Database tool

Comparison of sequence variants in transcriptomic control regions across 17 mouse genomes

Cao Nguyen[1,2], Abdul Baten[3] and Grant Morahan[1,2,*]

[1]Centre for Diabetes Research, The Western Australian Institute for Medical Research, Western Australia, Australia, [2]Centre of Medical Research, University of Western Australia, Perth, Western Australia, Australia and [3]Southern Cross Plant Science, Southern Cross University, Lismore, New South Wales, Australia

*Corresponding author: Tel: +61 86151 0756; Fax: +61 8 9224 0322; Email: grant.morahan@uwa.edu.au

The laboratory mouse is the most widely used mammalian model organism in biomedical research, so a thorough annotation of functional variation in the mouse genome would be of significant value. In this study, we compared sequence variation in a comprehensive list of functional elements (e.g. promoters, enhancers and CTCF binding sites) across 17 inbred mouse strains. Sequences were derived for ~300 000 functional elements experimentally identified by the mouse ENCODE project as regulating gene expression in 19 different tissue sources. We aligned sequences for each predicted *cis*-regulatory element to genomes of 17 mouse strains. This yielded a database comprising ~5 million aligned sequences, allowing interrogation of sequence variation of functional elements for each of the 19 tissues/cell types in commonly used mouse strains. We also developed an online tool to visualize the genome around each predicted *cis*-regulatory element in each tissue context and which allows efficient comparison of variation between any two sets of strains. This will be particularly useful in the context of the Collaborative Cross (CC), which was conceived as a powerful new systems genetics resource to accelerate gene discovery. Comprising a large number of inbred strains derived from eight genetically diverse founders, the CC offers rapid mapping and identification of genes that mediate complex traits. We show that, among the 17 sequenced strains, the set of CC founder strains captures the most variability in the ENCODE elements, further emphasizing the value of this resource.

Database URL: www.sysgen.org/ecco

Introduction

The laboratory mouse is the most widely used mammalian model organism for biomedical research owing to the many advantages it provides. Genetic discoveries based on investigation of the mouse allow insights into human traits because of the degree of conservation between the mouse and human genomes. With the assembly of the (almost) complete genome of the C57BL/6J strain, the ability to relate sequence to function was significantly enhanced (1, 2). This has enabled genetic screens in mice to be performed on an unprecedented scale (3), facilitated the creation of a set of null alleles for all genes (4, 5) and accelerated the definition of mouse sequence diversity (6, 7). Despite this great progress, the molecular basis for much morphological, physiological, biochemical and behavioral variation in laboratory mice remains largely unknown (8–10). We still lack a significant amount of knowledge regarding the molecular basis of the majority of genetically influenced phenotypes, from fully or partly penetrant Mendelian effects (11, 12) and nonadditive effects (12), to the quasi-infinitesimal genetic architecture that underlies

many quantitative traits (13). Furthermore, most of the functional sequences in the mouse genome have yet to be found; in particular, *cis*-regulatory sequences are still poorly annotated, while most *trans*-regulatory loci have not yet been defined.

To aid in the discovery of regulatory sequences, comparative genomics is a powerful tool (14). However, this method cannot resolve the temporal–spatial functions of such sequences. Recently, chromatin immunoprecipitation sequencing was applied to identify *cis*-regulatory elements in the genomes of several organisms including humans, *Drosophila melanogaster* and *Caenorhabditis elegans* (15–17). The ENCODE consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns (15). Using the same experimental approach, Shen *et al*. (18) produced a map of nearly 300 000 murine *cis*-regulatory sequences that represent active promoters, enhancers and CTCF-binding sites experimentally determined in a set of 19 diverse tissues and cell types. In our database, each of the *cis*-regulatory elements is mapped to a locus in chromosomes of the mouse genomes as defined by the ENCODE project. This provided a comprehensive resource for the annotation of functional elements in the mammalian genome and the study of regulatory mechanisms for tissue-specific gene expression.

Variation in such elements would contribute to differences in gene expression, and ultimately to differences in various phenotypic traits. Conventional approaches for identifying genetic contribution to variable traits use diverse gene mapping strategies that generally are limited by power, speed and precision. The Collaborative Cross (CC), a large set of recombinant inbred strains, was conceived as a powerful new systems genetics resource that could accelerate discovery (19). Among the many advantages it offers is the rapid mapping and identification of genes that mediate complex traits, especially those that can only be determined *in vivo*. A decade in planning and production in Australia, the USA and Israel (19–22), the CC is now reaching maturity (23).

