HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2009 October 01.

Published in final edited form as:

Nat Genet. 2009 April ; 41(4): 430–437. doi:10.1038/ng.350.

The impact of copy number variation on local gene expression in mouse hematopoietic stem/progenitor cells

Patrick Cahan, **Yedda Li**, **Masayo Izumi**, and **Timothy A. Graubert**

Department of Internal Medicine, Division of Oncology, Stem Cell Biology Section, Washington University, St. Louis, MO

Abstract

The extent to which differences in germ line DNA copy number contribute to natural phenotypic variation is unknown. We analyzed the copy number content of the mouse genome to a sub-10 kb resolution. We identified over 1,300 copy number variant regions (CNVRs), most of which are < 10 kb in length, are found in more than one strain, and, in total, span 3.2% (85 Mb) of the genome. To assess the potential functional impact of copy number variation, we mapped expression profiles of purified hematopoietic stem/progenitor cells, adipose tissue and hypothalamus to CNVRs *in cis*. Of the more than 600 significant associations between CNVRs and expression profiles, most map to CNVRs outside of the transcribed regions of genes. In hematopoietic stem/progenitor cells, up to 28% of strain-dependent expression variation is associated with copy number variation, supporting the role of germ line CNVs as major contributors to natural phenotypic variation in the laboratory mouse.

INTRODUCTION

Copy number variants (CNVs), currently defined as genomic sequences greater than one kilobase that are polymorphic in copy number, have been identified in diverse species including human, chimp, rat, mouse, and drosophila1–10. In the short interval since the discovery of wide-spread copy number variation in apparently healthy individuals, there has been rapid expansion of both CNV detection techniques and their application across a range of biological samples and species. From these studies, it is apparent that copy number variation exceeds single nucleotide polymorphisms (SNPs) as a source of genetic variation, and that many CNVs contain or overlap genes and, thereby, may have functional effects. However, the role of copy number variation in mediating both ‘normal’ phenotypic variation and disease susceptibility is only beginning to emerge11–14.

Fundamental questions about the nature and impact of CNVs remain unanswered, mainly due to methodological constraints. We set out to determine the copy number variable content of the mouse genome and estimate its functional impact, as measured by gene

Corresponding Author: Timothy Graubert, MD, Washington University School of Medicine, Division of Oncology, Stem Cell Biology Section, Campus Box 8007, 660 South Euclid Avenue, St. Louis, MO 63110, Phone: 314/747-4437, Fax: 314/362-9333, graubert@wustl.edu.

expression profiling *in vivo*. The inbred mouse is an ideal model organism for this study for several reasons, including its homozygous genome, the ease with which biological samples can be acquired, and the preeminent role of the mouse as a model for biomedically relevant traits and diseases. Gene expression variation is a trait amenable to genetic mapping because it is easily quantified *in vivo*, it is the phenotype most proximally related to genetics, and the expression of all genes can be measured simultaneously. Finally, it is reasonable to hypothesize that the effect size of structural variations on gene expression will be large, so that a genome-wide association study could be informative, even with modest sample sizes.

RESULTS

CNV detection, genotyping, and validation

To map the CNV content of the mouse genome, we selected 17 Tier 1–3 Mouse Phenome Project strains15 and three additional strains of biomedical interest (LG/J, NZB/BINJ, 129X1/SvJ), representing all major inbred lineages. We performed comparative genomic hybridization using a long-oligonucleotide array containing 2,149,887 probes evenly spaced across the reference genome with a median inter-probe spacing of 1,015 bases. We performed segmentation using wuHMM, a Hidden Markov Model algorithm that utilizes sequence-level information and can detect CNVs less than 5 kb in length (fewer than five probes) at a low false positive rate16. wuHMM scores CNVs based on the number and median log2-ratio of the probes comprising the prediction, such that calls with higher scores are more likely to represent true events. CNVs called in different strains that overlap can be assigned different boundaries due to technical or biological sources of variability. Because fine-mapping all putative CNVs is not feasible at present, a common approach to handling complexity and ambiguity in CNV boundaries is to treat overlapping CNVs as a unit, or, copy number variable region (CNVR)4. We merged overlapping wuHMM calls into CNVRs, some of which have complex architectures (Figure 1). We refer to CNVRs as ‘complex’ or ‘simple,’ as determined by wuHMM boundary concordance across strains (see Methods). To assign CNVR genotype calls to strains for QTL mapping and to improve upon the sensitivity of wuHMM, we clustered the log2-ratios of each CNVR (see Methods). The number of genotypes per CNVR was determined by selecting the cluster number that maximized the average silhouette function, which is a measure of clustering quality17. Genotypes were assigned according to the clusters in which strains were grouped. We refer to genotypes that differ from the reference strain’s genotype as ‘abnormal’ in complex CNVRs, and as ‘gain’ or ‘loss’ in simple CNVRs if the mean log2-ratio is greater or less than the reference sample, respectively.

Using initial parameters, wuHMM identified 10,681 putative CNVs which were merged into 3,359 CNVRs. To determine the false positive rate (FPR) of our CNV predictions, we randomly selected 61 short CNVRs for independent validation by qualitative (for losses) or quantitative (for gains) PCR (qPCR). The FPR approached 0 for CNVRs with average scores exceeding 1.5 and 2.5 for gains and losses, respectively (Supplementary Table 1). Therefore, we selected these score thresholds, resulting in an empirically estimated individual strain CNVR genotype FPR < 4.0%. For complex CNVRs, the same threshold was applied if the region contained either wuHMM gains or losses exceeding the

corresponding threshold. We called the 1,333 CNVRs that passed these thresholds 'high-confidence' CNVRs and retained them for further analysis and quantitative trait mapping (Supplementary Figure 1, available at http://graubertlab.dom.wustl.edu/downloads.html, and Supplementary Table 2).

Copy number variation in the inbred mouse genome

The 1,333 high-confidence CNVRs span 85 million non-redundant bases (3% of the genome) and are distributed across all 19 autosomes and the X chromosome (Figure 2). The CNVRs range in length from 1,871 bases to 3.84 Mb (mean length is 64 kb, median is 9 kb, over 50% are less than 10 kb) (Figure 3A). Although the length distribution of CNVRs is highly right-skewed, confirming previous estimates derived from CNVR mapping studies performed with lower resolution platforms18 and paired-end mapping19, the overall contribution of small CNVRs (i.e., less than 10 kb) to the total copy number variable content of the genome makes up only 3.3 Mb (0.13%) (Figure 3B), a finding consistent across all strains (Supplementary Figure 2). Complex CNVRs make up 23% of all CNVRs, but 63% of the CNV sequence content. The majority of small CNVRs are exclusively genotyped as losses (82%), probably reflecting the increased power to detect homozygous losses versus integral gains with a small number of aCGH probes. We detected a total of 663 gains, 2,854 losses, and 2,772 abnormal CNVR genotypes. 67% of CNVRs were called as gain, loss, or abnormal in more than one strain. The number of CNVR gains, losses, or abnormal genotypes ranges from 215 (C58/J) to 413 (KK/HIJ) per strain (mean = 331). The total CNV sequence per strain ranges from 26.4 (C58/J) to 48.3 (NOD/ShiLtJ) Mb (mean = 39.1 Mb); no single strain contributed disproportionately to the CNVR map (Figure 3C and Supplementary Figure 2).

Several previous reports have investigated the extent of copy number variation in inbred strains of mice1,2,5,20,21. If *de novo* events contribute only minimally to copy number variation among individuals within a strain22,23, then as detection technologies improve, studies assaying the same strains will have increasingly concordant results. We compared our CNVR map to previous reports that also used high-density oligonucleotide aCGH (see Methods). We found that when we compared CNVRs defined using strains in common with other studies, our map largely recapitulated the CNVRs found in the other studies: 64–84% of CNV content in the other studies was also detected in our high-confidence CNVRs (Supplementary Table 3). 48–87% of the copy number variable content that we report in the 19 strains is novel. However, when we compared CNVR maps regardless of strain we found that only 16% of the copy number variable content in our map was novel, suggesting that much of the total copy number variable sequence of the reference genome is known at the presently available detection limit.

