Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Deletions in the genomes of fifteen inbred mouse lines and their possible implications for fat accumulation*

SCHMITT Armin O.†[1], DEMPFLE Astrid[2], BROCKMANN Gudrun A.[1]

([1]*Institute for Animal Sciences, Humboldt-Universität zu Berlin, Invalidenstraße 42, 10115 Berlin, Germany*)

([2]*Institute for Medical Biometry and Epidemiology, Phillips-Universität Marburg, Bunsenstraße 3, 35037 Marburg, Germany*)

†E-mail: armin.schmitt@agrar.hu-berlin.de

Received Sept. 16, 2007; revision accepted Sept. 18, 2007

Abstract: Copy number variants (CNVs) are pieces of genomic DNA of 1000 base pairs or longer which occur in a given genome at a different frequency than in a reference genome. Their importance as a source for phenotypic variability has been recognized only in the last couple of years. Chromosomal deletions can be seen as a special case of CNVs where stretches of DNA are missing in certain lines when compared to the reference genome of the mouse line C57BL/6, for example. Based upon more than 8 million single nucleotide polymorphisms (SNPs) in the fifteen inbred mouse lines which were determined in a whole genome chip based resequencing project by Perlegen Sciences, we detected 20166 such long chromosomal deletions. They cover altogether between 4.4 million and 8.8 million base pairs, depending on the mouse line. Thus, their extent is comparable to that of SNPs. The chromosomal deletions were found by searching for clusters of missing values in the genotyping data by applying bioinformatics and biostatistical methods. In contrast to isolated missing values, clusters are likely the consequence of missing DNA probe rather than of a failed hybridization or deficient oligos. We analyzed these deletion sites in various ways. Twenty-two percent of these deletion sites overlap with exons; they could therefore affect a gene's functioning. The corresponding genes seem to exist in alternative forms, a phenomenon that reminds of the alternative forms of mRNA generated during gene splicing. We furthermore detected statistically significant association between hundreds of deletion sites and fat weight at the age of eight weeks.

Key words: Copy number variants (CNVs), Chromosomal deletions, Single nucleotide polymorphisms (SNPs), Resequencing, Cluster analysis, Association between genotype and phenotype

doi:10.1631/jzus.2007.B0777 **Document code:** A **CLC number:** Q78

INTRODUCTION

Copy number variants (CNVs) gained much attraction in the last couple of years. In two seminal papers 76 and 255 CNVs, respectively, were identified in a genome wide study in human (Sebat *et al*., 2004; Iafrate *et al*., 2004). On a chromosome wide scale, CNVs were identified using the technique of comparative genome hybridization (CGH) about a decade ago (Pinkel *et al*., 1998). CNVs were found to influence disease susceptibility, e.g., for AIDS if the number of copies of the chemokine receptor gene *CCL4L1* is too low (Gonzalez *et al*., 2005) or to cause disease directly. One example for the latter would be the well known Prader-Willi-syndrome, where a partially deleted chromosome 15 causes a range of severe physical and mental disablements in human. Deletions were also found to be the key driving force in the divergence between humans and chimpanzees (Britten, 2002). They should therefore be seen as one of the main factors driving evolution in general. Deletions in the genome can be considered as a special case of CNVs, where one individual has no copy of a certain piece of DNA and other individuals have one or several copies. Pieces of DNA can above all get lost through unequal crossing over during replication (Graur and Li, 1991). Other mechanisms

* Project supported by the German Ministry of Education and Research (BMBF) through the National Genome Research Network (NGFN) (Nos. 01GS0486 and 01GR0460) and the Deutsche Forschungsgemeinschaft (DFG) for a Travel Grant to Armin O. Schmitt

leading to deletions are presented by Stankiewicz and Lupski (2002). The deletions and other CNVs found in various studies are collected in the Database of Genomic Variants (Zhang *et al*., 2006).

While uncalled alleles, so called null genotypes, are usually excluded from further analysis in medical studies, it was shown that they can be exploited to derive deletions from single nucleotide polymorphism (SNP) data (McCarroll *et al*., 2006; Conrad *et al*., 2006). The basic idea is that uncalled alleles should cluster in an ordered genotype data set if they are caused by missing DNA probe, but they should be randomly scattered in the data set if they are caused by occasionally failed oligos or other sources of noise (Fig.1). Here, we apply a clustering algorithm to distinguish between the two types of uncalled alleles. Obvious questions that we address are: Can putatively functional DNA get lost? Can deletions affect the phenotpye?

MATERIALS AND METHODS

SNP data

We downloaded more than 8 million SNPs which were detected in a resequencing project at Perlegen Sciences (Frazer *et al*., 2007) (http://mouse.perlegen.com/mouse). In this project, all of the unique mouse genome was probed with oligomers derived from the line C57BL/6. Repetitive sequence cannot be examined this way, so 1500 megabases resulted. Unmapped SNPs and SNPs in amplificates whose primer pairs were marked as failed by Perlegen Scinences were excluded. The fifteen mouse lines whose genomes were resequenced are standard mouse lines (Fig.2).

