*Genes, Genomes and Genomics* ©2010 Global Science Books

# Electronic Sorting of SNP/Indel Sites in Expressed Sequence Tag Libraries of Cocoa (*Theobroma cacao* L.)

**Aykkal Riju[1] • Vadivel Arunachalam[2*]**

[1] Aikkal House, Kannur, Kerala - 670564, India
[2] Molecular Biology and Bioinformatics Laboratory, Central Plantation Crops Research Institute, Kudlu, Kasaragod, Kerala- 671 124, India

***Corresponding author***: * **v.arundevi@gmail.com** or **vadivelarunachalam@yahoo.com**

## ABSTRACT

The objective of this study was to explore the single nucleotide polymorphims (SNPs) in expressed sequence tags (ESTs) of cocoa (*Theobroma cacao* L.). We retrieved 6578 EST sequences consisting of seven tissues/libraries from dbEST of National Centre for Biotechnology Information. SNPs and small Indels (insertion/deletion) were located with the help of AutoSNP. We found a density of one SNP/166 bp and one Indel/360 bp in cocoa ESTs. Candidate SNPs were categorized according to nucleotide substitution as either transition (C/T or G/A) or transversion (C/G, A/T, C/A or T/G). We observed a relative increase in the proportion of transversions (1268) over transitions (950) in bean and leaves and defense related EST sequence libraries. Transversion of G/T (562) (25% of detected SNP) was predominant in cocoa ESTs. We worked out Shannon entropy to find out the distribution of ten different types of SNPs/Indels. An online database (http://www.riju.byethost31.com/cocoa/ccsnp.html) was created to enable cocoa workers to freely access the results of the study.

**Keywords:** database, Malvales, mutation, transition, transversion

## INTRODUCTION

Aim of the present study is to mine the expressed sequence tags (ESTs) of cocoa (*Theobroma cacao* L.) plant for single nucleotide polymorphisms/insertion and deletion (SNP/indel) sites and to work out the Shannon entropy (Shannon and Weaver 1949) for the 10 types of SNP/indels. We also intend to develop a user-friendly EST-SNP/indel information resource for cocoa researchers.

Cocoa is an important perennial tree of the tropics and an important ingredient of chocolates and confectionery dishes. It is a shade loving tree commonly grown as companion crop in orchards and plantations of coconut and areca. Cocoa is a diploid species (2n = 2X = 20) with a small genome size of 380 Mbp (Figueira *et al.* 1992) which is about 2.8 times the size of *Arabidopsis thaliana* (Couch *et al*. 1993). The development of DNA-based markers is important for selection and improvement of varieties and hybrids in plant breeding programs (Gupta *et al.* 2001; Kota *et al.* 2003). SNPs including insertion/deletions (indels) can provide a rich source of useful molecular markers in genetic analysis. ESTs are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene. The idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "tags" to fish a gene out of a portion of chromosomal DNA by matching base pairs. ESTs provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation, and for constructing genome maps. SNP, pronounced "snip", are one-letter variations in the DNA sequence which contribute to differences among individuals. They are the most common form of DNA sequence variation. They are useful as polymorphic markers to analyze the diversity and QTL mapping.

Majority of SNPs produce no effect when they occur in intronic or intergenic regions or as synonymous codon substitutions in exons. But even a single indel in coding region can cause frame shift mutations. A single non-synonymous SNP can convert an amino acid to another which in turn will lead to subtle differences in countless characteristics, like appearance, while some affect the risk for certain diseases. SNPs are molecular markers of choice in recent years for genome mapping and diversity analysis in many crop plants (soybean - Van *et al*. 2005; rye - Varshney *et al*. 2007; cassava - Kauwki *et al*. 2009). They are used in human genetics, such as for the detection of alleles associated with genetic diseases and the identification of individuals (Nikiforov *et al.* 1994). SNPs are invaluable as a tool for genome mapping, offering the potential for generating high-density genetic maps, which can be used to develop haplotyping system for genes or regions of interest (Rafalski 2002a). The low mutation rate of SNPs also makes them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (Syvanen 2001). Unlike random amplified polymorphic DNAs (RAPDs) and restriction fragment length polymorphisms (RFLPs), SNPs are direct markers because sequence information provides the exact nature of the allelic variation. They are far more prevalent than simple sequence repeats (SSR) and, therefore may provide a high density of markers near a locus of interest. One of the limitations of SNPs is the initial cost associated with their development. Many cost and time effective technologies have been developed in recent years for the identification of SNPs in plants including pyrosequencing (*Eucalyptus* - Novaes *et al*. 2008), resequencing and *in silico* methods (review by Ganal *et al.* 2009).