Recently, the sequences of genomes from 17 widely used mouse strains were obtained using next-generation sequencing (24). These included the eight founder strains of the CC (i.e. A/J, C57BL/6NJ, 129S1/SvlmJ, NOD, NZO, CAST/EiJ, PWK/PhJ and WSB/EiJ). Collectively, the sequences of these 17 strains allow study of genetic variation in the most commonly used strains of mice.

The availability of these *de novo* genome assemblies allowed us to analyze variation in the context of the CC. In particular, we aimed to make a comprehensive comparison of genetic variation in all of the 300 000 ENCODE genomic features by integrating these elements with the mouse genome sequences. We provide a database comprising the aligned sequences of the predicted *cis*-regulatory elements in each of 19 tissues and cell types. Our online database and search tool can provide visualization of the genome around each predicted element in each tissue context. A powerful application of this database will be comparing variation in ENCODE elements shared differentially between responder and nonresponder founder haplotypes because this can allow rapid identification of causal variants in candidate genes mapped using the CC. Similar web tools are available, for example, web-quantitative trait loci (QTL) (25), eQTL Viewer (26) and Genevar (27), but they are designed for conventional approaches such as gene expression and/or for analyzing human genome.

Methods

Data sources

The *de novo* genome assemblies of 17 mouse strains (24) were downloaded from the Sanger Institute's Web site (http://www.sanger.ac.uk/resources/mouse/genomes/). The mouse ENCODE project [(http://chromosome.sdsc.edu/mouse/; cf (18)] provided data for genomic localizations of RNA polymerase II (polII), the insulator-binding protein CCCTC-binding factor (CTCF) and three chromatin modification marks, histone H3 lysine 4 trimethylation (H3K4me3), histone H3 lysine 4 monomethylation (H3K4me1) and H3 lysine 27 acetylation (H3K27ac), in 13 adult tissues, 4 embryonic tissues and 2 primary cell lines.

Integration of functional and sequence data sets

A window of k base pair (max $k = 50$) either side of each *cis*-regulatory element was taken from the mouse genome reference NCBI Build 37 (UCSC mm9 database) and was aligned against the 17 mouse strains by BLAST search (28). We stored the aligned sequences resulting from BLAST into a MySQL database and developed an online tool to allow interactive visualization and comparison of the variations among the 17 sequenced strains. We used the following notations in our study:

Total single-nucleotide polymorphisms in a cis-regulatory element. The total number of single-nucleotide polymorphisms (SNPs) counted in the surrounding k base pair window.

Variable regulatory elements. These contain at least one SNP in the surrounding k base pair window. If there was at least one sequence variation around an element in any strains of a group then the group is considered *variable* at the element.

Figure 1. An example for the notations defined in our study. The aligned sequences at chr1: 51 952 487–51 952 587 around an *hek4me3* element in *liver* are shown. Of the total two SNPs (highlighted in black) around the selected element, both were found in the configured group I (red) and neither was found in the configured group II (blue). Group I is considered as variable at this element because there are three sequence variations associated with strains A/J, CAST/Ei, PWK/Ph and WSB/Ei. In contrast, group II is invariable because all the aligned sequences of the group are identical.

Invariable elements. These elements have no SNPs in the surrounding k base pair window. *A group of strains is an invariable group* for that element if all sequences of the group are invariable at that element.

Figure 1 shows an example of the above notations. Our web-based tool is written in HTML5/Java script, MySQL and PHP. The database we constructed, termed the Encode CC Omnibus (ECCO), is freely available for user access (www.sysgen.org/ecco).

Results

Sequence variations in ENCODE elements

The sequences deriving from *cis*-regulatory elements (polII, CTCF, H3K4me3, H3K4me1 and H3K27ac) for 17 mouse strains in 19 tissues and cell lines were extracted from the mouse reference genome. Using BLAST search against the other sequenced genomes, the relevant elements were selected for each strain. All elements were then aligned to the reference genome. The resulting ECCO database contains ∼5 million aligned sequences of *cis*-regulatory elements (polII, CTCF, H3K4me3, H3K4me1 and H3K27ac) for 17 mouse strains in 19 tissues/cell lines. All sequences are available for download via a button embedded in the Web site.

The SNPs and sequence variations in k base pair windows around any of the predicted *cis*-regulatory elements between the CC founder strains or between any of the 17 sequenced strains can be efficiently compared and visualized, as shown in Figure 2. Users can visualize sequence variations in a location of interest either by searching by gene name or by selecting a specific chromosome region. Users can also select the aligned database based on tissue/cell lines and/or *cis*-regulatory elements. The default k base pair window is set at 50, but users can specify other values ranging from 1 to 50 for this parameter.