Algorithm to identify deletion sites

We considered uncalled alleles as caused by missing DNA probe due to deletions if the following three criteria were met:

1. The same mouse lines have three or more uncalled alleles at consecutive SNP positions.

2. The uncalled alleles span 1 kb ore more.

3. The distance between neighbouring SNPs does not exceed 1 kb.

The last requirement had to be met in order not to bridge too much of non-genotyped DNA. We mapped the genotypes onto a two-letter alphabet. One symbol stands for the four nucleotides A, C, G, and T.

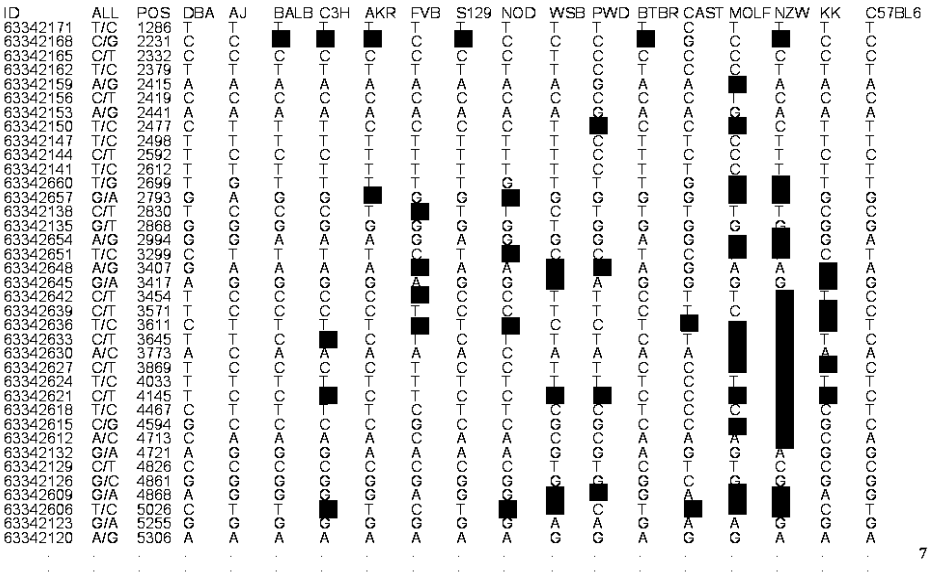

Fig.1 A section from a typical genotyping data file. Rows represent genotypes, columns represent mouse lines. Missing genotypes are marked with black squares. At first glance they appear to be pretty scattered over the data. After a closer look, a long sequence of missing genotypes can be recognized for the mouse line NZW (the third column from the right) and, to a lesser extent, for MOLF (the fourth column from the right). Missing genotypes clustering at certain lines are probably the consequence of missing DNA, whereas scattered missing genotypes are probably due to deficient oligos or other noise

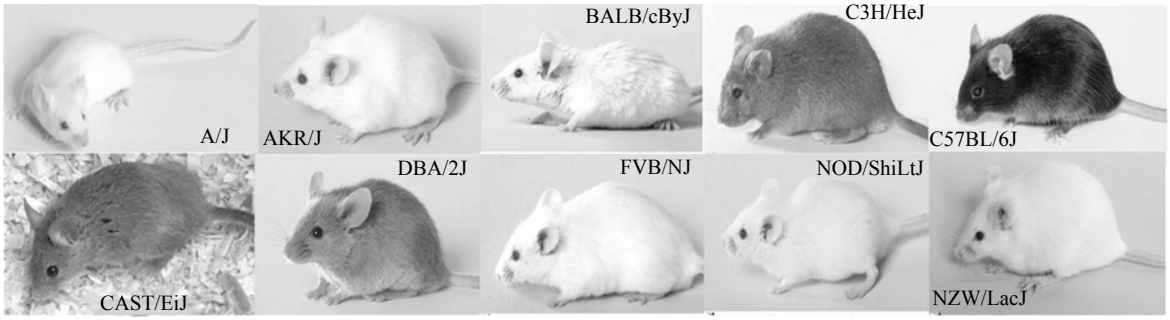

Fig.2 Ten of the inbred mouse lines whose genomes were resequenced by Perlegen Sciences. C57BL/6J was used as reference line, i.e. the genomes of all other mouse lines were compared against its genome. The six other lines for which no pictures are available are 129S1/SvImJ, WSB/EiJ, PWD/PhJ, BTBR T^{+}tf/J, MOLF/EiJ, and KK/HlJ. Reproduced with kind permission from the Mouse Phenome Project (the Jackson laboratory)

The other stands for uncalled alleles. We were then looking for stretches of at least three SNPs where the identical mouse lines have uncalled alleles. All such stretches were extracted and subsequently criteria 2 and 3 were applied to extract the deletion sites. In order to assess if deletion sites could also be observed in the case where the uncalled alleles were randomly scattered across the mouse lines, we generated a dataset of scrambled genotypes, i.e., we fixed the number of uncalled alleles that were found at a specific chromosomal position, but assigned them randomly to the mouse lines. Then the search algorithm was applied to this scrambled dataset.

