

SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation

Frank Panitz¹, Henrik Stengaard¹, Henrik Hornshøj¹, Jan Gorodkin², Jakob Hedegaard¹, Susanna Cirera², Bo Thomsen¹, Lone B. Madsen¹, Anette Høj¹, Rikke K. Vingborg¹, Bujie Zahn¹, Xuegang Wang¹, Xuefei Wang¹, Rasmus Wernersson³, Claus B. Jørgensen², Karsten Scheibye-Knudsen², Troels Arvin², Steen Lumholdt², Milena Sawera², Trine Green², Bente J. Nielsen², Jakob H. Havgaard², Søren Brunak³, Merete Fredholm² and Christian Bendixen^{1,*}

¹Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, University of Aarhus, DK-8830 Tjele,

²Department of Basic Animal and Veterinary Sciences, Faculty of Life Sciences, University of Copenhagen,

DK-1870 Frederiksberg C and ³Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, DK-2800 Lyngby, Denmark

ABSTRACT

Motivation: Single nucleotide polymorphisms (SNPs) analysis is an important means to study genetic variation. A fast and cost-efficient approach to identify large numbers of novel candidates is the SNP mining of large scale sequencing projects. The increasing availability of sequence trace data in public repositories makes it feasible to evaluate SNP predictions on the DNA chromatogram level. MAVIANT, a platform-independent Multipurpose Alignment Viewing and Annotation Tool, provides DNA chromatogram and alignment views and facilitates evaluation of predictions. In addition, it supports direct manual annotation, which is immediately accessible and can be easily shared with external collaborators.

Results: Large-scale SNP mining of polymorphisms bases on porcine EST sequences yielded more than 7900 candidate SNPs in coding regions (cSNPs), which were annotated relative to the human genome. Non-synonymous SNPs were analyzed for their potential effect on the protein structure/function using the PolyPhen and SIFT prediction programs. Predicted SNPs and annotations are stored in a web-based database. Using MAVIANT SNPs can visually be verified based on the DNA sequencing traces. A subset of candidate SNPs was selected for experimental validation by resequencing and genotyping. This study provides a web-based DNA chromatogram and contig browser that facilitates the evaluation and selection of candidate SNPs, which can be applied as genetic markers for genome wide genetic studies.

Availability: The stand-alone version of MAVIANT program for local use is freely available under GPL license terms at <http://snp.agrsci.dk/maviant>.

Contact: christian.bendixen@agrsci.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Single nucleotide polymorphism (SNP) analysis provides an important tool in applications as genetic linkage mapping, fine-mapping of candidate regions and to determine haplotypes associated with traits of interest, in order to understand the genetic basis of phenotypic diversity within and between populations. Recently, large-scale identification and characterization of SNPs has attracted much interest in connection with the sequencing projects of the human and vertebrate genomes (Guryev *et al.*, 2004; Sachidanandam *et al.*, 2001; Wiltshire *et al.*, 2003; Wong *et al.*, 2004). Due to the high abundance in the genome, thousands of potentially informative SNP markers can be identified for the development of high density SNP maps (Zimdahl *et al.*, 2004), which are an essential resource to identify the underlying genes responsible for the variation of complex traits or QTLs (Andersson, 2001; Andersson and Georges, 2004; Vignal *et al.*, 2002). A fast and cost-efficient approach to identify a large number of novel SNPs is the data mining of large-scale sequencing projects and sequence data from public repositories. However, despite porcine transcript sequences being available from public databases, only a very limited number of SNPs are identified in pigs (Fahrenkrug *et al.*, 2002; Sherry *et al.*, 2001; Uenishi *et al.*, 2004). To overcome the limited amount of SNP markers in candidate genes, we performed SNP mining analysis of the EST sequence resource generated as part of the Danish–Chinese Pig Genome Sequencing Initiative (Gorodkin *et al.*, 2007; Jørgensen *et al.*, 2005, Wernersson *et al.*, 2005). More than 7900 coding SNPs (cSNPs) were predicted from the sequences from 98 porcine cDNA libraries and analyzed for potential effects on protein function. However, due to the current lack of a complete porcine genome draft SNP mining is likely to be impaired by the occurrence of sequence paralogues and pseudogenes in the cluster assembly. In order to facilitate and improve the evaluation process by visual inspection of the polymorphic sites, we devised a Multipurpose Alignment Viewing and Annotation Tool (MAVIANT), an application

*To whom correspondence should be addressed.

that provides contig and DNA chromatogram views together with an integrated annotation tool in a platform-independent browser.