Non-allelic homologous recombination (NAHR) has been proposed as a mechanism of CNV formation24. The hypothesis that segmental duplications (sequences >1 kb and having > 90% similarity to at least one other genomic region) act as nurseries of CNV by promoting NAHR has been supported by the enrichment of segmental duplications within and around CNVRs20,25. By permutation testing (see Methods), we found that there is significantly more segmental duplication sequence within and directly bordering medium (10–100 kb)

and large (>100 kb) CNVRs (fold = 3.0 and 12.9, respectively, P < 0.01), but that segmental duplications are found less often than expected by chance within and near small (<10 kb) CNVRs (fold = 0.37, P < 0.01) (Figure 4 and Supplementary Table 4), consistent with a prior report of stronger association between segmental duplications and long CNVRs4. The pattern of enrichment of segmental duplication sequences near medium and large CNVRs extends to 2 Mb beyond the CNVR boundaries (fold from 2.25 - 1.43 and 7.80 - 2.45, respectively, P < 0.01) as does the pattern of depletion around small CNVRs (fold = 0.27 – 0.75, respectively, P < 0.01). Like segmental duplications, it has been suggested that repetitive elements may facilitate CNV generation through NAHR. Indirect evidence supporting this hypothesis has been presented in inbred mice where LINEs are enriched within segmental duplications20. We found that LINEs are enriched within medium and large CNVRs (fold = 1.61 and 1.50, P < 0.01), but are not enriched in small CNVRs (fold = 0.95, P = 0.81). We found an enrichment of LINE elements in sequences flanking all CNVRs types, although the association is less for small CNVRs (fold = 1.14 for small, 1.51 for medium, 1.43 for large, P< 0.01). Therefore, it is unlikely that small CNVRs are variations in the copy number of repetitive elements themselves26, but rather LINEs may facilitate the removal or expansion of neighboring sequence. Long terminal repeats (LTRs) are enriched within all CNVRs (fold = 1.3, 1.4, 1.53, P<0.01). This association persists for regions surrounding CNVRs to at least 10 kb for medium and large, but not small CNVRs. SINEs are depleted within and surrounding medium and large, but not small CNVRs (fold = 0.7, 0.45, P<0.01). Taken together, this analysis confirms that CNVRs greater than 10 kb frequently contain or directly border highly homologous elements of the genome that can facilitate NAHR and therefore CNVR generation. But, with the exception of the weak association between the regions surrounding small CNVRs and LINE sequences, there is no apparent genomic feature that could facilitate NAHR and give rise to the abundant, small, high-confidence CNVRs. Therefore, their origins will require detailed genomic analysis and further exploration.

We next determined the gene content of the high-confidence CNVRs, finding that 432 high-confidence CNVRs contain or partially contain 679 genes. Previous CNVR studies of the mouse genome have shown that CNVRs overlap coding sequence no more often than expected by chance, in contrast to CNVRs in human and rat genomes which appear to be enriched for gene content1,4,8,21. With a more comprehensive and finer-resolution map, we retested this hypothesis by permutation analysis. We found that small, medium and large CNVRs are found in genic regions less frequently than expected by chance (fold=0.86, 0.71, 0.90 respectively, P<0.01, 0.01, 0.05) (Figure 4).

Expression profiling

To estimate the overall impact of CNV on gene expression *in vivo*, we first performed expression profiling of hematopoietic stem/progenitors cells using the Illumina Mouse Beadchip-6v1 platform (see Methods). Among many cell types and tissues suitable for this study we chose to profile a population that has well-defined surface markers, enabling the enrichment of a highly purified cell population that is transcriptionally active27, increasing the number of genes that could be assessed for association with CNVRs. We pooled bone marrow cells from two individuals from each strain and analyzed 2–3 biological replicates

per strain (46 expression experiments). 29% of the probes on the array were detected as 'present' in at least three strains (see Methods). To validate the sort purity, we examined the expression profiles of the cell surface markers utilized in the sort strategy and found that they were consistent with the immunophenotype of the post-sort products (Supplementary Figure 3).

To determine the extent to which expression variation is associated with copy number variation, we first identified the genes that exhibit strain-specific expression. We identified 1,469 probes with significantly higher between- versus within-strain expression differences ($P < 0.01$, see Methods). We also determined the strain-specific expression profiles in epididymal adipose tissue and hypothalamus, as those data sets were publicly available28,29. We removed expression data for strains that were not profiled in our CNVR mapping work, leaving 15 strains from each study. Since no strain replicates were available in these studies, we identified strain-specific probes as those with a ratio of maximum to minimum expression > 3, the same threshold used to identify 'variable' expression traits in those studies (Table 1). It is impossible to determine if the differences in the number of 'Present' and strain-specific expression traits between tissues is due to fundamental differences in cross-tissue expression variation or, more likely, to the significant differences in the expression profiling platforms and analysis methods utilized in these studies.

Expression quantitative trait mapping

CNVRs may impact local gene expression through a variety of mechanisms, including gene dosage, removal or relocation of regulatory material, or 'neighborhood effects' that disrupt local chromatin structure30. We estimated the overall contribution of CNVRs on local expression by *in silico* eQTL mapping, in which gene expression profiles were treated as quantitative traits and CNVRs as genetic markers. We limited the analysis to CNVR-expression traits that are tightly linked (< 2 Mb apart) because of reduced power to detect *trans* effects with a small sample size. We calculated eQTL significance using a weighted permutation method that accounts for the complex ancestral relationship among inbred strains31,32, and controlled the family-wise error rate arising from testing the association between a trait and multiple CNVRs by applying the Holm multiple testing correction to each trait's p-values separately33.

We identified 672 significant associations between strain-specific expression traits and CNVRs in the hematopoietic stem/progenitor compartment. Because we used an alpha threshold of 0.05, after correcting for multiple tests we would expect to find only 113 associations by chance. The number of traits associated with a CNVR (degree of pleiotropy) ranged from 1–18 (mean=2.47, median=2); the number of CNVRs associated with a trait ranged from 1–9 (mean=1.65, median=1). While there were more eQTLs in which the Illumina probe sequence overlapped the CNVR than expected by chance ($P < 0.05$ by Fisher's Exact Test), most eQTLs (92.3%) map outside of the corresponding CNVR. If these intergenic CNVRs mediate expression variation, they do so via mechanisms other than changes in gene dosage. CNVRs of each categorization, either by size or complexity, were found to be eQTLs and each was as likely to be an eQTL as expected by chance. After selecting the most significant association per trait from the 672 eQTLs, we found that 408

strain-specific expression traits representing 391 genes (27.8% of 1,469 strain-specific traits) were associated with 214 CNVRs (16% of all 1,333 CNVRs and 44.2% of the 484 testable CNVRs) (Table 1 and Supplementary Table 5). The frequency of eQTLs dropped with increasing distance from CNVR boundaries to expression probe locations (proximity) (Supplementary Figure 4). Similarly, the fraction of expression variation explained by a trait's association with a CNVR decreased significantly with proximity (Supplementary Figure 4)

To validate the KL eQTLs, we queried the expression profiles of the 391 eQTL-associated genes in Kit+/Lineage−/Sca1+ (KLS) hematopoietic stem cells purified from BXD recombinant inbred mice34. Because the BXD mice are homozygous for either the C57BL/6J or DBA/2J genotype at most loci and SNP genotype data is publicly available, we were able to assign an inferred CNVR genotype based on the parental strain of origin of the SNP markers spanning each CNVR (Supplementary Table 6). Of the 160 KL eQTL-associated genes that were unambiguously annotated with a gene symbol, 74 genes (93 probe sets) were present on the Affymetrix U74A expression platform and 31 were detected as expressed in >80% of the RI lines. We found that 29% of these testable eQTL-associated genes had expression profiles that were also associated with the inferred CNVR genotype in the KLS BXD data (P-value < 0.05) (Supplementary Table 7).