Sequence variations around *cis*-regulatory elements

Next, we examined the total sequence variations and invariable elements: how many of the predicted elements were variable and invariable, respectively? For each of the aligned sequences in each strain, we determined the number of SNPs, the number of sequence variations in each element (variable elements) and the number of invariable elements.

Figure 3 shows the results of these statistics in all studied *cis*-regulatory elements for the 19 tissues and cell lines over 17 sequenced strains, for each of settings of $k = 50, 40, 30, 20, 10$ and 5 bp windows, respectively. For example, at $k = 50$ bp window, the average number of SNPs per strain for each type of *cis*-regulatory element is ∼567 000, ranging from on average 229 000 (polII) to 667 000 (H3K4me1). Consequently, the total predicted variable elements range from an average 48 000 (polII) to 184 000 (H3K4me1) elements. The full data including number of SNPs, number of variable elements, number of invariable elements for each strain in each studied *cis*-regulatory elements is available at Supplementary Table S1.

Validation of CC founder strains as source of functional genetic variation

Next, we compared the sequence variations around the predicted *cis*-regulatory elements that were different among the CC founder strains with the other strains (for convenience, referred to below as nonfounder strains). For each group, we counted how many times the group was variable or invariable at each of the *cis*-regulatory elements with k base pair windows of 50, 40, 30, 20, 10 and 5, respectively. Note that the definition of variable was liberal: a group at a predicted *cis*-regulatory element was considered as variable if at least one strain had at least one sequence variation in the specified k base pair windows. Among the eight founder strains, we found that at $k = 50$ the group
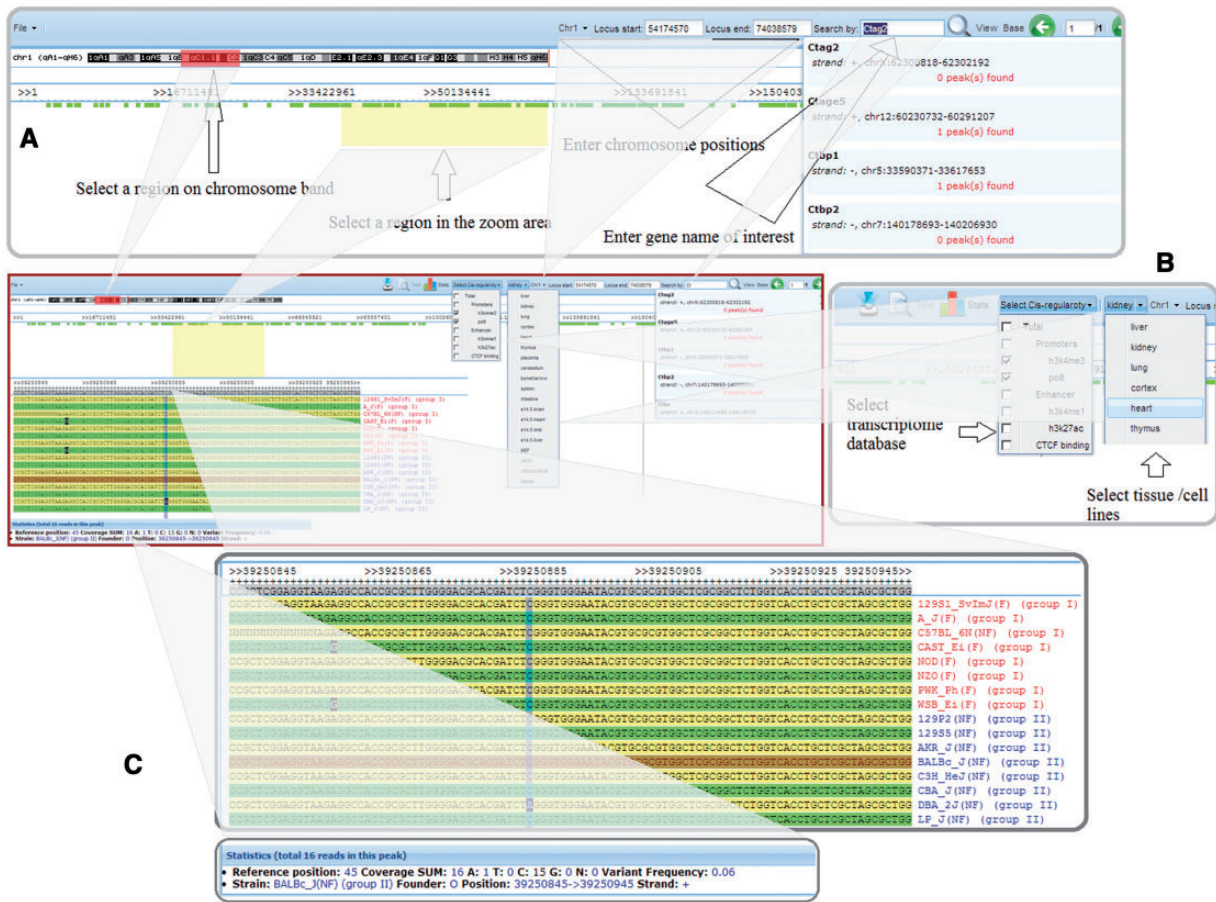
Figure 2. Visualization of *cis*-regulatory elements with sequence variations between the CC founder strains and the set of nine nonfounder strains. (**A**) Users can select a region of interest by four different methods: (i) selecting a region on a chromosome band, (ii) entering chromosome positions, (iii) inputting a gene name in the search box or (iv) highlighting a region in the zoom area; (**B**) Users can select any one or more of the *cis*-regulatory element types and tissue/cell lines for visualization; and (**C**) Strains' genomic sequences are aligned, showing SNPs and statistics.

