5-24-2009

Integrated Genomics Of Susceptiblity To Therapy-Related Leukemia

Patrick Cahan
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational Biology

Dissertation Examination Committee:
Timothy Graubert, Chair
Michael Brent
Timothy Ley
Howard McLeod
Rakesh Nagarajan
Nancy Saccone

INTEGRATED GENOMICS OF SUSCEPTIBLITY TO THERAPY-RELATED LEUKEMIA

by

Patrick Cahan

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2009

Saint Louis, Missouri

Acknowledgements

My thesis advisor, Dr. Tim Graubert, has been a constant source of motivation and scientific insight. Our many discussions, in which Tim would share his incisive logic in breaking down a problem or a faulty line of reasoning, have made me a better scientist. When my energy or enthusiasm for a particular project would begin to wane, Tim would intuit this and help, however possible, to move things forward. Everyone in his lab has benefited greatly from the unique learning environment that he has created. Thank you, Tim.

I am grateful for the time, effort and thought that members of the thesis committee have offered: Dr. Michael Brent, Dr. Tim Ley, Dr. Howard McLeod, Dr. Rakesh Nagarajan, and Dr. Nancy Saccone. Through committee meetings and discussions I have gained both insight into technical aspects of my project and an appreciation for the power of bringing together diverse expertise. I would especially like to thank Tim Ley for leading and sustaining the vibrant intellectual community that is the 6th floor, and for providing many opportunities for me to expand my experience beyond the immediate scope of my thesis work. I am grateful to Howard McLeod for initially taking on the mentorship of a computational biologist in a predominantly wet lab, and for subsequently looking out for my development and career in a selfless way. I am also grateful to both Michael Brent and Nancy Saccone for taking the time outside of committee meetings to discuss technical aspects of my thesis work. I would also like to thank Michael for serving as thesis committee chair.

I would like to express thanks to previous and current Graubert lab members, including Julie Fortier, Ryan Funk, Megan Janke, Masayo Izumi, Yedda Li, Elise Peterson Lu, Theresa (Treeza) Okeyo-Owuor, and Richard Walgren for their helpful comments, discussions, technical help, and tolerance and patience. You have made it engaging, educational, and fun to be in the lab. I would like to especially thank Masayo for her critical role in many of the experiments in this thesis work. Without her diligent consistence, much of this work would not have been possible. I would also like the thank Yedda Li for her significant and insightful contributions to the CNV validation work. I would also like to thank Julie Fortier for lending her expert knowledge and training me in several wet lab experiments. I would be remiss if I did not thank Bill Eades, Jackie

Hughes, Chris Holley, and Bill Lamberton in the High Speed Cell Sorter Core. They have been patient and accommodating when my preparations took longer than predicted (almost always), and their expertise was critical to the success of my thesis work. I also thank Bill Eades for the extended use of a computer.

I would like to acknowledge support from the NIH Genome Analysis Training Program and Kauffman Fellowship Pathway in Life Sciences Entrepreneurship.

My parents, Joe and Jane Cahan, are awesome. From the beginning, they have provided me with every opportunity to grow and learn. They have instilled in me a curiosity about the world, and a confidence to go and explore it. Thank you, Mom and Dad! I am also thankful to my brothers Josh and Matt, who have always looked out for me, have been supportive through this journey, and who never fail to make me laugh.

This dissertation is dedicated to my wife, Moira Nealon Cahan. Through the past 5+ years of graduate school, she has been a source of constant support. I thank Moira for the many sacrifices that she has made so that I could pursue this dream of becoming a scientist and for making sure that I always ‘keep it real’.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

ABSTRACT OF THE DISSERTATION

Integrated Genomics of Susceptibility to

Therapy-related Leukemia

by

Patrick Cahan

Doctor of Philosophy in Biology and Biomedical Sciences (Computational Biology)

Washington University in St. Louis, 2009

Professor Timothy Graubert, Chair

Therapy-related acute myeloid leukemia t-AML is a secondary, generally incurable, malignancy attributable to the chemotherapeutic treatment of an initial disease. Although there is a genetic component to susceptibility to therapy-related leukemias in mice, little is understood either about the contributing loci, or the mechanisms by which susceptibility factors mediate their effect. An improved understanding of susceptibility factors and the biological processes in which they act may lead to the development of t-AML prevention strategies.

In this thesis work, we identified expression networks that are associated with t-AML susceptibility in mice. These networks are robust in that they emerge from distinct methods of analysis and from different gene expression data sets of hematopoietic stem and progenitor lineages. These networks are enriched in genes involved in cell cycle and DNA repair, suggesting that these processes play a role in susceptibility. By integrating gene expression and genetic information we prioritized network nodes for experimental validation as contributors to expression networks and t-AML susceptibility.

Network analysis and node prioritization required a comprehensive map of genetic variation in mouse, which was not available at the outset of this thesis work. Specifically, DNA copy number variations (CNVs), defined as genomic sequences that are polymorphic in copy number and range in length from 1,000 to several million base pairs, were largely uncharacterized in inbred mice. We developed a computational approach, Washington University Hidden Markov Model (wuHMM), to identify CNVs from high-density array comparative genomic

hybridization data, accounting for the high degree of polymorphism that occur between mouse strains. Using wuHMM we analyzed the copy number content of the mouse genome (20 strains) to a sub-10-kb resolution, finding over 1,300 CNV-regions (CNVRs), most of which are < 10 kb in length, are found in more than one strain, and span 3.2% (85 Mb) of the reference genome. These CNVRs, along with haplotype blocks we derived from publicly available SNP data, were integrated into susceptibility expression network analysis. In addition to addressing questions regarding t-MDS/AML susceptibility, we also used this data to assess the potential functional impact of copy number variation by mapping expression profiles to CNVRs. In hematopoietic stem and progenitor cells, up to 28% of strain-dependent expression variation is associated with copy number variation, supporting the role of germline CNVs as key contributors to natural phenotypic variation.

Therapy-related Acute Myeloid Leukemia (t-AML)

Acute myeloid leukemia (AML) is a clonal malignancy characterized by the accumulation of immature leukocytes in the bone marrow. The associated disruption of hematopoiesis in AML patients reduces the number of red blood cells (anemia), neutrophils (neutropenia), and platelets (thrombocytopenia), and leads to complications arising from the loss of proper function of these cell types. Untreated AML is fatal, but with chemotherapy the survival rate for those under 65 is approximately 40%[1]. Frequent chromosomal abnormalities in AML include t(8;21), resulting in the AML1-ETO fusion gene, translocations involving chr11q23, which harbors the Mixed-Lineage Leukemia (MLL) gene, to a variety of other sites, t(15;17) resulting in the PML_RARα fusion gene, and structural re-arrangements of chr16. However, none of these events are sufficient to cause leukemia. The identification of cooperating mutations promises to lead to a better understanding of this heterogeneous disease and, eventually, to improved treatments. Important advances in this field are being made by whole genome re-sequencing efforts, where mutations in genes not previously linked to cancer have been identified[2,3].

Therapy-related acute myeloid leukemia (t-AML) is a secondary malignancy attributable to the chemotherapeutic and/or radiotherapeutic treatment of a variety of diseases, including hematological and solid tumors. Therapy-AML does not exclusively reflect a predisposition to sporadic AML because t-AML occurs in patients treated for autoimmune disorders such as rheumatoid arthritis and multiple sclerosis (Karran 2003). The incidence of t-AML ranges widely depending on study and primary disease. Of breast cancer survivors, 1.7% develop secondary bone marrow diseases[4]. Of lymphoma and Hodgkin disease patients, 5-20% go on to acquire t-AML[5]. Although arguably distinct diseases, 80% of t-AML cases are preceded by therapy-related myelodysplastic syndromes and in this thesis I will refer to them as a single entity (t-AML). Therapy-AML typically appears three to ten years after initial chemotherapy. Common cytogenetic events associated with t-AML are loss of all or part of chromosomes 5 and/or 7 (70%)[6]. Therapy-AML comprise 5-20% of all AML cases and their prevalence is increasing along with the population undergoing chemotherapy[7,8]. t-AML are generally incurable[9]. Median survival

time from diagnosis is eight months and survival time for associated with the combined chr5/chr7 loss karyotypes is 5 months[10]. It is 11 months for those with no karyotypic abnormalities[6]. Currently, hematopoietic stem cell transplant (HCT) the only cure but often is infeasible and risky as there is a 49% transplant-related mortality[11]. Complete remission of t-AML occurs in 28% of patients treated versus 65-80% in primary AML[12]. Differences in response are due to a variety of factors, including persistence of primary disease, tissue/organ damage by treatment (bone marrow stroma, depletion of HSC), immunosuppression and resulting infections. Because t-AML is a clinically induced malignancy, it is, by definition, preventable. Therefore, a long-term goal of t-AML research is to gain sufficient understanding of susceptibility factors in order to make worthwhile the personalization of chemotherapeutic regiments based on t-AML risk. Also, because t-AML shares many characteristics with its primary counterpart, an understanding of t-AML susceptibility may provide insight into the etiology of primary AML and progression from MDS.

Approximately 75% of t-AML cases are associated with prior alkylator treatment (i.e., melphalan, busulfan, thiotepa)[10]. The therapeutic effect of alkylator agents is believed to result from the formation of DNA adducts and single and double-strand breaks, which trigger apoptosis or growth arrest[13]. The precise mechanisms of action are unclear, as are the effects of alkylators on RNA and protein. Topoisomerase II inhibitors (i.e., etoposide, doxorubicin, mitoxantrone) also cause therapy-related leukemias distinct from those induced by alkylators: there is a shorter latencies (1-3 years), a preceding phase of MDS is infrequent, and tumors often contain chr11q23 translocations and other translocations, but not complete or partial loss of chr5/7. The focus of the current work is susceptibility to alkylator-induced AML. However, it is likely that many of the methods and resources developed here will be directly applicable to investigate the susceptibility to secondary malignancies due to topoisomerase II exposure.

Therary-AML is not due entirely to the stochastic nature of therapy-induced mutations. Contributing factors to this complex phenotype include the primary disease[8], the cumulative dose of chemotherapy[14], and genetic background[15]. There are rare, familial cancer predisposition syndromes with mutations in TP53[16], XPD[17], or NF1[18] that increase t-AML susceptibility. Beyond

these rare cases, it has been hypothesized susceptibility is a complex trait in that inherited polymorphisms in multiple genes each contribute a small amount to overall susceptibility status. Based on the presumed genotoxic mechanism of alkylators, genes involved in DNA repair[19], response to oxidative stress[20], and drug metabolism[21] have been investigated as mediators of susceptibility in candidate gene studies. While many studies have been performed, the results have been either conflicting, inconclusive, or find relative weak effect sizes[22]. A notable limitation of most candidate gene studies to date is that they have focused on polymorphisms believed to result in (non-conservative) changes in the protein sequence.

Perhaps the most promising approaches to identify genes and pathways involved in susceptibility are unbiased, genome-wide methods. One of the first genome-wide studies was performed leveraging genetic variation across an inbred panel of mice as a disease susceptibility model[23]. In this study, eight to twelve individual mice from each of 20 inbred strains were treated with the alkylating agent *N*-nitroso-*N*-ethylurea (ENU), a potent mutagen with a propensity to cause AT:TA transversions and AT:GC transitions[24]. Mice were monitored for the development of MDS and AML for up to 16 months post ENU exposure. Myeloid tumors varied by strain, supporting the hypothesis of a strong genetic component in t-AML susceptibility (estimated to be 0.10). This study also used the *in silico* mapping method to identify two genomic intervals associated with susceptibility. A follow-up study of an F2 cross of susceptible and resistant parental strains identified thirteen quantitative trait loci (QTLs) associated with the t-AML traits, including leukemia-free survival time, white blood cell count, and spleen weight[25]. These studies have demonstrated that susceptibility in inbred mice is not purely stochastic. Further, they have identified candidate loci. Although the QTLs do not coincide between the two studies, they have served as starting points to identify quantitative trait genes. A limitation of these approaches (and in mapping any QTL) is the difficulty in narrowing QTLs down to quantitative trait genes, which can take many years and has been successful in approximately 20 out of over 2,000 QTLs reported as of 2005[26].

Integrated genomics

Relatively unbiased, genome wide approaches such as genome-wide association studies (GWAS) hold great promise to reveal much about complex traits in human populations. They are similar to candidate gene studies in that they compare the relative frequencies of polymorphisms between case and control groups. Unlike candidate gene studies, GWAS use panels of markers that span the genome and capture a large fraction of SNP variation (this varies depending on the population assayed and platform used). Recent genome wide association studies have identified candidate susceptibility loci for several cancers: 29 in prostate cancer[27-33], 13 in breast cancer[34-38], 10 in colon cancer[39-41]. Most loci identified to date are non-overlapping between cancer types, suggesting that tissue specific forces are important in cancer susceptibility[42]. In contrast to sporadic cancers, for cancers associated with exposure (i.e., lung cancer) only a handful of loci have been found: 3 in lung cancer[43-45] and 3 in bladder cancer[46,47]. Whether this means that susceptibility to exposure-based cancers has a more modest genetic component remains to be determined. In a recent GWAS, more associations with t-AML were detected than would be expected by chance, even given the relatively small number of individuals in the study[48] (80 cases, 150 controls). None of the three validated candidate SNPs had previously been implicated in susceptibility previously. A drawback to GWAS studies, especially in light of the apparent complexity of cancer susceptibility, is the low power to detect weak effects. A second drawback is that even when association studies are successful, mechanisms linking candidate variants to susceptibility are not readily apparent. For example, the ApoE E4 genotype has been known as a risk factor for late-onset Alzheimer's for more than ten years and yet the mechanism by which it contributes to the disease remains unknown[49].

Evidence is accumulating that many genetic contributors to complex traits are not protein-coding changes[50]. If true, then the only other class of genetic events that can effect phenotype must, at some level, impact expression (i.e. eQTLs). Combining information from expression profiling experiments and genetic association studies can identify such events (i.e. eQTLs that contribute to disease/complex traits) involved in myocardial calcification[51], atherosclerosis[52] and

obesity (proposed in [53], candidates discovered in [54], causal genes validated in [55]). By augmenting these approaches with network analysis, it is possible to extend the insight of integrated studies to a better understanding of the molecular underpinnings of complex phenotypes[56,57]. These approaches can be further extended by comparing networks across species, which has practical benefits in terms of initial tests of candidate targets[58]. The work described in this thesis applies an integrative genomics approach to identify and prioritize genetic and transcriptional networks underlying t-AML susceptibility (Figure 1).

DNA copy number variation

In contrast to previous integrated genomics work, this thesis explicitly includes information on DNA copy number variations. Copy number variants (CNVs), currently defined as genomic sequences greater than one kilobase that are polymorphic in copy number[59], have been identified in diverse species including human, chimp, rat, mouse, and drosophila[60-79]. 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 emerge. Fundamental questions about the nature and impact of CNVs remain unanswered, mainly due to methodological constraints. In this thesis work, we set out to determine the copy number variable content of the mouse genome so that this information could be included in the integrated genomics study of t-AML susceptibility. Further, we estimated its functional impact, as measured by gene expression profiling *in vivo*.

At the time this thesis work began, the genome-wide discovery of CNVs was limited to large (>20 kb) events due to technological constraints. In order to accurately assess the impact of copy number variation on phenotype, as well as to learn more about their fine structure and

origins, it is necessary to reliably detect CNVs of all sizes and accurately determine their genomic boundaries. Currently, the most common genome-wide approaches to identify CNVs are array-based. These platforms include bacterial artificial chromosome (BAC) array comparative genomic hybridization (aCGH)[80,81], long oligonucleotide arrays [82-84] and single nucleotide polymorphism (SNP) genotyping arrays[85]. A critical aspect in selecting a platform for CNV detection is effective resolution, which we define as the length of the shortest CNV that is detectable at an acceptable false positive rate (FPR). A number of factors contribute to resolution, including probe density (i.e., the number of probes that interrogate a region of the genome), probe specificity and sensitivity. Due to their high probe density, long oligonucleotide arrays theoretically have the highest resolution and genome coverage of the three array-based platforms[86,87]. However, the higher level of noise of these platforms[86,88] has hampered efforts to mine these data for novel CNVs using available analytical tools, which were designed for BAC-array analysis. At the time this thesis work began, there was only one published account of a method designed specifically for detecting CNVs from such data[89] but there has been no comprehensive analysis of the achievable genome-wide resolution of these platforms. As a prerequisite to mapping common CNVs in inbred mice, estimating their impact on expression, and including them in integrated genomics studies of t-AML susceptibly, we first set out to develop a method for detecting CNVs specifically from long-oligo aCGH data. This CNV detection work constitutes the first phase of this dissertation project, as described in the next chapter

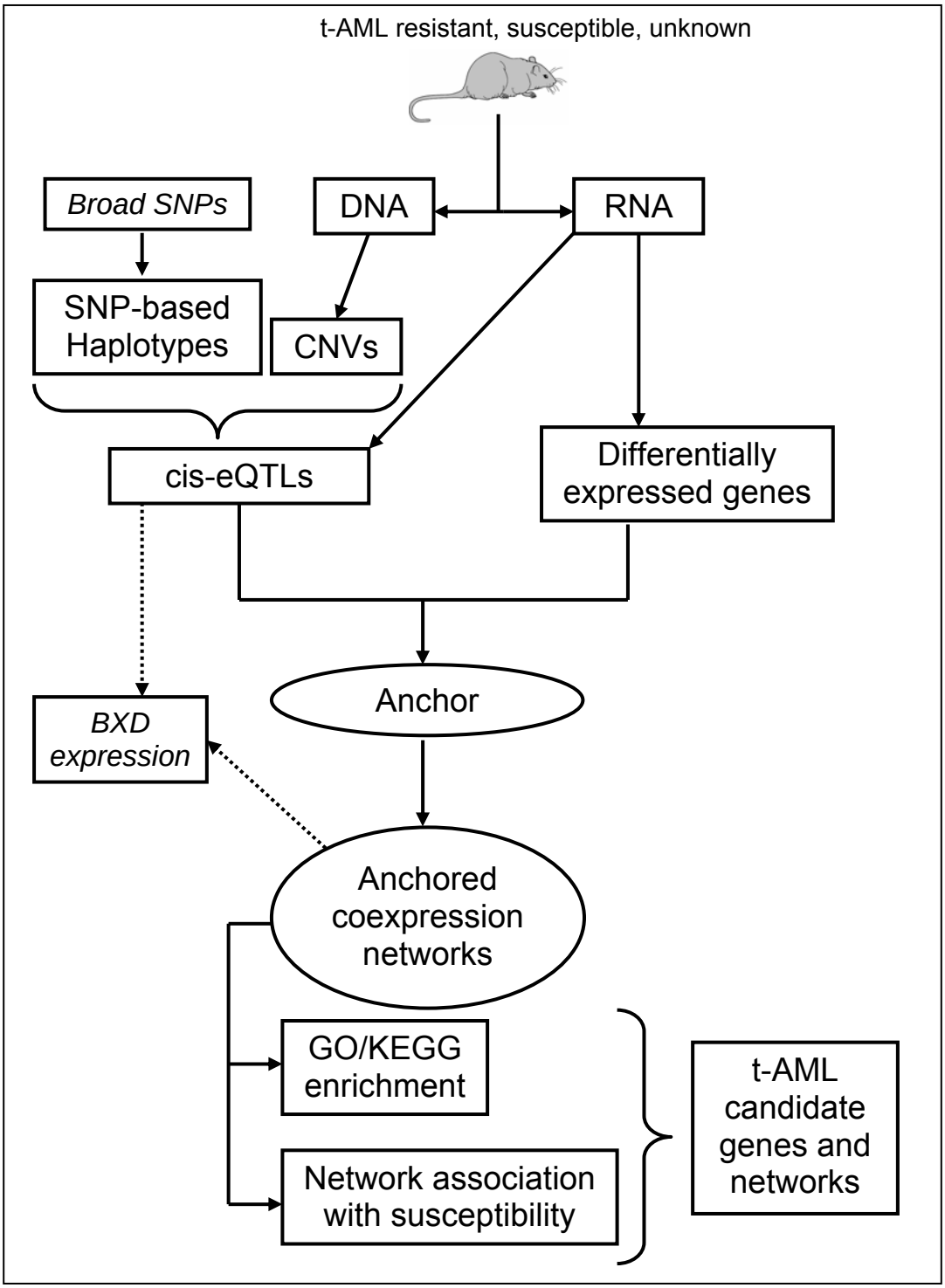

t-AML resistant, susceptible, unknown
Broad SNPs
DNA
RNA
SNP-based Haplotypes
CNVs
cis-eQTLs
Differentially expressed genes
Anchor
BXD expression
Anchored coexpression networks
GO/KEGG enrichment
Network association with susceptibility
t-AML candidate genes and networks

Figure 1: Flow chart of integrated genomics methodology for identification of candidate network underlying susceptibility. DNA and RNA are collected from 20 inbred strains of mice, 15 of which vary in susceptibility to alkylator induced AML. DNA is hybridized to aCGH arrays for detection of CNVs. SNPs genotypes downloaded from a public repository (Broad) are used to generate a haplotype map of 48 classical inbred strains of mice (superset of the 20 strains assayed for CNV). RNA is used for gene expression profiling to (1) determine genes that are differentially expressed between t-AML susceptible and resistant strains, and (2) to map expression traits to CNVs and haplotypes, in cis. Expression quantitative traits that do not replicate in independent data are removed from further analysis. Genes that are both differentially expressed between susceptible and resistant strains, and are linked to a validated eQTL are termed 'anchors'. Anchored coexpression networks are derived for each anchor by identifying all genes that have significantly correlated expression profiles to the anchor. Networks are trimmed of all genes that do not have reproducible association with anchor gene expression in independent data. Anchored coexpression networks are prioritized for downstream experimental assessment by GO/KEGG enrichment and association with t-AML susceptibility.

REFERENCES

1. Lowenberg, B., Downing, J.R. & Burnett, A. Acute myeloid leukemia. *N Engl J Med* **341**, 1051-62 (1999).

2. Ley, T.J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72 (2008).

3. Mardis, E.R. et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-66 (2009).

4. Crump, M. et al. Risk of acute leukemia following epirubicin-based adjuvant chemotherapy: a report from the National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol* **21**, 3066-71 (2003).

5. Brown, J.R. et al. Increasing incidence of late second malignancies after conditioning with cyclophosphamide and total-body irradiation and autologous bone marrow transplantation for non-Hodgkin's lymphoma. *J Clin Oncol* **23**, 2208-14 (2005).

6. Larson, R.A. & Le Beau, M.M. Therapy-related myeloid leukaemia: a model for leukemogenesis in humans. *Chem Biol Interact.* **153-154**, 187-195 (2005).

7. Leone, G., Pagano, L., Ben-Yehuda, D. & Voso, M.T. Therapy-related leukemia and myelodysplasia: susceptibility and incidence. *Haematologica* **92**, 1389-1398 (2007).

8. Leone, G., Voso, M.T., Sica, S., Morosetti, R. & Pagano, L. Therapy related leukemias: susceptibility, prevention and treatment. *Leuk Lymphoma* **41**, 255-76 (2001).

9. Schoch, C., Kern, W., Schnittger, S., Hiddemann, W. & Haferlach, T. Karyotype is an independent prognostic parameter in therapy-related acute myeloid leukemia (t-AML): an analysis of 93 patients with t-AML in comparison to 1091 patients with de novo AML. *Leukemia* **18**, 120-5 (2004).

10. Smith, S.M. et al. Clinical-cytogenetic associations in 306 patients with therapy-related myelodysplasia and myeloid leukemia: the University of Chicago series. *Blood* **102**, 43-52 (2003).

11. Larson, R.A. Etiology and management of therapy-related myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, 453-9 (2007).

12. Kantarjian, H.M., Estey, E.H. & Keating, M.J. Treatment of therapy-related leukemia and myelodysplastic syndrome. *Hematol Oncol Clin North Am* **7**, 81-107 (1993).

13. Meikrantz, W., Bergom, M.A., Memisoglu, A. & Samson, L. O6-alkylguanine DNA lesions trigger apoptosis. *Carcinogenesis* **19**, 369-72 (1998).

14. van Leeuwen, F.E. et al. Leukemia risk following Hodgkin's disease: relation to cumulative dose of alkylating agents, treatment with teniposide combinations, number of episodes of chemotherapy, and bone marrow damage. *J Clin Oncol* **12**, 1063-73 (1994).

15. Knoche, E., McLeod, H.L. & Graubert, T.A. Pharmacogenetics of alkylator-associated acute myeloid leukemia. *Pharmacogenomics* **7**, 719-29 (2006).

16. Felix, C.A. et al. The p53 gene in pediatric therapy-related leukemia and myelodysplasia. *Blood* **87**, 4376-81 (1996).

17. Allan, J.M. et al. Genetic variation in XPD predicts treatment outcome and risk of acute myeloid leukemia following chemotherapy. *Blood* **104**, 3872-7 (2004).

18. Maris, J.M. et al. Monosomy 7 myelodysplastic syndrome and other second malignant neoplasms in children with neurofibromatosis type 1. *Cancer* **79**, 1438-46 (1997).

19. Seedhouse, C. et al. The genotype distribution of the XRCC1 gene indicates a role for base excision repair in the development of therapy-related acute myeloblastic leukemia. *Blood* **100**, 3761-6 (2002).

20. Allan, J.M. et al. Polymorphism in glutathione S-transferase P1 is associated with susceptibility to chemotherapy-induced leukemia. *Proc Natl Acad Sci U S A* **98**, 11592-7 (2001).

21. Larson, R.A. et al. Prevalence of the inactivating 609C-->T polymorphism in the NAD(P)H:quinone oxidoreductase (NQO1) gene in patients with primary and therapy-related myeloid leukemia. *Blood* **94**, 803-7 (1999).

22. Seedhouse, C. & Russell, N. Advances in the understanding of susceptibility to treatment-related acute myeloid leukaemia. *Br J Haematol* **137**, 513-29 (2007).

23. Fenske, T.S. et al. Identification of candidate alkylator-induced cancer susceptibility genes by whole genome scanning in mice. *Cancer Res* **66**, 5029-38 (2006).

24. Noveroske, J.K., Weber, J.S. & Justice, M.J. The mutagenic action of N-ethyl-N-nitrosourea in the mouse. *Mamm Genome* **11**, 478-83 (2000).

25. Funk, R.K. et al. Quantitative trait loci associated with susceptibility to therapy-related acute murine promyelocytic leukemia in hCG-PML/RARA transgenic mice. *Blood* **112**, 1434-42 (2008).

26. Flint, J., Valdar, W., Shifman, S. & Mott, R. Strategies for mapping and cloning quantitative trait genes in rodents. *Nat Rev Genet* **6**, 271-86 (2005).

27. Eeles, R.A. et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-21 (2009).

28. Thomas, G. et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40**, 310-5 (2008).

29. Yeager, M. et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).

30. Gudmundsson, J. et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet* **41**, 1122-6 (2009).

31. Gudmundsson, J. et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-7 (2007).

32. Gudmundsson, J. et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* **40**, 281-3 (2008).

33. Yeager, M. et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet* **41**, 1055-7 (2009).

34. Thomas, G. et al. A multistage genome-wide association study in breast cancer identifies

two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**, 579-84 (2009).

35. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).

36. Hunter, D.J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870-4 (2007).

37. Stacey, S.N. et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39**, 865-9 (2007).

38. Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324-8 (2009).

39. Houlston, R.S. et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40**, 1426-35 (2008).

40. Zanke, B.W. et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39**, 989-94 (2007).

41. Tenesa, A. et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40**, 631-7 (2008).

42. Chung, C.C., Magalhaes, W., Gonzalez-Bosquet, J. & Chanock, S.J. Genome-wide Association Studies in Cancer - Current and Future Directions. *Carcinogenesis* (2009).

43. Hung, R.J. et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633-7 (2008).

44. McKay, J.D. et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404-6 (2008).

45. Wang, Y. et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**, 1407-9 (2008).

46. Kiemeney, L.A. et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* **40**, 1307-12 (2008).

47. Wu, X. et al. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat Genet* **41**, 991-5 (2009).

48. Knight, J.A. et al. Genome-wide association study to identify novel loci associated with therapy-related myeloid leukemia susceptibility. *Blood* **113**, 5575-82 (2009).

49. Strittmatter, W.J. et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A* **90**, 1977-81 (1993).

50. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).

51. Meng, H. et al. Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc Natl Acad Sci U S A* **104**, 4530-5 (2007).

52. Wang, S.S. et al. Mapping, genetic isolation, and characterization of genetic loci that determine resistance to atherosclerosis in C3H mice. *Arterioscler Thromb Vasc Biol* **27**, 2671-6 (2007).

53. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).

54. Schadt, E.E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**, 710-7 (2005).

55. Yang, X. et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet* **41**, 415-23 (2009).

56. Ghazalpour, A. et al. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2**, e130 (2006).

57. Plaisier, C.L. et al. A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet* **5**, e1000642 (2009).

58. Emilsson, V. et al. Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).

59. Freeman, J.L. et al. Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-61 (2006).

60. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75-81 (2006).

61. Cutler, G., Marshall, L.A., Chin, N., Baribault, H. & Kassner, P.D. Significant gene content variation characterizes the genomes of inbred mouse strains. *Genome Res* **17**, 1743-54 (2007).

62. Dopman, E.B. & Hartl, D.L. A portrait of copy-number polymorphism in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **104**, 19920-5 (2007).

63. Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O. & Long, M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**, 1629-31 (2008).

64. Graubert, T.A. et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* **3**, e3 (2007).

65. Guryev, V. et al. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40**, 538-45 (2008).

66. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38**, 82-5 (2006).

67. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).

68. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).

69. Korbel, J.O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-6 (2007).

70. Lee, A.S. et al. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* **17**, 1127-36 (2008).

71. Li, J. et al. Genomic segmental polymorphisms in inbred mouse strains. *Nat Genet* **36**, 952-4 (2004).

72. Perry, G.H. et al. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685-95 (2008).

73. Perry, G.H. et al. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* **103**, 8006-11 (2006).

74. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).

75. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).

76. She, X., Cheng, Z., Zollner, S., Church, D.M. & Eichler, E.E. Mouse segmental duplication and copy number variation. *Nat Genet* **40**, 909-14 (2008).