Smaller proportions of strain-specific expression variation were associated with CNVRs in the other two tissues that we were able to analyze: 181 of 4,083 (4.4%) and 78 of 2,879 (2.7%) strain-specific traits in adipose tissue and hypothalamus, respectively, after selecting the most significant associations per trait (Table 1). Similarly, fewer CNVRs were detected as eQTLs: 24.9% and 15.0% of testable CNVRs in adipose tissue and hypothalamus, respectively. While there is variability in the impact of CNV on expression variation between tissues, differences in the number of eQTLs we detected in adipose tissue and hypothalamus cells are likely due to the reduced power (25% fewer samples) and less robust methods used to identify strain-specific expression in these data. The relationships between eQTL frequency and proximity, and between eQTL effect size and proximity, were present to a lesser extent in adipose and were not present in the hypothalamus (Supplementary Figure 4). As we found in the hematopoietic compartment, few adipose and hypothalamus eQTLs overlapped their associated traits (6.0% and 6.4%, respectively), but this was more than expected by chance (P<1e-5 and P<0.01 in adipose and hypothalamus, respectively). CNVRs across all length and complexity ranges were observed as eQTLs; no categorization was enriched or depleted.

Next, we asked whether any eQTLs were shared across tissues. Because we utilized expression data from different platforms, we defined expression trait overlap at the level of gene annotation rather than probe sequence. We found twenty-three eQTLs present in more than one tissue, five of which were gene-dosage effects (Figure 5 and Table 2). A correlation between *Alad* gene dosage, mRNA abundance, and enzymatic activity was previously demonstrated35,36 and *Alad* expression variation was associated with a *cis*-eQTL reported in an F2 inter-cross37, demonstrating that our analysis was able to detect known gene dosage eQTLs. Further, we found that strain-specific *Glo1* over-expression is due to a large gain and that this gene-dosage effect is consistent across all three tissues that we tested

(Figure 6A). A strain-specific expression pattern of *Glo1* in hypothalamus was previously shown to be associated with and potentially casual for anxiety-related behavior38. Our analysis is the first, to our knowledge, to show that this expression variation is due to a CNV. Most eQTLs are found in only one tissue, indicating that tissue-specific factors compensate for CNVR-mediated gene expression variation. For example, the expression of guanylate-binding protein 1 (*Gbp1*) is associated with a CNVR containing its 3′-exon and 3′-UTR in hematopoietic and adipose cells, but not hypothalamus (Figure 6B). The expression pattern of *Gbp1* (highly expressed in both hematopoietic stem/progenitor cells and adipose tissue in strains that contain the CNV, but not expressed at detectable levels in strains without the CNV or in the hypothalamus regardless of CNV genotype) is consistent with a model of expression regulation where hypothalamus-specific down-regulation or alterative splicing of *Gbp1* overcomes the CNVR effect apparent in other tissues.

We reasoned that CNVRs that mediate expression variation by large scale disruption or modification of local chromatin structure rather than by gene dosage were likely to impact the expression of more than one gene. We tested one implication of this hypothesis using random permutations of the hematopoietic eQTL data. We calculated the probabilities of finding the observed number of CNVRs with a given degree of pleiotropy (defined as the number of expression traits associated with a CNVR). We found that there were more CNVRs with 7 and 8 associated expression traits than expected by chance ($P < 0.05$, 10,000 permutations). One CNVR (CNVR-ID 3014) with seven associated traits is a deletion located approximately 100 kb from the Major histocompatibility (*Mhc*) locus on chromosome 17 that removes highly conserved sequence with predicted regulatory potential. All of the associated traits are *Mhc* class Ib genes, many of which are expressed in multiple tissues and have unknown specific functions39. Genes at this locus have been speculated to undergo distal regulation via a chromosomal looping mechanism40 and, therefore, copy number changes that modify this looping structure would be expected to have pleiotropic effects on local expression. Alternatively, because the *H2-T* locus is known to have strain-specific duplications39, it is possible that the expression variation that we observed was due to gene dosage differences that are too complex for our computational methods to properly detect but are, in effect, tagged by the associated CNVR.

DISCUSSION

The central goal of our work was to estimate the functional impact of germ line copy number variation *in vivo*. To achieve this goal, we first identified CNVRs in twenty inbred strains at the highest resolution reported to date. We discovered 1,333 CNVRs spanning approximately 3% of the mouse genome. On average, there are over 300 CNVs per strain. As predicted, we found that the frequency of CNVRs increased with decreasing CNVR length, but that short CNVs account for only a small fraction of the total copy number variable sequence content of the mouse genome. We speculate that this trend will hold as higher resolution technologies are developed. Unexpectedly, we found that small CNVs (<10 kb) lack the enrichment of highly homologous sequences that frequently flank, and are presumed to contribute to the formation of medium (10–100 kb) and large (>100 kb) CNVs. Determining the mechanisms that generated these CNVs would facilitate the design of targeted assays to detect new CNVs and provide a better understanding of the forces that

shaped the mouse genome. We are aware of only one report documenting similar short deletions in a small number of human genomes and therefore a mouse-to-human CNVR comparison will be informative as high-resolution human data become available41. A caveat of our CNVR map is that, as is true for all comparative genomic hybridization experiments, we were limited to finding variants in comparison to a reference sequence; sequences that do not exist in the C57BL/6J genome but vary in copy number among other strains were not detected. Therefore, the total extent of copy number variation relative to the union of all inbred mouse genomes must await comprehensive sequencing of other strains. However, a reasonable estimate of the amount of mouse genomic sequence lost in the C57BL/6J strain is the amount of genomic material lost per strain relative to C57BL/6J, which ranged from 16.8 to 33.8 Mb (mean = 25.5 Mb).

Using a relatively small number of inbred mouse strains, we found that all classes of CNVs were associated with gene expression changes in a variety of tissues. We found that 28% of strain-specific expression traits were associated with copy number variation in the hematopoietic progenitor/stem compartment, consistent with the 18% previously reported in human lymphoblastoid cell lines42. To validate these eQTLs, we inferred the CNVR genotypes of the BXD RI panel and analyzed publicly available KLS expression data. Over 29% of the testable KL eQTLs were supported in the BXD data set, a striking concordance given the substantial experimental and biological differences between the studies. We also detected many CNVR eQTLs in adipose tissue and hypothalamus, even though these data were produced with different mice, using different expression platforms, and the eQTL analysis was performed with 25% fewer strains. Much of the recent speculation on the potential impact of CNVs on phenotypic variation has centered on gene-dosage effects43. However, we found that only 7.3% of CNVR eQTLs contain the associated expression probe and therefore were due to gene-dosage effects. Presumably, the remaining CNVR eQTLs reflect expression variation mediated by alteration of regulatory material or local chromatin structure. This would be consistent with a model where (subtle) alterations in expression patterns are better tolerated than complete or partial gene gains or losses.

Some of the CNVR eQTLs reported here may be in linkage disequilibrium with another allele causing the associated expression change, underscoring the need to characterize the relationship between CNVs and other genetic variants. It is likely that there are additional eQTLs not detected here: CNVRs that alter expression in only one or two strains, *trans* eQTLs, eQTLs that associate with genes expressed in tissues not sampled here, and eQTLs with weak effects. Increasing the number of strains and the tissues sampled would address some of these limitations. However, extending this work to a much larger population with greater genetic diversity (i.e., the Collaborative Cross44) would increase the power to detect *trans* and weaker effects and therefore enable a clearer understanding the overall impact of CNVR on expression variability. Future work must reach beyond identifying statistical associations to better characterize the mechanisms by which a CNVR affects phenotypic (including expression) variation. In addition to estimating the impact of CNVRs on expression variation, the CNVR eQTLs reported here may be of practical value in identifying the causal variants in traditional QTLs because they present plausible hypotheses linking genetic differences between inbred strains to complex traits.