was variable at a significantly higher number of elements than were invariable. In contrast, the proportion of variable elements was significantly smaller in the set of nine nonfounder strains. This observation was consistent for all five types of regulatory elements (polII, CTCF, H3K4me3, H3K4me1 and H3K27ac). In addition, we found that in each of the settings of $k = 50, 40, 30, 20, 10$ and 5, the variable elements in founder strains were consistently higher than those in nonfounder strains, suggesting the CC mice would provide a rich source of functional genetic variations (see Figure 4 and Supplementary Table S2 for more details).

Identification of useful polymorphisms

Another application of our database is to identify SNPs, which could provide useful polymorphisms for further characterization of CC strains, or to select specific strains or even outbred mice [such as the Diversity Outcross; (29)] having particular genotypes at elements of interest for further

study. To identify all candidate SNPs for a particular QTL, the database of founder strains' genome sequences can compare responder against nonresponder haplotypes. This allows identification of variable sequences that could be investigated further. These SNPs and microsatellites would be useful to increase map resolution by testing strains with recombinant haplotypes, or to identify individual outcross mice with particular alleles. These could then be tested to confirm and refine locations of QTLs. Users can select regions or genes of interest and configure the groups of responders and nonresponders from the 17 sequenced mouse strains. Our database provides two statistical tools for this comparison: a proportional *z*-test to see the significance level of any difference between variations and nonvariations within a group of strains, and a bar chart for visualization of total SNPs in each group of strains (Figure 5). The *z*-test *P*-value is visible at the bottom of each column as well as when mouse is over the bars in the chart.
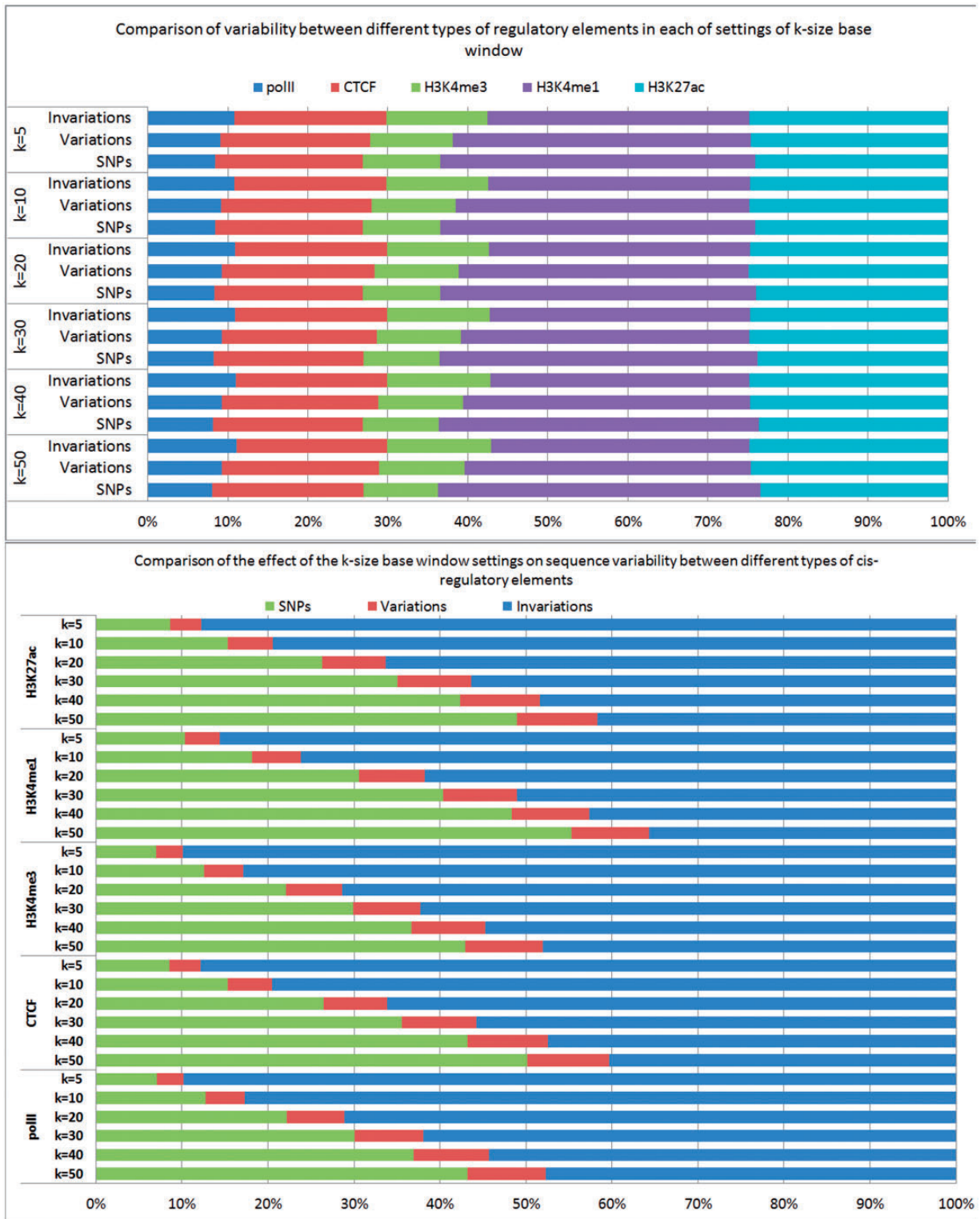
Figure 3. The average number of SNPs, variable *cis*-elements and invariable *cis*-elements in five studied *cis*-regulatory elements for the 19 tissues and cell lines over 17 sequenced strains, for each of settings of $k = 50$, 40, 30, 20, 10 and 5 bp windows, respectively.
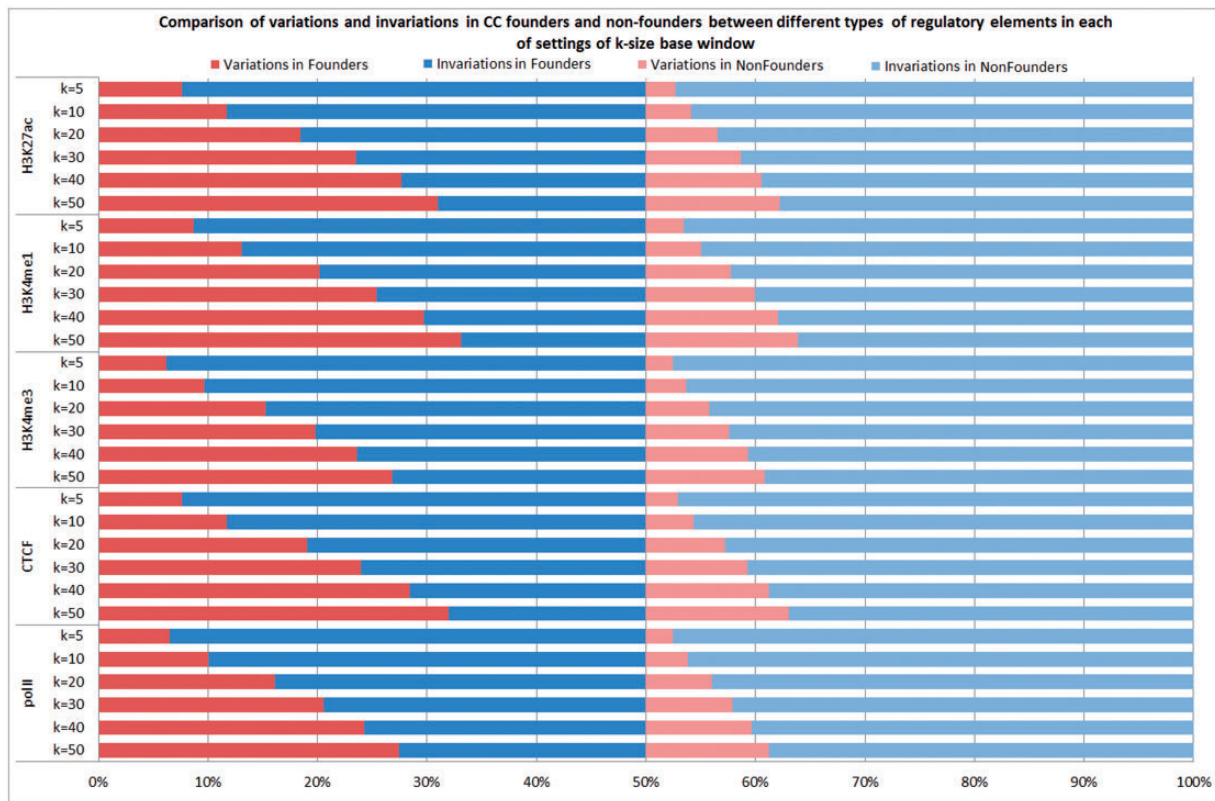
Figure 4. Comparison of variations and invariations in CC founders and nonfounders between different types of regulatory elements in each of settings of k-size base window.