The objectives of this SNP mining project were:

- (1) Development of a web-based SNP evaluation tool, MAVIANT, for inspection and validation of predicted SNPs by displaying precalculated contig alignments and sequencing chromatograms. Integration of an online annotation feature in order to allow collaboration on the same dataset.
- (2) Generation of a database of polymorphisms located in gene-coding regions by identifying novel cSNPs from more than 800 000 EST sequences. Characterization and annotation of non-synonymous SNPs for their potential deleterious effects on protein structure using PolyPhen (Ramensky *et al.*, 2002) and SIFT (Ng and Henikoff, 2001). Finally, experimental validation of a subset of candidate SNPs by genotyping and calculation of allele frequencies.

In a benchmark of the procedure, 190 of 322 randomly selected SNPs were confirmed by resequencing in an animal panel. The major incentive for large-scale SNP discovery efforts is to generate a genome-wide set of gene-based markers that can be tested and applied in high-throughput studies on genetic variation.

2 METHODS

2.1 EST sequencing and SNP discovery

Ninety-eight cDNA libraries have been constructed from various *Sus scrofa* tissues, organs and different developmental stages and represent one or more animals from one or more breeds (Duroc, Hampshire, Landrace, Yorkshire and ErHuaLin). The cDNA inserts average about 0.7–1.5 kb in length and were sequenced on MegaBACE 1000 using ET Dye Terminator chemistry (Amersham Pharmacia). Trace files were base called with Phred (Ewing and Green, 1998; Ewing *et al.*, 1998) using a quality cut-off of 0.05 and the—trim_alt parameter. Following removal of vector contamination using Crossmatch we retained 810 124 sequences of at least 50 bp length. For additional details see Gorodkin *et al.* (2007).

EST sequences were masked for artiodactyl- and mammalian-specific repeats and low-complexity regions using RepeatMasker (Smit, A.F.A. and Green, P., unpublished data) and RepBase 8.3 (RepeatMasker version; Jurka 2000) with the—pig option and default sensitivity. EST clustering was performed using the accelerated BLAST (Altschul *et al.*, 1997) based TeraClu algorithm on a DeCypher FPGA Computer (Timelogic/Active Motif) followed by cluster assembly using Phrap (Green, P., unpublished data). EST and contig consensus sequences were compared with a human reference database (human genome NCBI build 35 version1), *Sus scrofa* UniGene v.36 and the TIGR Pig Gene Index v.11 using TeraBlastN to obtain primary annotation. Large-scale SNP detection in the assembled contigs was performed using PolyBayes v3.0 (Marth *et al.*, 1999). SNPs predicted in sequences from Duroc, Hampshire, Landrace, Yorkshire were filtered using a minimum base quality of 30 and a probability threshold of at least 0.8, 0.95 and 0.99, respectively. Only single basepair substitutions were considered, as insertion/deletion (indel) type mismatches are more difficult to evaluate, since they are more likely attributed to alignment artifacts in the assembly process (Picoult-Newberg *et al.*, 1999).

For all biallelic polymorphic sites located in alignments SNP types (synonymous/nonsynonymous) were calculated. Based on the CDS regions annotated in the human reference database the coding sequence for each alignment was translated and the amino acid variation for each SNP calculated.