METHODS

Mice

Male mice were obtained from The Jackson Laboratory (Bar Harbor, ME), housed in a specific pathogen-free facility, and sacrificed at 8–10 weeks of age. The same individual mice were used for both DNA- and RNA-based analyses. All experiments were performed in compliance with the guidelines of the Animal Studies Committee at Washington University, St. Louis, MO.

DNA preparation

DNA was prepared from spleen, liver, kidney, and tail by phenol-chloroform extraction, and was quantified using UV spectroscopy (NanoDrop 1000, Thermo Scientific, Wilmington, DE). Kidney DNA for aCGH experiments were pooled in equal masses from 2–6 individuals per strain. Only individual samples passing NimbleGen quality control requirements were pooled.

aCGH analysis

A tiling-path CGH array for whole-genome analysis in mouse (mm8, NCBI Build 36) was utilized (http://www.nimblegen.com). Isothermal probes from 45–75 bp were selected with a median probe spacing of 1 kb. Labeling, hybridization, washing and array imaging were performed as previously described45. Previously, we demonstrated that regions of the mouse genome with high sequence divergence between the test and reference strains have lower aCGH probe signal intensities and can, therefore, potentially disrupt the identification of CNVs16. Using an imputed single nucleotide map46, we defined regions of high sequence divergence between the test and reference genomes for input to wuHMM, a Hidden Markov Model algorithm for CNV detection16. All putative wuHMM CNV calls with scores less than 1.5 or 1.9 (gains or losses, respectively) were discarded, as we have previously shown that they contain a high number of false positive predictions. CNVRs were defined by merging overlapping wuHMM calls across all individuals. To assess the complexity of the CNVRs, we calculated average boundary concordances (the average of the length of the intersection of a CNV and CNVR divided by the total CNVR length). CNVRs having average concordances <= 0.75 (Supplementary Figure 5) comprised less than 23% of the CNVRs detected in this study. We refer to these regions as ‘complex’ and all other CNVRs as ‘simple’. All microarray data, aCGH and expression, is available for download from GEO (http://www.ncbi.nlm.nih.gov/geo/) under series accession GSE10656.

CNVR genotyping

Clustering of CNVRs was performed using partitioning among medoids (PAM) as implemented in R17. The average silhouette function calculates the average between versus within group distances and ranges from −1 to 1, with 1 representing perfect clustering17. We modified this function to weight groupings by their agreement with wuHMM calls. We executed PAM, varying the number of clusters from 2–7, and calculated the weighted average silhouette. The number of clusters with the maximum, modified average silhouette was selected for the number of genotypes per CNVR. Sometimes a clustering would result

in a group of strains in which no wuHMM call had been made, representing a new gain, loss or abnormal genotype. These genotypes were disallowed and these strains were assigned into the same genotype label as the reference strains. CNVRs with both average silhouettes < 0.3 and average scores < 2.0 were discarded, as they were likely to represent spurious clusters.

CNVR validation

61 simple CNVRs were randomly selected for validation from the set with average scores between 1.3 and 3.3. These CNVRs ranged from 887 bases to 67 kb (2 to 47 aCGH probes) and scored from 1.3 – 2.3 for gains, and 1.9 – 3.3 for losses. For qualitative PCR validation (losses only), primers were designed to target reference sequence within the predicted boundaries of the CNVR, prioritizing amplicons near or overlapping the aCGH probes with the maximum log2-ratio magnitudes. One to three amplicons were designed per CNVR. A positive control amplicon was designed for a region with no predicted CNVs in any of the 20 strains (primer sequences in Supplementary Table 8). For quantitative PCR (gains only), relative copy numbers were determined by real-time PCR (qPCR) using TaqMan detection chemistry and the ABI Prism 7300 Sequence Detection System (Applied Biosystems, http://www.appliedbiosystems.org), as previously described1. A CNVR loss was validated if no amplicon was produced using primers targeted within predicted CNVR boundaries. A CNVR gain was validated when qPCR demonstrated a >2-fold increase in inferred relative copy number relative to the reference strain. We defined the false positive rate (FPR) as the number of false positives divided by the number of gain and loss genotypes at or exceeding a given score threshold. The FPR for putative copy number losses with scores between 2.0 and 2.5 was 25% (152/608 CNV calls tested). Nearly a third of these amplicons (50/152) exhibited altered electrophoretic mobility consistent with the CNV strain distributions predicted by aCGH analysis. To better understand this phenomenon, we cloned and sequenced two of the amplicons from four affected strains and discovered three novel SNPs in each amplicon which overlapped an aCGH probe sequence in the CNVR in each case. Sequence divergence can disrupt probe hybridization resulting in decreased signal intensity and, at times, false positive deletion calls. Further, we found a 14- and a 10-bp insertion near the probe sequence in the affected strains, which accounted for the altered size of the amplicons. The co-occurrence of SNPs and in/dels has previously been reported and their potential causal relationship is under investigation47. For CNVRs with average scores exceeding 1.5 and 2.5 for gains and losses, respectively, the FPR approached 0 (Supplementary Table 1). Therefore, only calls that exceeded these thresholds were retained for further analysis.

Comparison to other studies

CNVR coordinates were translated from mm6 to mm8 using liftOver, when necessary (http://genome.ucsc.edu/cgi-bin/hgLiftOver). We defined subsets of CNVRs by selecting only those CNVRs that have a gain, loss, or abnormal genotype in at least one of the strains in common with another study. Overlap between studies was reported as either the total shared sequence in the intersection of CNVRs or as the number of CNVRs that have overlapping boundaries. For comparisons of CNV content by CNVR size, sequence overlap

was determined by calculating the total sequence intersection between only small, medium or large, high-confidence CNVRs and all CNVRs from other studies.

Sequence may be reported as copy number variable exclusively in other studies due to differences in genome coverage20, *de novo* events22, or because lower resolution platforms tend to over-estimate CNVR boundaries1,21. The comparison to a study that mainly targeted segmental duplicated regions of the genome resulted in the lowest agreement (63.9%)20. Many of these regions have sparse probe coverage on the platform that we utilized and therefore are problematic regions in which to detect CNVs. The second lowest overlap (64.3%) was with a study that specifically targeted the identification of *de novo* events in C57BL/6-derived strains22. It is possible that the 36.7% of CNV content exclusive to that study was not detected here because those sequences did not exist in, or comprised an undetectable fraction of the samples used in our study. We also assessed the overlap between CNVRs in our study and others, defined across all strains, to determine the overall consensus of reported copy number variation in the inbred mouse genome. To perform this comparison, we first merged all CNVs from previous studies into a single set of CNVRs finding that the amount of novel CNV sequence content is relatively low (16%) (Supplementary Table 3).

Enrichment analysis

The association between CNVRs and genomic features was tested by randomly permuting the chromosome and position of each CNVR 100 times and determining the sequence content of the resulting region or flanking regions. Gene overlap enrichment was tested similarly, except that the test statistic was the number of CNVRs per permutation that overlapped at least one gene using UCSC's *knownGene* annotation (http://genome.ucsc.edu/cgi-bin/hgGateway?org=Mouse&db=mm8).