Molecular basis of transcriptional phenotypes

Systems genetics is a powerful technology to analyze effects of genome-wide genetic variants on transcriptome-wide variation in gene expression (30). Here we combined our resource database with the systems genetics approach for systematic investigation of transcriptional phenotypes. Gene expression data from liver, kidney, lung, cortex, cerebellum and spleen were analyzed from a large panel of isogenic recombinant inbred strains of BXD mice [(www. genenetwork.org, cf (28)]. The BXD family is a collection of recombinant inbred strains that were created by successive inbreeding of progeny generated from matings of C57BL/6J and DBA/2J mice. Using WebQTL (25), we extracted highly expressed genes (mean expression >10) likely to have variable *cis*-regulatory elements within 1 MB. These eQTL had peak likelihood ratio statistics (LRS) over 40.

Coordinates for each of these eQTLs were input to our database to further examine sequence variations in *cis*-regulatory elements. Table 1 shows the number of suggestive QTLs with at least one variable element in any of the five *cis*-regulatory elements in the tissues studied. These candidate eQTLs regulate expression of genes that have been associated with traits that are worth investigating in

further studies. For example, our analysis showed that *Prp19*, a gene essential for cell survival and DNA repair (31) and linked with tumorigenesis (32), is highly expressed in liver (mean expression at 10.95, max LRS at 129), lung (11.02, 157), kidney (11.03, 95) and cerebellum (11.6, 85.9) and is associated with variable H3K4me1 and H3K27ac elements. The full list of candidate QTLs including gene names, tissue, gene expression variation, the LRS and the number of variable, the total *cis*-regulatory elements and the total variable elements is available in Supplementary Table S3.

