovery using the maize genome database ZmDB. *Nucleic Acids Res* **2000**; 28: 94-96.

[105] Duvick J, Fu A, Muppirala U, *et al.* PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res* **2008**; 36: D959-D965.

[106] Dong QF, Lawrence CJ, Schlueter SD, *et al.* Comparative plant genomics resources at PlantGDB. *Plant Physiol* **2005***;* 139: 610-618.

[107] Dong QF, Schlueter SD, Brendel V. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res* **2004**; 32: D354-D359.

[108] Wheeler DL, Barrett T, Benson DA, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* **2008**; 36: D13-D21.

[109] Lim GAC, Jewell EG, Xi L, *et al.* A comparative map viewer integrating genetic maps for Brassica and Arabidopsis. *BMC Plant Biology*: 7 (July) 10pp. **2007**; 7.

[110] Liao W, Collins A, Hobbs M, Khatkar MS, Luo JH, Nicholas FW. A comparative location database (CompLDB): map integration within and between species. *Mammalian Genome* **2007**; 18: 287-299.

[111] Somers DJ, Isaac P, Edwards K. A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* **2004**; 109: 1105-1114.

[112] Jesubatham AM, Burow MD. PeanutMap: an online genome database for comparative molecular maps of peanut. *BMC Bioinform* **2006**; 7: 375.

[113] Jung S, Staton M, Lee T, *et al.* GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res* **2008**; 36: D1034-D1040.

[114] Gonzales MD, Archuleta E, Farmer A, *et al.* The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res* **2005**; 33: D660-D665.

[115] Flicek P, Aken BL, Beal K, *et al.* Ensembl 2008. *Nucleic Acids Res* **2008**; 36: D707-D714.

[116] Birney E, Andrews D, Bevan P, *et al.* Ensembl 2004. *Nucleic Acids Res* **2004**; 32: D468-D470.

[117] Love CG, Batley J, Lim G, *et al.* New computational tools for Brassica genome research. *Comparative Funct Genomics* **2004**; 5: 276-280.

[118] Karolchik D, Kuhn RM, Baertsch R, *et al.* The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **2008**; 36: D773-D779.

Received: July 16, 2008 Revised: October 23, 2008 Accepted: October 24, 2008