Cell sorting and RNA extraction

Bone marrow cells were harvested from mouse femurs and stained with FITC-conjugated lineage markers (Gr-1, CD19, B220, CD3, CD4, CD8, TER119, and IL-7Rα) and APC-conjugated c-kit (BD Biosciences, San Diego, CA). Lineage-negative, c-kit positive cells were enriched using a modified MoFlo high speed sorter (Cytomation, Fort Collins, CO). Total RNA was prepared using Trizol LS (Invitrogen, Carlsbad, CA) and its concentration quantified using UV spectroscopy (Nanodrop). Total RNA quality was then determined by Agilent 2100 Bioanalyzer (Agilent Technologies) according to the manufacturer's recommendations.

Expression profiling

RNA transcripts were amplified by T7 linear amplification (MessageAmp TotalPrep amplification kit; ABI-Ambion). First strand synthesis was primed with oligo-dT, followed by *in vitro* transcription to generate amplified RNAs (aRNA). The aRNAs were then quantitated on a spectrophotometer, and quality determined by Agilent 2100 bioanalyzer according to the manufacturer's recommendations. Hybridization to the MouseWG-6 v1.1 Expression Beadchip (Illumina), washing, and signal detection were performed using standard protocols. Quantitated data were imported into Beadstudio software (Illumina). On-

slide spot replicates were averaged by Beadstudio and individual spot data was reported. Probes were defined as 'present' in a sample when the signal was significantly higher than in a set of negative control probes, ($P < 0.05$ after correcting for multiple tests). A probe was defined as present in a strain if it was called 'present' in all replicate samples of that strain. The correlation of within-strain expression profiles exceeded between-strain correlations in all but two strains (average within strain correlation = 0.9782, average between-strain correlation = 0.9528), demonstrating that the expression profiles reflect biological variation and not technical artifacts (i.e., due to differences in cell staining, sorting, RNA labeling, or hybridization).

eQTL Mapping

Expression quantitative trait mapping was implemented as previously described28,31,32 with the exception that CNVR instead of SNP genotypes were used as predictor variables. Null distributions of F-statistics for CNVR-expression trait tests were generated by 10,000 random permutations of expression values. The permutations were weighted according to strain-relatedness as defined using an imputed SNP map46 (exponent = 3) such that closely related strains more frequently replaced each other than distantly related strains. All permutation analyses were implemented on custom software and executed on a compute cloud (http://aws.amazon.com/ec2). Often a single trait was tested against multiple CNVRs therefore the permutation-derived P-values were corrected by applying the Holm multiple testing correction separately for each trait.

BXD RI SNP genotype data was downloaded from: http://www.genenetwork.org/dbdoc/BXDGeno.html. A CNVR genotype of 'B', 'D', or 'U' was assigned for each CNVR to each strain if the two markers spanning the CNVR were both C57BL/6J, both DBA/2J, or discordant, respectively. BXD KLS expression data was downloaded from GEO, accession number GSE2031. Of the genes identified as having significant associations with CNVRs in *cis* in the KL expression data set, only those that were detected in at least 80% of the samples from either or both CNVR genotype groups were assessed for concordant expression in the BXD KLS data. Association between KLS expression and inferred CNVR genotype was performed as for KL expression data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by a grant from the NIH/NCI (CA101937). P.C was supported in part by the National Human Genome Research Institute (T32 HG000045) and a Kauffman Fellowship. Mice were kindly provided through a collaboration with the Mouse Phenome Project (The Jackson Laboratory, Bar Harbor, ME). Additional mice were provided by Ming You. We thank Tim Ley, Dan Link, and Matt Walter for helpful discussions. Cell sorting was performed by the High Speed Cell Sorter Core in the Alvin J. Siteman Cancer Center at Washington University School of Medicine. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant (P30 CA91842).

References

1. Graubert TA, et al. A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome. PLoS Genetics. 2007; 3:e3. [PubMed: 17206864]
2. Li J, et al. Genomic segmental polymorphisms in inbred mouse strains. Nat Genet. 2004; 36:952–954. [PubMed: 15322544]
3. Perry GH, et al. Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A. 2006; 103:8006–11. [PubMed: 16702545]
4. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444:444–54. [PubMed: 17122850]
5. Snijders A, et al. Mapping segmental and sequence variations among laboratory mice using BAC array CGH. Genome Res. 2005; 15:302–311. [PubMed: 15687294]
6. Dopman EB, Hartl DL. A portrait of copy-number polymorphism in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2007; 104:19920–5. [PubMed: 18056801]
7. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science. 2008; 320:1629–31. [PubMed: 18535209]
8. Guryev V, et al. Distribution and functional impact of DNA copy number variation in the rat. Nat Genet. 2008; 40:538–45. [PubMed: 18443591]
9. Iafrate AJ, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36:949–51. [PubMed: 15286789]
10. Sebat J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004; 305:525–8. [PubMed: 15273396]
11. Aitman TJ, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006; 439:851–5. [PubMed: 16482158]
12. McCarroll SA, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet. 2008
13. Singleton AB, et al. alpha-Synuclein locus triplication causes Parkinson's disease. Science. 2003; 302:841. [PubMed: 14593171]
14. Walsh T, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008; 320:539–43. [PubMed: 18369103]
15. Bogue MA, Grubb SC. The Mouse Phenome Project. Genetica. 2004; 122:71–4. [PubMed: 15619963]
16. Cahan P, et al. wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. Nucleic Acids Res. 2008; 36:e41. [PubMed: 18334530]
17. Kaufman, L.; Rousseeuw, PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley; New York: 1990.
18. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006; 38:75–81. [PubMed: 16327808]
19. Korbel JO, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007; 318:420–6. [PubMed: 17901297]
20. She X, Cheng Z, Zollner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. Nat Genet. 2008; 40:909–14. [PubMed: 18500340]
21. Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD. Significant gene content variation characterizes the genomes of inbred mouse strains. Genome Res. 2007; 17:1743–54. [PubMed: 17989247]
22. Egan CM, Sridhar S, Wigler M, Hall IM. Recurrent DNA copy number variation in the laboratory mouse. Nat Genet. 2007; 39:1384–9. [PubMed: 17965714]
23. Watkins-Chow DE, Pavan WJ. Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. Genome Res. 2008; 18:60–6. [PubMed: 18032724]
24. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. 1998; 14:417–22. [PubMed: 9820031]

25. Sharp AJ, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005; 77:78–88. [PubMed: 15918152]
26. Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE. Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res. 2008; 18:869–80. [PubMed: 18381897]
27. Chambers SM, et al. Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny. Cell Stem Cell. 2007; 1:578–591. [PubMed: 18371395]
28. McClurg P, et al. Genomewide Association Analysis in Diverse Inbred Mice: Power and Population Structure. Genetics. 2007; 176:675–683. [PubMed: 17409088]
29. Wu C, et al. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS Genet. 2008; 4:e1000070. [PubMed: 18464898]
30. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet. 2005; 76:8–32. [PubMed: 15549674]
31. Pletcher MT, et al. Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. PLoS Biol. 2004; 2:e393. [PubMed: 15534693]
32. McClurg P, Pletcher MT, Wiltshire T, Su AI. Comparative analysis of haplotype association mapping algorithms. BMC Bioinformatics. 2006; 7:61. [PubMed: 16466585]
33. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979; 6:65–70.
34. Bystrykh L, et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. Nat Genet. 2005; 37:225–32. [PubMed: 15711547]
35. Bishop TR, Cohen PJ, Boyer SH, Noyes AN, Frelin LP. Isolation of a rat liver delta-aminolevulinate dehydrase (ALAD) cDNA clone: evidence for unequal ALAD gene dosage among inbred mouse strains. Proc Natl Acad Sci U S A. 1986; 83:5568–72. [PubMed: 3502704]
36. Bishop TR, Miller MW, Wang A, Dierks PM. Multiple copies of the ALA-D gene are located at the Lv locus in Mus domesticus mice. Genomics. 1998; 48:221–31. [PubMed: 9521876]
37. Schadt EE, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422:297–302. [PubMed: 12646919]
38. Hovatta I, et al. Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature. 2005; 438:662–6. [PubMed: 16244648]
39. Ohtsuka M, Inoko H, Kulski JK, Yoshimura S. Major histocompatibility complex (Mhc) class Ib gene duplications, organization and expression patterns in mouse strain C57BL/6. BMC Genomics. 2008; 9:178. [PubMed: 18416856]
40. Kumar PP, et al. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. Nat Cell Biol. 2007; 9:45–56. [PubMed: 17173041]
41. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet. 2006; 38:82–5. [PubMed: 16327809]
42. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–53. [PubMed: 17289997]
43. Korbel JO, et al. The current excitement about copy-number variation: how it relates to gene duplications and protein families. Curr Opin Struct Biol. 2008; 18:366–74. [PubMed: 18511261]
44. Churchill GA, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet. 2004; 36:1133–1137. [PubMed: 15514660]
45. Selzer RR, et al. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer. 2005; 44:305–19. [PubMed: 16075461]
46. Szatkiewicz JP, et al. An imputed genotype resource for the laboratory mouse. Mamm Genome. 2008; 19:199–208. [PubMed: 18301946]
47. Tian D, et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature. 2008; 455:105–8. [PubMed: 18641631]