Discussion

A large fraction of *cis*-regulatory elements act in a cell-type specific manner and are involved in regulating tissue-specific gene expression (18). The ECCO database incorporates flexible statistical methods to analyze ~300 000 *cis*-regulatory elements in 19 different tissues and cell lines. We were able to show that CC founder strains contained significantly more sequence variations in *cis*-regulatory elements in any of the studied tissues and cell lines than did nonfounder strains. Our web-based tool is also available, so researchers
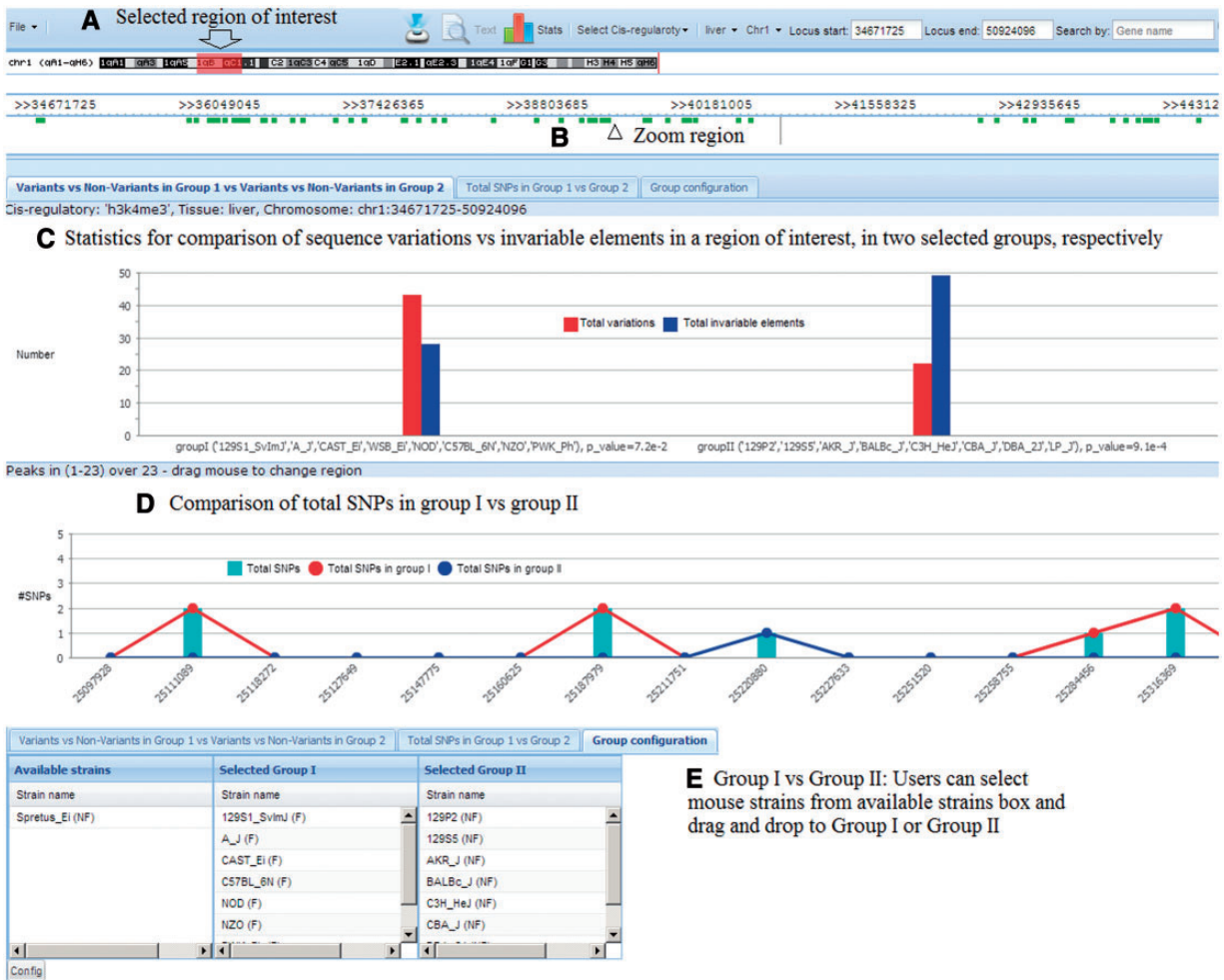
Figure 5. Tool to identify useful polymorphisms. **(A)** Users select region of interest; **(B)** Click 'View' to zoom into the region; **(C)** Statistics show the difference between total variations and nonvariations in each group; **(D)** Statistics show the total SNPs in each group and **(E)** Users can select mouse strains for each group.

Table 1. Number of genes associated with transcriptional phenotypes found by webQTL and the ECCO database using the default $k=50$ bp window

Tissue	Genes	Mean expr	Max LRS	Total elements					Total variable elements				
				polII	CTCF	H3K4me3	H3K4me1	H3K27ac	polII	CTCF	H3K4me3	H3K4me1	H3K27ac
Liver	28	11.33 (10.02–15.66)	66.11 (41.5–129)	29	15	23	52	81	12	6	12	25	39
Kidney	59	11.44 (10.07–15.33)	66 (40.1–174)	57	83	62	208	199	24	24	26	70	60
Lung	85	10.89 (10.03–12.95)	70.2 (40.3–196)	64	92	88	331	318	31	29	28	120	115
Cortex	33	11.4 (10.03–14.89)	70.07 (43.3–131)	30	33	22	153	45	13	18	11	63	15
Cerebellum	11	11.92 (10.13–14.31)	60.2 (41.5–89.4)	10	10	13	33	13	6	2	9	21	7
Spleen	35	10.81 (10.01–13.17)	68.01 (41.7–182)	22	23	30	63	78	10	7	13	32	28

Mean expr: the mean gene expression (normalized log2 expression values); Max LRS: maximum likelihood ratio statistics (LOD values / 4.61).

may select specific strains for further investigation and comparison of these genetic variants.

A powerful use of our database is in the study of the relationship between genotype and phenotype. A key challenge in complex trait genetics is how to identify the sequence variations that underlie complex traits. In addition to the phenotypic information and gene expression data on inbred strains provided by the CC, the aligned sequence of the 17 mouse genomes and the associated catalog of variants in transcriptomes will serve as a basic tool for understanding trait differences. Collectively, our presented database will allow further insights into the nature of functional variants and will help examine the path from genotype to phenotype.