A variety of approaches have been adopted for the discovery of novel SNP markers. Limited work has been carried out to examine the occurrence of SNPs in plants, those results indicated that SNPs appear to be even more abundant in plant systems than the human genome. Very high DNA marker densities are needed for identifying DNA polymorphisms linked to phenotypic and quantitative trait

**Received: 1 June, 2009. Accepted: 1 April, 2010.**

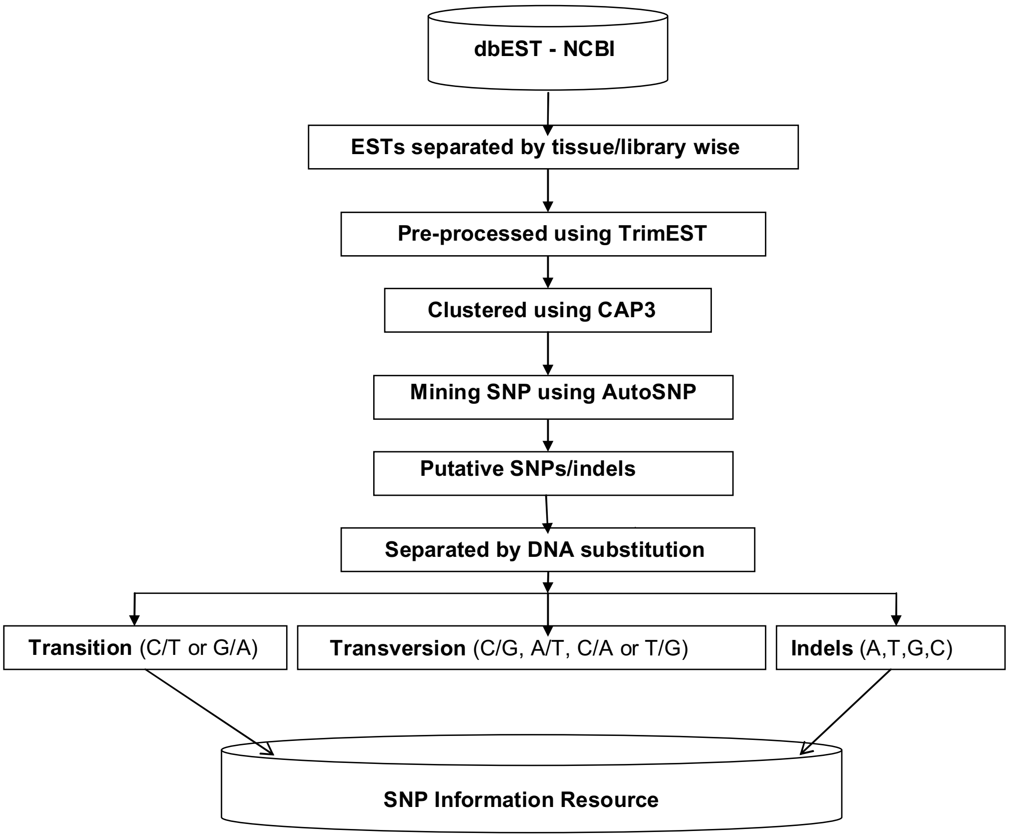

**Fig. 1 Flowchart of *in silico* SNP/indel discovery in cocoa.**

loci through whole-genome association mapping approaches and can only be achieved using SNPs, the most abundant class of DNA polymorphisms (Collins *et al.* 1998; Aquadro *et al.* 2001; Wiltshire *et al.* 2003). While SSR and indel markers are versatile and highly portable, and have been mainstays in molecular breeding and genomics applications (Taramino and Tingey 1996; Bhattramakki *et al.* 2002), SNPs are significantly more common than either and critical for massively parallel array-facilitated genotyping (Lindblad-Toh *et al.* 2000; Syvanen 2001; Rafalski 2002a, 2002b; Buckler and Thornsberry 2002; Syvanen 2005).

## MATERIALS AND METHODS

The GenBank accession numbers AM 117760 – AM 117768, DN 237949 – DN237957, CK 144293 – CK 144298, CF 972636 – CF 974749 and CA 794213 – CA 798660 were retrieved from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) of National Centre for Biotechnology Information (NCBI). These 6581 (dbEST release 012006) were represented seven tissue/condition libraries such as bean and leaves, defense related (cocoa leaves), differential display, immature zygotic embryo, mature zygotic embryo, young red leaves and somatic embryo. EST sequences were trimmed and clustered using CAP3 (Contig Assembly Program) server (Huang and Madan 1999) and the result is given as the input for SNP detecting perl script, Auto_snp version 1.0 (Barker *et al.* 2003). A cocoa SNP database was constructed using MySQL and the details are given as **Fig. 1**.