77. Snijders, A.M. et al. Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res* **15**, 302-11 (2005).

78. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).

79. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* (2009).

80. Pinkel, D. et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207-11 (1998).

81. Solinas-Toldo, S. et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**, 399-407 (1997).

82. Brennan, C. et al. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res* **64**, 4744-8 (2004).

83. Barrett, M.T. et al. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A* **101**, 17765-70 (2004).

84. Selzer, R.R. et al. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* **44**, 305-19 (2005).

85. Zhao, X. et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* **64**, 3060-71 (2004).

86. Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R.H. & Meijer, G.A. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic

hybridization (array CGH). *Nucleic Acids Res* **34**, 445-50 (2006).

87. Wicker, N. et al. A new look towards BAC-based array CGH through a comprehensive comparison with oligo-based array CGH. *BMC Genomics* **8**, 84 (2007).

88. Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**, S16-21 (2007).

89. Korbel, J.O. et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci U S A* **104**, 10110-5 (2007).

wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data

Patrick Cahan[1], Laura E Godfrey[1], Peggy S Eis[2], Todd A Richmond[2], Rebecca R Selzer[2], Michael Brent[1], Howard L McLeod[3], Timothy J Ley[1], Timothy A Graubert[1]

[1]Departments of Internal Medicine and Genetics, Division of Oncology, Stem Cell Biology Section, Washington University, St. Louis, MO.

[2]Roche NimbleGen, Inc., Madison, WI

[3]Institute for Pharmacogenomics and Individualized Therapy, University of North Carolina, Chapel Hill, NC.

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
email: graubert@medicine.wustl.edu

Running head: CNV Detection

ABSTRACT

Copy number variants (CNVs) are currently defined as genomic sequences that are polymorphic in copy number and range in length from 1,000 to several million base pairs. Among current array-based CNV detection platforms, long-oligonucleotide arrays promise the highest resolution. However, the performance of currently available analytical tools suffers when applied to these data because of the lower signal:noise ratio inherent in oligonucleotide-based hybridization assays. We have developed wuHMM, an algorithm for mapping CNVs from array comparative genomic hybridization (aCGH) platforms comprised of 385,000 to more than 3 million probes. wuHMM is unique in that it can utilize sequence divergence information to reduce the false positive rate (FPR). We apply wuHMM to 385K-aCGH, 2.1M-aCGH, and 3.1M-aCGH experiments comparing the 129X1/SvJ and C57BL/6J inbred mouse genomes. We assess wuHMM's performance on the 385K platform by comparison to the higher resolution platforms and we independently validate 10 CNVs. The method requires no training data and is robust with respect to changes in algorithm parameters. At a FPR of less than 10%, the algorithm can detect CNVs with five probes on the 385K platform and three on the 2.1M and 3.1M platforms, resulting in effective resolutions of 24 kb, 2-5 kb, and 1 kb, respectively.

INTRODUCTION

DNA copy number variation comprises a significant component of total genetic variation in human (1-4), chimpanzee (5), and mouse (6-9) populations. CNVs have been associated with disease susceptibility (10-16) and underlie variation in gene expression (17). To date, the genome-wide discovery of CNVs has been limited to large (>20 kb) events due to technological constraints. In order to accurately assess the impact of copy number variation on phenotype, as well as to learn more about their fine structure and origins, we must first be able to reliably detect CNVs of all sizes and accurately determine their genomic boundaries.

The most common genome-wide approaches to identify CNVs are array-based. These platforms include bacterial artificial chromosome (BAC) array comparative genomic hybridization (aCGH) (18,19), long oligonucleotide arrays (20-22) and single nucleotide polymorphism (SNP) genotyping arrays (23). A critical aspect in selecting a platform for CNV detection is effective resolution, which we define as the length of the shortest CNV that is detectable at an acceptable false positive rate (FPR). A number of factors contribute to resolution, including probe density (i.e., the number of probes that interrogate a region of the genome), probe specificity and sensitivity. Due to their high probe density, long oligonucleotide arrays theoretically have the highest resolution and genome coverage of the three platforms (24,25). However, the higher level of noise of these platforms (24,26) has hampered efforts to mine these data for novel CNVs using available analytical tools, which were designed for BAC-array analysis. To date, there has been only one published account of a method designed specifically for detecting CNVs from such data (27) but there has been no comprehensive analysis of the achievable genome-wide resolution of these platforms.

The goal of our work was to develop a method for detecting CNVs specifically from long-oligo aCGH data, characterize its sensitivity, FPR and effective resolution, and compare it to other CNV detection algorithms. Our focus is the detection of homozygous changes in the inbred mouse genome. Detection of heterozygous germline changes or somatic changes in mixed

cellular populations may present additional challenges due to diminished signal intensity. However, existing computational tools detect even homozygous CNVs with relatively low sensitivity and unacceptably high false positive rates. Although sequence divergence between a probe and its target impacts hybridization, no existing CNV detection algorithm has addressed this problem in the context of oligo-aCGH. Here we show that there is a strong association between regions of sequence divergence and hybridization signal in high resolution aCGH data from inbred strains of mice. We present a method that optionally incorporates sequence information into a Hidden Markov Model (HMM)-based calling algorithm. We assess its sensitivity and precision, and compare its performance to other algorithms, three of which are commonly used for lower resolution platforms and one recently developed for dense microarrays.

MATERIALS AND METHODS

Sample preparation and array comparative genomic hybridization

DNA was extracted from the spleens and kidneys of healthy, young adult (age 8-12 week) 129X1/SvJ and C57BL/6J mice (The Jackson Laboratory, Bar Harbor, ME). Different DNA samples were used for each aCGH platform (385K, 2.1M, and 3.1M). Array comparative genomic hybridization (aCGH) studies were performed using long oligonucleotide arrays designed and manufactured by Roche NimbleGen (Madison, WI). The aCGH experiments were performed using a single array (385K-aCGH) with a median probe spacing of 5.2 Kb (MM6, NCBI Build 34), a single array (2.1M-aCGH) with a median probe spacing of 1.015 Kb (MM8, NCBI Build 36) or an 8-array set (3.1M-aCGH) with median probe spacing of 0.49 Kb (MM7, NCBI Build 35). Labeling, hybridization, washing, and array imaging were performed as previously described (9,22). All mouse genome coordinates are based on NCBI Build 36 (MM8). Roche NimbleGen probe coordinates were re-mapped using liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Data is available at GEO (http://www.ncbi.nlm.nih.gov/geo/index.cgi) under accession GSE10511.

Algorithm overview

We developed Washington University HMM (wuHMM) specifically to maximize CNV detection on high density, long oligonucleotide arrays. wuHMM is comprised of several stages: clustering log2-ratios, finding regions more likely to contain CNVs, performing local CNV segmentation, and scoring (Figure 1A). The clustering stage bins log2-ratios for input to the HMM, which facilitates the incorporation of sequence information. There is an optional stage in which each chromosome is partitioned according to sequence divergence between the probe and target genomes based on independently derived genotype data. Segmentation is achieved by first searching for seeds consisting of short runs of probes with large magnitude log2-ratios. Seeded regions are then input to an HMM for segment boundary detection and scoring. The HMM (Figure 1B) is comprised of 5 states that represent normal and abnormal DNA copy number. The model requires a minimum length of stay in abnormal states in order to prevent singleton outliers from

being called as CNVs. CNVs are scored based on log2-ratio magnitude, number of probes, and local noise.

wuHMM can be downloaded from: http://groups.google.com/group/wuhmm. Default parameters (seed length, number of clusters, and noise penalty) are set to optimized values based on the sensitivity and FPR of wuHMM applied to data of known copy number. These parameters and the use of sequence divergence data can be specified by the user.

Sequence divergence

In this optional pre-processing step, partitioning of a chromosome is accomplished by utilizing a three-state HMM, in which the states represent regions of sequence divergence or similarity compared to a reference genome, or runs of no genotype calls (Supplementary Figure 1). The reference is the C57BL/6J inbred mouse genome. The observations in the model are determined by the genotypes of 138,608 known SNPs (28,29). Specifically, an observation is coded as '0' when the genotype differs between the test and reference genomes, as a '1' when the genotypes agree, and 'n' when there is no call in either strain. This model is appropriate for pair-wise comparisons between inbred mouse strains containing genomic regions of high pair-wise polymorphism rates. We required that the system remain within a state for at least five observations, yielding an average minimum block size of 87 kb, which lies within the estimated size range of ancestral block sizes in inbred mice (mean: 58 kb, range:1 kb to 3 Mb) (30). The HMM is trained by expectation maximization.

Clustering

We clustered probes by log2-ratios to achieve two aims. First, clustering facilitated the normalization of log2-ratios between regions of sequence divergence and similarity. Second, binning probes by log2-ratios provided a convenient means of linking the decoded states of probes, as determined by the HMM, to biologically meaningful DNA copy number states (normal, gain, or loss). The following procedure assigned cluster labels to each probe, ensuring that there

is the expected number of clusters for input to the HMM:

1. Divide probes in two groups:

    Group A: Probes with log2-ratios >= 0

    Group B: all other probes

2. Cluster probes in each group into floor(*n/2)* + *1* groups.

    n = number of clusters

3. Merge the cluster in Group A having the minimum magnitude mean log2-ratio and the cluster in Group B with the minimum magnitude mean log2-ratio into one cluster, resulting in *n* clusters.

4. Rank clusters by mean log2-ratio.

5. Label each probe by the rank of its cluster.

We used Partitioning Among Medoids (PAM), as implemented in R's 'cluster' package using the *clara* function (31). When sequence divergence information is utilized, probes are separated according to sequence divergence state first, then clustered and labeled as described above (Supplementary Figure 2). Probe cluster labels are treated as observations by the HMM.

Seeding

It was necessary to target regions of the genome that were likely to contain CNVs prior to executing a more sensitive CNV-detection algorithm. Without the seeding step we found that training the HMM on whole chromosomes periodically led to reduced power to detect short CNVs and misclassification of large regions of chromosomes as CNVs. We identified regions likely to harbor CNVs by the presence of consecutive probes with large magnitude log2-ratios. This was achieved using a stringent HMM in which the emissions from abnormal states were restricted to corresponding clusters. We trained the stringent HMM and performed decoding on each chromosome separately, producing a set of seeds. A seeded region, which was used as input to the more sensitive CNV detection algorithm, was defined as the seed-spanning region plus 100 probes on either side. Overlapping seeded regions were merged.

Hidden Markov Model

Our HMM generally follows the approach to decoding copy number from aCGH data as first described by Fridyland, et al (32) with several notable exceptions. The true, unobserved DNA copy number of a given probe is treated as a hidden state and probe cluster labels are the observed emissions from the model (Figure 1B). The initial emissions of abnormal states are weighted most heavily to the highest and lowest cluster ranks. Emissions from abnormal states cannot be from clusters with oppositely signed means. The initial transition probabilities are set such that most of the chromosome is assumed to be in a normal state. 'Joiner' states, which have an initial emission distribution weighted toward the corresponding abnormal state but permit emissions from all states, exist in order to prevent CNV call fragmentation. Final emission and transition probabilities are determined by the Baum and Welch expectation maximization algorithm for each seeded region until convergence of the model likelihood, which is typically achieved in fewer than 10 iterations. Training is repeated for each seeded region, varying the minimum length of stay in an abnormal state from 3 to 10. The model with the greatest likelihood is then used to determine copy number with the Viterbi decoding algorithm (33). The GHMM library (http://ghmm.sourceforge.net/software) was used to implement the HMMs.

Scoring function and permutation

We devised a scoring function that uses local noise, number of probes, and log2-ratios to ascertain the quality of CNV calls. This score, S_{cnv}, is defined as:

$S_{cnv} = \ln(n_{cnv}) * | \text{median}(\text{log2-ratio}_{cnv}) | - \text{SD}(\text{log2-ratio}_{cnv\_nps}) * W$, where:

n = number of probes comprising the CNV

cnv_nps = index probes within a distance of 5 * length of the call that share the same sign as the $\text{mean}(\text{log2-ratio})_{cnv}$

W = noise weight term

In attempting to determine the significance of a CNV score, probe locations were randomized for each chromosome, the segmentation method was applied, and the best score was stored. We

repeated these steps one hundred times to generate a null distribution of CNV scores for each chromosome. P-values were computed using R's 'quantile' function, which uses linear interpolation to estimate the given quantile (34).

Validation

Two methods were used to validate CNV calls. First, we used replicate aCGH experiments at increasing probe density to identify probes on the 385K array that have reproducible log2-ratio shifts. This information was used to assess the performance of wuHMM and other CNV detection algorithms, as described below (see Sensitivity and False Positive Rate). We performed three replicate aCGH experiments at increasing probe densities: two 2.1M-aCGH (each comprised of a single 2.1M feature array) experiments and one 3.1M-aCGH (eight-385K arrays) experiment. We included probes for assessment analysis only if there were at least four probes in the 6 kb centered at a 385K probe (median inter-probe distance on the 385K array is 6 kb) on both the 2.1M and 3.1M platforms. We termed these 'informative probes'. The gold standard is the copy number status (i.e. gain, loss, or neutral) of the informative probes. The copy number status of an informative probe was defined according to the $|\text{mean log2-ratio}_{region}|$ on the replicate arrays. Specifically, an informative probe was considered to represent a DNA copy number change if the $|\text{mean log2-ratio}_{region}| >$ threshold on all replicates, where the threshold varied between arrays and regions of sequence similarity and divergence. If an informative probe was in a divergent region and its log2-ratio < 0, then it was considered to represent a DNA copy number change if $|\text{mean log2-ratio}_{region}| > SD_{divergent\_blocks}$ for all replicate arrays, where $SD_{divergent\_blocks}$ is the standard deviation of probes in divergent regions. For all other informative probes, the threshold is the standard deviation of the sequence similar regions. The SD cutoffs for the similar regions were 0.2416, 0.2176 and 0.2200 for the 385K, 2.1M and 3.1M platforms, respectively. SD cutoffs for the divergent regions were 0.4115, 0.3457 and 0.3142.

Independent validation of 10 CNVs (all deletions) was achieved by attempting to amplify by PCR regions within CNV boundaries. PCR primers (Supplementary Table 1) were designed to localize

within a CNV. Amplification reactions contained 10 µl of Jumpstart Ready Mix Taq (Sigma, http://www.sigmaaldrich.com), 100 ng of each primer, and 10 ng of genomic DNA in a final volume of 20 µl. Amplifications were performed on a PTC-225 Peltier Thermal Cycler (MJ Research) at standard conditions for 30 cycles and the product was run on a 2% agarose gel, stained with ethidium bromide, and visualized on a GelDoc (BioRad).

Sensitivity and False Positive Rate

We calculated sensitivity and FPR of CNV detection algorithms on the 385K platform based on the gold standard. We calculated the sensitivity of CNV calls as the number of probes representing a true copy number change within predicted CNVs divided by the total number of probes representing true copy number changes in the gold standard. We defined the FPR as one minus the proportion of CNVs that are significantly enriched for probes representing a true copy number change. The enrichment of a CNV was determined by randomly selecting equally sized regions of the chromosome and recording the proportion of probes representing true copy number changes that they contain. We repeated this step one hundred times, generating a null distribution of enrichment values. We designated an observed call as a true positive if its enrichment value exceeded 95% of the random enrichment values. We observed that due to differences in probe design between platforms, some high-scoring calls on the 385K-aCGH were not sufficiently covered on the higher resolution platforms. Therefore, we excluded calls that were comprised of fewer than 25% informative probes in any performance analysis for wuHMM and other segmentation algorithms. Also, singletons and doubleton calls were not considered in any performance analysis.

Other segmentation algorithms

We applied GLAD (35), CBS (36), and BioHMM (37) to the 385K-aCGH data using BioConductor's *snapCGH* package (38). To reduce the amount of processing time required by GLAD and DNACopy, we divided each chromosome into blocks of approximately 50 Mb. These methods do not explicitly define segments as amplified or deleted. Segments were classified as

'abnormal' if the predicted log2-ratio was greater than 0.35 or less than -0.35. We used BreakPtr (27) version 1.0.5 downloaded from http://tiling.mbb.yale.edu/BreakPtr/. We trained the data using known gains and losses in 129X1/SvJ. We used the Finder-Core module with the default transition probabilities.

Other statistical tests

To test the association between sequence divergence and signal intensity, probes were partitioned according to sequence divergence state as described. A t-test, using R's *t.test* function not assuming equal variances, was applied to the raw, linear-scale signal intensities of the129X1/SvJ channel.

RESULTS

Sequence divergence affects probe hybridization signal

There are long regions of the 129X1/SvJ aCGH data that exhibit a dispersed but pronounced negative log2-ratio (Figure 2). These regions differ from true deletions, which are comprised almost entirely of negative log2-ratios. It was previously hypothesized that a similar phenomenon observed in BAC arrays was a result of decreased hybridization efficiency due to sequence polymorphism between the test and reference genomes (8). There are regions of classical inbred mouse genomes that exhibit pair-wise polymorphism rates exceeding 1/400 base pairs, reflecting divergent subspecies ancestry (30). We tested the hypothesis that the regions of dispersed negative log2-ratios represent blocks of different ancestry in C57BL/6J versus 129X1/SvJ by partitioning the 129X1/SvJ genome into blocks of sequence similarity and divergence relative to the C57BL/6J sequence using approximately 140,000 genotype calls**.** We found 1,826 sequence-similar blocks and 1,790 sequence-divergent blocks (median length 190 and 262 kb, respectively). As predicted, the signal intensity of 129X1/SvJ in regions of sequence divergence is significantly lower than in regions of sequence similarity in all experiments in the majority (18/19, 17/19, and 13/19, on 385K, 2.1M, and 3.1M arrays, respectively) of autosomes (Table 1). Similarly, the test channel intensity is lower in divergent blocks of 385K-aCGH data from 18 other inbred mouse strains, suggesting that the association between blocks of sequence divergence and aCGH signal is not an idiosyncrasy of a single strain comparison but represents a general phenomenon (data not shown). In order to determine the impact of sequence divergence on segmentation algorithms we attempted to validate by PCR five deletions in divergent regions called by a variety of algorithms on 385K-aCGH data. All five putative deletions failed to validate (Supplementary Figure 3 and data not shown), indicating that they do not represent true deletions but are instead artifacts of sequence polymorphism affecting hybridization. This underscores the importance of incorporating methods to differentiate between CNVs and blocks of high polymorphism rates in order to reduce the number of false positive segment calls.

Gold standard

In order to assess the FPR and sensitivity of wuHMM and other segmentation methods we needed to determine the true copy number state of each assayed region of the 129X1/SvJ genome. Replication by independent methods (e.g., PCR, qPCR, and FISH) is the accepted standard by which CNV predictions are considered validated. It would not be practical to use any of these methods to systematically validate the thousands of predictions made by all algorithms tested. Instead, we determined the 129X1/SvJ copy number of the 6 kb region spanning each 385K-aCGH probe (approximately equal to the median spacing of the platform) by comparison to replicate experiments at higher resolutions (two 2.1M-aCGH, one 3.1M-aCGH). We reasoned that if the signal from a 385K-aCGH probe represents a true copy number change, then the log2-ratio shift will be reproducible on higher density platforms with more probes reflecting the variation. The higher density platforms contain, on average, 5.6 and 8.7 probes per 6 kb window on the 2.1M and 3.1M platforms, respectively. 336,470 probes on the 385K array are informative (i.e., there were at least 4 probes in the 6 kb region spanning the probe on both the 2.1M and 3.1M platforms). Of the informative probes, we found that 1,886 represented true copy number changes since they had reproducible log2-ratio shifts on all three replicate arrays. 1,226 informative probes were singletons (i.e., probes representing a copy number change that are adjacent to informative probes that do not represent true copy number change). Two hundred fifty-two probes were doubletons, similarly defined as an adjacent pair of validated probes surrounded by informative probes not representing true copy number change.

We next asked if it would be feasible to detect singletons or doubletons using only log2-ratio thresholds. Standard deviation (SD) multipliers were used to identify probes as potential CNVs. Even when the SD multiplier threshold > 5 was applied, 89% of the called probes were false positives and less than 5% of the called probes were true positives (Table 2). These results demonstrate that attempting to detect singletons or doubletons from a single experiment will result in unsatisfactory sensitivity and FPR. For this reason, we removed singletons and doubletons from both the gold standard and CNV predictions prior to the calculation of sensitivity

and FPR. Four hundred and eight probes representing true copy number changes remained after removing singletons and doubletons.

We calculated the sensitivity and FPR of all CNV detection algorithms based on the 385K gold standard, which is defined as the copy number status of the informative probes. CNV predictions were considered correct if they contained a significantly enriched number of informative probes that represented a true copy number change. The FPR was calculated as one minus the ratio of the number of correct CNV predictions to the total number of CNV predictions. In this way, the FPR is presented at a CNV-level. However, the sensitivity could only be calculated at the level of individual probes because the total number of 'correct' CNVs remains unknown in our gold standard. The sensitivity is calculated as the ratio of the number of informative probes contained within predicted CNVs that represented a true copy number change to the total number of probes representing true copy number changes.

Scoring function

It is common practice to prioritize or rank CNV predictions for downstream analysis and experiments such as validation and evaluation of functional significance. We view this prioritization in terms of a scoring function that relates aspects of the call (e.g., the amplitude of deviation from a log2-ratio of 0, the number of probes within a segment) to the quality of the call. A well-designed scoring function will generate high scores for true positive calls and low scores for false positive calls. We first asked which choice of threshold acted as a better scoring function: the number of probes per segment, or the |mean log2-ratio| of the segment. We calculated the sensitivity and FPR of wuHMM across a range of parameter settings and reported the maximum sensitivity when the FPR < 15% (Supplementary Table 2 and Supplementary Figure 4). The |mean log2-ratio| performed poorly (mean sensitivity = 8.5%). The number of probes per segment threshold performed substantially better (mean sensitivity = 40.6%), but we speculated that a scoring function that uses both parameters would provide further improvement. A combined scoring function (see Methods) had the best performance at all parameter settings

(mean sensitivity = 47.8%).

Next, we hypothesized that we could assign a statistical significance to CNV calls by generating a null distribution of scores for calls made on randomized data. On a per-chromosome basis, we randomized probe locations, executed wuHMM and stored the highest score. We repeated this process 100 times to generate a null distribution of scores. We calculated p-values for each observed call based on comparison of its score to the null distribution of scores. We found that the FPR of scores with p-values < 0.01 remained above 47%, indicating that this permutation approach to determining CNV call quality did not achieve an acceptable FPR. Therefore, the scoring function can be used to evaluate algorithm performance, but significance thresholds for the scores must be determined empirically.

Algorithm parameters

An important goal in developing wuHMM was to make it tunable such that changes in initial parameter settings would have predictable effects on performance and therefore could be adjusted to meet the needs of each individual analysis. We evaluated the effect on wuHMM's sensitivity and FPR of varying: the number of clusters, the minimum number of probes required in the seeding step (seed length), use of sequence information, and the scoring function noise penalty. First, we investigated the effect of varying only seed length and the number of clusters. We expected that increasing the seed length would decrease the overall sensitivity and FPR because larger values of the seed length would increase the likelihood that the algorithm would skip regions containing small CNVs. We executed wuHMM using a range of seed lengths and number of clusters, calculated the sensitivity and FPR at increasing score thresholds, and generated Receiver Operating Curves (Figure 3). As expected, we found that increasing the seed length reduced the maximum sensitivity (from 70% to 34%) and the maximum FPR (86% to 35%). The best performance (sensitivity = 53% at FPR < 10%) was achieved when seed length was 2, although a value of 3 performed nearly as well. There was no clear performance trend with increasing the number of clusters. The best performance (sensitivity = 50%, FPR < 10%),

achieved with the number of clusters = 5, was substantially better than other numbers of clusters. These results demonstrate that seed length can be increased to decrease the maximum FPR at the expense of a much reduced sensitivity. Further, they show that a combination of seed length = 2 and number of clusters = 5 produces the optimal performance tradeoff. To determine if wuHMM would be generally applicable with these parameter settings (i.e., that it is not over-trained) we applied it to previously described data from 19 other inbred strains at the 385K resolution (9). Of the 72 previously discovered 'high-confidence' CNVs, 71 (98.6%) were detected with wuHMM using identical parameter settings (e.g. seed length = 2, number of clusters = 5, using sequence divergence information). Additionally, the range of call lengths and number of calls per genome are consistent with the 129X1/SvJ calls (length range: 9 kb - 4 Mb, median length = 138 kb, mean length = 460 kb). The calls per genome range from one (C57BL/6Tac) to 75 (Molf/EiJ), with a mean of 36 +/- 17.

We next analyzed the effect of incorporating sequence divergence on wuHMM's performance. We calculated the difference between the sensitivity and FPR of wuHMM with or without sequence divergence at increasing score thresholds. As predicted, utilizing sequence information reduced both the FPR and the probe-level sensitivity (Figure 4). These effects were greatest for calls scoring between 0.8 and 1.4, a score range which includes validated gains and losses. We next calculated sensitivity and FPR using a range of values for the noise penalty, W, which decreases the score of calls in regions of greater noise (see Methods). We found that increasing the noise penalty resulted in equalizing the FPRs between wuHMM with sequence information and without sequence information. At the same time, the sensitivity did not substantially improve, demonstrating that the use of a noise penalty with sequence divergence information results in worse overall performance.

Genotype information is not readily available for all aCGH experiments that may contain noise due to sequence divergence. We asked if using a noise penalty would improve FPR at an acceptable loss of sensitivity when sequence information is not available. We executed wuHMM

without sequence information using a range of penalty values and calculated the sensitivity and FPR at increasing score thresholds (Supplementary Figure 5). We found that there was no performance improvement when using any non-zero penalty. We concluded that for the range of values tested, the noise penalty does not enable the score function to differentiate between real calls and noise. Therefore, we recommend the use of conservative score thresholds when there is substantial noise in the data.

Effective resolution

Using parameter values that optimized sensitivity and FPR (seed length = 2, number of clusters = 5, noise penalty = 0) we applied wuHMM to all data sets. We selected a score threshold that yielded a FPR < 7% and sensitivity of 56% on the 385K platform. We attempted to independently validate ten calls made from the 2.1M and 3.1M experiments by PCR. We considered a call to be validated when we were able to detect an amplified product in the C57BL/6J sample but not in the 129X1/SvJ sample. All ten calls confirmed the wuHMM predictions, independently demonstrating that wuHMM can reliably detect calls comprised of as few as three probes on 2.1M-aCGH and seven probes on 3.1M-aCGH (Figure 5).

We estimated the effective resolution of the 385K platform by determining the length of the call with the fewest probes with a score exceeding 1.9 (i.e. at a FPR < 7%) (Table 3). Assuming that the relationship between CNV score and the FPR remains relatively constant across aCGH densities, we estimated the effective resolutions of the 2.1M and 3.1M platforms by averaging the lengths of the calls comprised of the fewest probes with scores exceeding 1.9 (Table 3).

Comparison to other methods

We compared the performance of our approach to four other segmentation algorithms: Gain and Loss Analysis of DNA (GLAD), BioHMM, DNACopy, and BreakPtr. The performances of GLAD and DNACopy, as well as other HMM implementations have been compared previously using well-characterized BAC array and simulated data (39,40). Using default parameters, we applied

each algorithm to the 385K-aCGH data, scored CNV calls, removed singletons, doubletons, and calls comprised of less than 25% informative probes (see Methods), and computed sensitivity and FPR based on the gold standard. In order to ensure an unbiased comparison of algorithms, we determined the lowest score cutoff at which each method reached a FPR < 10%. For all methods this score threshold was 1.9. wuHMM reached the highest sensitivity, followed closely by DNACopy and more distantly by BreakPtr and GLAD (Table 4). All HMM-based methods required less than an hour of execution time. Although input data was partitioned prior to input to DNACopy and GLAD, these methods still had the longest executions times at 1.4 and 12.4 hours, respectively. BreakPtr appeared to be critically dependent on its training set. We initially trained the 'no-change' state with data from a self-self hybridization, but this resulted in BreakPtr calling over 10% of the informative probes, resulting in a 99% FPR. Among currently available methods, wuHMM achieves the highest sensitivity while maintaining an acceptable FPR.

DISCUSSION

Prior to this report, the selection of tools for the analysis of long oligonucleotide aCGH data was limited largely to software originally designed for other aCGH platforms, such as BAC-based or SNP genotyping arrays. We developed wuHMM to improve CNV detection from long oligonucleotide aCGH data that may be confounded by sequence divergence. wuHMM addresses sequence divergence by increasing the call stringency in sequence divergent regions of the genome. The effect of this strategy is to lower the FPR and, to a lesser extent, the sensitivity. In order to assess the algorithm, we developed a validated data set that should be a useful resource for the evaluation of other segmentation methods. By applying wuHMM to the validated data set, we demonstrated that it reaches the highest sensitivity among currently available methods at a FPR of less than 10%.

There are two caveats that apply to this analysis. First, in the current version of wuHMM, sequence divergent regions were estimated using only 140,000 SNPs. Therefore, small regions of sequence divergence may be missed. When more sequence data becomes available it can be incorporated into our method to better define the divergent regions, perhaps even down to the single aCGH probe level. Second, we expect that all existing CNV detection algorithms will exhibit reduced sensitivity when applied to aCGH data from outbred populations or samples with mixtures of somatic and germline copy number changes.

We estimate that effective resolutions of the 2.1M and 3.1M probe aCGH platforms, extrapolated based on a score threshold that yielded a FPR < 10% on the 385K probe platform, are 2-5 kb and 1 kb, respectively. However, although we independently validated several CNVs shorter than 5 kb, the overall confidence in resolution estimates for the 2.1M and 3.1M probe arrays will require additional evaluation. The first genome-wide studies of normal copy number variation in the mouse genome, based on BAC-aCGH platforms, were limited to a resolution of approximately 1 Mb (6-8). In 385K-aCGH data sets using a single whole-genome array (median probe spacing of 5.2 kb) and CNV analysis algorithms available at the time, we previously reported a total of five

CNVs in the 129X1/SvJ genome (9). Applying wuHMM to the 385K-aCGH data, we can now detect 15 CNVs in the 129X1/SvJ genome at an empirical FPR < 10%. Applying wuHMM to 3.1M-aCGH (an 8-fold increase in resolution) yields 167 CNVs. Theoretically, another 10-fold increase in probe density to a median probe spacing of approximately 87 bases for the mouse genome will enable the resolution of 'sub-CNV' events (i.e. insertion-deletions). Comprehensive tools such as the ones presented here are necessary to accurately assess the phenotypic impact of CNVs, improve our understanding of CNV origins, and facilitate integrated quantitative trait locus (QTL) mapping, linkage, and association studies.