2.2 MAVIANT

MAVIANT is a web-based SNP evaluation tool written in Perl using GD graphics and in-house libraries. The program generates views build from html, png image and javascript files to support a cross-platform environment [Internet Explorer, Mozilla Firefox, Netscape, Opera, Konquerer (Linux), Safari (Mac)]. MAVIANT output is pregenerated for faster browser loading and makes it completely independent from data source. Alignment overview of sequences and features can be displayed in three different modes: data (consensus and sequence overview), filter (consensus and sequences filter overview) and quality (consensus and sequence quality overview). Alignment fragments are generated for each 50 bases in three different modes: data (sequence bases colored with transparent background), background (sequence base color as background) and quality (sequence bases colored and background indicating sequence quality). All three modes are also generated in a filtered version, which indicates base differences between consensus and sequences. The program reads alignments from Cap3, Phrap and PolyBayes ace files and generates chromatogram images using phd files and .abi, .abd or .abl chromatogram files. Features like SNPs, repeats and custom annotation can be defined by means of a feature file containing sequence name, annotation type (SNP, Repeat), start and end position and color for each individual feature. The annotations are entered via a dynamic website and stored directly using flat files or databases.

2.3 SNP evaluation

DNA sequencing traces of candidate SNPs of interest were visually inspected with MAVIANT and validated by the user. From these verified SNPs we selected subsets for resequencing and genotyping analysis. A few EST contigs (11) showing obvious clustering artifacts were removed after manual inspection

2.4 Estimation of candidate non-synonymous SNPs on protein function

To predict potential effects of non-synonymous SNPs on protein function, we assumed that the conservation of coded protein(structure) within the mammalian genome is sufficiently high to allow the analysis of human protein sequences having the site of interest substituted with the polymorphic alleles identified in the pig. The prediction of amino acid variants was performed using PolyPhen v1.11 (Ramensky *et al.*, 2002) and SIFT v2.0 (Ng and Henikoff, 2001). Both programs were run as stand-alone applications using the same protein databases (UniProt-SWISSPROT, UniProt-TREMBL) and the same BLAST parameters (expectation cut-off = 1E-4), in order to obtain comparable results.

2.5 SNP verification and genotyping

322 SNPs selected at random from candidates in alignment regions were validated by resequencing and yielded 190 confirmed SNPs. From the SNPs that were successfully verified by resequencing, 138 were selected for genotyping in the breed panel using TaqMan assays (Livak, 1999). After analyzing for allelic discrimination allele frequencies were calculated. See Supplementary Material for additional information. Protocols are available on request.

3 RESULTS

3.1 MAVIANT

MAVIANT is a contig and DNA chromatogram browser developed for the visual evaluation of predicted SNPs with respect to collaboration on SNP mining projects using predictions based on contig clusters originating from a few hundred thousand sequences. It was implemented as a server application using precalculated data. The user can access individual contig views from a link in the database or via search in MAVIANT.

MAVIANT output is generated using html, png and JavaScript files, thus making it independent from the data source. By pregenerating the output files the application is optimized for fast loading speed. From a navigation panel DNA sequencing traces in the alignment views can be collapsed and expanded, in order to provide overviews and quick navigation (Fig. 1). As input sequencing trace files (ab1, abd, ab1 trace files), base quality information (phd files) and clustering and alignment information (ace files) are used. Alignment overviews of sequences and features are available in three different modes: data (consensus and sequence overview), filter (overview of differences between consensus and sequences) and quality (phd-based quality overview). Sequence alignments views are depicted as 50 bp fragments in three different coloring schemes for easier visualisation: data (colored base), background (colored base background) and quality. Additional features like SNPs, repeats or other custom annotations can be added by using a feature file containing sequence identifier, annotation type (SNP, repeat, etc.), start and end position and a specific color for each individual feature.

The visual inspection of base quality at the polymorphic site as well as the neighboring bases allows the user to validate and annotate SNPs in a single step. Apart from being a mere viewing tool for DNA chromatograms and alignments MAVIANT is designed to facilitate manual annotation of predicted polymorphic sites by direct user input into an annotation box (Fig. 1D). These are stored in an integrated flat file using a dynamic website. The annotations are immediately accessible and can be easily shared with external collaborators. However, collaboration on large sets of contigs would require transferring trace/quality files or the complete analysis output, in order to evaluate sequence alignments and *in silico* SNP predictions on the DNA chromatogram level. As many standard large-scale sequence assembly and SNP detection programs as Cap3 (Huang and Madan, 1999), Phrap (Green, P., unpublished data), PolyBayes (Marth *et al.*, 1999) and PolyPhred (Nickerson *et al.*, 1997) that generate assembly and SNP information (ace files) run in a UNIX-based environment, including the contig viewing program Consed (Gordon *et al.*, 1998), we also provide a platform independent stand-alone browsing tool for contig assembly (ace files) and trace files (scf, abd, abi, ab1 files).

