Tissue related in-silico mining of single nucleotide polymorphisms (SNPS) from expressed sequence tags (ESTS) in livestock species

Neeraj Kumar Dhaliwal, Aruna Pandey, Birham Prakash and Avnish Kumar Bhatia*
ICAR - National Bureau of Animal Genetic Resources, Karnal – 132001 (Haryana) India

ABSTRACT

Tissue specific Single nucleotide polymorphisms (SNPs) hold significance as potential expressed Quantitative Trait Loci (eQTL). Although there have been numerous studies for mining SNPs in livestock species, there is little focus on discovery of tissue-specific SNPs. We performed tissue related *in-silico* SNP mining from Expressed Sequence Tags (ESTs) in two livestock species - pig and cattle. EST data for tissues such as skin, liver, spleen, intestine and mammary gland for the two species were downloaded from NCBI website. ESTs were pre-processed using the online tool EGassembler and assembled into contigs using CAP3 program. SNPs were predicted from contigs using QualitySNP tool. Contigs were searched in the genome assembly of a species using Blat tool in UCSC genome browser. Perl scripts were written to find genomic position of SNPs from the alignment of contigs with the genomic segments, and to find availability of predicted SNPs in the dbSNP database. A database of tissue related SNPs was developed.

Keywords: EST, SNP, tissue-specific, livestock, eQTL
***Corresponding author:** avnish@lycos.com

INTRODUCTION

Expressed Sequence Tags (ESTs) are partial sequences of complementary DNA (cDNA) clones measuring several hundred nucleotides [Baxevanis and Ouellette, 2001]. There have been voluminous increases in EST data generation and submission, especially for livestock species, to the primary databases such as NCBI, DDBJ and EMBL. Single nucleotide polymorphisms (SNPs) are the simplest type of genomic variation. Over the past decade, SNPs have been the genetic markers of choice due to their high density, stability and the highly automated techniques for their detection [Kerstens et al., 2009]. Thousands of potentially informative SNP markers can be identified for development of high density SNP maps [Zimdahl et al., 2004], which are an essential resource to identify genes responsible for variation of complex traits or Quantitative Traits Loci (QTL) [Andersson, 2001; Andersson and Georges, 2004]. SNP analysis provides an important tool in applications such as genetic linkage mapping, fine mapping of candidate regions and to determine haplotypes associated with traits of interest [Panitz et al., 2007]. With availability of genome sequence assembly of a number of livestock species like cow, sheep, chicken, pig and horse, mining of sequence data for identification of SNPs is a major task for researchers. Huge amount of EST data for livestock species- pig and cattle on different tissues like skin, mammary gland, spleen, liver, intestine etc. are available in public databases. ESTs data allow discovery of SNPs in the transcribed regions [Marth, 2003].

There are few studies showing importance of tissue-Related SNPs, particularly in species of economic importance such as livestock species. In a study of Tissue Specific Temporal (TST) exome capture, presence of tissue (muscle) specific genes and SNPs in *Bubalus bubalis* has been revealed [Jakhesara et al., 2012]. Recent Genome Wide Association Studies (GWAS) in humans have revealed that the genetic variants may be operating in tissue dependent manner. Subsets of genetic polymorphisms show a statistical association with transcript expression levels, and have therefore been called as expression quantitative trait loci (eQTLs) [Nicolae et al., 2010].

In this study we have used a bioinformatics pipeline for mining of *in-silico* tissue-related SNPc from EST data in pig and cattle. Discovered SNPs have been validated by revealing their availability in dbSNP database. A database of tissue wise SNPs for livestock

species have also been developed with additional information on genes.

METHODS

EST data processing : EST data for different tissues of pig and cattle were downloaded from dbEST database available at NCBI (www.ncbi.nlm.nih.gov/nucest). ESTs needs processing to remove low quality DNA sequences, contaminating sequences such as vector sequences and repetitive sequences [Chou and Holmes, 2001]. The preprocessing was performed using online tool EGassembler [Masoudi-Nejad et al., 2006], which performs sequence cleaning, repeats masking and vector cleaning as a single operation. Repbase repeats library for the respective species was selected as repeats library to remove repetitive sequences and NCBI core vector library was chosen for vector masking process. Processed ESTs were free from low quality sequences, poly A/ poly T tail, repetitive sequences and the vector sequences.

ESTs were assembled in contigs using the cluster assembly program CAP3 [Huang and Madan, 1999] using default parameters, which performs clustering or assembly of the sequences into contigs by pairwise sequence similarity searches between sequences.