FUNDING

Supported by the National Cancer Institute (P01 CA101937) and the National Human Genome Research Institute (T32 HG000045). The Siteman Cancer Center is supported in part by NCI Cancer Center Support Grant #P30 CA91842.

ACKNOWLEDGMENTS

Mice were kindly provided through a collaboration with the Mouse Phenome Project (The Jackson Laboratory, Bar Harbor, ME). We thank Alexander Schliep for advice concerning the gHMM library and Matthew Walter and Richard Walgren for critical reading and helpful suggestions on the manuscript.

Table 1. Relationship between sequence identity and aCGH signal.

	385K-aCGH					2.1M-aCGH					3.1M-aCGH				
	Probe count		Test signal			Probe count		Test signal			Probe count		Test signal		
Chr	M	MM	M	MM	p-value	M	MM	M	MM	p-value	M	MM	M	MM	p-value
1	13,188	16,205	4655	4489	8.90E-15	72,049	91,734	3550	3331	1.56E-82	101,963	124,620	3266	3120	2.60E-44
2	13,909	13,990	4612	4455	1.30E-12	76,513	77,565	3516	3463	5.04E-06	111,127	114,313	2999	2854	2.80E-48
3	11,306	11,954	4682	4485	7.70E-16	64,400	67,515	3427	3316	9.24E-20	85,703	90,427	2494	2422	2.00E-12
4	8,747	13,700	4628	4434	4.00E-15	48,344	78,238	3612	3351	3.07E-81	72,101	111,243	2508	2427	4.50E-16
5	9,935	12,574	4646	4505	1.20E-07	55,891	69,525	3601	3497	2.35E-14	79,773	102,559	2703	2629	1.70E-15
6	12,095	10,437	4661	4414	1.60E-27	67,247	57,338	3205	3113	1.49E-14	94,546	82,567	2865	2772	3.20E-19
7	8,542	11,126	4626	4359	8.40E-21	48,250	63,624	3334	3029	6.48E-116	74,732	93,915	3606	3450	1.60E-21
8	7,962	11,620	4662	4379	2.40E-23	43,465	65,059	3303	2986	3.81E-124	65,419	94,888	3646	3422	1.20E-39
9	7,903	11,389	4617	4422	2.20E-14	43,317	62,739	3295	3137	9.58E-32	65,014	94,193	2385	2095	7.00E-149
10	13,865	5,569	4670	4562	2.10E-04	77,515	32,036	3139	3183	1.49E-03	106,880	43,868	2011	2000	2.60E-01
11	11,058	8,255	4567	4438	5.40E-06	62,019	44,518	3311	3210	3.13E-13	95,181	69,851	2445	2453	5.00E-01
12	8,686	7,689	4660	4417	1.60E-17	50,261	43,365	3057	2972	6.60E-10	69,487	61,473	3193	3220	1.60E-01
13	9,250	8,121	4671	4507	2.60E-08	51,745	45,216	3062	2916	7.69E-28	75,538	63,704	3269	3193	8.30E-06
14	7,982	9,259	4682	4389	4.90E-23	46,043	51,318	2918	2820	2.87E-13	60,075	73,647	2674	2683	5.10E-01
15	7,888	7,931	4637	4388	4.60E-16	43,200	43,898	3073	2814	1.34E-71	63,517	62,254	2512	2365	1.90E-42
16	7,768	6,861	4616	4563	7.70E-02	44,036	37,931	2967	2856	4.09E-15	60,921	51,324	2474	2465	4.20E-01
17	5,464	8,188	4642	4486	1.10E-05	30,042	46,894	3025	2958	1.72E-05	42,824	67,144	2851	2708	1.80E-23
18	6,324	7,615	4707	4538	4.80E-08	34,739	41,374	2999	2980	1.93E-01	48,748	60,966	3124	3053	9.70E-07
19	7,336	1,773	4645	4490	1.10E-03	40,292	9,748	3111	3008	2.57E-05	60,703	14,985	3078	3055	3.20E-01

MM: Regions of high polymorphism between C57BL/6J and 129X1/SvJ (“mismatched”); M: Non-polymorphic regions (“matched”). Probe count columns contain the number of probes within M and MM regions. Test signal columns contain the mean, single channel, linear-scale aCGH intensities of the M and MM regions. The p-value is the result of a t-test, testing the difference of the mean signals of M and MM probes.

Table 2. Detection of singletons and doubletons on 385K-aCGH.

SD multiplier	Singleton Sensitivity	Doubleton Sensitivity	FPR	Number of probes (percent of total)
0.25	0.869	0.881	0.993	251166 (74.6)
0.50	0.742	0.782	0.992	176851 (52.6)
0.75	0.631	0.698	0.989	118501 (35.2)
1.00	0.553	0.631	0.985	77423 (23)
1.25	0.487	0.560	0.979	50327 (15)
1.50	0.431	0.496	0.971	33049 (9.8)
1.75	0.376	0.425	0.963	22172 (6.6)
2.00	0.336	0.381	0.953	15630 (4.6)
2.25	0.300	0.329	0.942	11409 (3.4)
2.50	0.259	0.298	0.933	8594 (2.6)
2.75	0.234	0.282	0.923	6698 (2)
3.00	0.206	0.214	0.916	5270 (1.6)
3.25	0.177	0.183	0.910	4226 (1.3)
3.50	0.152	0.159	0.905	3384 (1)
3.75	0.127	0.139	0.902	2718 (0.8)
4.00	0.108	0.115	0.896	2200 (0.7)
4.25	0.090	0.091	0.894	1786 (0.5)
4.50	0.074	0.079	0.888	1415 (0.4)
4.75	0.064	0.052	0.885	1109 (0.3)
5.00	0.048	0.040	0.891	906 (0.3)
5.25	0.037	0.036	0.897	735 (0.2)
5.50	0.031	0.016	0.898	598 (0.2)
5.75	0.024	0.012	0.899	467 (0.1)
6.00	0.020	0.008	0.903	393 (0.1)

Table 3. Effective resolution of aCGH platforms analyzed by wuHMM.

Platform	Resolution (kilobases)	Standard deviation	Segment Length (base pairs)		Segment Length (probes)	
			Minimum	Median	Minimum	Median
385K	23.7	0.2629	23,577	191,594	5	23
2.1M-a[1]	5.2	0.3336	1,872	7,618	3	7
2.1M-b[2]	2.2	0.2846	1,906	7,067	3	7
3.1M	1.1	0.2690	909	6,156	3	9

[1]First technical replicate. [2]Second technical replicate.

Table 4. Performance of segmentation algorithms on 385K-aCGH data

Method	Probe sensitivity	Execution time (hours)	Additional input
wuHMM	56.1	0.17	Genotype data
DNACopy	54.4	1.4	Partition input
BreakPtr	43.9	0.02	Supervised training
GLAD	43.1	12.4	Partition input
BioHMM	21.6	0.1	None

REFERENCES

1. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat Genet*, **36,** 949-951.

2. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305,** 525-528.

3. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444,** 444-454.

4. Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E. *et al.* (2007) A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet*, **80,** 91-104.

5. Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Caceres, A.M., Iafrate, A.J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E. *et al.* (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A*, **103,** 8006-8011.

6. Li, J., Jiang, T., Mao, J.H., Balmain, A., Peterson, L., Harris, C., Rao, P.H., Havlak, P., Gibbs, R. and Cai, W.W. (2004) Genomic segmental polymorphisms in inbred mouse strains. *Nat Genet*, **36,** 952-954.

7. Adams, D.J., Dermitzakis, E.T., Cox, T., Smith, J., Davies, R., Banerjee, R., Bonfield, J., Mullikin, J.C., Chung, Y.J., Rogers, J. *et al.* (2005) Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat Genet*, **37,** 532-536.

8. Snijders, A.M., Nowak, N.J., Huey, B., Fridlyand, J., Law, S., Conroy, J., Tokuyasu, T., Demir, K., Chiu, R., Mao, J.H. *et al.* (2005) Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res*, **15,** 302-311.

9. Graubert, T.A., Cahan, P., Edwin, D., Selzer, R.R., Richmond, T.A., Eis, P.S., Shannon, W.D., Li, X., McLeod, H.L., Cheverud, J.M. *et al.* (2007) A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome. *PLoS Genet*, **3,** e3.

10. Fellermann, K., Stange, D.E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C.L., Reinisch, W., Teml, A., Schwab, M., Lichter, P. *et al.* (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet*, **79,** 439-448.

11. Szatmari, P., Paterson, A.D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X.Q., Vincent, J.B., Skaug, J.L., Thompson, A.P., Senman, L. *et al.* (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet*, **39,** 319-

328.

12. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316,** 445-449.

13. Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R. *et al.* (2003) alpha-Synuclein locus triplication causes Parkinson's disease. *Science*, **302,** 841.

14. Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E. *et al.* (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, **439,** 851-855.

15. Mullighan, C.G., Goorha, S., Radtke, I., Miller, C.B., Coustan-Smith, E., Dalton, J.D., Girtman, K., Mathew, S., Ma, J., Pounds, S.B. *et al.* (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*, **446,** 758-764.

16. Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J. *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436,** 117-122.

17. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315,** 848-853.

18. 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,** 207-211.

19. Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. and Lichter, P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20,** 399-407.

20. Brennan, C., Zhang, Y., Leo, C., Feng, B., Cauwels, C., Aguirre, A.J., Kim, M., Protopopov, A. and Chin, L. (2004) High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res*, **64,** 4744-4748.

21. Barrett, M.T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P.S. *et al.* (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A*, **101,** 17765-17770.

22. Selzer, R.R., Richmond, T.A., Pofahl, N.J., Green, R.D., Eis, P.S., Nair, P., Brothman, A.R. and Stallings, R.L. (2005) Analysis of chromosome breakpoints in neuroblastoma

at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer*, **44,** 305-319.

23. Zhao, X., Li, C., Paez, J.G., Chin, K., Janne, P.A., Chen, T.H., Girard, L., Minna, J., Christiani, D., Leo, C. *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, **64,** 3060-3071.

24. Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R.H. and Meijer, G.A. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*, **34,** 445-450.

25. Wicker, N., Carles, A., Mills, I.G., Wolf, M., Veerakumarasivam, A., Edgren, H., Boileau, F., Wasylyk, B., Schalken, J.A., Neal, D.E. *et al.* (2007) A new look towards BAC-based array CGH through a comprehensive comparison with oligo-based array CGH. *BMC Genomics*, **8,** 84.

26. Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*, **39,** S16-21.

27. Korbel, J.O., Urban, A.E., Grubert, F., Du, J., Royce, T.E., Starr, P., Zhong, G., Emanuel, B.S., Weissman, S.M., Snyder, M. *et al.* (2007) Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci U S A*, **104,** 10110-10115.

28. Wade, C.M. and Daly, M.J. (2005) Genetic variation in laboratory mice. *Nat Genet*, **37,** 1175-1180.

29. Bogue, M.A., Grubb, S.C., Maddatu, T.P. and Bult, C.J. (2007) Mouse Phenome Database (MPD). *Nucleic Acids Res*, **35,** D643-649.

30. 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*.

31. Rousseeuw, L.K.a.P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

32. Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N.A.N. (2004) Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, **90,** 132-153.

33. Rabiner, L. (1990) A tutorial on hidden Markov models and selected applications in speech recognition. 267-296.

34. Hyndman, R.J. and Fan, Y. (1996) Sample Quantiles in Statistical Packages. *The American Statistician*, **50,** 361-365.

35. Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F. and Barillot, E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20,** 3413-3422.

36. Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5,** 557-572.

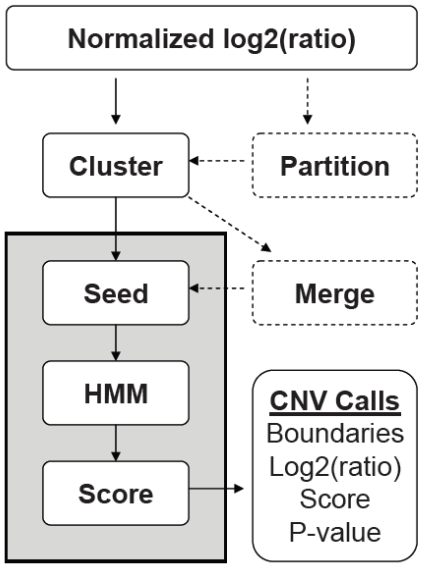
37. Marioni, J.C., Thorne, N.P. and Tavare, S. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22,** 1144-1146.

38. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5,** R80.

39. Willenbrock, H. and Fridlyand, J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21,** 4084-4091.

40. Lai, W.R., Johnson, M.D., Kucherlapati, R. and Park, P.J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21,** 3763-3770.

A.


Normalized log2(ratio)
Cluster
Partition
Seed
Merge
HMM
Score
CNV Calls
Boundaries
Log2(ratio)
Score
P-value

B.

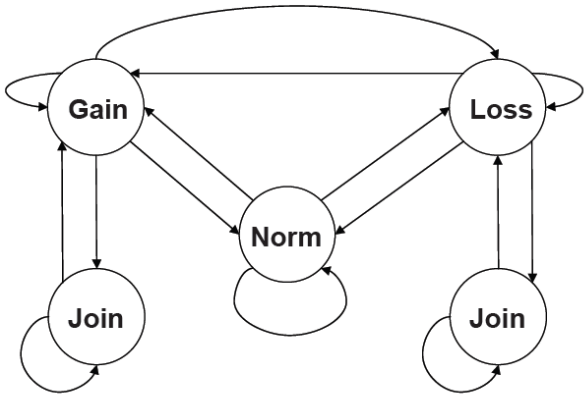

Gain
Loss
Norm
Join
Join

Figure 1. (A) Flow diagram of the wuHMM algorithm. Dashed processes are optional and are executed when the sequence divergence information is utilized. Processes in gray are repeated on permuted probe locations to generate null score distributions for each chromosome. (B) Hidden Markov Model. 'Norm', 'Gain, and 'Loss' indicate states representing normal, increased, and reduced DNA copy number, respectively. Not shown, but implemented, are multiple states per abnormal state that enforce a minimum number of probes per abnormal state. This minimum is automatically selected for each seeded region as described in Methods. Transitions are permitted between normal, increased, and reduced states. A 'Join' state can transition to itself or back to the corresponding abnormal state.

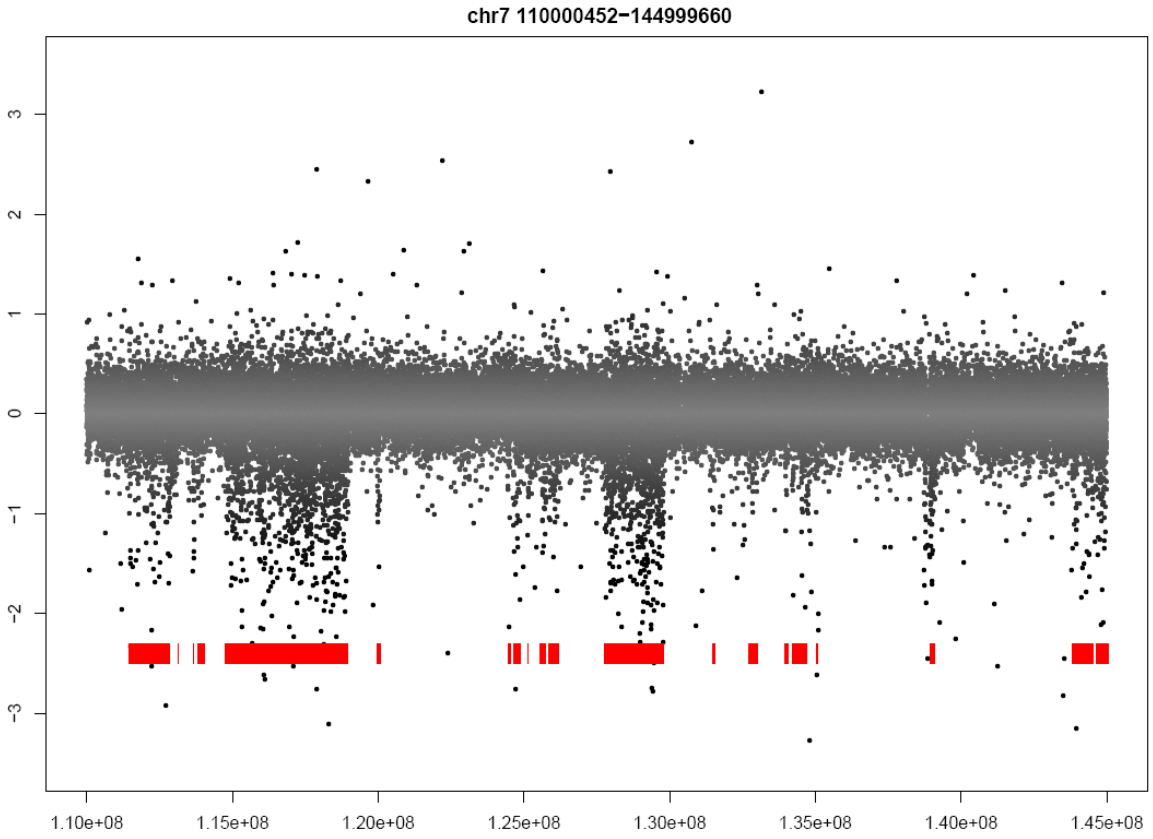

chr7 110000452−144999660
3
2
1
0
−1
−2
−3
1.10e+08
1.15e+08
1.20e+08
1.25e+08
1.30e+08
1.35e+08
1.40e+08
1.45e+08

Figure 2. 3.1M-aCGH log2-ratio plot of 129X1/SvJ chromosome 7. Blocks of sequence divergence are shown in red. Blocks of divergence correspond to aCGH probes with lower log2-ratios and can potentially confound CNV calling algorithms.

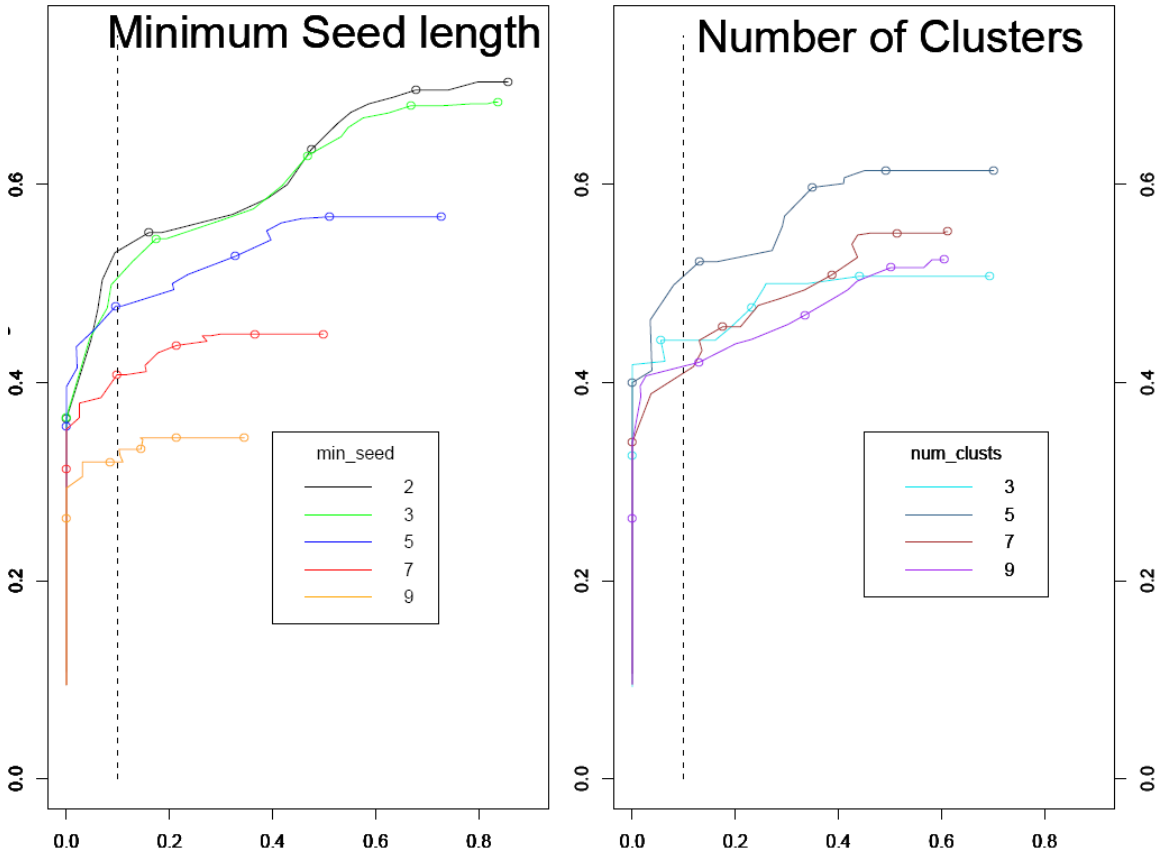

Minimum Seed length
min_seed
2
3
5
7
9
Number of Clusters
num_clusts
3
5
7
9
0.0
0.2
0.4
0.6
0.8

Figure 3. Receiver Operating Curves characterize the performance of wuHMM. (A) Each curve represents the performance of wuHMM at a given minimum seed length. Score cutoffs ranging from 0 to 2.5 were used to calculate sensitivities and false positive rates averaged across executions of wuHMM with different numbers of clusters. Circles represent score cutoffs of 0.0, 0.5, 1.0, 1.5, and 2.0, from right to left. The vertical dashed line represents a FPR = 10%. (B) The performance of wuHMM varying the number of clusters in the clustering stage. Score cutoffs ranging from 0 to 2.5 were used to calculate sensitivities and false positive rates averaged across executions of wuHMM with different seed lengths. As in (A), circles represent score cutoffs of 0.0 ,0.5, 1.0, 1.5, and 2.0, from right to left, and the vertical dashed line represents a FPR = 10%.

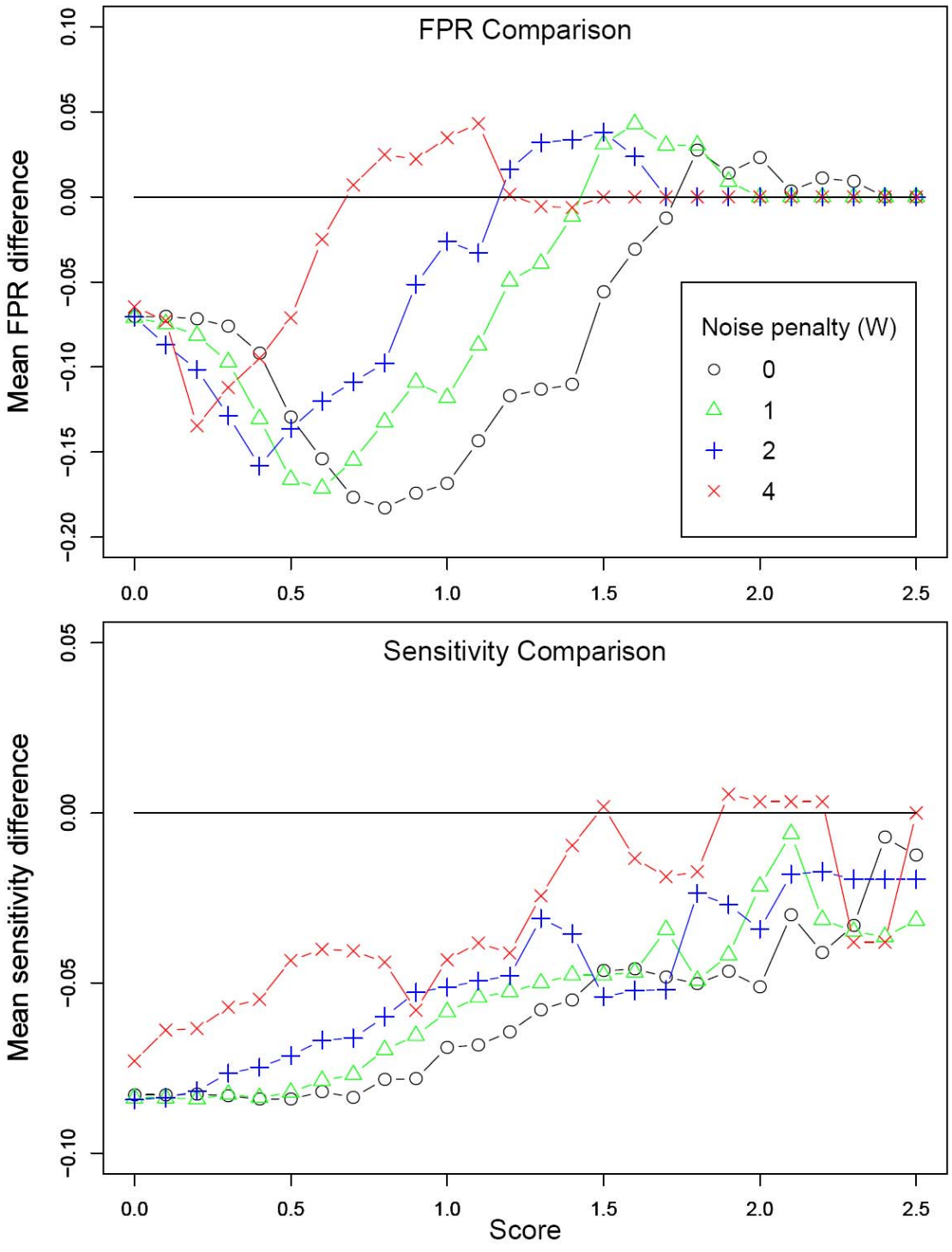

FPR Comparison
Mean FPR difference
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0.0
0.5
1.0
1.5
2.0
2.5
Noise penalty (W)
0
1
2
4
Sensitivity Comparison
Mean sensitivity difference
0.05
0.00
−0.05
−0.10
0.0
0.5
1.0
1.5
2.0
2.5
Score

Figure 4. Performance differences between wuHMM with sequence divergence and without sequence divergence. (A) FPR difference. Y-axis is the difference between the average false positive rates at the given score cutoff. A value below the y=0 line represents an improvement in the FPR when sequence divergence is utilized. (B) Sensitivity difference. Y-axis is the difference between the average sensitivities at the given score cutoff. In (A) and (B) each curve represents the performance difference with varying noise penalties (W). FPRs and sensitivities are averaged across a range of values for the number of clusters and minimum seed length.

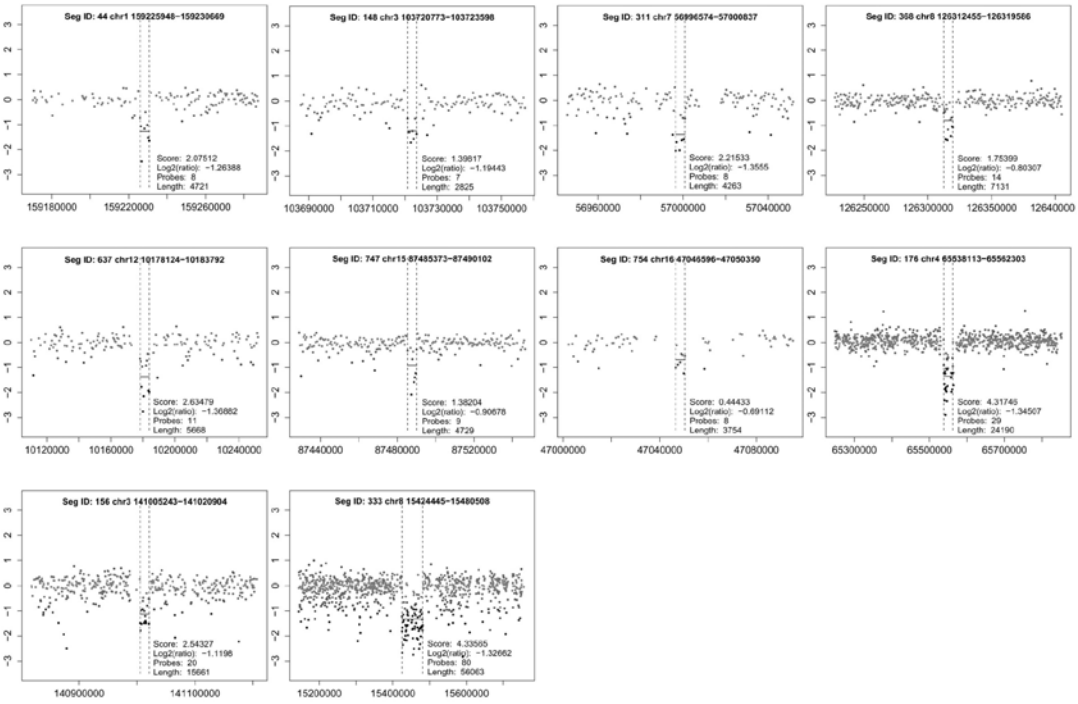
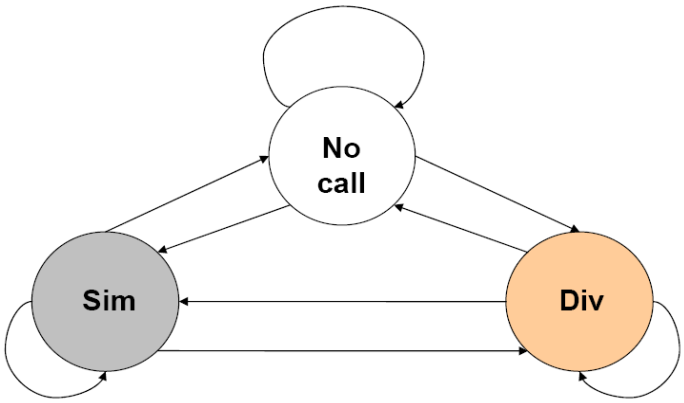

A.
Seg ID: 44 chr1 159225948-159230669
Score: 2.07512
Log2(ratio): -1.26388
Probes: 8
Length: 4721
159180000
159220000
159260000
Seg ID: 148 chr3 103720773-103723598
Score: 1.39817
Log2(ratio): -1.19443
Probes: 7
Length: 2825
103690000
103710000
103730000
103750000
Seg ID: 311 chr7 56996574-57000837
Score: 2.21533
Log2(ratio): -1.3555
Probes: 8
Length: 4263
56960000
57000000
57040000
Seg ID: 368 chr8 126312455-126319586
Score: 1.75399
Log2(ratio): -0.80307
Probes: 14
Length: 7131
126250000
126300000
126350000
12640000
Seg ID: 637 chr12 10178124-10183792
Score: 2.63479
Log2(ratio): -1.36882
Probes: 11
Length: 5668
10120000
10160000
10200000
10240000
Seg ID: 747 chr15 87485373-87490102
Score: 1.38204
Log2(ratio): -0.90678
Probes: 9
Length: 4729
87440000
87480000
87520000
Seg ID: 754 chr16 47046596-47050350
Score: 0.44433
Log2(ratio): -0.69112
Probes: 8
Length: 3754
47000000
47040000
47080000
Seg ID: 176 chr4 65538113-65562303
Score: 4.31746
Log2(ratio): -1.34567
Probes: 29
Length: 24190
65300000
65500000
65700000
Seg ID: 156 chr3 141005243-141020904
Score: 2.54327
Log2(ratio): -1.1196
Probes: 20
Length: 15661
140900000
141100000
Seg ID: 333 chr8 15424445-15480508
Score: 4.33565
Log2(ratio): -1.32662
Probes: 80
Length: 56063
15200000
15400000
15600000
B.
NT
B6
129
CNV ID
044
148
311
368
637
747
754
176
156
333
cont.