3.2 SNP discovery and evaluation

This study describes the development and integration of a SNP database and contig/sequencing trace browser as means to

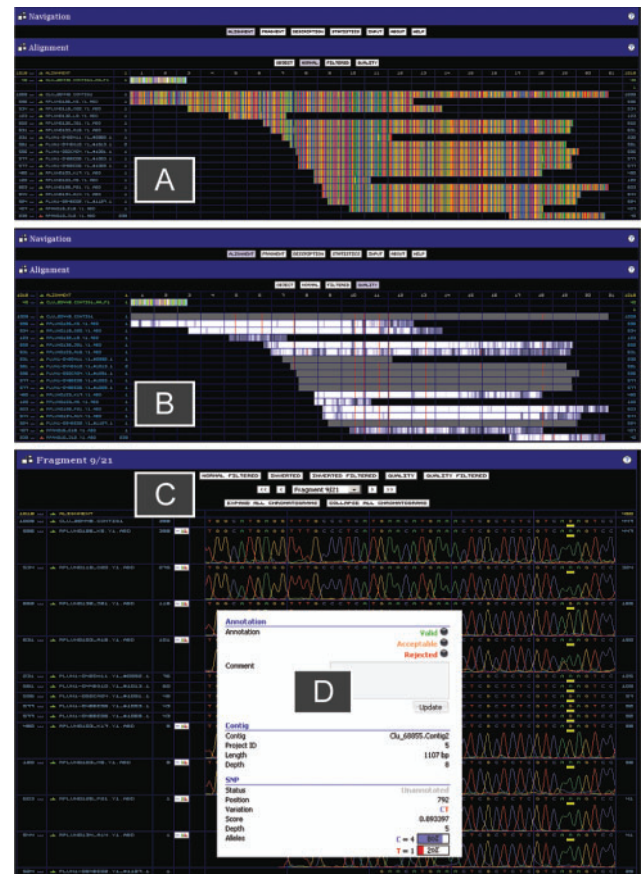


Fig. 1. Maviant SNP viewer. Contig alignment overview displaying colored bases (A) or quality information (B). In the sequence alignment view, SNPs are highlighted and chromatogram traces can be exploded or collapsed individually (C). For each SNP evaluation status and comments can be added directly by means of an annotation box (D).

evaluate predicted SNPs, in order to support the selection of SNPs for genotyping and mapping applications. The analysis for polymorphic sites was performed on contigs which were derived from clustering of 810 124 trimmed and repeat-masked EST sequences. 24 921 clusters were assembled with Phrap yielding 44 266 contigs, which were screened for SNPs using PolyBayes (Marth *et al.*, 1999), a program that scores each SNP position, measuring the probability of a site being polymorphic, dependent on the sequence quality, alignment depth and allele frequency. Data mining of the EST clusters identified 7979 biallelic SNPs (7924 unique) in coding regions (SNP probability ≥ 0.8), corresponding to 3894 porcine EST contigs (which could be aligned to 3428 unique human reference sequences).

The EST contigs were annotated according to the gene/exon information of the human genome assembly as a pig genome assembly is currently not available. Primary annotation was obtained by comparing cluster consensus sequences against human and porcine transcript databases. 16 290 Blast hits of at least 100 bp alignment length and a minimum alignment identity of 80% were detected against a human exon reference database. By comparison with two public pig transcript

Table 1. SNP discovery statistics

Sequence data			
ESTs from 98 cDNA libraries		810 124	
Clustering and database comparison ^a			
Contigs		44 266	
Hits to human reference database ^b		16 290	
Hits to NCBI Pig UniGene v36		15 836	
Hits to TIGR Pig Gene Index v11		16 775	
Prediction of biallelic SNPs ^c			
Score (PolyBayes)	0.8	0.95	0.99
cSNPs	7979 (5688)	6009 (4895)	4846 (4263)
Synonymous	4707 (3317)	3687 (2981)	3064 (2683)
Non-synonymous	3272 (2371)	2322 (1914)	1779 (1580)
Non-synonymous SNPs predicted to have potential effect on protein ^d			
PolyPhen	826 (637)	564 (478)	416 (382)
SIFT	824 (654)	609 (519)	474 (427)
Combined (unique)	1166	821	622