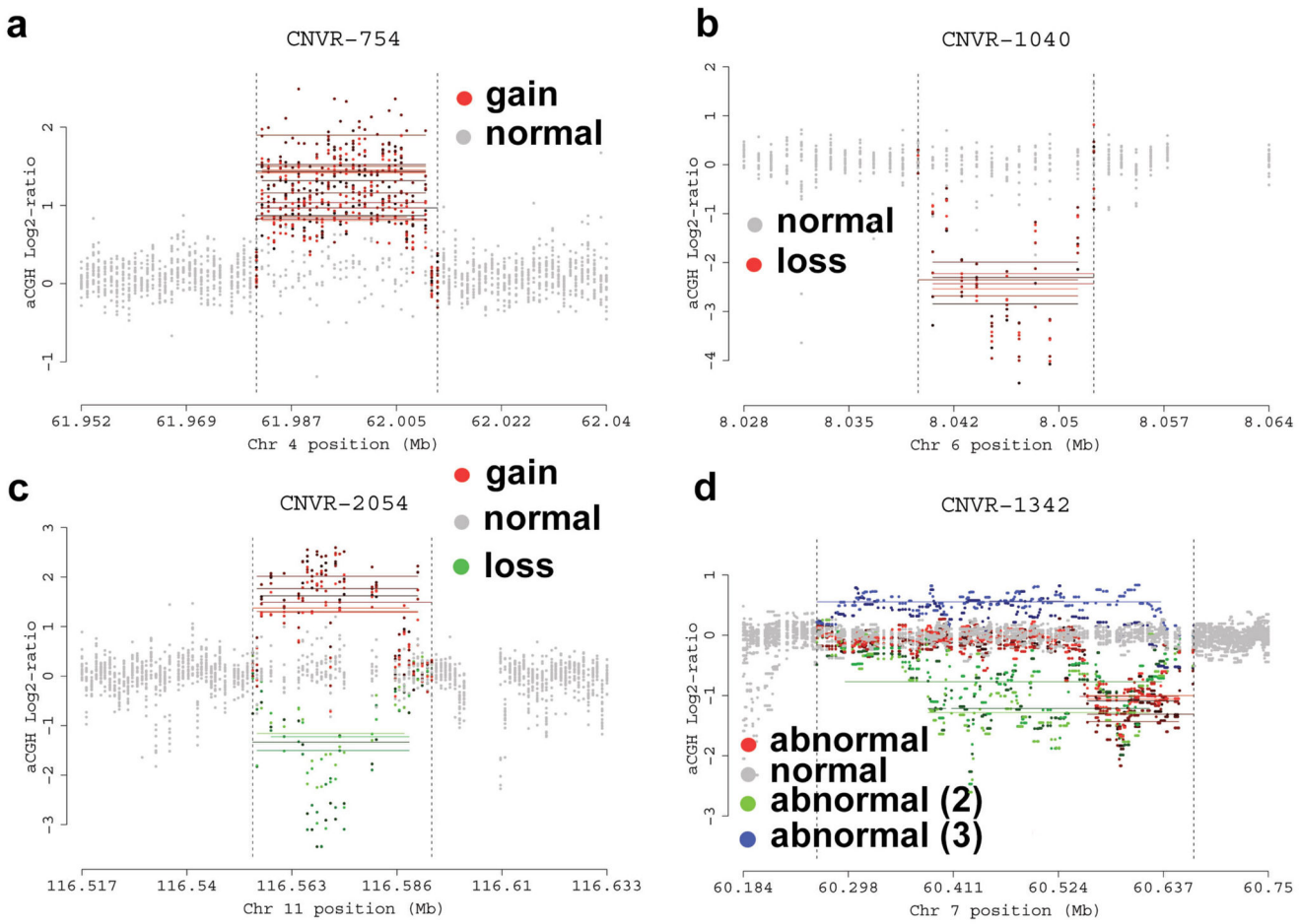

Figure 1. CNVR genotyping

Log2-ratio plots of the test versus reference (C57BL/6J) aCGH signal intensities. All twenty strains are shown in each plot. Horizontal lines represent wuHMM segmentation calls, which are made independently for each strain. CNVs are merged into CNV-regions (CNVRs), represented as vertical dotted lines. CNVR genotypes (see Methods) are indicated by probe coloring and strains are indicated by probe shading. (a) A 30 kb simple CNVR gain present in 16 strains. wuHMM call boundaries largely agree with the CNVR boundaries, resulting in a high average concordance (91.6%). (b) A 12 kb simple CNVR loss occurring in 8 strains. (c) A 39 kb simple gain/loss CNVR called as a ‘gain’ in 7 strains and as a ‘loss’ in 3 strains. (d) A 416 kb complex CNVR assigned 5 different genotype groups.

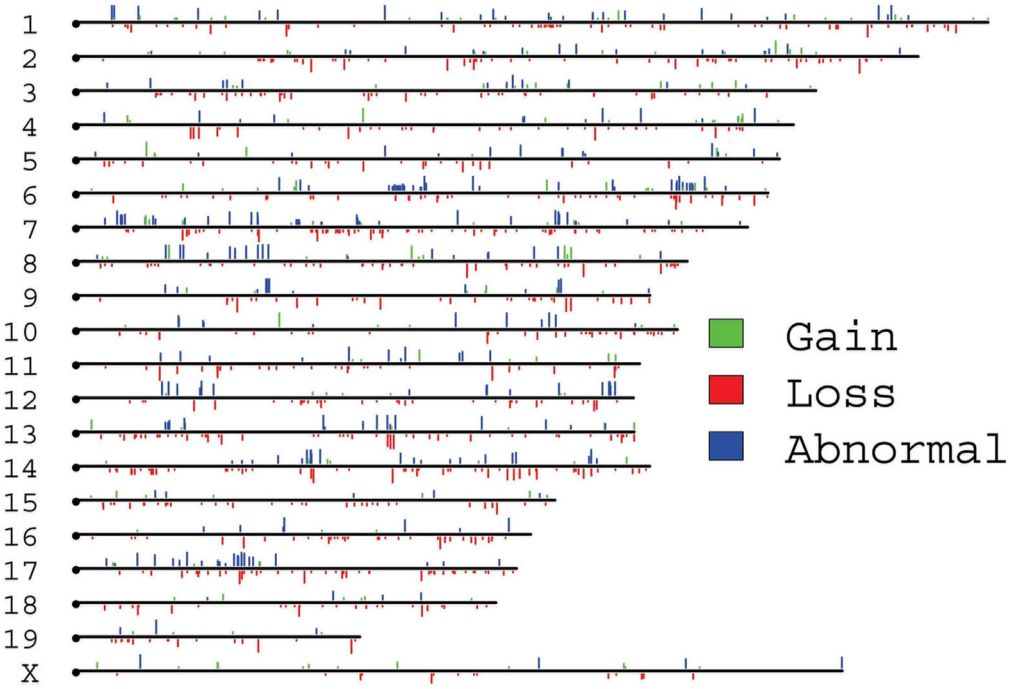

Figure 2. Location of CNVRs in the inbred mouse genome

The ideograms depict chromosomal locations of CNVRs in the autosomes and X chromosome from 20 inbred strains. Gains relative to the reference genome (C57BL/6J) are green lines, losses are red, and complex CNVRs are blue. The height of the lines reflects the number of strains in which the genotype call is made.