Figure 5. Validation of selected 3.1M-aCGH CNV calls in 129X1/SvJ. (A) Log2-ratio plots of validated 3.1M-aCGH CNV calls. The genomic position is plotted on the x-axis and the log2 (129X1/SvJ signal / C57BL/6J signal) is plotted on the y-axis. CNVs are annotated with a unique identifier (Seg ID), boundaries, mean log2-ratio, and score. Dotted lines indicate CNV boundaries as determined by wuHMM. (B) PCR validation. All ten deletions were validated by PCR, as demonstrated by a visible product using C57BL/6J, but not 129X1/SvJ genomic DNA. The marker is a 100 bp ladder. A region not deleted in 129X1/SvJ serves as a positive control.

SUPPLEMENTARY FIGURES AND TABLES

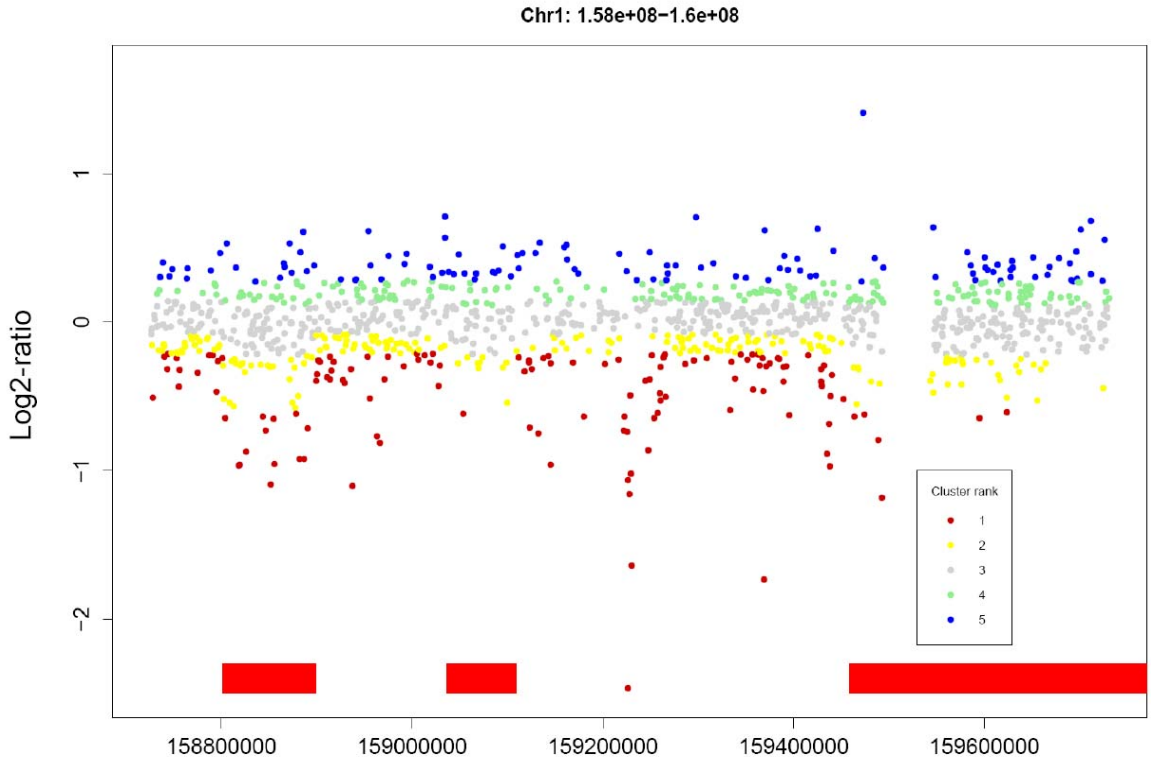
A.



B.

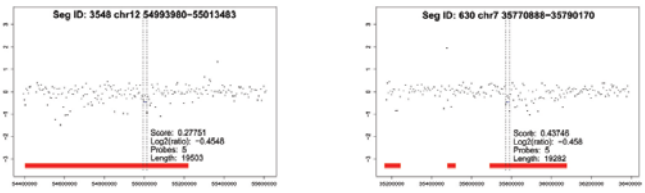
129X1/SvJ	A	T	C	C	A	T	T	C	G	G	C	T	T	C	C	C
C57BL/6J	A	T	C	C	A	T	T	C	A	G	G	C	C	T	T	T

Supplementary Figure 1. Genome partitioning by sequence divergence as determined using SNP genotype calls. (A) An HMM for determining regions of sequence divergence (Div) or similarity (Sim), compared to a reference genome, or runs of no genotype calls (No call). (B) Example of a transition from a similar (grey) to a divergent (tan) sequence block based on SNP genotype calls.

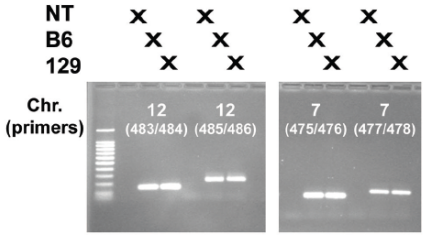

Chr1: 1.58e+08−1.6e+08
Log2-ratio
1
0
−1
−2
158800000
159000000
159200000
159400000
159600000
Cluster rank
1
2
3
4
5

Supplementary Figure 2. Ranking probes by cluster. After probes are divided according to sequence divergence, log2-ratios are separately clustered using PAM. Clusters are ranked by mean log2-ratio and probes are assigned their respective cluster rank. The goal of clustering probes separately is to normalize signals across regions of sequence divergence. Probes are colored by their cluster rank. Note that in regions of sequence divergence (indicated by red blocks) a larger magnitude log2-ratio is required for a probe to be included in extreme clusters.

A.

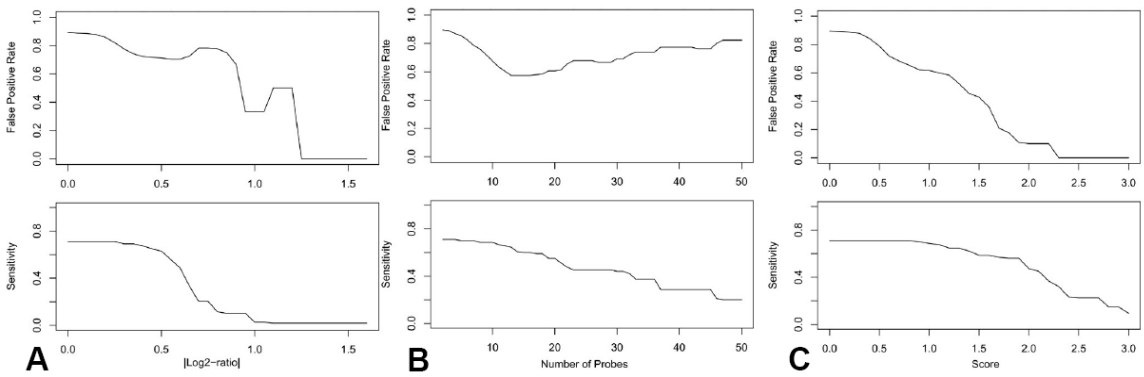


B.

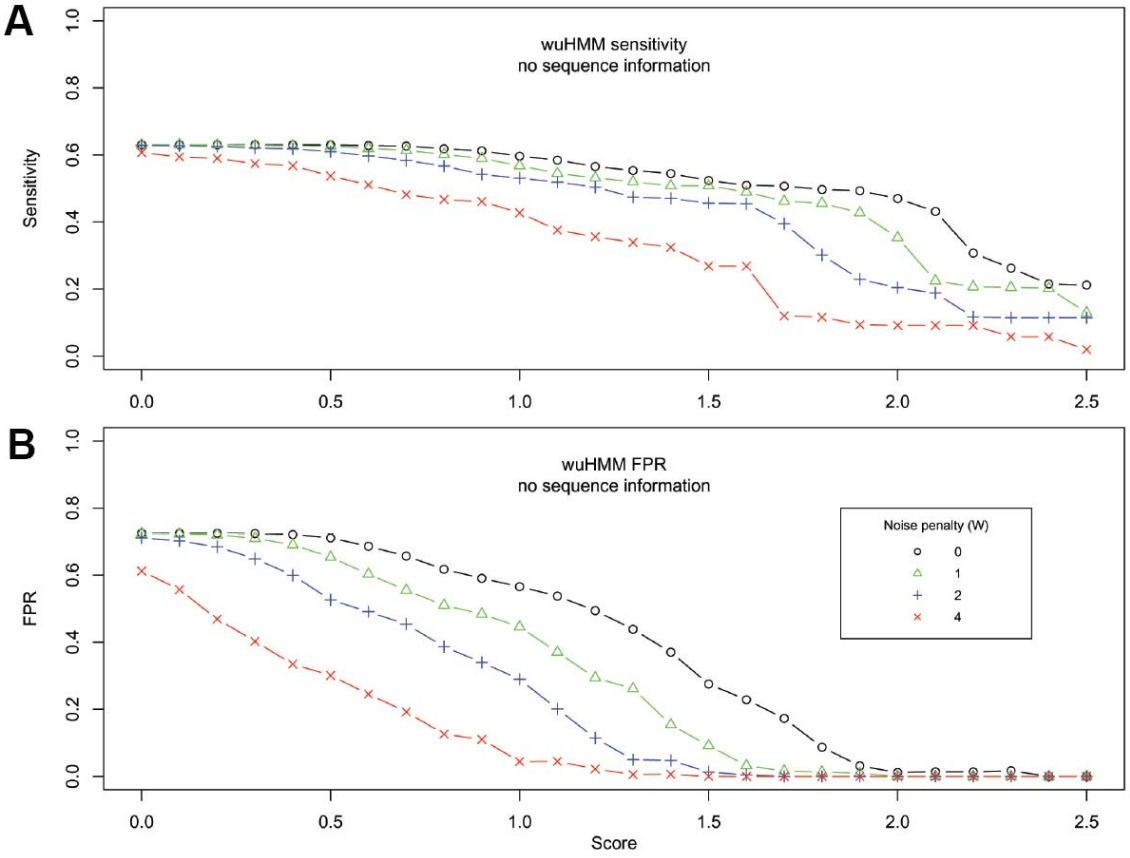


Supplementary Figure 3

Supplementary Figure 3. (A) Log2-ratio plots of the invalidated calls. Red blocks indicate blocks of sequence divergence. Seg ID 3548 was only called by wuHMM when sequence divergence information was not used. (B) PCR validation of two putative deletions called by several segmentation algorithms. For each call, primer pairs for two non-overlapping amplicons were designed. Primer pair IDs are displayed and sequences are available in Supplementary Table 1. In both cases, the putative CNV is invalidated. NT, no template.


False Positive Rate
Sensitivity
|Log2−ratio|
Number of Probes
Score
A
B
C

Supplementary Figure 4. Comparison of threshold criteria. Different criteria were applied to predictions made with wuHMM at optimal algorithm parameters (seed length = 2, clusters = 5, using sequence divergence information). (A) FPR (top) and sensitivity (bottom) of a threshold using only log2-ratio amplitude, (B) only number of probes, and (C) score function.


A
wuHMM sensitivity
no sequence information
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
B
wuHMM FPR
no sequence information
FPR
Noise penalty (W)
0
1
2
4
Score

Supplementary Figure 5. Noise penalty performance comparison of wuHMM without using sequence information for score penalties ranging from 0 to 4. (A) The sensitivity of calls made at increasing score cutoffs. (B) The FPR of calls made at increasing score cutoffs.

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.

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
email: graubert@wustl.edu

Running head: CNV eQTL mapping

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 drosophila[1-10]. 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 emerge[11-14].

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 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 strains[15] 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 rate[16]. 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 quality[17]. 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 platforms[18] and paired-end mapping[19], 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/HlJ) 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 mice[1,2,5,20,21]. If *de novo* events contribute only minimally to copy number variation among individuals within a strain[22,23], 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 formation[24]. 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 CNVRs[20,25]. 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 CNVRs[4]. 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 duplications[20]. 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 themselves[26], 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 content[1,4,8,21]. 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 active[27], 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 available[28,29]. 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 structure[30]. 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 strains[31,32], 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 separately[33].

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 mice[34]. 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 demonstrated[35,36] and *Alad* expression variation was associated with a *cis*-eQTL reported in an F2 inter-cross[37], 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 behavior[38]. 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 functions[39]. Genes at this locus have been speculated to undergo distal regulation via a chromosomal looping mechanism[40] 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 duplications[39], 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 available[41]. 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 lines[42]. 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 effects[43]. 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 Cross[44]) 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 described[45]. 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 CNVs[16]. Using an imputed single nucleotide map[46], we defined regions of high sequence divergence between the test and reference genomes for input to wuHMM, a Hidden Markov Model algorithm for CNV detection[16]. 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 R[17]. The average silhouette function calculates the average between versus within group distances and ranges from -1 to 1, with 1 representing perfect clustering[17]. 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 described[1]. 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 investigation[47]. 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 sub-sets 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 coverage[20], *de novo* events[22], or because lower resolution platforms tend to over-estimate CNVR boundaries[1,21]. 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 strains[22]. 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 described[28,31,32] 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 map[46] (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.

ACKNOWLEDGEMENTS

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, T.A. et al. A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome. *PLoS Genetics* **3**, e3 (2007).
2. Li, J. et al. Genomic segmental polymorphisms in inbred mouse strains. *Nat Genet* **36**, 952-954 (2004).
3. Perry, G.H. et al. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* **103**, 8006-11 (2006).
4. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
5. Snijders, A. et al. Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res* **15**, 302-311 (2005).
6. Dopman, E.B. & Hartl, D.L. A portrait of copy-number polymorphism in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **104**, 19920-5 (2007).
7. Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O. & Long, M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**, 1629-31 (2008).
8. Guryev, V. et al. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40**, 538-45 (2008).
9. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).
10. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).
11. Aitman, T.J. et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851-5 (2006).
12. McCarroll, S.A. et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* (2008).
13. Singleton, A.B. et al. alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
14. Walsh, T. et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-43 (2008).
15. Bogue, M.A. & Grubb, S.C. The Mouse Phenome Project. *Genetica* **122**, 71-4 (2004).
16. Cahan, P. et al. wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. *Nucleic Acids Res* **36**, e41 (2008).
17. Kaufman, L. & Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*, (Wiley, New York, 1990).
18. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75-81 (2006).
19. Korbel, J.O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-6 (2007).
20. She, X., Cheng, Z., Zollner, S., Church, D.M. & Eichler, E.E. Mouse segmental duplication and copy number variation. *Nat Genet* **40**, 909-14 (2008).
21. Cutler, G., Marshall, L.A., Chin, N., Baribault, H. & Kassner, P.D. Significant gene content variation characterizes the genomes of inbred mouse strains. *Genome Res* **17**, 1743-54 (2007).
22. Egan, C.M., Sridhar, S., Wigler, M. & Hall, I.M. Recurrent DNA copy number variation in

the laboratory mouse. *Nat Genet* **39**, 1384-9 (2007).

23. Watkins-Chow, D.E. & Pavan, W.J. Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome Res* **18**, 60-6 (2008).
24. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**, 417-22 (1998).
25. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).
26. Akagi, K., Li, J., Stephens, R.M., Volfovsky, N. & Symer, D.E. Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res* **18**, 869-80 (2008).
27. Chambers, S.M. et al. Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny. *Cell Stem Cell* **1**, 578-591 (2007).
28. McClurg, P. et al. Genomewide Association Analysis in Diverse Inbred Mice: Power and Population Structure. *Genetics* **176**, 675-683 (2007).
29. Wu, C. et al. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet* **4**, e1000070 (2008).
30. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* **76**, 8-32 (2005).
31. Pletcher, M.T. et al. Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol* **2**, e393 (2004).
32. McClurg, P., Pletcher, M.T., Wiltshire, T. & Su, A.I. Comparative analysis of haplotype association mapping algorithms. *BMC Bioinformatics* **7**, 61 (2006).
33. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70 (1979).
34. Bystrykh, L. et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**, 225-32 (2005).
35. Bishop, T.R., Cohen, P.J., Boyer, S.H., Noyes, A.N. & Frelin, L.P. 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* **83**, 5568-72 (1986).
36. Bishop, T.R., Miller, M.W., Wang, A. & Dierks, P.M. Multiple copies of the ALA-D gene are located at the Lv locus in Mus domesticus mice. *Genomics* **48**, 221-31 (1998).
37. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
38. Hovatta, I. et al. Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature* **438**, 662-6 (2005).
39. Ohtsuka, M., Inoko, H., Kulski, J.K. & Yoshimura, S. Major histocompatibility complex (Mhc) class Ib gene duplications, organization and expression patterns in mouse strain C57BL/6. *BMC Genomics* **9**, 178 (2008).
40. Kumar, P.P. et al. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nat Cell Biol* **9**, 45-56 (2007).
41. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38**, 82-5 (2006).
42. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
43. Korbel, J.O. et al. The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol* **18**, 366-74 (2008).
44. Churchill, G.A. et al. The Collaborative Cross, a community resource for the genetic

analysis of complex traits. *Nat Genet* **36**, 1133-1137 (2004).

45. Selzer, R.R. et al. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* **44**, 305-19 (2005).

46. Szatkiewicz, J.P. et al. An imputed genotype resource for the laboratory mouse. *Mamm Genome* **19**, 199-208 (2008).

47. Tian, D. et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105-8 (2008).

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 ($P < 0.05$).

Table 2: Subset of CNVR-eQTLs in hematopoietic stem/progenitor cells, hypothalamus, and adipose tissues.

				Hematopoietic		Hypothalamus		Adipose	
Gene symbol	CNVR ID	Chr	Position (Mb)	R^2	Proximity (Kb)	R^2	Proximity (Kb)	R^2	Proximity (Kb)
Alad	754	4	62	0.76	0	0.66	0	0.41	0
Glo1	3001	17	30	0.81	0	0.87	0	0.86	0
Sox13	216	1	135	0.42	5	0.30	5		0
2310009E04Rik	766	4	97	0.35	1,149	0.48	1149		0
Thumpd1	1383	7	119	0.28	423	0.26	422		0
Ifi205	127	1	177	0.39	1,181			0.42	1,308
Cstf3	420	2	104	0.24	61			0.29	5
Hdc	432	2	128	0.26	1,761			0.26	1,761
Gbp1	640	3	143	0.96	0			0.91	0
Hdhd3	754	4	62	0.49	0			0.43	0
Trim56	925	5	137	0.78	37			0.37	37
Gtf3a	931	5	146	0.50	764			0.28	763
Capg	1077	6	72	0.94	473			0.69	468
Mir16	1383	7	119	0.43	592			0.58	598
Hemk1	1749	9	107	0.94	233			0.30	233
Pbx1	232	1	171			0.47	794	0.31	794
Trim34	1372	7	104			0.54	263	0.76	263
4833420G17Rik	2405	13	121			0.75	1	0.38	1
Paip1	2405	13	121			0.35	36	0.51	36
Zfr	2719	15	12			0.49	474	0.36	474
Cxadr	2872	16	79			0.68	477	0.47	477
Sytl3	2978	17	6			0.47	0	0.37	0
H2-T23	3014	17	36			0.47	114	0.55	114

All eQTLs listed in the table are significant at an alpha < 0.05 after correcting for multiple tests. 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 Sytl3).

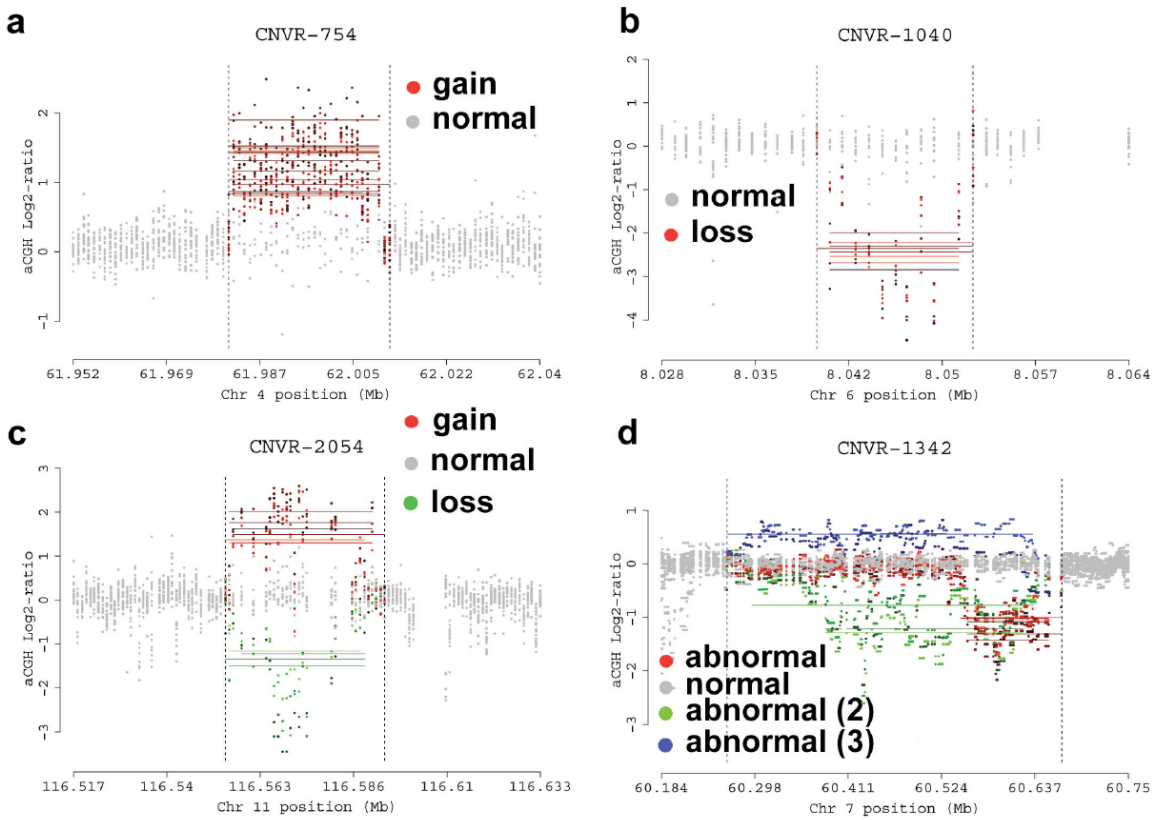

a
CNVR-754
gain
normal
aCGH Log2-ratio
Chr 4 position (Mb)
b
CNVR-1040
normal
loss
aCGH Log2-ratio
Chr 6 position (Mb)
c
CNVR-2054
gain
normal
loss
aCGH Log2-ratio
Chr 11 position (Mb)
d
CNVR-1342
abnormal
normal
abnormal (2)
abnormal (3)
aCGH Log2-ratio
Chr 7 position (Mb)

Figure 1

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.

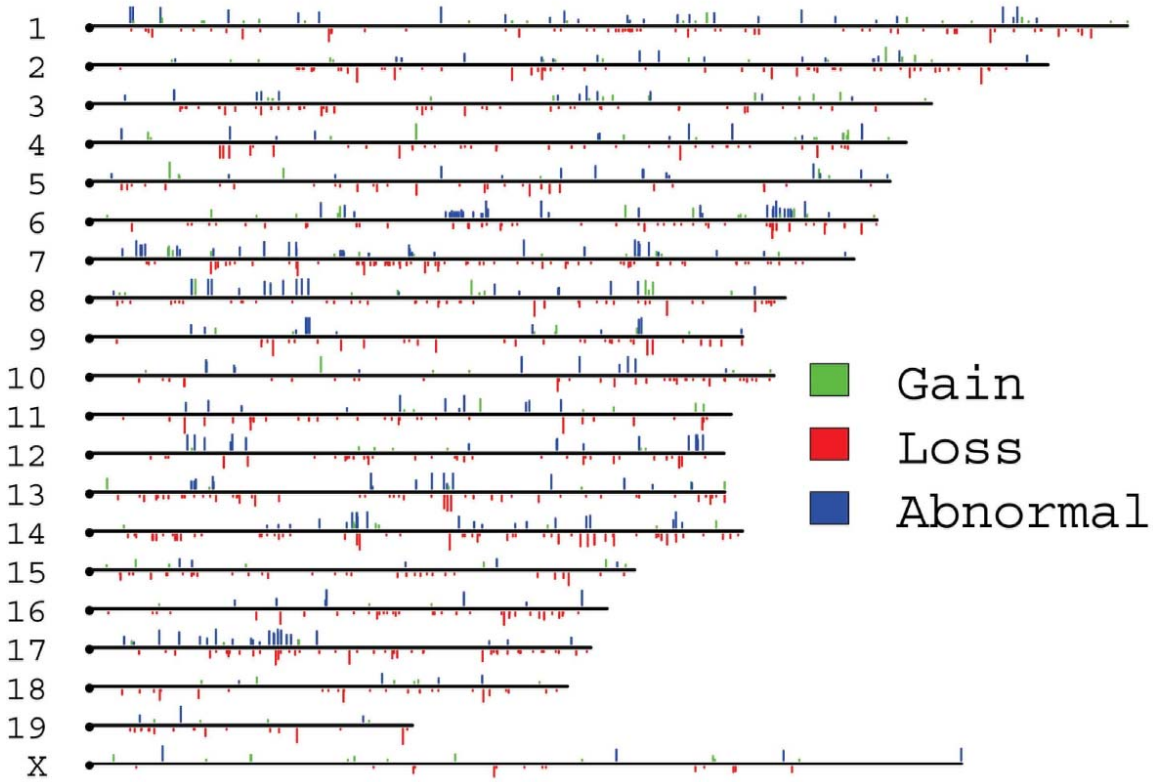


Figure 2

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.

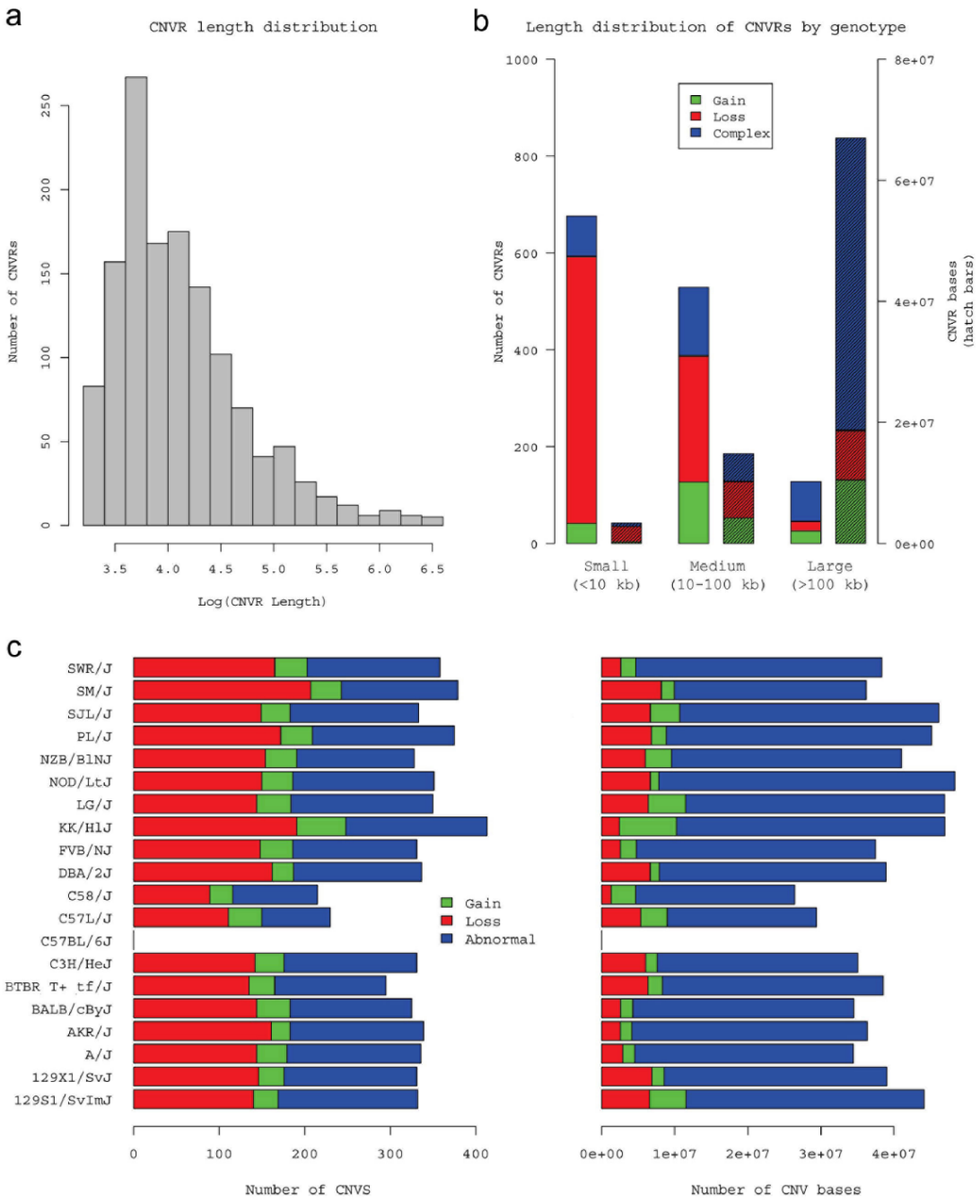

a
CNVR length distribution
Number of CNVRs
Log(CNVR Length)
b
Length distribution of CNVRs by genotype
Gain
Loss
Complex
Number of CNVRs
CNVR bases (hatch bars)
Small (<10 kb)
Medium (10-100 kb)
Large (>100 kb)
c
SWR/J
SM/J
SJL/J
PL/J
NZB/BlNJ
NOD/LtJ
LG/J
KK/HlJ
FVB/NJ
DBA/2J
C58/J
C57L/J
C57BL/6J
C3H/HeJ
BTBR T+ tf/J
BALB/cByJ
AKR/J
A/J
129X1/SvJ
129S1/SvImJ
Gain
Loss
Abnormal
Number of CNVS
Number of CNV bases

Figure 3

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.