^aNon-redundant BLAST hits; threshold: alignment length \geq 100 bp, identity \geq 80%, ^bReference database based on human genome NCBI build 35, version1, ^cNumber of SNPs with PolyBayes allele depth of ≥ 2 (≥ 5) sequences, ^dTotal number of protein effecting predictions.

databases, the TIGR porcine Gene Index (Release 11.0; 38 781 clusters) and pig UniGene (Build 36; 35 620 clusters), 34 653 EST contig hits to 16 775 unique target loci and 30 375 hits to 15 836 targets were identified, respectively. The results are summarized in Table 1. The SNPs were classified as coding based on pig-human alignments as the human gene annotations currently provide a more comprehensive annotation than the pig gene cluster databases UniGene and TIGR Gene Index. The alignment information was used to determine the relative position and type (synonymous/nonsynonymous) for each SNP based on the GenBank annotation for human coding regions; 41% of the cSNPs were found to represent non-synonymous substitutions at protein level.

As no sufficient information was available on whether the EST sequences in a given cluster originated from the same genomic location or if they belonged to a different but similar, duplicated region in the genome, a stringent paralogue filtering against genomic anchor regions could not be performed. Therefore, the number of SNPs will potentially include false-positive predictions due to the possible occurrence of closely related gene family members or pseudogenes in the contig assemblies. While a higher SNP score threshold yields a higher confirmation rate by reducing the number of false-positives, it conversely disregards more true SNPs. This might pose a problem especially when mining for rare SNPs, as applying lower detection thresholds potentially increases the number of false positives. An evaluation of different stringency settings, however, indicated clearly that the number of identified candidate SNPs could be increased by using relaxed threshold criteria coupled with visual inspection of DNA sequencing traces for polymorphic sites. In order to obtain a more stringent selection of SNPs, the sequence depth at the polymorphic site can be increased (Table 1). The recent availability of inexpensive high-density SNP-genotyping arrays for high-throughput genetic analysis of large populations, however, has made it feasible to utilize large-scale SNP discoveries.

3.3 Estimation of effects of nonsynonymous SNPs on protein function

The characterization of the 7924 cSNPs showed that a substantial fraction of candidate SNPs represent non-synonymous variants. We evaluated the likely effect of amino acid substitutions on protein function for more than 3200 non-synonymous cSNPs using PolyPhen (Ramensky *et al.*, 2002) and SIFT (Ng and Henikoff, 2001). These programs predict possible deleterious effects of an amino acid variant on the structure and function of a protein based on sequence homology, phylogenetic and structural information. Fifty percent of the EST derived non-synonymous substitutions were predicted to be relevant to protein function by PolyPhen or SIFT (1650 of 3272). 484 (42%) polymorphic sites of the EST-based candidates were found to be relevant to protein function by both programs. The resulting amino acid changes can potentially affect the structure and hence the function of the encoded protein. As a result nonsynonymous SNPs might have direct influence on the traits of interest and contribute supposably to phenotypic effects. cSNPs resulting in structure or function altering variants or nonsense mutations have been associated with disease alleles and are therefore good candidates to study disease-association and phenotypic differences.

3.4 SNP database

The results from the large-scale SNP mining using EST trace sequences from 98 porcine cDNA libraries are collected in a database which contains the predicted cSNPs, contig sequences and the matching human transcript information as determined by Blast. The SNP predictions and annotations can be accessed and searched in a web-based database (<http://snp.agrsci.dk/>).

3.5 Experimental validation of candidate SNPs

Determining allele frequencies in a population will provide useful information for association mapping and QTL projects.