SNP identification : Contigs were analyzed for SNPs using SNP prediction tool QualitySNP [Tang et al., 2006] with default parameters. This tool takes the contigs.ace file generated by CAP3 tool as input and predicts SNPs on contigs. SNPs predicted by the tool are differentiated into three categories, viz., potential SNPs, High Quality SNPs and Reliable SNPs. It provides position of SNPs on contigs, number of SNPs on each contig, major and minor alleles and d-value denoting the standard deviation of normalized number of SNPs per haplotype, which identify clusters that probably contains paralogs.

SNPs position on chromosome and availability in dbSNP Database : Contigs with potential SNPs were searched in the genome assembly of a species in UCSC Genome Browser using BLAT tool (http://genome.ucsc .edu/cgi-bin/hgBlat? command=start). This tool provides alignment of the contig with particular chromosome segment of a species.

SNPs data of both the species were downloaded from Ensembl Genome Browser (http://asia.ensembl.org/biomart/martview/), which included SNP name, SNP type, chromosome name, chromosome position, chromosome strand, SNP alleles, Ensembl gene id, gene name, and gene start and gene end positions. SNP data were retrieved separately for each chromosome for these two species for ease of data handling on a computer.

Perl script was written to find position of SNPs in the best alignment of contigs and genomic sequences as obtained using the BLAT tool. These SNPs were searched for their presence in the dbSNP database available in the Ensembl Biomart (http://asia.ensembl.org/biomart/martview/). The perl script is available for download within online database described in the section 3.

Database development : A database on Tissue-wise SNPs mined from ESTs was developed using MySQL and PHP to provide user-friendly interface on tissue wise SNPs and their availability in dbSNP database. SNPs data in text files was inserted into the database using a perl script. Information on chromosomes and genes were also inserted in the database.

RESULTS

Table-1 displays information of ESTs and SNPs discovered in pig and cattle. There were 50410 ESTs in dbEST database for skin tissue of pig. Assembly of these ESTs provided 6476 contigs and a total of 3444 SNPs were identified out of which 1550 were high quality SNPs and 834 were reliable SNPs. Only 421 out of 3444 detected SNPs were found in dbSNP database.

The database developed on tissue-wise SNPs provides a user-friendly web interface for data query and visualization on search terms such as species name, tissue name, chromosome name and gene name. Output field settings include dbSNP name, gene name, chromosome name, contig name and sequence. SNPs are retrived with information on high quality SNPs, reliable SNPs, Blat strand, dbSNP strand, dbSNP id, dbSNP allele, contig allele, contig name and sequence. The database also provides genomic information such as Ensembl gene id, Ensembl gene name, gene start position, gene end

Table 1. Tissue related SNPs mined from ESTs in pig and cattle species

Livestock species	Tissue	ESTs	Contigs	Contigs with SNPs	SNPs	High Quality SNP	Reliable SNP	Matches in dbSNP database
Pig	Skin	50410	6476	698	3444	1550	834	421
	Spleen	73312	7575	761	5032	2389	1463	595
	Mammary Gland	23061	1545	146	1687	1027	311	74
	Liver	129323	17941	2761	17082	8466	4499	2246
Cattle	Mammary gland	106150	8189	1158	10150	6031	2468	43
	Intestine	3010	394	32	192	75	55	0
	Spleen	25781	4890	808	3690	979	754	15
	Liver	178249	19664	4310	31023	16361	7141	157

position, chromosome name and position of SNP on chromosome. Information on tissue-wise SNPs is linked to other databases such as Ensembl genome browser and dbSNP database. The database is accessible through a link 'Database' available at the website http://www.nabgr.res.in/.

DISCUSSION

Significance of tissue specific SNPs, genes and eQTLs has been highlighted recently in humans [Dimas et al., 2009; Nica et al., 2011; Hernandez et al., 2012]. Tissue specific SNPs have also been studied in *Bubalis bubalus* [Jakhesara et al, 2012]. In these studies, effects of genetic variants have been reported in which 69 to 80% of the regulatory variants are operating in cell-specific manner. Also eQTLs have been identified, which may be unique or shared among cell types or tissues. Many tissues and cell types have specific gene expression patterns and so it is not clear how frequently eQTLs found in one tissue type will be replicated in others. Therefore, tissue specific studies using SNPs to detect eQTLs have been taken up. A unique set of tissue Related eQTLs have been identified in blood and brain tissues in humans [Nica et al., 2011]. Nicolae et al., 2010 showed that SNPs associated with complex traits are more likely to be eQTLs comparted to minor-allele-frequency matched SNPs from GWAS. In view of these studies, tissue specific SNP mining should be of considerable value in the livestock species.