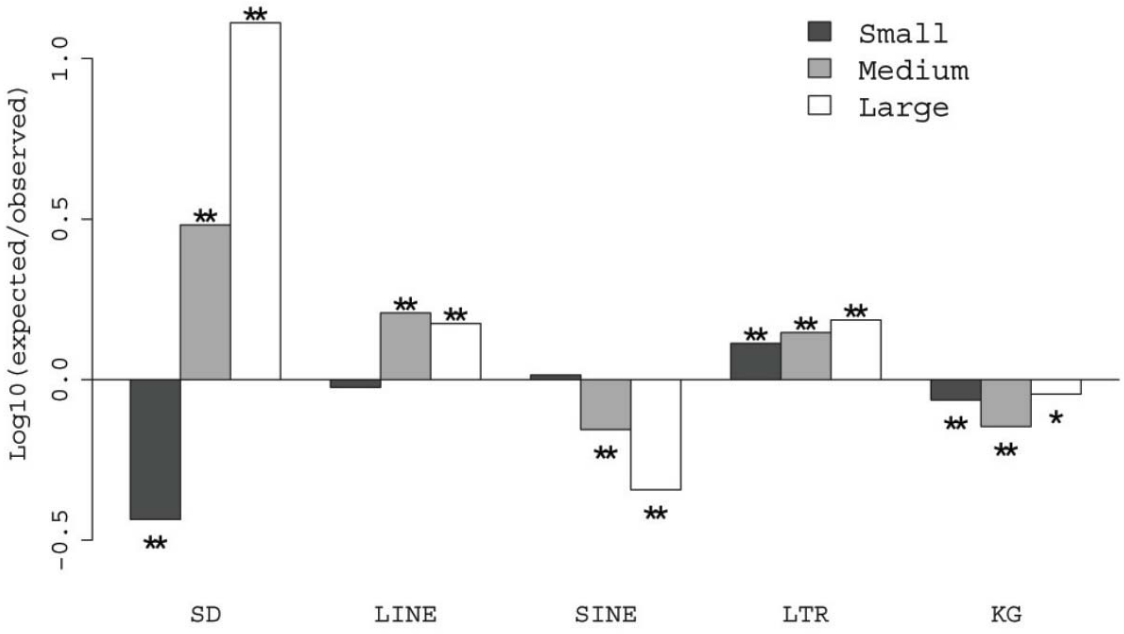

Small
Medium
Large
Log10(expected/observed)
1.0
0.5
0.0
-0.5
SD
LINE
SINE
LTR
KG

Figure 4

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. * $P < 0.05$. ** $P < 0.01$.

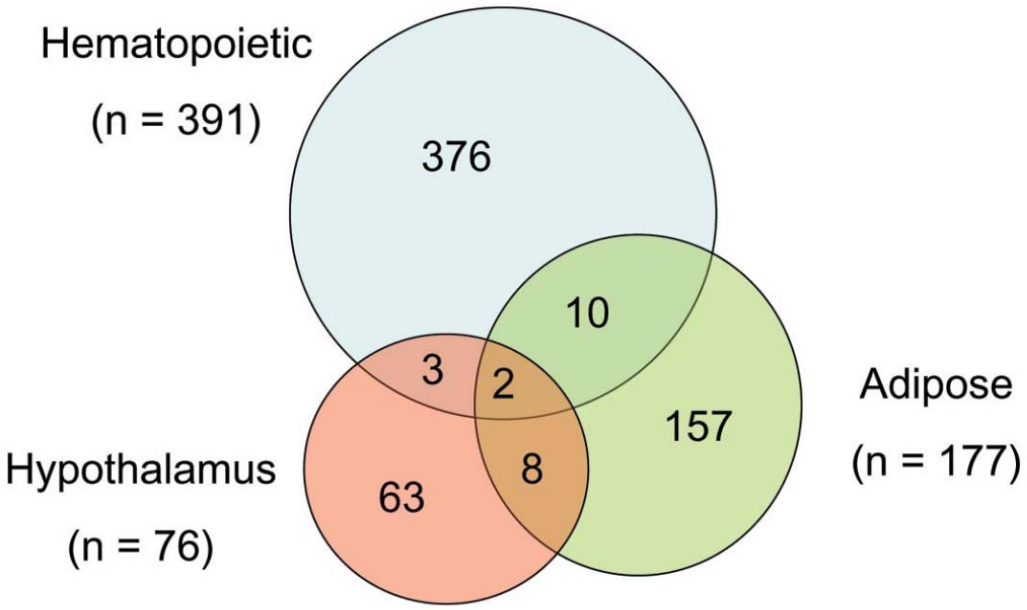

Hematopoietic
(n = 391)
376
10
3
2
157
Adipose
(n = 177)
Hypothalamus
(n = 76)
63
8

Figure 5

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.

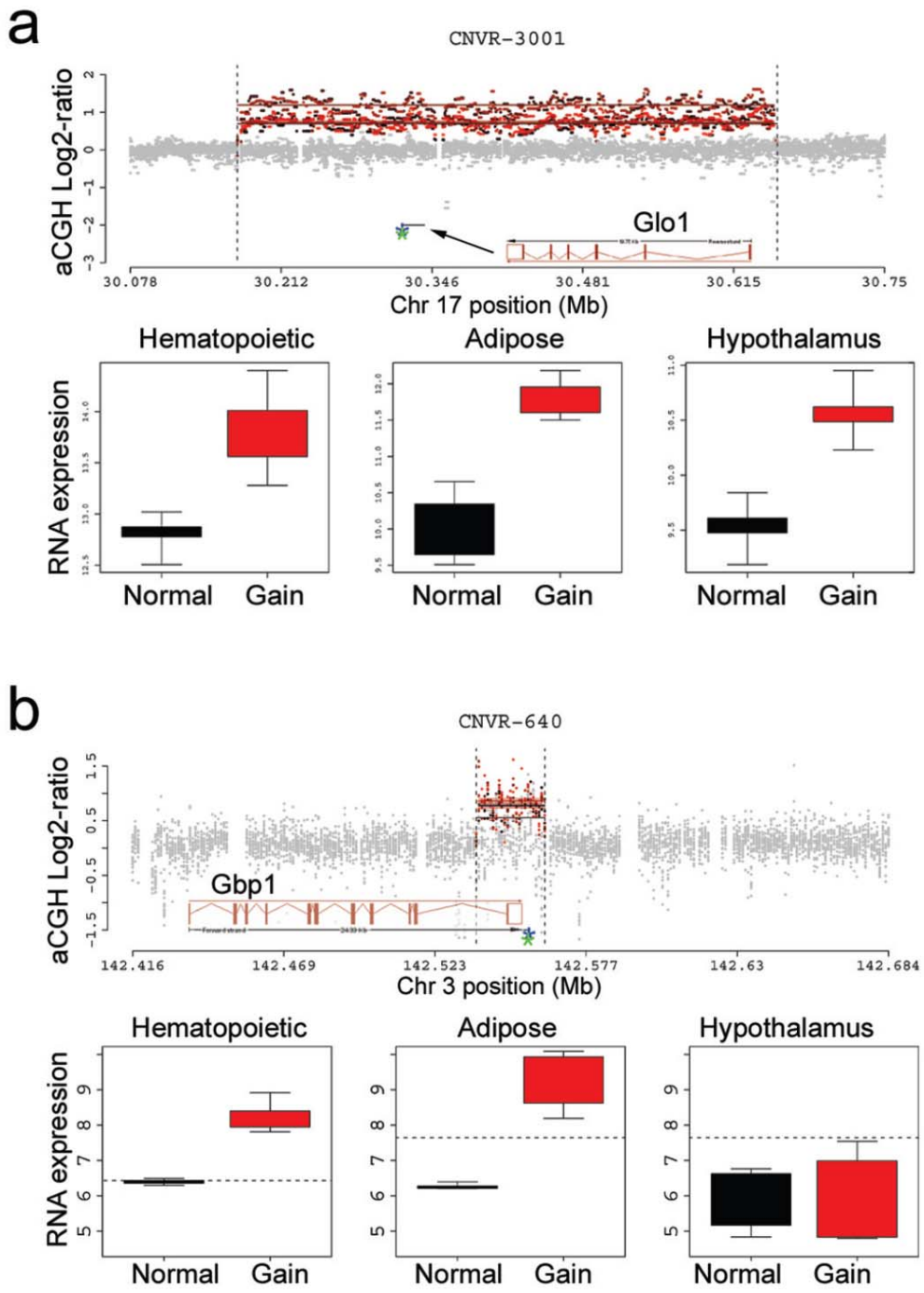
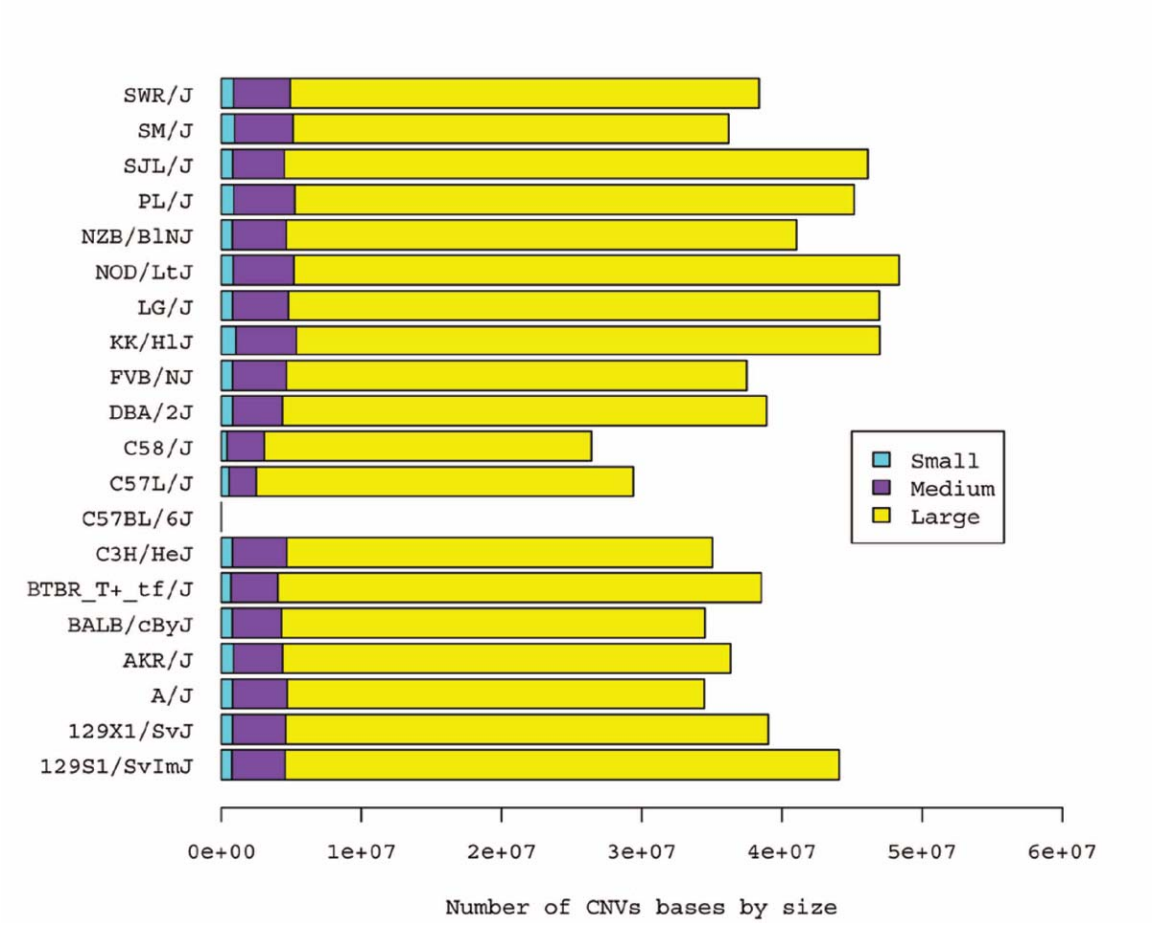

a
CNVR-3001
aCGH Log2-ratio
Glo1
30.078
30.212
30.346
30.481
30.615
30.75
Chr 17 position (Mb)
Hematopoietic
Adipose
Hypothalamus
RNA expression
Normal
Gain
b
CNVR-640
aCGH Log2-ratio
Gbp1
142.416
142.469
142.523
142.577
142.63
142.684
Chr 3 position (Mb)
Hematopoietic
Adipose
Hypothalamus
RNA expression
Normal
Gain

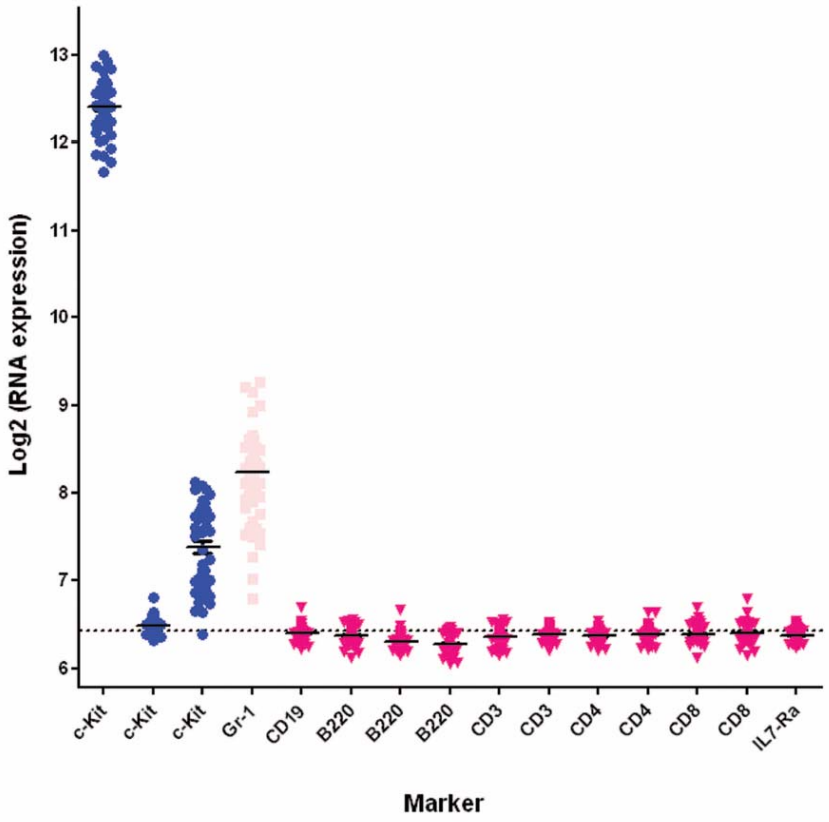
Figure 6

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.

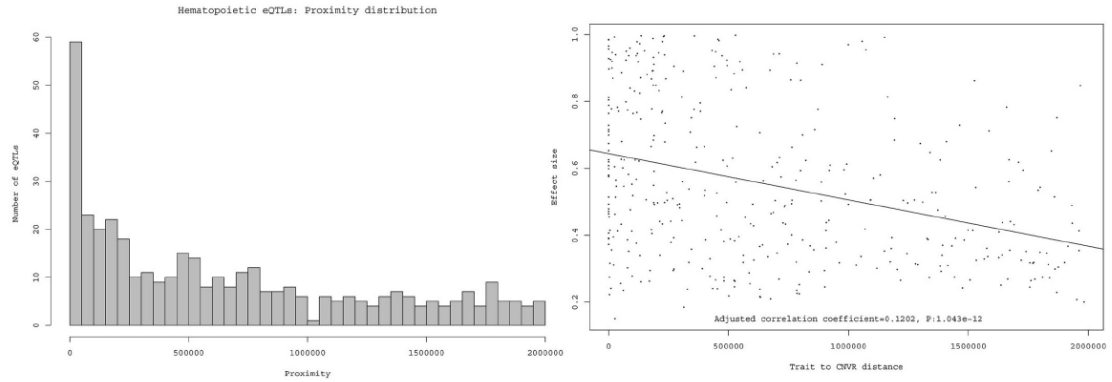
Supplementary Figure 1. Log2 plots for all CNVRs. Log2-ratio plots of the test versus reference (C57BL/6J) aCGH signal intensities for all 1,329 high-confidence CNVRs. All twenty strains are shown in each plot. Horizontal lines represent wuHMM segmentation calls, which are made independently for each 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. Available online at: http://graubertlab.dom.wustl.edu/PDF_Docs/Supplementary%20Figure.pdf

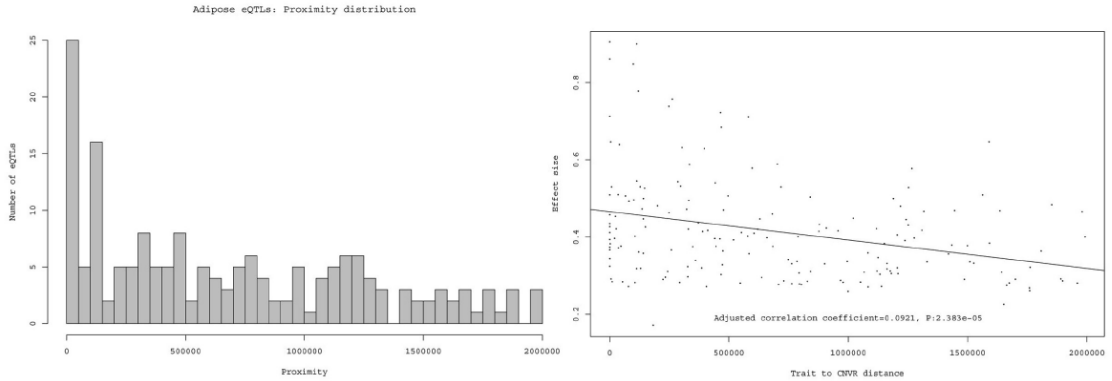

SWR/J
SM/J
SJL/J
PL/J
NZB/BlNJ
NOD/LtJ
LG/J
KK/HlJ
FVB/NJ
DBA/2J
C58/J
C57L/J
C57BL/6J
C3H/HeJ
BTBR_T+_tf/J
BALB/cByJ
AKR/J
A/J
129X1/SvJ
129S1/SvImJ
Small
Medium
Large
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Number of CNVs bases by size

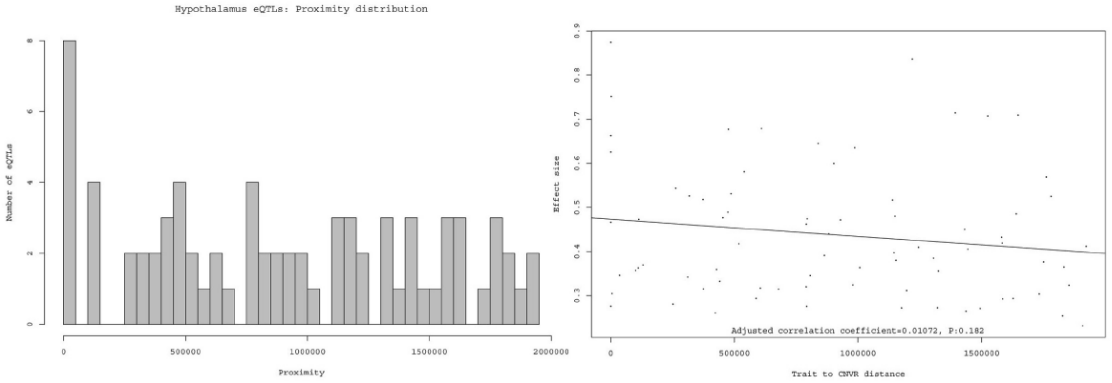
Supplementary Figure 2. Distribution of copy number variable sequence content by CNVR size across strains. CNVs are separated by size. Most CNVs are small or medium in all strains.


Log2 (RNA expression)
13
12
11
10
9
8
7
6
c-Kit
c-Kit
c-Kit
Gr-1
CD19
B220
B220
B220
CD3
CD3
CD4
CD4
CD8
CD8
IL7-Ra
Marker

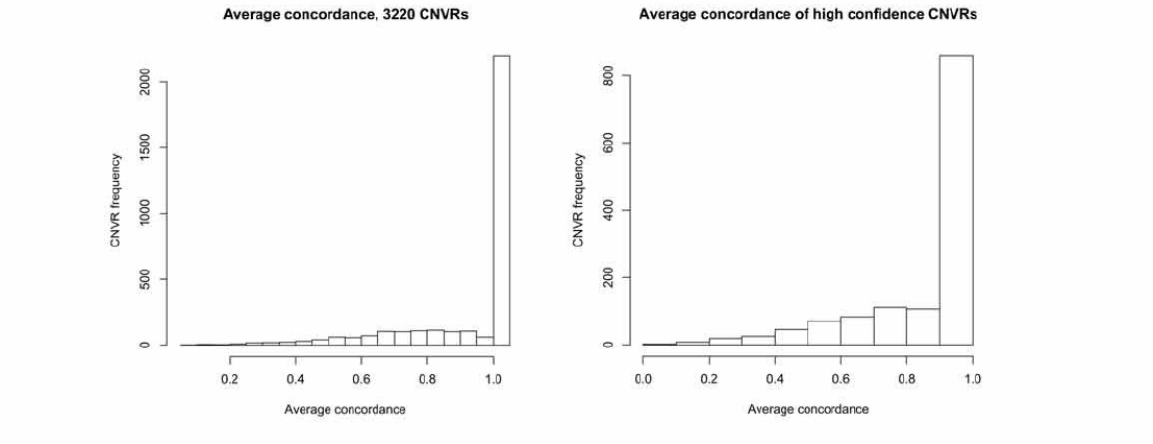
Supplementary Figure 3. Validation of cell sort purity by gene expression profile. The log2 gene expression values of the cell surface markers utilized in the hematopoietic progenitor/stem cell sort strategy from all 48 samples. c-Kit, a primitive hematopoietic marker, is highly expressed in the samples. The lineage makers genes (Gr-1, CD19, B220, CD3, CD4, CD8, and IL-7Rα) are either expressed at low levels (Gr-1) or below or near the level of detection (all CD19, B220, CD3, CD4, CD8, and IL-7Rα). No probe for *Ly76* (Ter119) was on the Illumina expression array.


Hematopoietic eQTLs: Proximity distribution
Number of eQTLs
Proximity
Effect size
Trait to CNVR distance
Adjusted correlation coefficient=0.1202, P:1.043e-12


Adipose eQTLs: Proximity distribution
Number of eQTLs
Proximity
Effect size
Trait to CNVR distance
Adjusted correlation coefficient=0.0921, P:2.383e-05


Hypothalamus eQTLs: Proximity distribution
Number of eQTLs
Proximity
Effect size
Trait to CNVR distance
Adjusted correlation coefficient=0.01072, P:0.182

Supplementary Figure 4. Relationship between CNVR and eQTL by distance and effect size. Top left: Frequency of hematopoietic eQTLs as a function of distance between the expression probe and CNVR. Most eQTLs do not overlap the associated expression probe. Top right: Hematopoietic eQTL effect size (correlation coefficient) as a function of distance between the expression probe and CNVR. The effect size significantly decreases with increasing distance. Middle row: Same as above for adipose tissue. Bottom row: same as above for hypothalamus.


Average concordance, 3220 CNVRs
CNVR frequency
0
500
1000
1500
2000
0.2
0.4
0.6
0.8
1.0
Average concordance
Average concordance of high confidence CNVRs
CNVR frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
Average concordance

Supplementary Figure 5. Concordance between CNVRs and individual CNV boundaries.

Average concordance is the average of wuHMM call lengths divided by the CNVR length and is used to distinguish between 'complex' (average concordance ≤ 0.75) and 'simple' (average concordance > 0.75) CNVRs. Left: average concordance histogram prior to filtering calls based on empirically derived score thresholds. Right: average concordance histogram after applying the score thresholds (>1.5 for gains, >2.5 for losses).

Integrated genomics of susceptibility to therapy-related leukemia

Patrick Cahan and Timothy A. Graubert

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

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
email: graubert@wustl.edu

Running head: Systems genetics of susceptibility to t-AML

INTRODUCTION

Therapy-related acute myeloid leukemia (t-AML) is a secondary malignancy attributable to the chemotherapeutic treatment of an initial disease. Therapy-AML comprise 5-20% of all AML cases and its prevalence is increasing along with the population undergoing chemotherapy[1,2]. While there is evidence that chemotherapy regimen[3] and genetic background[4] contribute to t-AML, little else is known definitively regarding susceptibility. Gaining a better understanding of t-AML susceptibility factors is a pressing concern as it may lead to prevention strategies and provide insight into the genesis of *de novo* AML.

One class of chemotherapeutic associated with t-AML is the alkylators (i.e. melphalan, busulfan, thiotepa). The therapeutic effect of alkylator agents is believed to result from the formation of DNA adducts and single and double-strand breaks, which trigger apoptosis or growth arrest[5]. Based on this presumed mechanism of alkylator action, genes involved in DNA repair[6], response to oxidative stress[7], and drug metabolism[8] have been investigated as mediators of susceptibility in candidate gene studies, with largely inconclusive results. A recent study in our lab investigated the genetic basis of t-AML susceptibility using inbred mice[9]. In this study, eight to twelve individual mice from each of 20 inbred strains were treated with the alkylating agent *N*-nitroso-*N*-ethylurea (ENU), a potent mutagen with a propensity to cause AT:TA transversions and AT:GC transitions[10]. Mice were monitored for the development of MDS and AML for up to 16 months post ENU exposure. Myeloid tumors varied by strain, supporting the hypothesis of a strong genetic component in t-AML susceptibility. Although much has been learned from the combined efforts of candidate gene and genome-wide studies to elucidate the basis of t-AML susceptibility, major contributing factors to t-AML susceptibility have yet to be identified.

We hypothesized that the pre-exposure transcriptional state of hematopoietic stem and progenitor cells, the putative target of leukemogenesis[11], underlies variation in susceptibility to t-AML. A pre-exposure transcriptional basis of susceptibility would be expected if a rapid response is critical in determining a cell's ultimate fate upon mutagen exposure. This hypothesis is consistent with the observations that genes critical to surviving genotoxic stress in yeast are not

differentially expressed upon exposure[12]. A similar situation has been reported in human lymphoblastoid, in which the pre-exposure transcriptional state of the cell more accurately predicts survival than the post-exposure state[13].

In this study, we apply an integrative genomics approach[14] to identify and prioritize genetic and transcriptional networks underlying t-AML susceptibility in mice. By linking expression profiles and complex traits to common genomic loci, this method can ameliorate some of the limitations inherent in genetic association and expression profiling studies[15-19]. When combined with network analysis, this methodology has proven useful in elucidating the molecular networks underlying several complex traits[20,21].

RESULTS

Expression quantitative trait loci in hematopoietic stem and progenitor cells

Previously, we reported the association 408 expression traits (391 genes) with 214 DNA copy number variant regions (CNVRs) in kit+/lineage- (KL) cells[22]. However, this cis-eQTL map does not include genetic variation that is not captured by the CNVRs reported therein (i.e. SNPs). To derive a more complete map of cis-eQTLs in this population, we used publicly available SNP data from 48 classical inbred strains to map SNP-based eQTLs[23]. The SNP resource includes 132,285 SNPs per genome, of which 115,009 we considered informative (as defined in Methods). Prior to eQTL mapping, we used a simple merging algorithm to iteratively join adjacent SNPs into haplotype blocks. This algorithm results in haplotype blocks in which the genotypes of a complete set of SNPs is predictable to a given level of accuracy. We selected a threshold such that for a given block, we can accurately predict the genotype of every SNP in all 48 strains with at most one error (Figure 1A). The 23,884 resulting haplotype blocks are comprised of 1 to 62 SNPs (mean=4.82, median=4) (Figure 1B). There are 2 to 6 haplotypes per block (mean=3.92, median=4) (Figure 1C). 9,324 blocks have one error, and the remaining 14,560 have zero. Of the 7,405 haplotype blocks within 250 Kb of a CNVR boundary, only 39 have genotypes that perfectly correlate with CNVR genotypes (Figure 1D). We speculated that the low correlation is due to the fact that using all 48 classical inbred strains in the haplotype block construction resulted in higher numbers of haplotype labels. Therefore, we also derived a haplotype block map using only the 20 strains from the CNVR study. However, this analysis resulted in highly similar results in terms of the map and haplotype correlation with CNVR genotypes (data not shown). This suggests in current data sources, CNVRs and SNPs represent distinct sources of genetic variation in the mouse genome. We used the 48-strain haplotype resource to map KL expression traits to SNP-based haplotypes, as previously described[22]. We considered only cis-eQTL-associated genes, as it has been shown that trans-eQTLs contain a large proportion of false positives[24]. We found 127 associations between expression traits and haplotypes, after selecting the most significant association per trait.