We have employed Shannon entropy (Shannon and Weaver 1949) for working out an index to compare distribution of 10 possible categories of SNP/indels in different EST libraraies such as bean and leaves, red leaves and differentially displayed ESTs. Frequency of each of nucleotide substitutions as either transition (C/T or G/A), transversion (C/G, A/T, C/A or T/G) and indels (A, T, G, C) were scored. From this value, proportion (Pi) of occurrence of each type (nature of transition/transversion/indel) to the total SNP/indels in each tissue library was worked out. Shannon index (H) estimates (Shannon and Weaver 1949) have been worked out using the formula:

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

where S is the total number of SNP/indel states (10) and pi = proportion of ESTs in the $i^{th}$ type of SNP/indel state.

The calculated value is divided by $\log_2 10$ to get uniformity.

## RESULTS AND DISCUSSION

We retrieved 6581 EST sequences of cocoa from dbEST. Among these, 4505 ESTs sequences were found to have sequence similarity with at least one sequence and were grouped as 784 cluster sequences or contigs. Out of seven libraries only three (bean and leaves, defense related, and young red leaves) libraries contains redundant set of ESTs. We have predicted SNP and indel sites from those libraries. A total of 2218 SNPs and 1021 indels were discovered in the present study (**Table 1**). Candidate SNPs were categorized according to nucleotide substitution as either transition (C↔T or G↔A) or transversion (C↔G, A↔T, C↔A or T↔G). We found transversions (1268) as more predominant than transitions (950) in the cocoa genome. Among the four types of transversions, the G↔T transversion (562) was found to be abundant followed by the C↔T transition. Indel sites occurred at very high frequencies in all the libraries of cocoa analyzed. Shannon entropy of 10 types of SNP/indel types varied from 0.73 to 0.98 among three tissues in cocoa with a mean value of 0.97. ESTs of red young leaves recorded high density of SNPs (71/kb) and indels (26/kb). The summary of the cocoa SNPs and indels discovered in EST libraries is given as **Table 2**.

With the development of high-throughput sequencing technology, large amount of data is being submitted to the various DNA databases offering scope for data mining for SNP discovery. Our studies indicate that the density of SNP in cocoa is 1 SNP/166 bp and that of Indel is 1 indel/360 bp. Earlier, Coryell *et al.* (1999) identified 2 SNPs in approximately 400 bp of sequences in soybean (*Glycine max*) and

**Table 1** Frequency and type of SNPs/indels in expressed sequences of cacao.

| Tissue name | No of clusters | SNP sites | Transitions (Ts) | Transversions (Tv) | Indels | Ts / Tv | Frequency of indels per Kb | Frequency of SNP per Kb |
|---|---|---|---|---|---|---|---|---|
| Bean and leaf | 424 | **1261** | 556 | 705 | **943** | 0.79 | 4.37 | 5.84 |
| Defense Related ESTs | 359 | **935** | 382 | 553 | **70** | 0.69 | **2.17** | **6.13** |
| Young Red leaves | 1 | **22** | 12 | 10 | **8** | 1.20 | **25.64** | **71.42** |
| Total | 784 | 2218 | 950 | 1268 | **1021** | 0.75 | **2.78** | **6.02** |

**Table 2** SNP and Indels in cacao ESTs.

| Nucleotide Substitution | Bean and Leaves | Defense Related ESTs | Young red leaves | Total |
|---|---|---|---|---|
| C/T | 293 | 194 | 11 | 498 |
| G/A | 263 | 188 | 1 | 452 |
| Total(Transition) | 556 | 382 | 12 | 950 |
| A/T | 206 | 61 | 6 | 273 |
| C/G | 118 | 71 | 0 | 189 |
| G/T | 272 | 286 | 4 | 562 |
| A/C | 109 | 135 | 0 | 244 |
| Total(Transversion) | 705 | 553 | 10 | 1268 |
| A | 233 | 10 | 4 | 247 |
| C | 181 | 12 | 3 | 196 |
| G | 264 | 21 | 1 | 286 |
| T | 265 | 27 | 0 | 292 |
| Total(Indel) | 943 | 70 | 8 | 1021 |
| Shannon index | 0.98 | 0.82 | 0.73 | 0.97 |