Association between deletion sites and quantitative traits

We downloaded the measurements for average fat content at the age of eight weeks from the Mouse Phenome Database (http://www.jax.org/phenome) (Bogue *et al*., 2007). The symbol for this trait was 'tissuemass_fat8' and the filename containing all the measurements was 'summstats.zip'. We extracted the corresponding measurements for male animals. Phenotypic measurements for all genotyped lines were available except for the line PWD/PhJ. For each deletion site two groups of mouse lines were formed, one group with the deletion and the other without the deletion. We tested all ~20000 deletion sites for association using the *t*-test for clustered data (Donner *et al*., 1981), where each line was considered as one cluster. To this end, the R-function *t*.test.cluster in library Hmisc by Frank Harrell was adopted for our needs. The *P*-values were corrected using the method of Bonferroni.

Analysis of the chromosomal location of deletion sites

In order to examine if exons were affected by deletions, we compared their positions with the positions of all exons in the genome database of Ensembl release 46 (http://www.ensembl.org) using bioperl modules. The position of the deletion sites are defined by the chromosomal position of their first and their last SNP. We confirmed these positions by sequence comparison of the SNP flanking sequence of 100 bp upstream and 100 bp downstream of the SNP against the mouse genome (National Center for Biotechnology Information (NCBI) build 36, Ensembl database release 46) using BLAST (Altschul *et al*., 1997). When the first and the last SNP flanking sequences could be mapped to the same chromosomal locations as indicated in the original data file with an *E*-value of 10^{-50} or better, we accepted a deletion site for further analysis.

RESULTS

Applying the algorithm sketched in the section MATERIALS AND METHODS we identified 20885 deletion sites. The position of 20166 deletion sites with a total length of 31.8 Mb could be confirmed by mapping the sequence of the first and the last SNPs of each deletion site onto the current assembly of the mouse genome (NCBI build 36, Ensembl database release 46). In our scrambled data set we only found 52 deletion sites with a total length of 65 kb. This is only a minute proportion of the results obtained from real data. There is, thus, good evidence that the

deletion sites that we found are no artefacts. On average, 5.4 Mb were lost in a mouse line, with a minimum of 4.4 Mb in the line A/J and a maximum of 9.3 Mb in the line PWD/PhJ. Next, we were interested if exclusively so called junk DNA gets lost or if also presumably functional DNA is affected. To this end, we compared the positions of the deletions sites, as defined by the position of the first and the last SNPs in it, with the positions of the exons in the genome. Surprisingly, 2594 (12.9%) of the deletion sites overlap with at least one exon. Altogether, exons of 2104 out of 28545 confirmed and predicted mouse genes (7.3%) and a total of 4121 murine exons (1.75% of all 235412 exons, including predicted ones), are affected by deletions in at least one mouse line. We tested all 20166 deletion sites with confirmed location for association with the fat weight at the age of eight weeks. Two thousand two hundred and sixty-six deletion sites had a significant correlation with this trait (significance level α=0.001 after Bonferroni correction). As an example, we show the fat distribution for the two groups of lines, one with the deletion, the other without the deletion, at the chromosomal location between 93339270 bp and 93340305 bp on chromosome 13 (Fig.3). The first SNP of this deletion site is ss52681234, the last ss52681232. This deletion site overlaps the gene *Msh3*, which was previously associated with DNA mismatch repair (Edelmann *et al*., 2000).

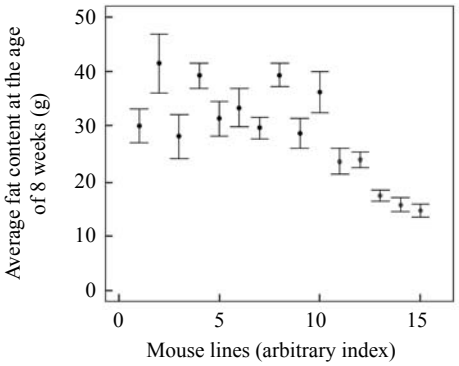

Fig.3 At position 93340 kb of chromosome 13, five of the fifteen genotyped and phenotyped mouse lines have a deletion (indices 11 to 15) of about 1 kb, and ten do not (indices 1 to 10). These two groups segregate significantly with respect to fat weight at the age of eight weeks ($P\approx10^{-18}$; *t*-test for clustered data; Bonferroni corrected). Measurements were taken from between 7 (MOLF/EiJ) and 17 (AKR/J) male animals. The error bars represent the standard deviation