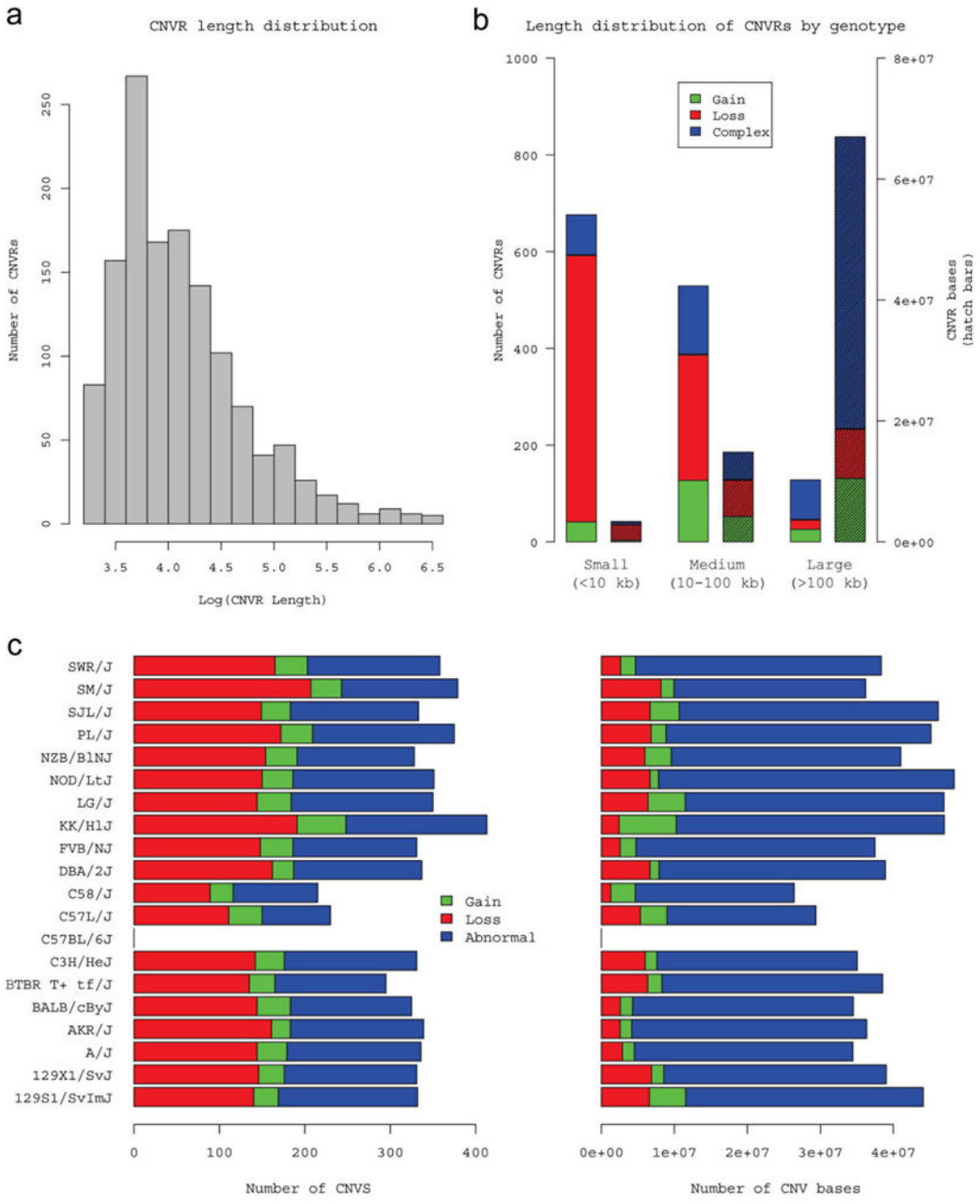

Figure 3. Distribution of CNVR sizes

(a) Length distribution of CNVRs (on log10 scale). Most CNVRs are shorter than 10 kb. (b) Length distribution of CNVRs separated by CNVR genotype. CNVRs are divided into small (<10 kb), medium (10 kb ≥ CNVR length < 100 kb), or large (≥100 kb). Frequency is indicated by solid bars (left axis) and sequence content by hatched bars (right axis). Most CNVRs are small losses, but most of the copy number variable sequence in the mouse genome is in large, complex CNVRs. (c) The number of gain, loss, or abnormal CNVR genotypes and the copy number variable sequence per strain. C57/J and C58/J, the most closely related strains to C57BL/6J, have fewer CNVs than more distantly related strains.

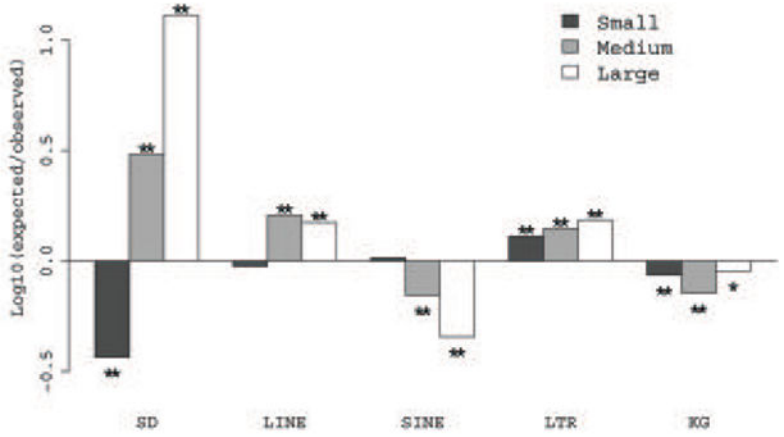

Figure 4. Co-localization of CNVRs with other genomic elements

The enrichment or depletion of segmental duplications (SD), LINEs, SINEs, LTRs, and genes as annotated in UCSC's knownGene track (KG) in CNVRs was tested by permuting the location of CNVRs. The percent of the CNVR sequence comprised of SD, LINE, SINE, and LTR was compared to the permuted background, as was the number of CNVRs that overlapped at least one gene. The ratio of permuted to observed results (log10 scale) are shown, where a negative value indicates depletion and positive indicates enrichment. * $P < 0.05$. ** $P < 0.01$.