the SNP occurrence was even more frequent in maize (*Zea mays*), one SNP approximately every 48 bp and every 130 bp in 3′ untranslated regions and coding regions, respectively (Rafalski 2002a). SNPs occurred at 1.36 SNP/100 bp in oil palm ESTs (Riju *et al.* 2007), 1 SNP/706 bp in apple (*Malus domestica*) ESTs (Newcomb *et al.* 2006), 1 SNP/130 bp in beet root (Schneider *et al.* 2001), 1 SNP/45.7 bp and 1 indel/277 bp in sunflower (Kolkman *et al.* 2007) and 1 SNP/62 bp in ESTs of clonally propagated, predominantly out-crossing cassava (Lopez *et al.* 2005). The density of SNP in ESTs of cocoa is similar to soybean and maize but less frequent than sun flower and oil palm. The transition to transversion ratio (Ts/Tv) was found to be 0.75 (**Table 1**) in the cocoa genome. It shows that the nucleotide substitution is happening towards purine to pyramidine or pyramidine to purine than transition in bean and leaves and defense related libraries of cocoa. In general transitions occur at higher frequencies than transversions such as beet root (Schneider *et al.* 2001), maize (Batley *et al.* 2003) and oil palm (Riju *et al.* 2007). But, Ts/Tv ratio < 1 (more transversions than transitions) was seen in regulatory genes such as endonuclease reverse transcriptase and Tc1-like transposase (Hale *et al.* 2009). A recent study on grasshopper genome reveals that the majority of transitions of cytosine residues are at methylated sites (CpG dinucletoide). After accounting for this methylation effect, there was no significant difference between transition and transversion rates (Keller *et al.* 2007). Transversions were seen at higher frequencies than transitions in MA (mutation accumulated) lines of genomes of nematode *C. elegans*. The much lower Ts/TV ratios observed in MA-line genomes suggest that, genome wide, transversions might be more susceptible to selective purging than transitions in *C. elegans* natural populations (Denvera *et al.* 2009).

In cocoa, indels occur at high frequency of almost half the frequency of SNP sites. In case of oil palm, indels occurred at very lower frequency than SNPs (Riju *et al.* 2007). Indels may be produced by errors in DNA synthesis, repair, recombination or be due to the insertion and excision of transposable elements that often leaves a characteristic DNA footprint of several nucleotide bases. SNPs mined from ESTs of cocoa plants affected by witches' broom disease show potential to tag disease resistance in cocoa (Lima *et al.* 2009). Our study identifies additional set of SNP markers in cocoa for genome mapping and population genetics applications, marker assisted selection (MAS), cultivar identification and characterization of genetic resources.

## CONCLUSIONS

Our findings suggest a nonrandom nucleotide substitution pattern with a strong bias toward transversion mutations. High shannon's index estimates (>0.7) indicate the ESTs of cocoa to undergo almost equal probability of all 10 types of point mutations (SNP/Indels). The results of the study are given as an online database 'CEMID' to help cocoa researchers (http://www.riju.byethost31.com/cocoa/ccsnp.html). The result of this study has practical implications for cocoa breeders as a source of potential SNP /Indel markers after validation by Polymerase Chain Reaction. The report also gives location of many putative point mutations in expressed sequences of cocoa.

## ACKNOWLEDGEMENTS

We thank David Edwards (Australian Centre for Plant Functional Genomics, Institute for Molecular Biosciences and School of Land, Crop and Food Sciences, University of Queensland, Australia) for providing the license to use Auto_SNP and web resources providers for data collection and development. This work was supported by a grant from Department of Biotechnology (BTISnet), Government of India, New Delhi, India.

## REFERENCES

**Aquadro CF, Bauer DuMont V, Reed FA** (2001) Genome-wide variation in the human and fruitfly: a comparison. *Current Opinion in Genetics and Development* **11**, 627-634

**Barker G, Batley J, O'sullivan H, Edwards KJ, Edwards D** (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* **19**, 421-422

**Batley J, Barker G, Helen O'Sullivan, Edwards KJ, Edwards D** (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology* **132**, 84-91

**Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A** (2002) Insertion-deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Molecular Biology* **48**, 539-547

**Buckler ES, Thornsberry JM** (2002) Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology* **5**, 107-111

**Collins FS, Brooks LD, Charkravarti A** (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research* **8**, 1229-1231