DISCUSSION AND CONCLUSION

We have shown that uncalled alleles obtained in a genome-wide SNP experiment lend themselves to the identification of genomic deletions. A straightforward algorithm which is largely based upon the coherence between uncalled alleles in mouse lines, is capable of discerning clusters of uncalled alleles from uncalled alleles which are scattered across the data. As a first validation, we scrambled the genotype data and obtained only a tiny fraction of the deletion sites that we obtained in the real data. The extent of the deletions that we extracted in 15 inbred mouse lines, on average 5.4 Mb, is comparable to that of the SNPs, which was 8 million. Of course, we are aware that a rigorous validation of the deletion sites that we identified would include sequencing of mouse lines with and without the predicted deletion. A further improvement of this analysis would consist in the merging of neighbouring deletion sites. Since we wanted to be on the safe side, we did not bridge regions that were not genotyped. This leads to a strong fragmentation of the putatively longer deletion sites. Another validation would be the comparison of the deletion sites that we detected with those that Li *et al*.(2004) found in 14 inbred lines using whole-genome BAC arrays. We found furthermore that a significant proportion of the deletion sites overlaps with exons, which should lead to functional consequences. As an example, we were looking for association between the deletion sites and the fat content of mice at the age of eight weeks. About 11% of the deletion sites showed a significant association with this trait. There is, therefore, evidence that chromosomal deletions play an important role for quantitative traits in general. Next to SNPs and epigenetic changes, their careful analyses should gain more attention.

ACKNOWLEDGEMENT

We thank Ivo Große (University of Halle, Germany), Richard Mott (Wellcome Trust Centre for Human Genetics, Oxford, UK), Geoff Nilsen and Brian Karlak (both at Perlegen Sciences, Mountain View, USA), Bert Overduin and Xose Fernandez (both at European Bioinformatics Institute, Hinxton,

UK) for stimulating discussions and valuable assistance.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic. Acids Res.*, **25**(17):3389-3402. [doi:10.1093/nar/25.17.3389]

Bogue, M.A., Grubb, S.C., Maddatu, T.P., Bult, C.J., 2007. Mouse Phenome Database (MPD). *Nucleic. Acids Res.*, **35**(Database issue):643-649. [doi:10.1093/nar/gkl1049]

Britten, R.J., 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *PNAS*, **99**(21):13633-13635. [doi:10.1073/pnas.172510699]

Conrad, F.D., Andrews, T.D., Carter, N.P., Hurles, M.E., Pritchard, J.K., 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**(1):75-81. [doi:10.1038/ng1697]

Donner, A., Birkett, N., Buck, C., 1981. Randomization by cluster. *Am. J. Epidemiol.*, **114**(6):906-914.

Edelmann, W., Umar, A., Yang, K., Heyer, J., Kucherlapati, M., Lia, M., Kneitz, B., Avdievich, E., Fan, K., Wong, E., *et al.*, 2000. The DNA mismatch repair genes *Msh3* and *Msh6* cooperate in intestinal tumor suppression. *Cancer Res.*, **60**(4):803-807.

Frazer, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R.V., Montgomery, J., Morenzoni, M.M., Nilsen, G.B., *et al*., 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, **448**(7157):1050-1053. [doi:10.1038/nature06067]

Gonzalez, C., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., *et al*., 2005. The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**(5714):1434-1440. [doi:10.1126/science.1101160]

Graur, D., Li, W.H., 1991. Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, MA, USA, p.15.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.D., Qi, Y., Scherer, S.W., Lee, C., 2004. Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**(9):949-951. [doi:10.1038/ng1416]

Li, J., Jiang, T., Mao, J.H., Balmain, A., Peterson, L., Harris, C., Rao, P.H., Havlak, P., Gibbs, R., Cai, W.W., 2004. Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.*, **36**(9):952-954. [doi:10.1038/ng1417]

McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dalaire, S., Gariel, S.B., Lee, C., Daly, M.J., Altshuler, D.M., 2006. Common deletion polymorphims in the human genome. *Nat. Genet.*, **38**(1):86-92. [doi:10.1038/ng1696]

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., *et al*., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**(2):207-211. [doi:10.1038/2524]

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., *et al*., 2004. Large-scale copy number polymorphism in the human genome. *Science*, **305**(5683):525-528. [doi:10.1126/science.1098918]

Stankiewicz, P., Lupski, J.R., 2002. Genome architecture, rearrangements and genomic disorders. *TRENDS in Genetics*, **18**(2):74-82. [doi:10.1016/S0168-9525(02)02592-1]

Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., Scherer, S.W., 2006. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.*, **115**(3-4):205-214. [doi:10.1159/000095916]