In addition, the frequency data will also be valuable for large QTL studies as the number of markers needed can be reduced when avoiding SNPs with low minor allele frequencies. Therefore, a sample set of SNPs was experimentally tested, in order to assess allele frequencies as these will be of general use for linkage analysis in populations. Of 322 randomly selected SNPs that were analyzed by resequencing in an animal panel 59% could be confirmed. However, in order not to misinterpret the rate it has to be taken into account that the selection of SNPs for validation was performed using MAVIANT. The data selection is thus biased by MAVIANT evaluation, depending on the person performing the prediction and not directly comparable to completely unbiased validation rates. A total of 138 segregating SNPs (108 transcript based and 30 non-coding) were selected for genotyping analysis by TaqMan assays. A total of 10,013 genotype analyses were performed on a breed panel of six western breeds. The calculated average minor allele frequencies, which are important characteristics in defining their utility for genetic applications, are shown in Table S2 (see Supplementary Material). The highest minor allele frequency within the analyzed breeds was observed for Duroc (0.26), while Hampshire and Yorkshire showed the lowest frequencies (0.16 and 0.17, respectively). Figure S1 (Supplementary Material) shows the distribution of SNP allele frequencies among the six breeds where Duroc and Old Danish Landrace have a similar allele frequency distribution, with over 74% of the SNPs having a common minor allele frequency of >0.10. In general, the SNP information generated will extend the number of available genetic markers that can be applied for large-scale genotyping, improving comparative SNP maps and will also aid developing a dense map of SNP markers in or close to candidate genes.

4 CONCLUSION

This study provides a gene-transcript-based collection of SNPs to be used as potential genetic markers for genome wide studies. With the development of MAVIANT as a web-based contig and DNA sequencing trace browser predicted SNPs could be visually evaluated based on the underlying sequence chromatograms. SNPs can be annotated directly online thus allowing collaboration on SNP evaluation, annotation or selection of candidates.

ACKNOWLEDGEMENTS

This study was financially supported by the Danish Slaughterhouses/Danish National Committee for Pig Production and the Danish Ministry of Food, Agriculture and Fisheries.

Conflict of Interest: none declared.

REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Andersson,L. (2001) Genetic dissection of phenotypic diversity in farm animals. *Nat. Rev. Genet.*, **2**, 130–138.
- Andersson,L. and Georges,M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. *Nat. Rev. Genet.*, **5**, 202–212.
- Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
- Ewing,B. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
- Fahrenkrug,S.C. *et al.* (2002) Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. *Anim Genet.*, **33**, 186–195.
- Gordon,D. *et al.* (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
- Gorodkin,J. *et al.* (2007) Porcine transcriptome analysis based on 97 non-normalized cDNA libraries and assembly of 1,021,891 expressed sequence tags. *Genome Biol.*, **8**, R45.
- Guryev,V. *et al.* (2004) Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res.*, **14**, 1438–1443.
- Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
- Jorgensen,F.G. *et al.* (2005) Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. *BMC Biol.*, **3**, 2.
- Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
- Livak,K.J. (1999) Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet. Anal.*, **14**, 143–149.
- Nickerson,D.A. *et al.* (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.*, **25**, 2745–2751.
- Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
- Marth,G.T. *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.
- Picault-Newberg,L. *et al.* (1999) Mining SNPs from EST databases. *Genome Res.*, **9**, 167–174.
- Ramensky,V. *et al.* (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
- Sachidanandam,R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
- Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
- Uenishi,H. *et al.* (2004) PEDE (Pig EST Data Explorer): construction of a database for ESTs derived from porcine full-length cDNA libraries. *Nucleic Acids Res.*, **32**, D484–D488.
- Vignal,A. *et al.* (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.*, **34**, 275–305.
- Wernersson,R. *et al.* (2005) Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing. *BMC Genomics*, **6**, 70.
- Wiltshire,T. *et al.* (2003) Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl Acad. Sci.*, **100**, 3380–3385.
- Wong,G.K. *et al.* (2004) International Chicken Polymorphism Map Consortium. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, **432**, 717–722.
- Zimdahl,H. *et al.* (2004) A SNP map of the rat genome generated from cDNA sequences. *Science*, **303**, 807.

WEB SITE REFERENCES

- <http://www.phrap.org/>; Phred/Phrap.
- <http://repeatmasker.org/>; RepeatMasker
- <http://pede.dna.affrc.go.jp/>; PEDE (Pig EST Data Explorer).
- <ftp://ftp.ebi.ac.uk/pub/databases/swissprot/release/>; UniProt: Swissprot, TrEMBL