Global pre-exposure transcriptional state of hematopoietic stem and progenitor cells is associated with t-AML susceptibility

We performed gene expression profiling on 20 inbred strains listed as Priority 1-4 from the Mouse Phenome Database [25]. Two-to-three biological replicates arrays were analyzed per strain. This GEP data was previously published[22]. We excluded wild-derived strains from this analysis because the extent of genetic differences makes difficult the interpretation of aCGH, GEP, and eQTL mapping analysis. Fifteen of the strains were previously assayed for susceptibility to t-AML after exposure to ENU[9]. Unsupervised clustering of gene expression profiles largely separated susceptible from resistant strains (Figure 2A). The probability that the unsupervised clustering of expression profiles would reflect susceptibility status to the extent observed is < 0.01 (10,000 permutations, see Methods and Supplementary Figure 1A). Further, this clustering is not observed in other tissues that are highly unlikely to be involved in leukemogenesis, the hypothalamus and adipose tissue, nor does it reflect SNP-based strain distances (Supplementary Figure 1B-D). Taken together, this supports the notion that the KL clustering of susceptible strains is not due to sequence polymorphisms effecting target hybridization[26] but rather reflects tissue-specific differences in transcript abundance between inbred strains[27]. Additionally, this observation suggests that the pre-exposure expression differences of many genes, rather than only a few, segregate the KL cells of susceptible versus resistant strains.

Next, we sought those genes that are differentially expressed between susceptible and resistant strains in KL cells. We identified 917 differentially expressed genes (976 probes) at an FDR threshold of 5% (Supplementary Figure 2 and Supplementary Table 1). The differentially expressed genes are enriched in several GO-annotated biological processes (Table 1), including the GO terms ‘apoptotic program’ and ‘nucleotide metabolic process’. The Kegg pathways ‘Pyrimidine metabolism’ and Colorectal cancer’ were also enriched. ‘Acute myeloid leukemia’, ‘Apoptosis’, and ‘p53 signaling’ are biologically plausible pathways that were enriched at least two-fold in the differentially expressed genes, however none of these pathways passed the FDR

< 25% threshold. GO-apoptosis-annotated genes included both cell-intrinsic and extrinsic factors (Figure 2B).

Integrated cis-eQTL mapping identifies candidate drivers of susceptibility

There are 45 candidate driver genes (45 probes) that are both differentially expressed and linked to at least one eQTL. We refer to these genes as anchors throughout the text. 37 are linked to CNVR-eQTLs; the remaining 8 are linked to haplotype-eQTLs. To validate the cis-eQTL associations, we mined publicly available expression data representing hematopoietic stem, progenitor, erythroid and myeloid populations from the BXD recombinant inbred panel[28]. Because this data was generated using the same GEP platform as our KL data, we were able to ask how our kit/lineage population is related to these more purified populations (Supplementary Figure 3). As expected, our KL expression profiles cluster most closely with progenitor profiles and are distinct from both erythroid and myeloid lineages. For each candidate driver, we tested the association between BXD genotypes of SNPs within 2 Mb and driver expression. We found that 30 of the 45 drivers were significantly associated with at least one SNP within 2 Mb in at least one of the hematopoietic compartments (26 in either Stem or Progenitor), supporting the hypothesis that expression differences of the anchor are caused by locally encoded genetic [29,30]variation. Out of the total of 480 testable eQTLs-transcript associations, 300 (62.5%) were replicated in at least one of the hematopoietic data sets.

Anchored network analysis identifies t-AML susceptibility expression modules

Next, we hypothesized that expression differences of anchors would have multiple, downstream transcriptional effects. For each anchor, we identified correlated expression profiles (FDR < 1%) , resulting in 30 sets of co-expressed genes or modules. The number of targets per module ranged from 3 to 607 (mean=113, median=72). We reasoned that true response genes will exhibit association with driver expression even when the remaining genome is randomly shuffled, as is true in the BXD recombinant inbred cross. For each module, we tested the association between expression of the anchor and each response transcript in each of the BXD

hematopoietic populations. We removed target genes from modules that were not associated with driver expression in at least one compartment (FDR < 25%) (Table 2). We also used a previously described co-expression network algorithm to derive modules of correlated genes[29,30] independent of linkage to eQTLs. We filtered these modules on the basis of their reproducibility in the GdH datasets and compared the resulting modules with the anchored expression networks. The WGCNA modules are highly similar to the anchored modules in gene content, suggesting that the discovered co-expression structure is robust to different algorithms (data not shown).

The expression of each anchor gene is, by definition, associated with susceptibility status. However, the strength of the association between the target genes of an anchored module and susceptibility is unknown. To determine these values, we first computed eigengenes from each module[30]. Then, we ranked anchored modules according to differential expression of the module's eigengene and susceptibility status. Using both KEGG and GO annotations, we found that 8 anchored modules were enriched in at least one annotation. We visualized the anchored susceptibility modules as networks (Figure 3A), displaying the correlation between anchored modules and the strength of association between anchored modules and susceptibility status. We also visualized a subset of the anchored susceptibility network, focusing on biologically compelling modules (Figure 3B and 3D).

DISCUSSION

There is accumulating evidence that many genetic contributors to complex traits are not protein-coding changes[31]. If true, then the only other class of genetic events that can effect phenotype must, at some level, impact expression (i.e. eQTLs). Hypothesizing that such events contribute to t-AML susceptibility, we took an integrative genomics approach to identify and prioritize candidate genetic and transcriptional networks. The first step in this approach was to identify eQTLs in hematopoietic stem and progenitor cells, the likely target of leukemic transformation. Previously, we described a CNVR-eQTL map in classical inbred mice. In the current work, we expanded this map to include SNP-based haplotype-eQTLs. In deriving the mouse haplotype map, we found surprisingly little correlation between haplotypes and neighboring CNVRs. This is in contrast to human studies, where nearly 75% of common CNVRs are estimated to be in linkage disequilibrium with neighboring SNPs[32]. This suggests that at the currently available resolution and coverage (and genotyping accuracy), mouse haplotypes and CNVRs represent distinct sources of genetic information. We found two-fold more CNVR-eQTLs than haplotype-based eQTLs (401 vs. 167). It is tempting to speculate that this difference in eQTL types is because CNVRs have a stronger impact on expression *in cis* and therefore are more likely to be detected as eQTLs. However, the difference could largely be due to the reduced power to detect haplotype-eQTLs because of the exacerbated multiple testing problem that comes with performing approximately 20 times more statistical tests. Greater than 60% of the eQTLs were reproducible in an independent dataset. This is a conservative estimate of the true validation rate because only genetic differences between C57BL/6J and DBA/2J are present in the validation data.

The second step in the integrative approach was to find genes differentially expressed between susceptible and resistant strains. Because unsupervised clustering of all expressed transcripts grouped strains largely by susceptibility status, we expected to find a large number of genes associated with susceptibility status. Greater than 7% of the expressed transcripts are differentially expressed (976/13,496). These genes are enriched in several, independent

biological processes, most notably apoptosis. Among the intrinsic apoptosis genes are Caspase 9 (Casp9), B-cell leukemia/lymphoma 2 (Bcl-2), BCL2-associated agonist of cell death (Bad), BCL2-associated X protein (Bax), and mutS homolog 6 (Msh6). Msh6 is a member of the mutSα DNA mismatch recognition complex that has been shown mediate apoptosis in certain contexts[33,34]. Notably, the absence of mutSα activity in myeloid progenitors results in the complete loss of *O*6-methylguanine (*O*6MeG)-mediated cytotoxicity[35]. That resistant strains have higher expression of Msh6 suggests that upon alkylator exposure, resistant strains may recognize DNA damage and respond appropriately (i.e. die) whereas the KL cells of susceptible strains may tend to live, accumulate mutations, and transform. In KL cells, almost all susceptible strains have no detectable expression of Casp9, a critical initiator of programmed cell death, suggesting that these cells (low-to-no Casp9 expression) are less primed for Casp9-dependent apoptosis. However, susceptible strains had decreased expression of Bcl-2, an anti-apoptotic gene, and increased expression of Bad and Bax, both pro-apoptotic genes. This would suggest that the KL compartment of susceptible strains is more 'primed' for cell death, consistent with the observation that the SWR/J allele of Bcl2 confers increased survival in an F2 cross model of t-AML[36]. However, this pattern of expression is contrary to the prediction based on Casp9 expression, possibly indicating a regulatory feedback loop to compensate for the apparent absence of Casp9 in susceptible strains. Taken together, these results illustrate the complexity in assessing the relative functional activity of a cell population (i.e. readiness to commit to apoptosis) given a snapshot of the population's static transcriptional state. Experiments to test variation in alkylator-induced apoptosis will help to resolve this apparent paradox.

Differential expression and gene enrichment analysis highlighted several biologically plausible pathways that may underlie t-AML susceptibility. However, it remained unclear which pathway members, if any, are causal contributors to the phenotype, as illustrated by the complex expression patterns of the intrinsic apoptosis genes. More broadly, the role and relative importance of each of the 917 differentially expressed genes in susceptibility remained undetermined. We hypothesized that important transcriptional regulators of susceptibility affect the expression of multiple downstream genes. Therefore, as the third step in the integrated

genomics approach, we identified networks of genes that are significantly correlated with candidate susceptibility drivers. Drivers are those genes that are both differentially expressed and associated with eQTLs. We trimmed the networks of response genes whose expression was not reproduced in independent data sets.

One of the benefits of the integrative genomics approach is that it can implicate biological processes that would not have been detected using differential expression alone. Susceptibility networks are enriched genes involved in DNA repair, base excision repair, apoptosis, and cell cycle, among other annotations. A second benefit of the integrated approach is that it differentiates between upstream (drivers) and response genes. This is proving useful in prioritizing apoptosis-related genes for experimental validation. Although Casp9 and Bcl2 are differentially expressed, Casp9 is also the candidate driver of module A_33, the module most strongly associated with susceptibility status. We speculate that perturbation of candidate drivers, such as Casp9, are more likely to be informative in elucidating susceptibility than response genes (i.e. Bcl2).

Network analysis allowed us to predict the function of uncharacterized genes. For example, A630001G21Rik is expressed primarily in primitive hematopoietic and B-cells (ref GNF), yet its function is undetermined. Our analysis places it as the driver of module A_12, which is enriched in apoptosis-related genes including Bcl2. Therefore, A630001G21Rik may play previously unknown role in regulation of Bcl2 expression and apoptosis activity. Similarly, Cytoskeleton-associated protein-like 2 (Ckap2l) is the driver of the largest module, A_16, enriched in both cell cycle and DNA repair genes (Figure 3B). Although Ckap2l is highly expressed in hematopoietic progenitors[37], its functions are unknown (GNF). Its closest ortholog, Ckap2, is highly expressed in mouse stem cell lines and has detectable expression in hematopoietic progenitors, bone marrow, osteoclasts, osteoblasts, and macrophages[37]. There is a growing body of literature suggesting that Ckap2 (also known as Tumor-associated microtubule-associated protein) is involved in cell cycle progression[38-40]. It has long been recognized that disruptions to normal cell cycle parameters can impact cancer susceptibility[41]. It is possible that Ckap2l contributes to cell cycle regulation in HSCs and progenitors, and that

genetic disturbances to its expression alter t-AML susceptibility. Experiments that perturb expression of driver genes such as Casp9 and Ckap2l to assess their impact on module expression and activity are the next logical steps in determining the role of candidate networks in susceptibility. If such experiments demonstrate a causal link between driver genes and module expression, then moving forward to more definitive transplantation experiments will be warranted.

A drawback to the anchored network approach, as currently implemented, is that it assumes there is only a single anchor per module. In cases where CNVRs disrupt local chromosome structure, its is possible that a single genetic event impacts the expression of multiple neighboring genes (Figure 3C). In module A_37, we found that 10 response genes are located with 7 Mb of this CNVR (Figure 3D). This module warrants special attention because it includes poly (ADP-ribose) polymerase family member 2 (Parp2, the anchor) and apurinic/apyrimidinic endonuclease 1 (Apex1), both members of the base excision repair pathway[42,43]. Both genes have lower expression in susceptible strains, again suggesting that lowered overall DNA damage response promotes susceptibility.

A caveat to the current work is that maps of genetic variation in the mouse genome are incomplete. It is possible that un-captured genetic variants may be the ultimate cause of the observed co-expression networks. And they may mediate their impact through mechanisms other than altering the expression of drivers. In the extreme case, all modules may not be controlled by driver expression, but by undetected causes. Nevertheless, the modules themselves are still informative in that they describe sets of coordinately regulated genes that, collectively, are associated with both susceptibility and biologically plausible processes and pathways.

To our knowledge, this is the first report of an integrative genomics approach to dissect the role of the pre-exposure transcriptional state in t-AML susceptibility. From a clinical perspective, t-AML are important because they are generally incurable and median survival time from diagnosis is eight months[3]. But because t-AML are clinically induced malignancies, they are by definition preventable. Therefore, a long-term goal of t-AML research is to gain sufficient understanding of susceptibility factors in order to make worthwhile the personalization of

chemotherapeutic regiments based on t-AML risk. The transcriptional networks and their candidate drivers described here are an important early step towards gaining such an understanding.

METHODS

Genomic coordinates of 1,333 CNVRs were mapped from mm8 to mm9 using *liftOver*. 31 CNVRs were unmapped and dropped from further analysis. To derive haplotype blocks, SNPs for the haplotype map construction were downloaded from Broad Institute[23]. Only SNPs from 48 non-wild-derived strains were used. SNPs that were contained within CNVRs, had minor allele frequencies < 5%, or were not genotyped in 25% or more of strains were considered to be uninformative and were excluded from further analysis. The following steps were performed to simultaneously group SNPs into blocks and to assign haplotype to strains:

(1) Begin with the first informative SNP on a chromosome.

(2) If the number of SNPs in the current block is 1 then go to (3). Otherwise, go to (4).

(3) Group strains by genotype and add the next consecutive SNP to the current block.

(4) Cluster strains by SNP-based distance using PAM (number of clusters = 2 to 6).

(5) Assign haplotype labels to strains based in the clustering with the maximum average silhouette.

(6) Derive consensus haplotypes. For each haplotype cluster, a consensus haplotype is defined as the string comprised of the most frequent genotype at each SNP position.

(7) Compare the consensus haplotypes to the actual SNP genotypes.

(8) If the number of errors is greater than 1 then go to (9), otherwise go to (10).

(9) Remove the most recently added SNP from the current block. Store the haplotyping results from the previous iteration. Start a new block with the current SNP. Go to (3).

(10) Add the next consecutive SNP to the current block. Go to (4). If there are no more SNPs on the current chromosome, select a new chromosome and go to (2). The computation is complete when all chromosomes have been analyzed.

SNP-based distances between strains are computed as the sum of SNP differences between strains. The range of number of haplotypes per block to allow was selected based on the estimated number of ancestral haplotypes[44].

GEP expression profiling was previously described[22] and is available at GEO under accession GSE10656. This data is referred to as kit+/lineage- (KL) throughout the text. Hypothalamus and adipose were obtained from GEO (accessions GSE5961 and GSE8028, respectively). For clustering and network analysis, probes were first filtered based on detection. In the KL data, a probe was considered detected in a sample if its signal was greater than a set of negative controls on the Illumina array. 13,496 probes were detected in all biological replicates of at least three strains (excluding C3H, for which only one array was analyzed). Only the 14,871 and 10,040 probes that were detected as present in at least 25% of the strains in the hypothalamus and adipose data sets, respectively, were kept for clustering analysis. Unsupervised hierarchical clustering was performed with R's *hclust* function, using 1-Pearson correlation as the distance metric and the complete linkage method for node merging. To assess the non-randomness of the strains clustering according to susceptibility status, we computed the ratio of the mean distances of among susceptible strains to the mean of the distances between all susceptible and resistant strains. Then, we permuted the strain labels 10,000 times, and recomputed the ratio of distances. The P-value of the observed clustering is the number of random permutations in which the distance test statistic >= observed distance test statistic divided by 10,000. This analysis was performed on the median expression profiles of strain replicates, only in those strains in which the susceptibility status is known. SNP clustering was based on strain-strain pair-wise distances computed by counting the number of SNPs that differ between each the strains divided by the total number of SNPs that are typed in both strains.

Strains with unknown susceptibility status were not included in the differential expression analysis. We used the *limma* package in R to model the expression of each gene with coefficients representing strain replicates and susceptibility status[45,46] and the false discovery rate (FDR) was estimated using q-value[47]. All of the 976 significant probes were detected as present in at least 50% of either the susceptible or resistant strains. Association of module eigengenes

with susceptibility was tested in the same way as differential expression. Enrichment analysis was performed using DAVID[48]. Only the GO annotations Biological Process 5 and KEGG pathways were assessed. We only report annotations that pass an FDR threshold < 25%. Expression data from all 20 strains previously profiled were used in expression network analysis. Anchored expression networks were identified by searching for probes that exhibited expression profiles that were significantly correlated with driver gene expression at an FDR threshold < 1%.

Normalized gene expression data used for validation of eQTLs and anchored modules was downloaded from GEO (GSE18067). This data set includes profiling on sorted (purified) hematopoietic stem, progenitor, myeloid and erythroid populations from female BXD recombinant inbred mice[28]. This data is referred to as Gerald de Haan (GdH) throughout the text. Only detection calls, coded as 0 for absent or 1 for present, were used to globally compare our KL data to GdH. Clustering was performed using the same parameters as described above for the KL data. KL eQTLs were validated by testing the association between the genotypes of SNPs within 2 Mb of driver genes and driver gene expression in each compartment separately. Genotypes were treated as factors in a linear model of driver gene expression. P-values of the resulting F-statistics were adjusted for multiple testing using Holm's method[49]. Drivers that had corrected P-values < 0.05 in at least one compartment were considered validated. Assessing the reproducibility of the association between driver and response gene expression was performed in a similar manner. A linear model of response gene expression was fit with driver gene expression as the dependent variable (one model per driver-response gene pair per compartment). In this case, Benjamini and Hochberg's method to control the false discovery rate was applied to the resulting p-values[50]. WGCNA analysis was performed as previously described using the R package *WGCNA*[30]. Briefly, β values for calculating the weighted network adjacency were selected based on the power at which the scale law R^2 exceeded 0.9. Weighted adjacency matrices were computed, modules were defined using the *cutTreeDynamic* function (which selects good dendrogram cutoffs) and similar modules were merged using *mergeCloseModules* (which compensates for the high sensitivity of WGCNA). Eigengenes were computed as the first

principal component of a module's expression matrix. Eigengenes were tested for differential expression between susceptible and resistant as described above for individual genes.

Table 1: Functional Enrichment of Differentially Expressed Genes

Annotation	Annotation name	Count	P-Value (nominal)	Fold Enrichment	FDR (%)
GO:0008632	apoptotic program	11	9.58E-05	4.71	0.17
GO:0006464	protein modification process	88	1.77E-04	1.46	0.31
GO:0019318	hexose metabolic process	15	0.00107053	2.76	1.88
GO:0005996	monosaccharide metabolic process	15	0.00129867	2.71	2.28
GO:0046907	intracellular transport	43	0.00182315	1.63	3.18
mmu00240	Pyrimidine metabolism	11	0.00325706	2.98	3.99
GO:0031324	negative regulation of cellular metabolic process	25	0.00230736	1.95	4.01
GO:0009117	nucleotide metabolic process	18	0.00246768	2.27	4.29
GO:0009142	nucleoside triphosphate biosynthetic process	9	0.00286827	3.68	4.96
GO:0045934	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	22	0.00335044	2.00	5.78
GO:0006915	apoptosis	39	0.00643571	1.56	10.81
mmu05210	Colorectal cancer	10	0.00953927	2.74	11.28
GO:0008637	apoptotic mitochondrial changes	5	0.00751505	6.23	12.52
GO:0006396	RNA processing	25	0.00852521	1.76	14.08
GO:0009064	glutamine family amino acid metabolic process	6	0.00927287	4.57	15.22
GO:0015031	protein transport	39	0.01063622	1.51	17.27
GO:0019362	pyridine nucleotide metabolic process	5	0.01370052	5.27	21.69
GO:0008219	cell death	39	0.01444197	1.48	22.73
GO:0016481	negative regulation of transcription	19	0.01461908	1.85	22.98

Count: Number of differentially expressed genes with given annotation.

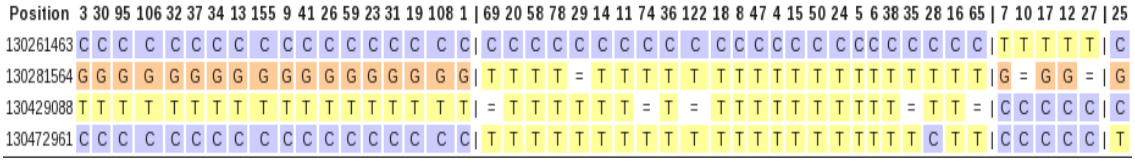
Table 2: Anchored Susceptibility Modules

Module	Anchor Gene	KL	HSC	Progenitor	Myeloid	Erythroid	Replicated	Association with Suscetibility	Top GO	Top Kegg
A_1	LOC634046	460	44	3	24	200	237	2.25	secretion by cell	
A_2	scl41743.2_361	402	141	273	208	20	330	1.09	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	Ubiquitin mediated proteolysis
A_3	GI_38089999	38	5	2	1	3	8	1.72		
A_4	A330106M24Rik	4	4	4	4	2	5	1.82		
A_6	Ociad2	132	64	4	17	11	76	1.5	regulation of phosphoprotein phosphatase activity	
A_7	GI_46852192-I	97	2	16	2	51	55	1.95	interphase of mitotic cell cycle	
A_9	Zfp862	106	14	20	2	1	27	2.3		
A_12	A630001G21Rik	112	80	51	39	49	105	2.38	regulation of transcription	
A_14	Aste1	102	8	49	1	2	53	2.67		
A_16	Ckap2l	607	5	117	290	5	357	1.28	DNA repair	Phosphatidylinositol signaling system
A_17	H2-Ke6	238	13	110	3	82	153	1.93		
A_20	Dusp16	91	46	5	44	31	73	2.08		
A_21	scl0217069.13_16	58	4	5	22	5	30	2.72		
A_22	Atf7ip	39	2	2	10	11	20	2.61		
A_23	Snrpn	4	3	3	1	2	4	0.94		
A_24	Atp6v0e2	78	5	6	2	1	6	2.91		
A_25	Gimap7	30	4	3	1	7	10	2.35		
A_26	Pdzk1ip1	79	40	32	1	21	58	2.35		
A_27	Polr1b	27	4	11	5	3	13	3.11		
A_28	Magohb	65	56	36	25	48	62	0.97	cellular lipid catabolic process	
A_30	Sox13	34	19	26	4	2	30	1.45	fatty acid metabolic process	
A_32	Ptcd3	18	8	15	6	5	18	2.68		
A_33	Casp9	37	2	1	7	2	7	3.22		
A_34	Ctsf	223	54	124	9	72	170	2.34		
A_36	scl46617.10.1_4	13	5	4	9	6	11	2.5		
A_37	Parp2	88	22	22	18	20	43	2.05		
A_38	Hdhd3	178	73	3	71	2	103	1.78		
A_39	5830417I10Rik	5	2	4	2	2	4	2.07		
A_41	Prcp	3	3	2	3	3	4	2.11		
A_43	Ggcx	7	7	8	6	4	8	2.53		

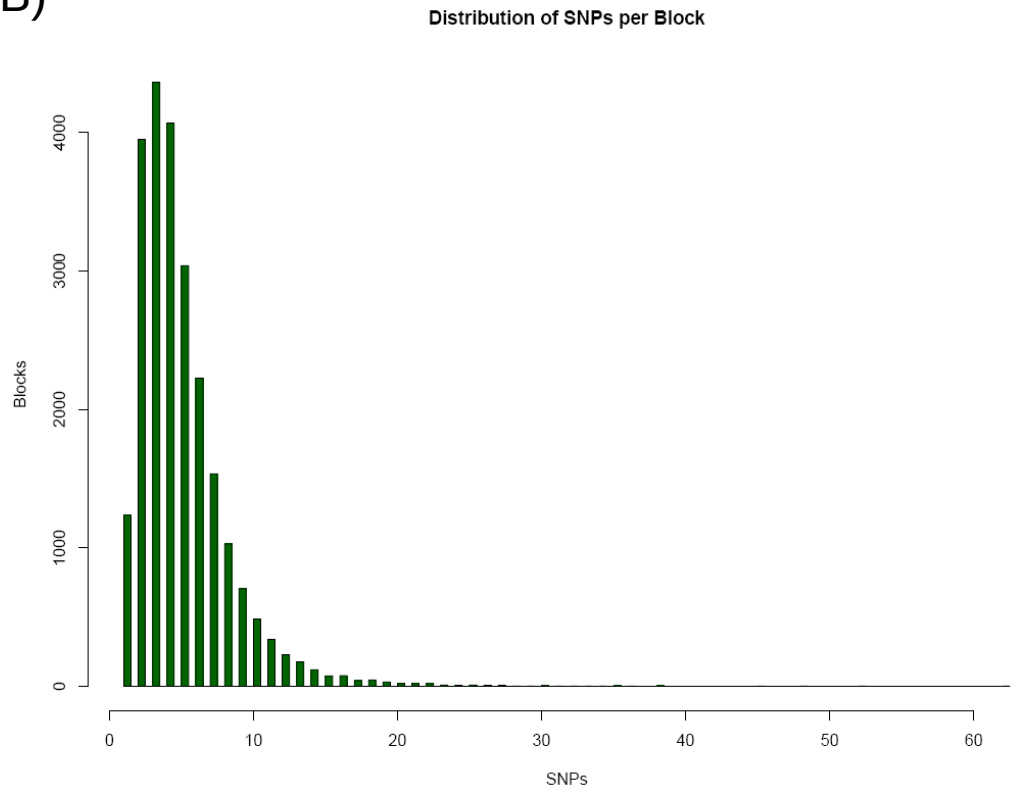
KL: Number of Illumina expression probes significantly assoicated with anchored gene expression in kit+/lineage- (KL) cells
HSC: Number of Illumina expression probes in preliminary anchored module significantly assoicated with anchored gene expression in GdH Sca+/kit+/lineage- (HSC) cells
Progenitor: Number of Illumina expression probes in preliminary anchored module significantly assoicated with anchored gene expression in Sca-/kit+/lineage- (Progenitor) cells
Myeloid: Number of Illumina expression probes in preliminary achored module significantly assoicated with anchored gene expression in Gr-1+ (Myeloid) cells
Erythroid: Number of Illumina expression probes in preliminary achored module significantly assoicated with anchored gene expression in TER-119+ (Erythroid) cells
Replicated: Number of Illumina expression probes in preliminary achored module significantly assoicated with anchored gene expression in at least one GdH data set.
Association with Suscetibility: -Log10(P-value)

FIGURES

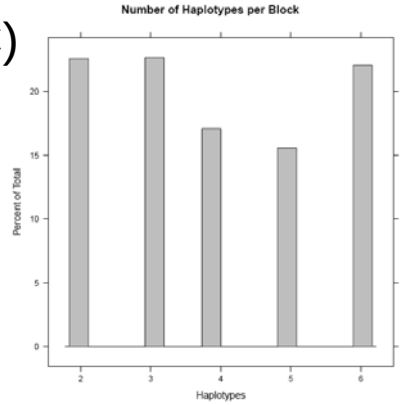
(A)



(B)



(C)



(D)