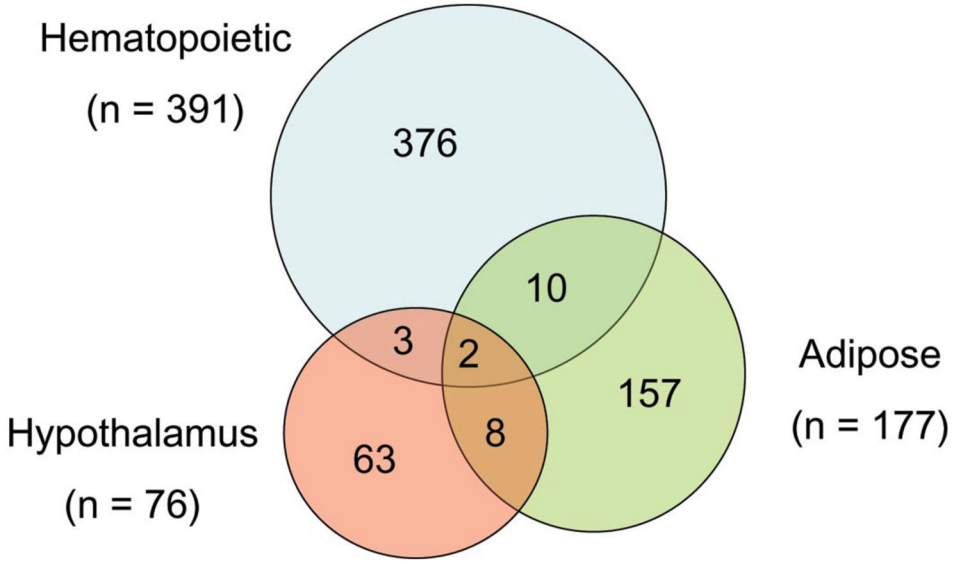

Figure 5. Tissue-specific CNVR eQTLs

Overlap of eQTL genes in hematopoietic stem/progenitors, adipose, and hypothalamus. Most eQTL genes are tissue- specific, implying that other factors can influence these expression traits.

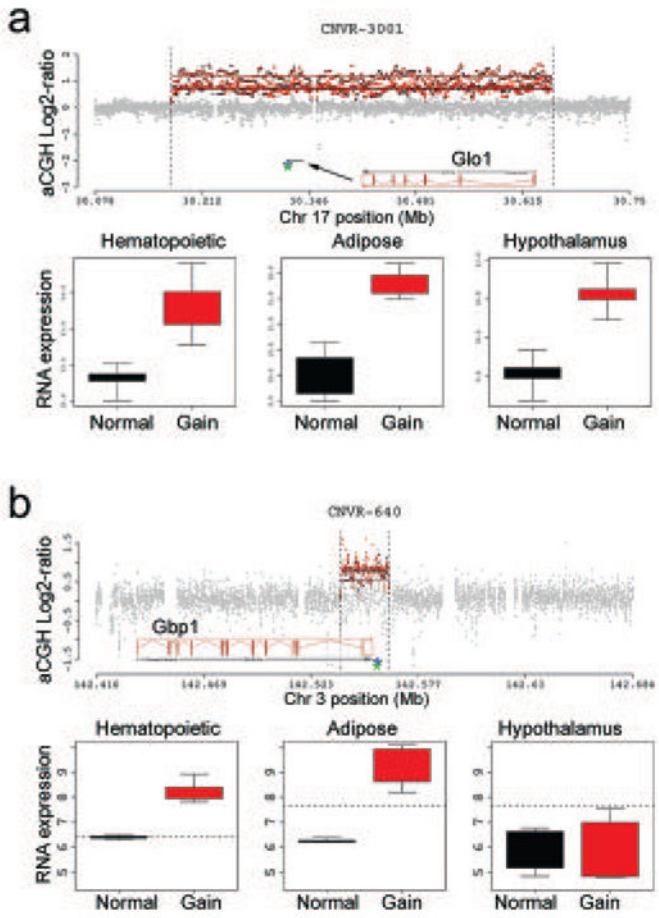

Figure 6. CNVR eQTLs

(a) Top: Log2-ratio plot of a 481 kb CNVR containing the complete coding sequence of *Glo1* (location is indicated by horizontal line). Positions of Illumina (blue asterisk) and Affymetrix (green asterisk) expression probes are shown. Bottom: *Glo1* expression in hematopoietic stem/progenitors, adipose tissue, and hypothalamus. Expression is significantly correlated with the CNVR gain. (b) Top: Log2-ratio plot of a 24 kb CNVR containing the 3′ exon and UTR of *Gbp1*. A gain is called in 8 strains. Bottom: *Gbp1* expression in the same tissues; expression is significantly correlated with the CNVR gain in hematopoietic stem/progenitors and adipose tissue. Dotted line represents the mean detection threshold across all arrays.

Table 1

CNVR eQTL characteristics.

	Expression Probes					CNVRs	
Tissue	**Probes**	**Present**	**Strain-specific**	**In *cis***	**eQTL**	**Testable CNVRs**	**eQTL CNVRs**
Hematopoietic	46,629	13,588	1,469	958	408 (391)	484	214
Adipose	32,533	10,040	4,083	2,056	181 (177)	466	116
Hypothalamus	32,533	14,871	2,879	789	78 (76)	440	66

"In *cis*" is the number of expression probes within 2 Mb of a CNVR.

Only CNVRs that have greater than two strains per genotype group are considered for eQTL mapping ("Testable CNVRs").

eQTL is the number of expression probes (genes) that are significantly associated with a CNVR ($P < 0.05$).

Nat Genet. Author manuscript; available in PMC 2009 October 01.

Table 2

Subset of CNVR-eQTLs in hematopoietic stem/progenitor cells, hypothalamus, and adipose tissues.

					Hematopoietic		Hypothalamus		Adipose	
Gene symbol	**CNVR ID**	**Chr**	**Str**	**Stp**	**R^2**	**Proximity**	**R^2**	**Proximity**	**R^2**	**Proximity**
Alad	754	4	61,981,136	62,011,544	0.76	0	0.66	0	0.41	0
Glo1	3001	17	30,172,971	30,654,177	0.81	0	0.87	0	0.86	0
Sox13	216	1	135,174,821	135,205,210	0.42	4,833	0.30	4,844		
2310009E04Rik	766	4	96,568,138	96,578,356	0.35	1,149,288	0.48	1,149,218		
Thumpd1	1383	7	119,077,924	119,084,035	0.28	422,561	0.26	422,467		
Ifi205	127	1	177,044,560	177,048,644	0.39	1,181,250			0.42	1,308,417
Cstf3	420	2	104,470,986	104,479,145	0.24	60,989			0.29	4,814
Hdc	432	2	128,046,370	128,053,520	0.26	1,760,692			0.26	1,760,686
Gbp1	640	3	142,537,542	142,566,929	0.96	0			0.91	0
Hdhd3	754	4	61,981,136	62,011,544	0.49	0			0.43	0
Trim56	925	5	137,433,034	137,442,249	0.78	36,914			0.37	36,632
Gtf3a	931	5	146,453,420	146,501,196	0.50	764,371			0.28	763,081
Capg	1077	6	72,014,338	72,019,334	0.94	472,988			0.69	468,015
Mir16	1383	7	119,077,924	119,084,035	0.43	592,227			0.58	597,792
Hemk1	1749	9	106,947,671	106,952,630	0.94	233,352			0.30	233,219
Pbx1	232	1	171,057,081	171,060,008			0.47	794,444	0.31	794,444
Trim34	1372	7	104,398,954	104,424,326			0.54	262,996	0.76	262,996
4833420G17Rik	2405	13	120,606,370	120,613,637			0.75	1,436	0.38	1,436
Paip1	2405	13	120,606,370	120,613,637			0.35	35,652	0.51	35,652
Zfr	2719	15	11,647,294	11,651,207			0.49	474,359	0.36	474,359
Cxadr	2872	16	78,700,106	78,704,117			0.68	476,795	0.47	476,795
Sytl3	2978	17	6,175,338	6,569,177			0.47	0	0.37	0
H2-T23	3014	17	35,831,648	35,840,806			0.47	114,162	0.55	114,162

R^2 is the correlation coefficient for the CNVR-to-eQTL association.

Proximity is the number of bases between the nearest boundaries of the expression probe and CNVR.

Gene dosage eQTLs have a proximity = 0 (Alad, Glo1, Gbp1, Hdhd3, and Syrl3).