We have performed tissue-related *in-silico* SNP mining from EST data in two agriculturally important livestock species - pig and cattle. These SNPs were mined from EST data retrieved for a tissue and therefore the SNPs discovered should be related to the tissue. These SNPs were then located on the genome assembly of the respective species and validated by searching in the dbSNP database.

The present study reports a bioinformatics pipeline for *in-silico* tissue-related SNP mining from EST data in pig and cattle, locating these SNPs on the genome assembly of the respective species, and searching their availability in dbSNP database using perl script. A number of potential SNPs were discovered in the two investigated livestock species-pig and cattle using the approach. A database of tissue-Related SNPs has also been developed for use by researchers.

ACKNOWLEDGMENTS

Authors thankfully acknowledge funding by Indian Council of Agricultural Research through National Agricultural Innovation Project.

REFERENCES

Andersson L. 2001. Genetic dissection of phenotypic diversity in farm animals.

Nature Reviews Genetics 2:130-38.

Andersson L and Georges M. 2004. Domestic animal genomics: Deciphering the genetics of complex traits. *Nature Reviews Genetics* 5: 202-12.

Baxevanis AD and Ouellette BFF. 2001. Bioinformatics: A Practical Guide to Analysis of Genes and Proteins. Second edition. John Wiley & Sons Inc. New York.

Chou HH and Holmes MH. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics* 17(12): 1093-1104.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Cohen HA, Ingle C, Beazley C, Arcelus MG, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET and Antonarakis SE. 2009. Common regulatory variation impacts gene expression in a cell type dependent manner. *Science* 325 (5945): 1246-50.

Hernandez DG, Nalls MA, Moore M, Chong S, Dillman A, Trabzuni D, Gibbs JR, Ryten M, Arepalli S, Weale ME, Zonderman AB, Troncoso J, O'Brien R, Walker R, Smith C, Bandinelli S, Traynor BJ, Hardy J, Singleton AB, and Cookson MR. 2012. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*, 47(1): 20-28.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Research, 9: 868-77.

Jakhesara SJ, Ahir VB, Padiya KB, Koringa PG, Rank DN, and Joshi CG. 2012. Tissue-specific temporal exome capture revealed muscle-specific genes and SNPs in Indian buffalo *Bubalus bubalis*. *Genomics, Proteomics & Bioinformatics* 10(2):107-13.

Kerstens HD, Kollers S, Kommadath A, Rosario MD, Dibbits B, Kinders SM, Crooijmans RP and Groenen M. 2009. Mining for single nucleotide polymorphisms in pig genome sequence data. *BMC Genomics 10: 4*.

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T and Goto S. 2006. EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic Acids Res* 34: W459-62.

Marth GT. 2003. Computational SNP discovery in DNA sequence data. *Methods in Molecular Biology* 212: 85-110.

Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman AK, Bataille V, Bell JT, Surdulescu G, Dimas AS, Ingle C, Nestle FO, Meglio PD, Min JL, Wilk J, Hammond CJ, Hassanali N, Yang TP, Montgomery SB, O'Rahilly S, Lindgren CM, Zondervan KT,Soranzo N, Barroso I, Durbin R, Ahmadi K, Deloukas P, McCarthy MI. Dermitzakis ET and Spector TD. 2011. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genetics* 7(2): e1002003.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, *et al.* 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* 6(4): e1000888.

Panitz F, Stengaard H, Hornshoj H, Gorodkin J, Hedegaard J, Cirera S, Thomsen B, Madsen LB, Hoj A, Vingborg RK, Zahn B, Wang X, Wang X, Wernersson R, Jørgensen CB, Knudsen KS, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Brunak S, Fredholm M and Bendixen C. 2007. SNP mining porcine EST with MAVIANT, a novel tool for SNP evaluation and annotation. *Bioinformatics* 23 (13): i387 - i391.

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA. 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* 7: 438.

Zimdahl H, Nyakatura G, Brandt P, Schulz H, Hummel O, Fartmann B, Brett D, Droege M, Monti J, Lee YA, Sun Y, Zhao S, Winter EE, Ponting CP, Chen Y, Kasprzyk A, Birney E, Ganten D and Hubner N. 2004. A SNP map of the rat genome generated from cDNA sequences. *Science* 303 (5659): 807.