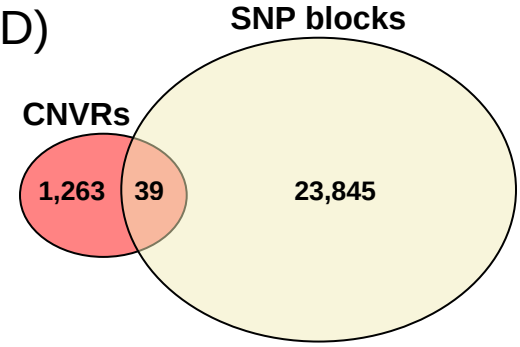
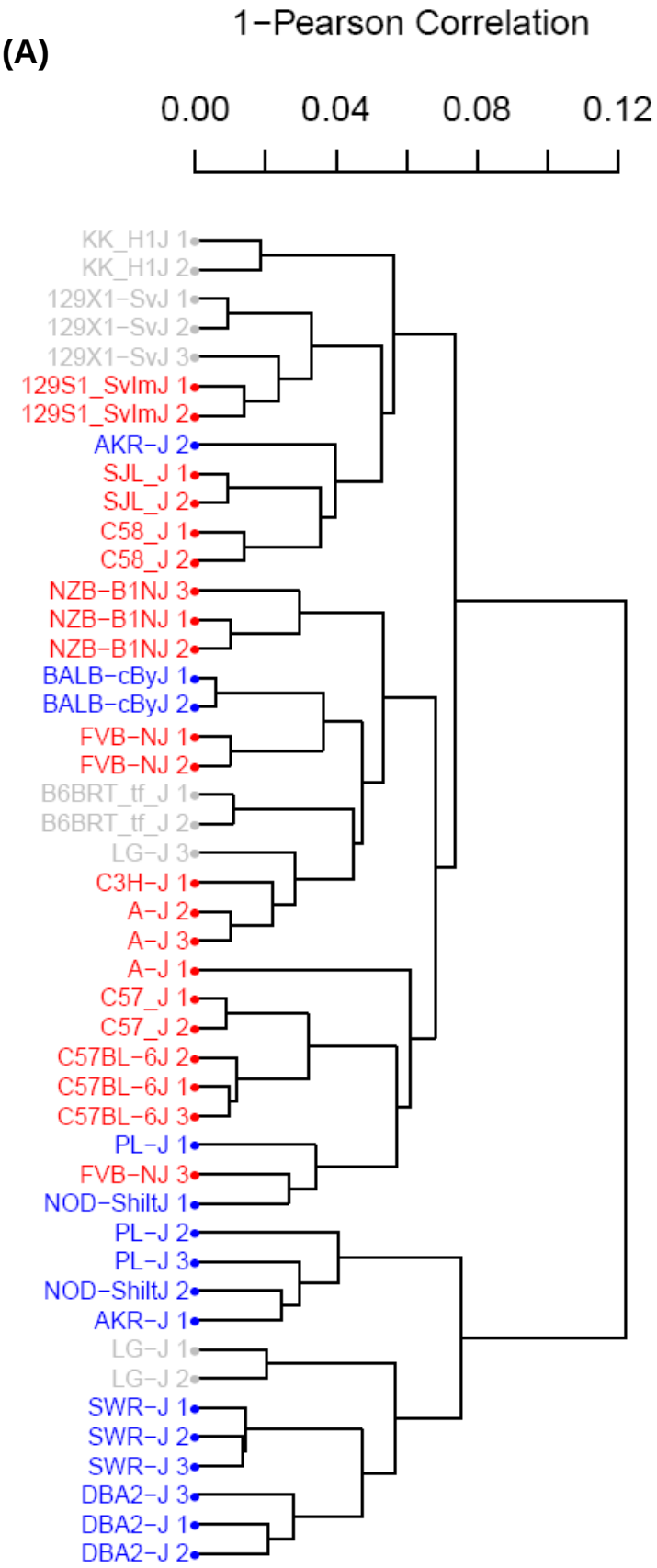
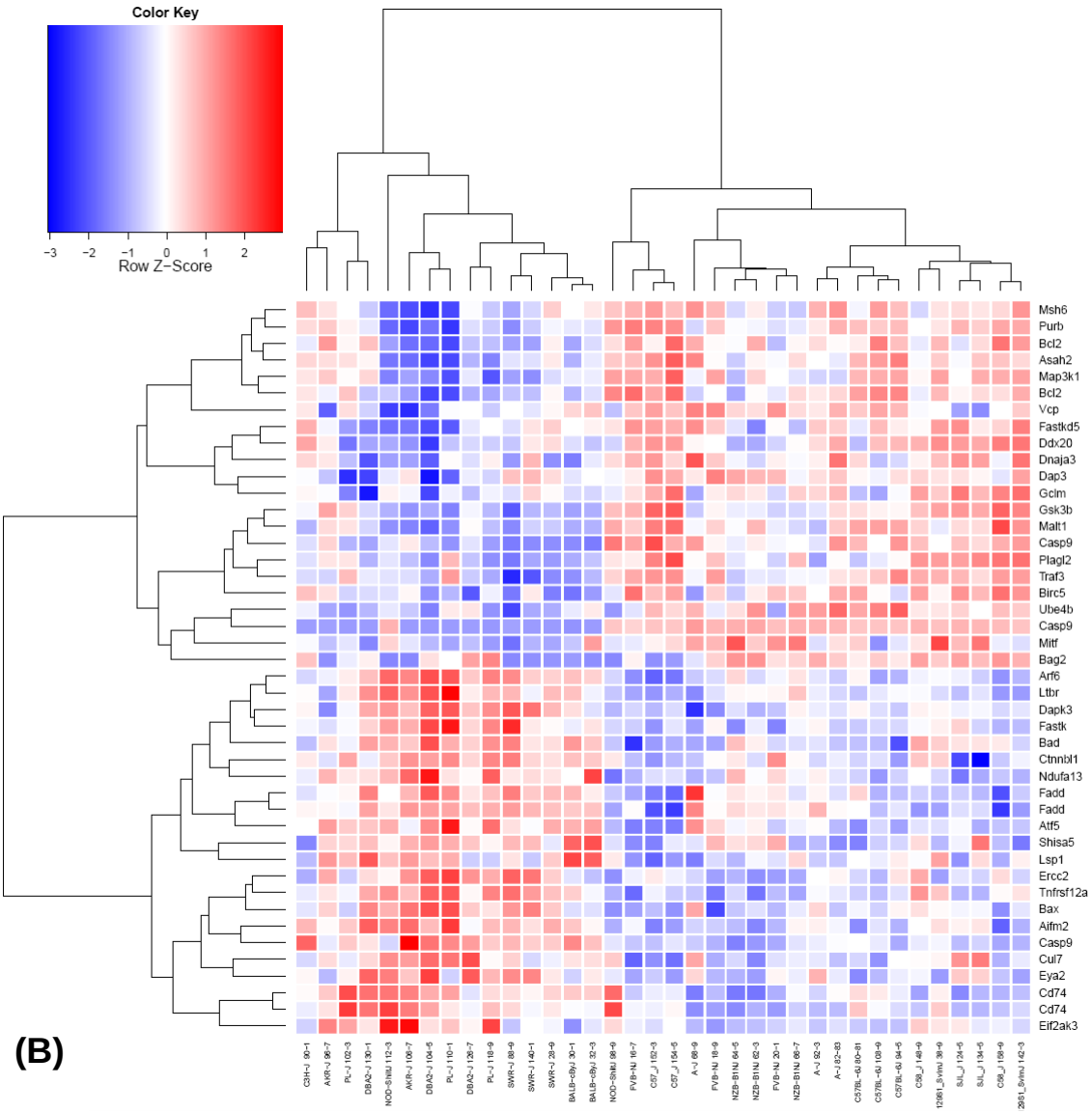


Figure 1: Mouse Haplotype Map. (A) Typical haplotype block derived from Broad SNP data. Rows represent SNPs, '=' are untyped. Columns represent 48 classical inbred strains. Strains sharing the same haplotype are grouped together and are separated from strains of other haplotypes by '|'. Given the strain haplotypes, it is possible to predict the all typed genotypes with at most a single error. The distribution of the number of SNPs **(B)** and haplotypes **(C)** per block. The number of CNVRs that are accurately predicted by neighboring haplotype blocks is relatively low **(D)**.

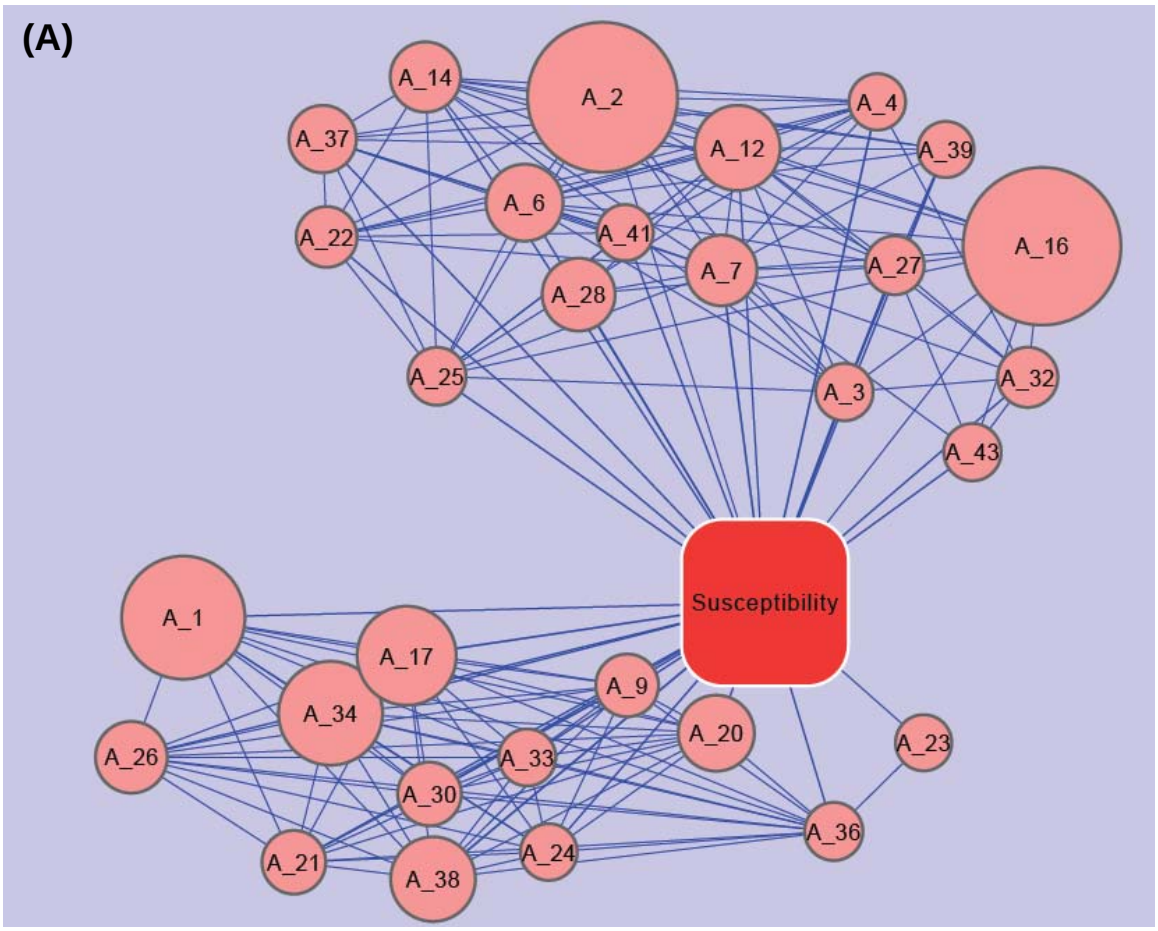
(A)


1−Pearson Correlation
0.00
0.04
0.08
0.12
KK_H1J 1
KK_H1J 2
129X1−SvJ 1
129X1−SvJ 2
129X1−SvJ 3
129S1_SvImJ 1
129S1_SvImJ 2
AKR−J 2
SJL_J 1
SJL_J 2
C58_J 1
C58_J 2
NZB−B1NJ 3
NZB−B1NJ 1
NZB−B1NJ 2
BALB−cByJ 1
BALB−cByJ 2
FVB−NJ 1
FVB−NJ 2
B6BRT_tf_J 1
B6BRT_tf_J 2
LG−J 3
C3H−J 1
A−J 2
A−J 3
A−J 1
C57_J 1
C57_J 2
C57BL−6J 2
C57BL−6J 1
C57BL−6J 3
PL−J 1
FVB−NJ 3
NOD−ShiltJ 1
PL−J 2
PL−J 3
NOD−ShiltJ 2
AKR−J 1
LG−J 1
LG−J 2
SWR−J 1
SWR−J 2
SWR−J 3
DBA2−J 3
DBA2−J 1
DBA2−J 2

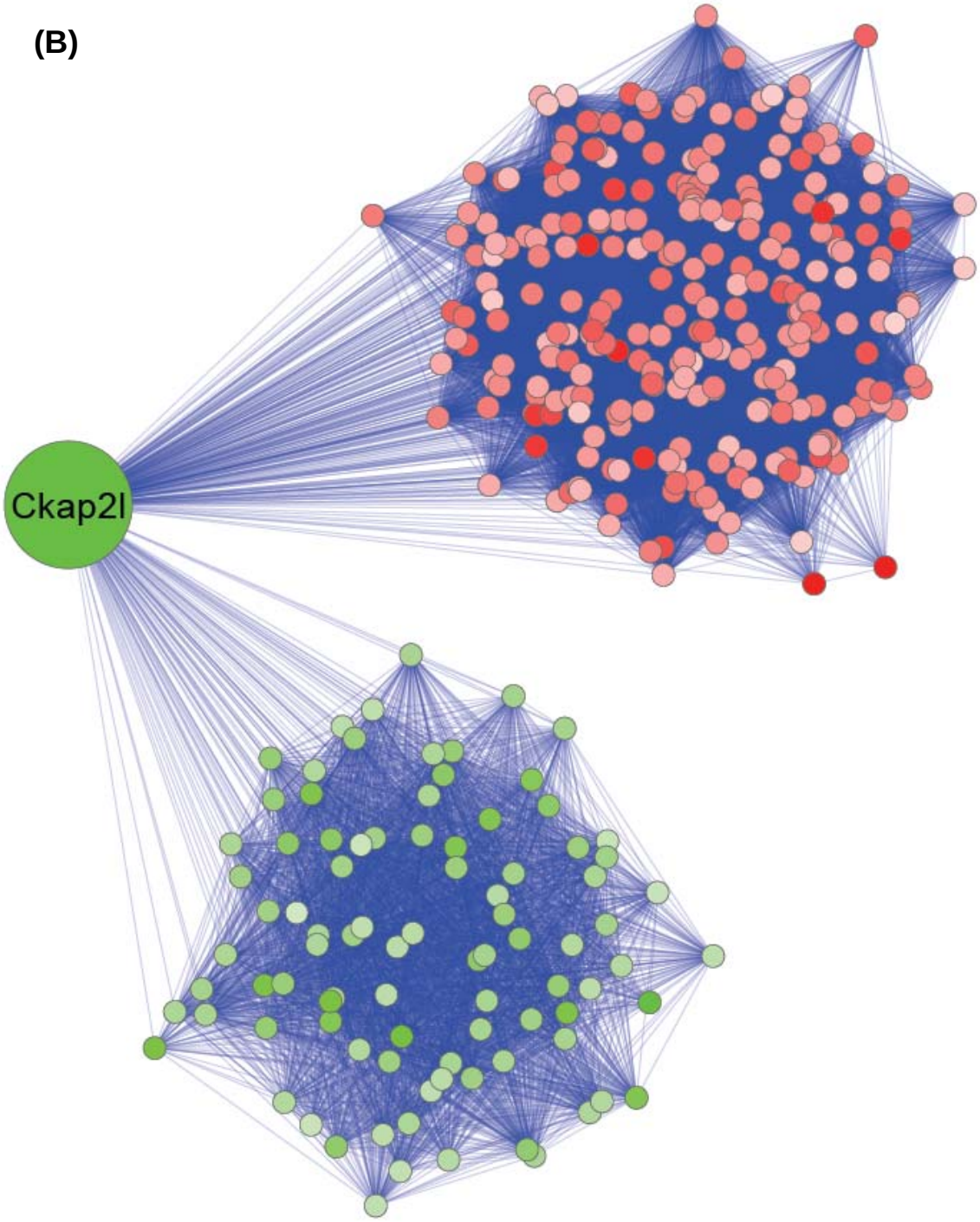

Color Key
−3 −2 −1 0 1 2
Row Z−Score
Msh6
Purb
Bcl2
Asah2
Map3k1
Bcl2
Vcp
Fastkd5
Ddx20
Dnaja3
Dap3
Gclm
Gsk3b
Malt1
Casp9
Plagl2
Traf3
Birc5
Ube4b
Casp9
Mitf
Bag2
Arf6
Ltbr
Dapk3
Fastk
Bad
Ctnnbl1
Ndufa13
Fadd
Fadd
Atf5
Shisa5
Lsp1
Ercc2
Tnfrsf12a
Bax
Aifm2
Casp9
Cul7
Eya2
Cd74
Cd74
Eif2ak3

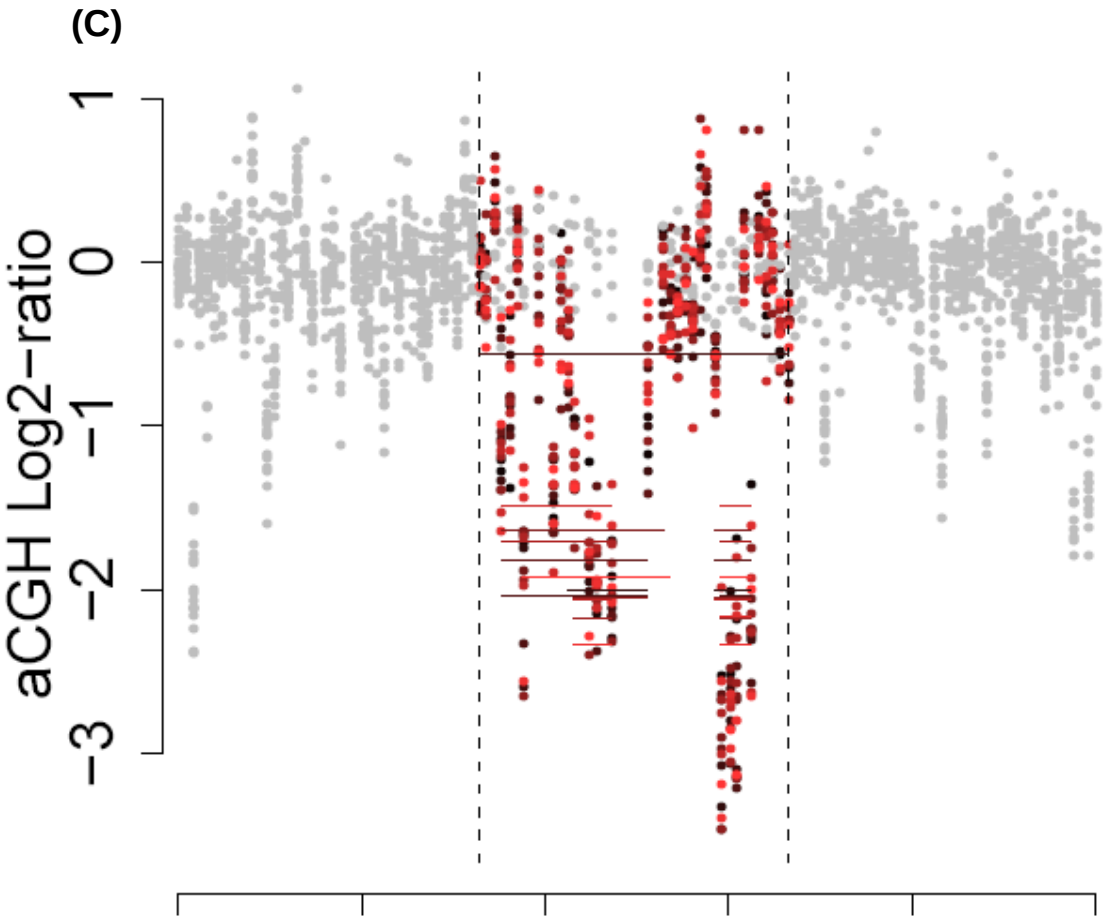
(B)

Figure 2: Gene Expression Profiling of Hematopoietic Stem and Progenitor Cells in t-AML Resistant and Susceptible Strains of Mice. (A) Unsupervised clustering of Illumina probes that are present in at least 3 strains largely separates susceptible (blue) from resistant (red) strains. Susceptibility status of some strains is undetermined (grey). **(B)** Differentially expressed genes are enriched in apoptosis-related genes. Heatmap of differentially expressed genes involved in apoptosis.


(A)
A_14
A_2
A_4
A_37
A_12
A_39
A_6
A_16
A_22
A_41
A_7
A_27
A_28
A_25
A_3
A_32
A_43
Susceptibility
A_1
A_17
A_9
A_34
A_20
A_23
A_26
A_33
A_30
A_36
A_21
A_24
A_38

(B)




(C)
aCGH Log2−ratio
−3
−2
−1
0
1

(D)

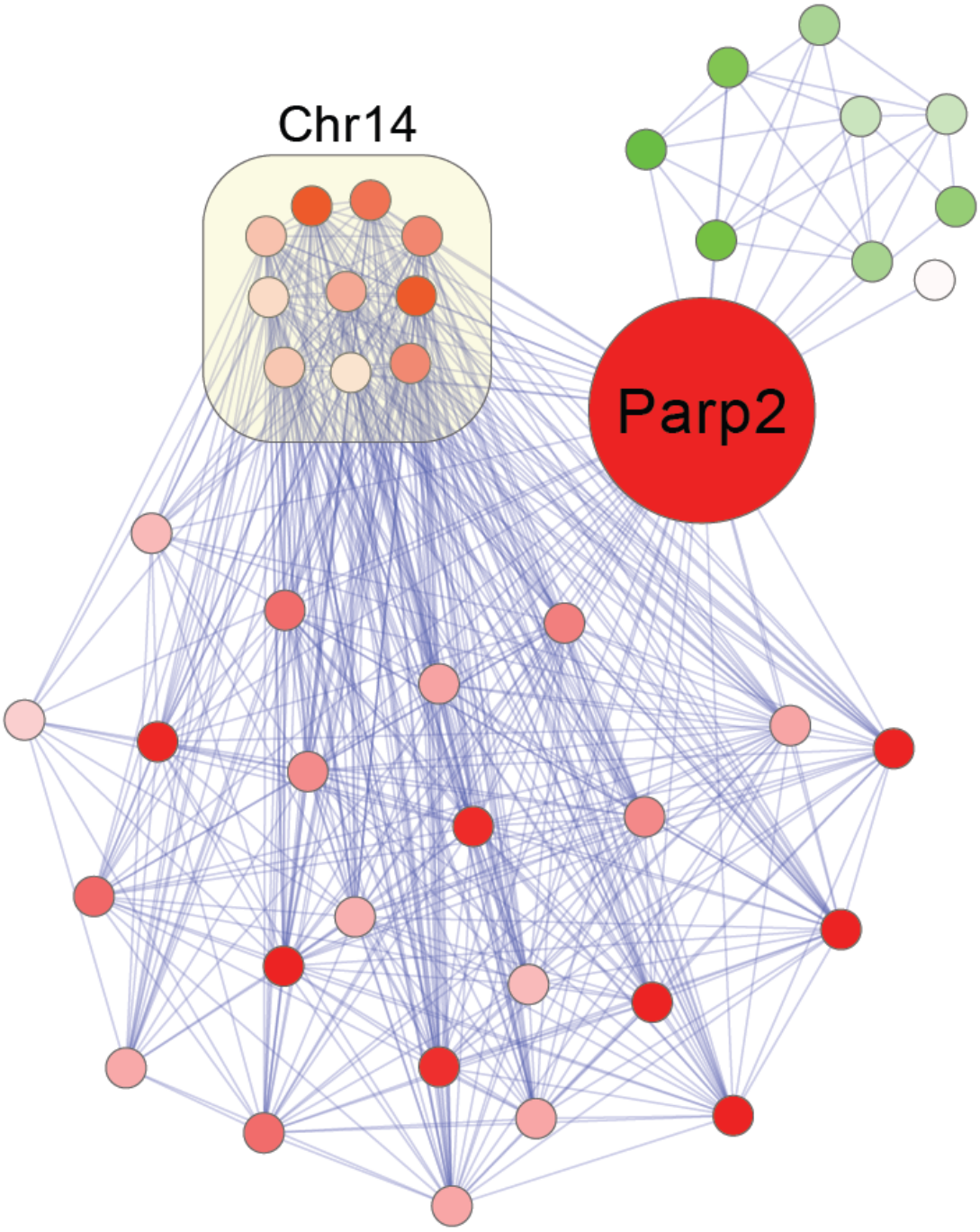
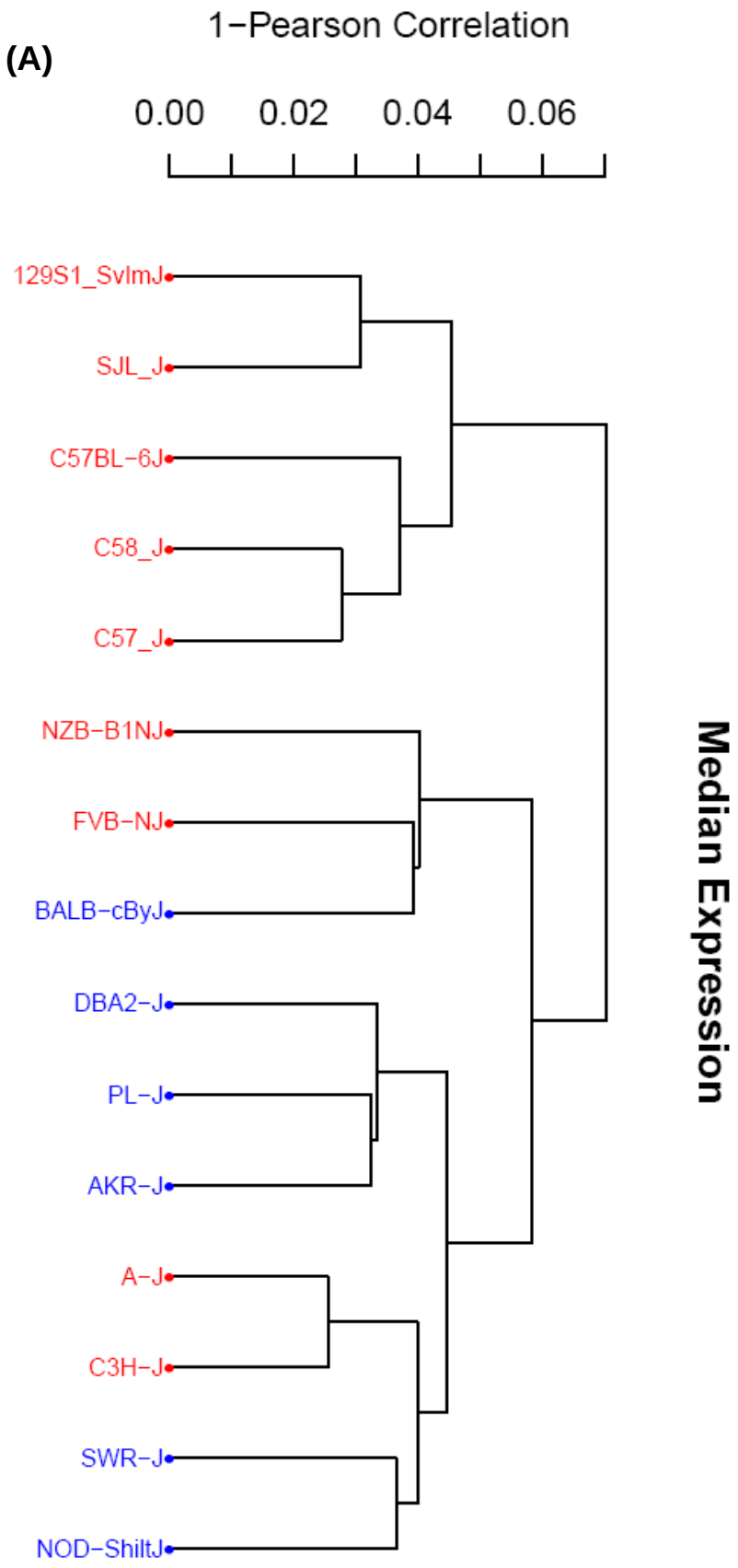
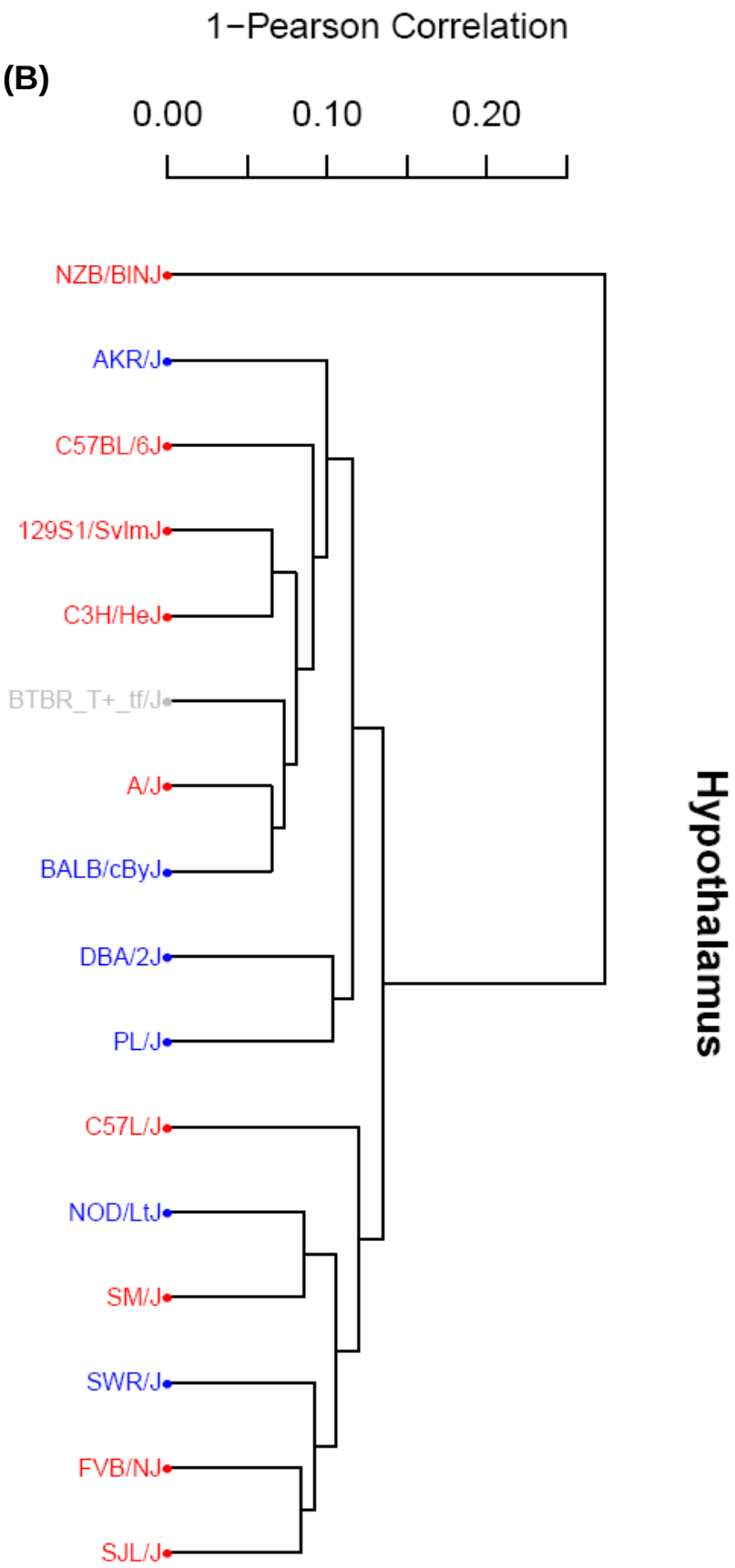
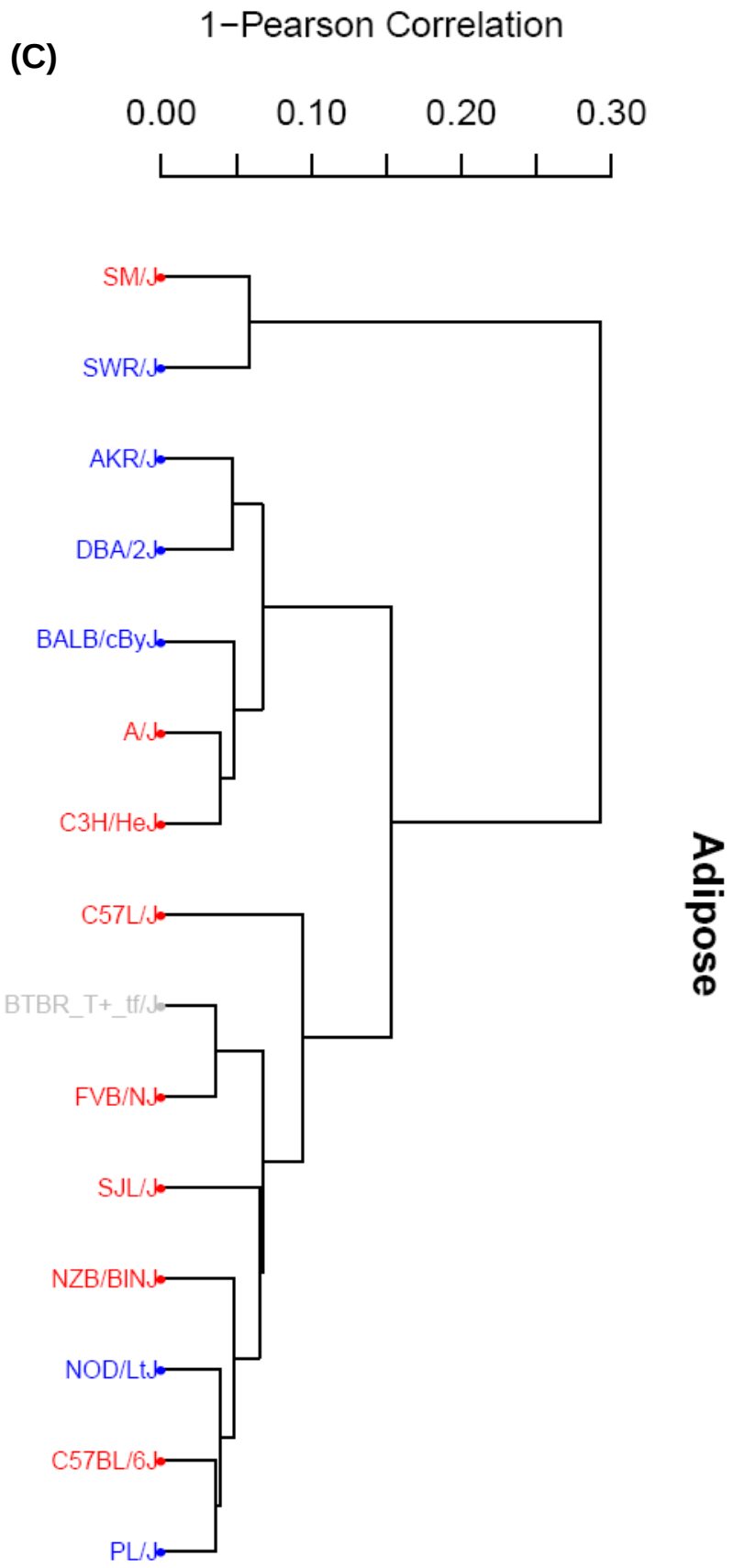