**Coryell VH, Jessen H, Schupp JM, Webb D, Keim P** (1999) Alele-specific hybridization markers for soybean. *Theoretical and Applied Genetics* **101**, 1291-1298

**Couch JA, Zintel HA, Fritz PJ** (1993) The genome of the tropical tree *Theobroma cacao* L. *Molecular and General Genetics* **237 (1-2)**, 123-128

**Denvera DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledó JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, Baer CF** (2009) A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences USA* **106 (38)**, 16310-16314

**Figueira A, Janik J, Goldsbrough P** (1992) Genome size and DNA polymorphism in *Theobroma cacao*. *Journal of the American Society for Horticultural Science* **117**, 673-677

**Ganal MW, Thomas Altmann T, Roder MS** (2009) SNP identification in crop plants. *Current Opinion in Plant Biology* **12**, 1-7

**Gupta PK, Roy JK, Prasad M** (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science* **80**, 524-535

**Hale MC, McCormick CR, Jackson JR, DeWoody JA** (2009) Next-generation pyrosequencing of gonad ranscriptomes in the polyploid lake sturgeon

(*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics* **10**, 203

**Huang X, Madan A** (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877

**Kawuki RS, Ferguson M, Labuschagne M, Herselman L, Kim DJ** (2009) Identification, characterisation and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). *Molecular Breeding* **23**, 669-684

**Keller I, Bensasson D, Nichols RA** (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genetics* **3 (2)**, e22

**Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao W, Shintani DK, Burke JM, Knapp SJ** (2007) Single nucleotide polymorphisms and linkage disequilibrium in sunflower. *Genetics* **177**, 457-468

**Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A** (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Molecular Genetics and Genomics* **270**, 24-33

**Lima LS, Gramacho KP, Carels N, Novais R, Gaiotto FA, Lopes UV, Gesteira AS, Zaidan HA, Cascardo JCM, Pires JL, Micheli F** (2009) Single nucleotide polymorphisms from *Theobroma cacao* expressed sequence tags associated with witches' broom disease in cacao. *Genetics and Molecular Research* **8 (3)**, 799-808

**Lindblad-Toh K, Winchester E, Daly M, Wang D, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan JB, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES** (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics* **24**, 381-386

**Lopez C, Piègu B, Cooke R, Delseny M, Tohme J, Verdier V** (2005) Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). *Theoretical and Applied Genetics* **110**, 425-431

**Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EHA, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ, Laing WA, McArtney A, Nain B, Ross GS, Snowden KC, Souleyre EJF, Walton EF, Yauk YK** (2006) Analyses of expressed sequence tags from apple. *Plant Physiology* **141**, 147-166

**Nikiforov TT, Rendle RB, Goelet P, Rogers YH, Kotewicz ML, Anderson S, Trainor GL, Knapp M** (1994) Genetic bit analysis: A solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Research* **22**, 4167-4175

**Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M** (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**, 312

**Rafalski A** (2002a) Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5**, 94-100

**Rafalski A** (2002b) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science* **162**, 329-333

**Riju A, Chandraseker A, Arunachalam V** (2007) Mining for single nucleotide polymorphisms and insertions/deletions in expressed sequence tag libraries of oil palm. *Bioinformation* **2 (4)**, 128-131

**Schneider K, Weisshaar B, Borchardt DC, Salamini F** (2001) SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. *Molecular Breeding* **8**, 63-74

**Shannon CE, Weaver W** (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, pp 29-125

**Syvanen AC** (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* **2**, 930-942

**Syvanen AC** (2005) Toward genome-wide SNP genotyping. *Nature Genetics* **37**, S5-S10

**Taramino G, Tingey S** (1996) Simple sequence repeats for germplasm analysis and mapping in maize. *Genome* **39**, 277-287

**Van K, Hwang EY, Kim YM, Park HJ, Lee SH, Cregan PB** (2005) Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. *Journal of Heredity* **96**, 529-535

**Varshney RK, Beier U, Khlestkina EK, Kota R, Korzun V, Graner A, Börner A** (2007) Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency and application for genome mapping and diversity studies. *Theoretical and Applied Genetics* **114**, 1105-1116

**Wiltshire T, Pletcher MT, Batalov S, Whitney Barnes S, Tarantino LM, Cooke MP, Wu H, Smylie K, Santrosyan A, Copeland NG, Jenkins NA, Kalush F, Mural RJ, Glynne RJ, Kay SA, Adams MD, Fletcher CF** (2003) Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proceedings of the National Academy of Sciences USA* **100**, 3380-3385