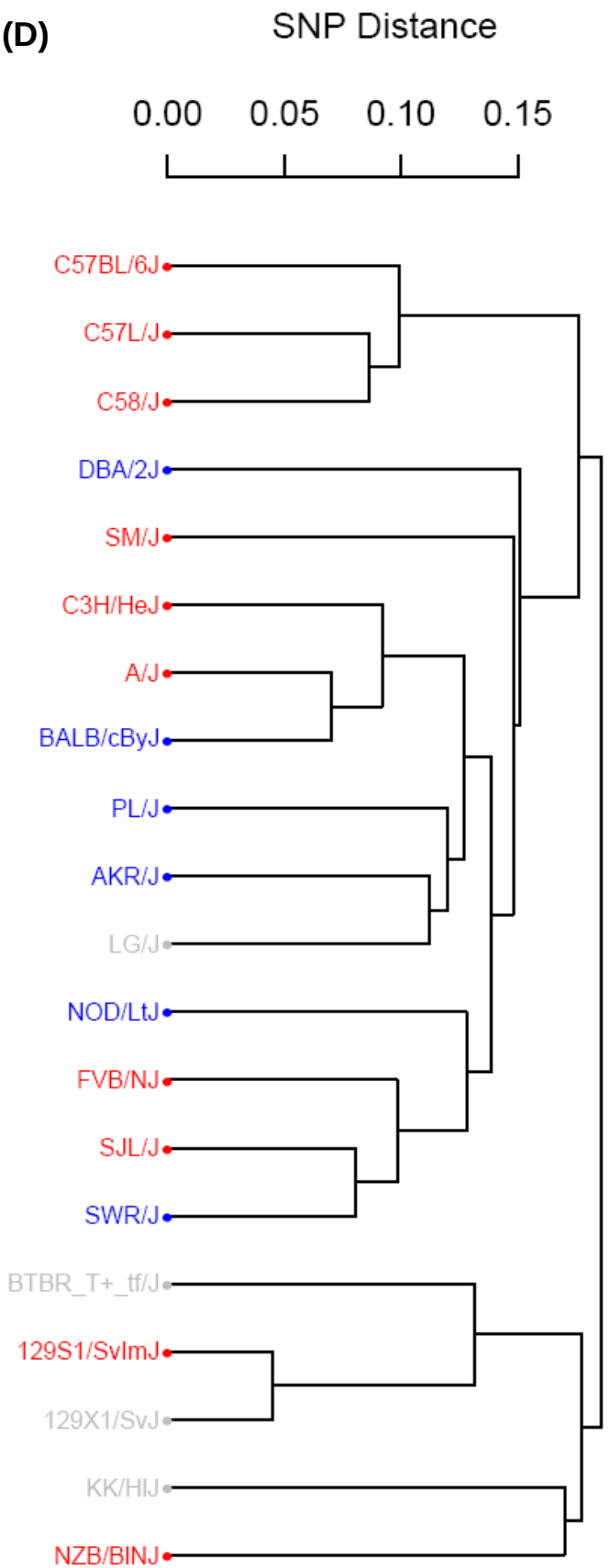
Figure 3: Anchored Susceptibility Networks. (A) Anchored modules are represented as nodes. Edges between modules represent network eigenegene correlation. Low and negative correlations are not shown for clarity. Edges between the 'Susceptiblity' and anchored network nodes represent association between network eigengenes and susceptibility status. Node size indicates the number of response genes in the anchored network. **(B)** Module A_16, anchored by Ckap2l, is enriched in cell cycle- and DNA damage-annotated genes. Green nodes represent genes with lower expression in susceptible strains, red nodes represent genes with higher expression in susceptible strains. Correlations among response genes, represented as edges, are only display for those relationships where the Pearson correlation > 0.5. **(C)** Log2-ratio plot indicating the presence of a CNVR approximately 150 kb from Parp2, the anchor gene for module A_37. **(D)** Module A_37, includes 10 genes located with 7 Mb of the CNVR depicted in Figure 3C.


(A)
1−Pearson Correlation
0.00
0.02
0.04
0.06
129S1_SvImJ
SJL_J
C57BL−6J
C58_J
C57_J
NZB−B1NJ
FVB−NJ
BALB−cByJ
DBA2−J
PL−J
AKR−J
A−J
C3H−J
SWR−J
NOD−ShiltJ
Median Expression

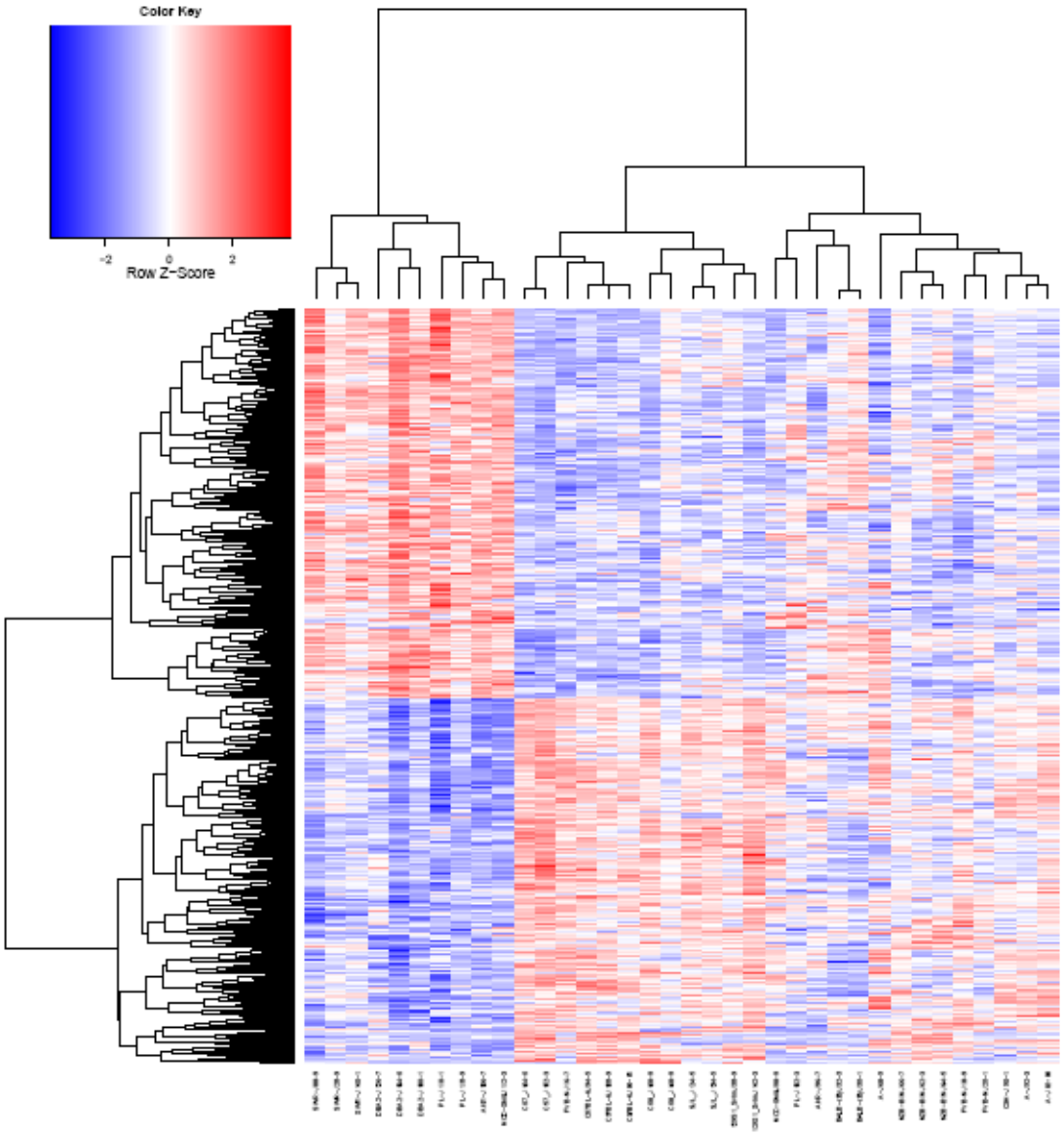

(B)
1−Pearson Correlation
0.00
0.10
0.20
NZB/BINJ
AKR/J
C57BL/6J
129S1/SvImJ
C3H/HeJ
BTBR_T+_tf/J
A/J
BALB/cByJ
DBA/2J
PL/J
C57L/J
NOD/LtJ
SM/J
SWR/J
FVB/NJ
SJL/J
Hypothalamus


(C)
1−Pearson Correlation
0.00
0.10
0.20
0.30
SM/J
SWR/J
AKR/J
DBA/2J
BALB/cByJ
A/J
C3H/HeJ
C57L/J
BTBR_T+_tf/J
FVB/NJ
SJL/J
NZB/BINJ
NOD/LtJ
C57BL/6J
PL/J
Adipose

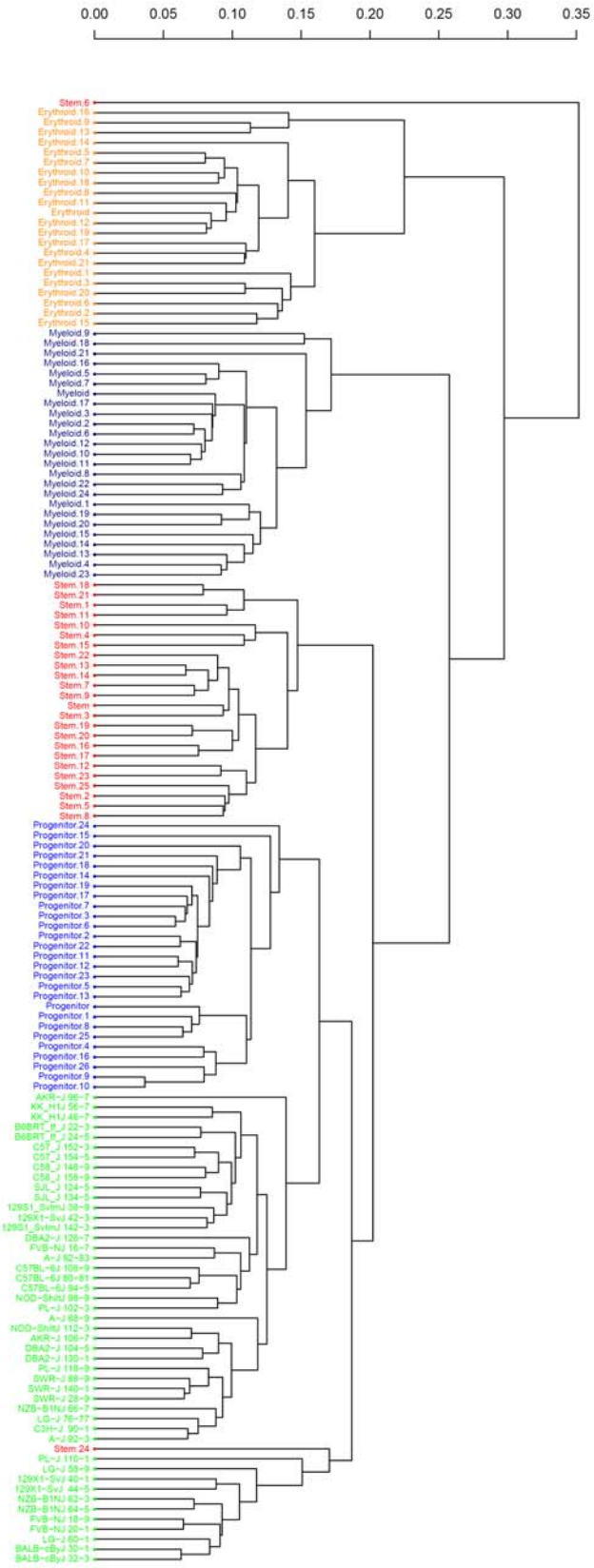
(D)


SNP Distance
0.00
0.05
0.10
0.15
C57BL/6J
C57L/J
C58/J
DBA/2J
SM/J
C3H/HeJ
A/J
BALB/cByJ
PL/J
AKR/J
LG/J
NOD/LtJ
FVB/NJ
SJL/J
SWR/J
BTBR_T+_tf/J
129S1/SvImJ
129X1/SvJ
KK/HlJ
NZB/BINJ

Supplementary Figure 1: Strain Dendrograms. Unsupervised clustering of strains using the strain median expression profile **(A)** groups strains by susceptibility status to an extent greater than expected by chance (see text), and differently than when clustering gene expression profiles of the hypothalamus **(B)**, adipose **(C)**, and when clustering based on SNP-based distance **(D)**.


Color Key
-2
0
2
Row Z-Score

Supplementary Figure 2: Differential gene expression. There are approximately 1,000 genes differentially expressed between susceptible and resistant strains.


0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35

Supplementary Figure 3: Gene Expression Profiling of Hematopoietic Cells from BXD Mice. Unsupervised clustering of Illumina detection calls groups expression profiles according to compartment (Erythroid: orange, Myeloid: navy blue, Hematopoietic stem cells (HSC): red, Progenitors: royal blue, KL cells: green), and correctly places KL samples between Progenitors and Stem cell samples. Two HSC samples are outliers.

REFERENCES

1. Leone, G., Voso, M.T., Sica, S., Morosetti, R. & Pagano, L. Therapy related leukemias: susceptibility, prevention and treatment. *Leuk Lymphoma* **41**, 255-76 (2001).

2. Leone, G., Pagano, L., Ben-Yehuda, D. & Voso, M.T. Therapy-related leukemia and myelodysplasia: susceptibility and incidence. *Haematologica* **92**, 1389-1398 (2007).

3. Larson, R.A. & Le Beau, M.M. Therapy-related myeloid leukaemia: a model for leukemogenesis in humans. *Chem Biol Interact.* **153-154**, 187-195 (2005).

4. Knoche, E., McLeod, H.L. & Graubert, T.A. Pharmacogenetics of alkylator-associated acute myeloid leukemia. *Pharmacogenomics* **7**, 719-29 (2006).

5. Meikrantz, W., Bergom, M.A., Memisoglu, A. & Samson, L. O6-alkylguanine DNA lesions trigger apoptosis. *Carcinogenesis* **19**, 369-72 (1998).

6. Seedhouse, C. et al. The genotype distribution of the XRCC1 gene indicates a role for base excision repair in the development of therapy-related acute myeloblastic leukemia. *Blood* **100**, 3761-6 (2002).

7. Allan, J.M. et al. Polymorphism in glutathione S-transferase P1 is associated with susceptibility to chemotherapy-induced leukemia. *Proc Natl Acad Sci U S A* **98**, 11592-7 (2001).

8. Larson, R.A. et al. Prevalence of the inactivating 609C-->T polymorphism in the NAD(P)H:quinone oxidoreductase (NQO1) gene in patients with primary and therapy-related myeloid leukemia. *Blood* **94**, 803-7 (1999).

9. Fenske, T.S. et al. Identification of candidate alkylator-induced cancer susceptibility genes by whole genome scanning in mice. *Cancer Res* **66**, 5029-38 (2006).

10. Noveroske, J.K., Weber, J.S. & Justice, M.J. The mutagenic action of N-ethyl-N-nitrosourea in the mouse. *Mamm Genome* **11**, 478-83 (2000).

11. Thirman, M.J. & Larson, R.A. Therapy-related myeloid leukemia. *Hematol Oncol Clin North Am* **10**, 293-320 (1996).

12. Birrell, G.W. et al. Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proc Natl Acad Sci U S A* **99**, 8778-83 (2002).

13. Fry, R.C. et al. Genomic predictors of interindividual differences in response to DNA damaging agents. *Genes Dev* **22**, 2621-6 (2008).

14. Schadt, E.E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218-23 (2009).

15. Meng, H. et al. Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc Natl Acad Sci U S A* **104**, 4530-5 (2007).

16. Wang, S.S. et al. Mapping, genetic isolation, and characterization of genetic loci that determine resistance to atherosclerosis in C3H mice. *Arterioscler Thromb Vasc Biol* **27**, 2671-6 (2007).

17. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).

18. Schadt, E.E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**, 710-7 (2005).

19. Yang, X. et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet* **41**, 415-23 (2009).

20. Ghazalpour, A. et al. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2**, e130 (2006).

21. Plaisier, C.L. et al. A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet* **5**, e1000642 (2009).

22. Cahan, P., Li, Y., Izumi, M. & Graubert, T.A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet* **41**, 430-7 (2009).

23. http://www.broadinstitute.org/mouse/hapmap/.

24. Breitling, R. et al. Genetical genomics: spotlight on QTL hotspots. *PLoS Genet* **4**, e1000232 (2008).

25. Bogue, M.A., Grubb, S.C., Maddatu, T.P. & Bult, C.J. Mouse Phenome Database (MPD). *Nucleic Acids Res* **35**, D643-9 (2007).

26. Cahan, P. et al. wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. *Nucleic Acids Res* **36**, e41 (2008).

27. Nadler, J.J. et al. Large-scale gene expression differences across brain regions and inbred strains correlate with a behavioral phenotype. *Genetics* **174**, 1229-36 (2006).

28. Gerrits, A. et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet* **5**, e1000692 (2009).

29. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).

30. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

31. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).

32. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74 (2008).

33. Young, L.C. et al. DNA mismatch repair protein Msh6 is required for optimal levels of ultraviolet-B-induced apoptosis in primary mouse fibroblasts. *J Invest Dermatol* **121**, 876-80 (2003).

34. Roos, W.P., Christmann, M., Fraser, S.T. & Kaina, B. Mouse embryonic stem cells are hypersensitive to apoptosis triggered by the DNA damage O(6)-methylguanine due to high E2F1 regulated mismatch repair. *Cell Death Differ* **14**, 1422-32 (2007).

35. Klapacz, J. et al. O6-methylguanine-induced cell death involves exonuclease 1 as well as DNA mismatch recognition in vivo. *Proc Natl Acad Sci U S A* **106**, 576-81 (2009).

36. Funk, R.K. et al. Quantitative trait loci associated with susceptibility to therapy-related acute murine promyelocytic leukemia in hCG-PML/RARA transgenic mice. *Blood* **112**, 1434-42 (2008).

37. http://biogps.gnf.org.

38. Seki, A. & Fang, G. CKAP2 is a spindle-associated protein degraded by APC/C-Cdh1 during mitotic exit. *J Biol Chem* **282**, 15103-13 (2007).

39. Hong, K.U. et al. Cdk1-cyclin B1-mediated phosphorylation of tumor-associated microtubule-associated protein/cytoskeleton-associated protein 2 in mitosis. *J Biol Chem* **284**, 16501-12 (2009).

40. Jeon, S.M. et al. A cytoskeleton-associated protein, TMAP/CKAP2, is involved in the proliferation of human foreskin fibroblasts. *Biochem Biophys Res Commun* **348**, 222-8 (2006).

41. Hartwell, L.H. & Kastan, M.B. Cell cycle control and cancer. *Science* **266**, 1821-8 (1994).

42. Schreiber, V. et al. Poly(ADP-ribose) polymerase-2 (PARP-2) is required for efficient base excision DNA repair in association with PARP-1 and XRCC1. *J Biol Chem* **277**, 23028-36 (2002).

43. Raffoul, J.J. et al. Apurinic/apyrimidinic endonuclease (APE/REF-1) haploinsufficient mice display tissue-specific differences in DNA polymerase beta-dependent base excision repair. *J Biol Chem* **279**, 18425-33 (2004).

44. Szatkiewicz, J.P. et al. An imputed genotype resource for the laboratory mouse. *Mamm Genome* **19**, 199-208 (2008).

45. Smyth, G.K. Limma: linear models for microarray data. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (ed. R. Gentleman, V.C., S. Dudoit, R. Irizarry, W. Huber) 397-420 (Springer, New York, 2005).

46. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

47. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).

48. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

49. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70 (1979).

50. Benjamini Y, H.Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* **57**, 289–300 (1995).

Conclusion

The achievements of this thesis work include the development of novel software for the detection and genotyping of CNVs, determining the copy number content CNVs to a 10-kb resolution in 20 inbred mouse strains commonly used in biomedical research, the development of an inbred mouse haplotype map, the mapping of both CNV- and SNP-based hematopoietic stem and progenitor cis-eQTLs, and the identification of candidate coexpression networks that underlie t-AML susceptibility. There are several findings that are of interest in the fields of cancer susceptibility and genetics.

DNA Copy number variation

We developed wuHMM to detect CNVs and demonstrated its performance characteristics. A novel aspect of this software is that it can use SNP information to improve its sensitivity in detecting small CNVs when abundant sequence divergence exists between the genomes under comparison. wuHMM has been utilized in other studies[1] (other manuscripts in preparation), and has been cited in other studies as a novel application of HMMs in CNV detection[2-4]. The advent of next-generation provide a new means of identifying CNVs[5] by read depth coverage. Theoretically, re-sequencing at a high coverage will improve CNV resolution down to a base pair level. Due to technical and cost constraints at present, it is unlikely that aCGH will become obsolete in the near-term. Further, because wuHMM uses emission distributions of discrete variables, it is theoretically possible to directly apply wuHMM on read-count data for CNV detection. The extent to which it would need to be modified and optimized for use on re-sequencing read data is unknown.

To determine the copy number content of the mouse genome, we performed comparative genomic hybridization using a long-oligonucleotide array containing approximately 2.1 million probes evenly spaced across the reference C57BL6/J genome (median inter-probe spacing of 1,015 bases). We applied wuHMM to this data to identify 1,333 CNVRs (82 Mb) at an empirically estimated false positive rate of less than 5%. Most CNVRs are less than 10 kb in

length, are found in more than one strain, and, in total, span 3.2% (85 Mb) of the genome. There are several pressing questions regarding structural variation in mice genomes. Second, are there tissue-specific CNVs? Our analysis included the comparison of DNA from four tissues of an individual C57BL6/J mouse. Although we detected no tissue specific copy number alterations, it does not preclude the possibility that they would be detected in more samples, different tissues, or using a higher resolution platform. Third, what is the age of CNVs relative to the SNP-based haplotypes? Are CNVs coming and going in the genomes of these putatively identical and homozygous strains of mice? These questions will need to be addressed by genotyping CNVRs and neighboring SNPs in a large number of individual mice from identical strains across multiple generations.

To assess the potential functional impact of copy number variation, we mapped expression profiles of purified hematopoietic stem and progenitor cells (data which we generated), adipose tissue and hypothalamus (data in public domain) 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. 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 alterations in expression patterns are better tolerated than complete or partial gene gains or losses. This observation refutes the prior prediction that the major impact of CNVs on expression would be through gene dosage effects[6]. The distant impact of CNVs on local expression variation was corroborated by an independent study of CNVs and expression variation in multiple mouse tissues[7]. Multiple studies have cited this mechanism to explain associations between CNVRs and phenotypes: GSTT2B expression variation[8], hypertrichosis[9], and Pea-comb phenotype[10]. It is difficult to prove that a non-gene dosage CNV causes an expression change. The development of a general framework for testing the link between CNV and expression would be beneficial. It would help to define the characteristics of the CNV sequences (or structures) that play a role in expression regulation, and could facilitate the development of genetic modifications for altering gene expression.

We found that in hematopoietic stem and 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. 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 Cross[11]) would increase the power to detect *trans* and weaker effects and therefore enable a clearer understanding the overall impact of CNVR on expression variability.

Therapy-related AML

The overarching goal of the thesis, to expand what is known about the processes underlying susceptibility to t-AML, has been met by the integrated genomics study presented in Chapter 4. We identified novel candidate expression networks associated with susceptibility and the putative upstream regulators of these modules. The biological processes implicated include apoptosis, DNA repair (including base excision repair), and cell cycle regulation. Each of these annotations are biologically plausible, given what is known about t-AML susceptibility, and warrant further experimental exploration. The networks were validated at several levels. First, the association between cis-markers and gene expression was assessed in independent data sets. eQTLs not reproduced were dropped from further analysis. Second, since we hypothesized that anchor genes drive expression of response genes, we also tested this association in independent data sets. Again, we removed those genes where an association was not reproduced, resulting in well-validated coexpression networks. We did not validate the role of expression networks in t-AML susceptibility as these sets of experiments will be long term-projects that extend beyond the

scope of this thesis. However, we have initiated the validation of the driver status of the Ckap2l module (enriched in cell cycle and DNA repair genes). In these studies, we are knocking down the expression of Ckap2l in purified KL cells and assessing the expression of the network after 24 and 48 hours. We predict that knockdown of Ckap2l in C57BL/6J (t-MDS/AML resistant) cells will recapitulate the susceptible strain expression pattern of the Ckap2l module, proving that Ckap2l regulates (directly or indirectly) the cell cycle/DNA repair network. If successful, then this paradigm will serve as a powerful method both to validate drivers of networks and to modulate network activity. Ultimately, the contribution of these networks to t-AML susceptibility will need to be tested formally. The ability to modulate network activity by altering driver gene expression will serve as a powerful tool in the costly and lengthy experiments that assess causality in t-AML susceptibility.

The development of a t-AML susceptibility classifier (or predictor) based on pre-exposure transcriptional profiles and anchored modules would be a valuable extension of the current work. Predictors could be tested by assessing the susceptibility of the BXD cross, for which extensive SNP and expression profiling data already exist. Ultimately, the development of highly accurate classifiers for human t-AML susceptibility would be valuable in a clinical setting. But in the near term, an accurate mouse classifier would be beneficial because it would enable the use of pre-exposure transcriptional profiles as a biomarker. This is practically important because the latency of t-AML can be up to 16 months. This imposes significant cost and time constraints on these in vivo experiments.

REFERENCES

1. Graubert, T.A. et al. Integrated genomic analysis implicates haploinsufficiency of multiple chromosome 5q31.2 genes in de novo myelodysplastic syndromes pathogenesis. *PLoS One* **4**, e4583 (2009).

2. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**, 1586-92 (2009).

3. Wu, L.Y., Chipman, H.A., Bull, S.B., Briollais, L. & Wang, K. A Bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics* **25**, 1669-79 (2009).

4. Li, W., Lee, A. & Gregersen, P.K. Copy-number-variation and copy-number-alteration region detection by cumulative plots. *BMC Bioinformatics* **10 Suppl 1**, S67 (2009).

5. Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).

6. Korbel, J.O. et al. The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol* **18**, 366-74 (2008).

7. Henrichsen, C.N. et al. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* **41**, 424-9 (2009).

8. Zhao, Y., Marotta, M., Eichler, E.E., Eng, C. & Tanaka, H. Linkage disequilibrium between two high-frequency deletion polymorphisms: implications for association studies involving the glutathione-S transferase (GST) genes. *PLoS Genet* **5**, e1000472 (2009).

9. Sun, M. et al. Copy-number mutations on chromosome 17q24.2-q24.3 in congenital generalized hypertrichosis terminalis with or without gingival hyperplasia. *Am J Hum Genet* **84**, 807-13 (2009).

10. Wright, D. et al. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. *PLoS Genet* **5**, e1000512 (2009).

11. Churchill, G.A. et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet* **36**, 1133-7 (2004).