Epigenomic approaches have recently been studied genome wide to tackle unanswered questions in human diseases, including cancer. It is known that enhancers in the human and mouse genomes are associated with active chromatin marks in a cell type–specific manner, whereas promoter and insulator elements tend to be ubiquitously occupied in multiple cell lines (18). During histone modifications, both lysine and arginine residues may be methylated. Methylated lysines are the best understood marks of the histone code, as specific methylated lysine marks match well with gene expression states. Methylation of lysines H3K4 and H3K36 is correlated with transcriptional activation, while demethylation of H3K4 is correlated with silencing of the genomic region. Methylation of lysines H3K9 and H3K27 is correlated with transcriptional repression (33, 34). The ECCO database enables us to map the three chromatin modification marks, histone H3 lysine 4 trimethylation (H3K4me3), histone H3 lysine 4 monomethylation (H3K4me1) and H3 lysine 27 acetylation (H3K27ac), in 13 adult tissues, 4 embryonic tissues and 2 primary cell lines for all the 17 mouse genomes. Using a similar approach, our method can also be applied to human ENCODE databases to connect genetics to epigenetics and to investigate further DNA methylation.

Many complex diseases result from gene–gene and gene–environment interactions that are not effectively modeled by isolated studies on fixed genetic backgrounds in mouse. The large panel of inbred CC mouse lines derived from eight genetically diverse strains captures almost 90% of the known variation present in laboratory mice (23), so it can provide better models for the genetic variation found in the human population. After 10 years of development, the CC project is now available for researchers. Research programs using the CC can be combined with our ECCO database for determining the molecular basis of complex diseases. For example, QTLs of complex phenotypes found by genetic mapping analysis can be further investigated using the ECCO web services. Polymorphisms associated with variable *cis*-regulatory elements in the target region can be considered as candidates and tested for biological variation that could determine the phenotypes of interest.

Supplementary Data

Supplementary data are available at *Database* Online.

References

1. Church,D.M., Goodstadt,L., Hillier,L.W. *et al*. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, **7**, e1000112.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. van der Weyden,L., Adams,D.J. and Bradley,A. (2002) Tools for targeted manipulation of the mouse genome. *Physiol. Genomics*, **11**, 133–164.
4. Ringwald,M., Iyer,V., Mason,J.C. *et al*. (2011) The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium. *Nucleic Acids Res.*, **39**, D849–D855.
5. Skarnes,W.C., Rosen,B., West,A.P. *et al*. (2011) A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, **474**, 337–342.
6. Frazer,K.A., Eskin,E., Kang,H.M. *et al*. (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, **448**, 1050–1053.
7. Quinlan,A.R., Clark,R.A., Sokolova,S. *et al*. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, **20**, 623–635.
8. Paigen,K. (2003) One hundred years of mouse genetics: an intellectual history. I. The classical period (1902–1980). *Genetics*, **163**, 1–7.
9. Paigen,K. (2003) One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981–2002). *Genetics*, **163**, 1227–1235.

10. Beckers,J., Wurst,W. and de Angelis,M.H. (2009) Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling. *Nat. Rev. Genet.*, **10**, 371–380.

11. Wilke,M., Buijs-Offerman,R.M., Aarbiou,J. *et al.* (2011) Mouse models of cystic fibrosis: phenotypic analysis and research applications. *J. Cyst. Fibros.*, **10**, S152–S171.

12. Shao,H., Burrage,L.C., Sinasac,D.S. *et al.* (2008) Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. USA*, **105**, 19910–19914.

13. Flint,J. and Mackay,T.F. (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.*, **19**, 723–733.

14. Visel,A., Rubin,E.M. and Pennacchio,L.A. (2009) Genomic views of distant-acting enhancers. *Nature*, **461**, 199–205.

15. The ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

16. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L. *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, **330**, 1775–1787.

17. Roy,S., Ernst,J., Kharchenko,P.V. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

18. Shen,Y., Yue,F., McCleary,D.F. *et al.* (2012) A map of cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–20.

19. Churchill,G. and The Complex Trait Consortium (2004) The collaborative cross: a community resource for the genetic analysis of complex traits. *Nat. Genet.*, **36**, 1133–1137.

20. Iraqi,F.A., Churchill,G. and Mott,R. (2008) The collaborative cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort. *Mamm. Genome*, **19**, 379–381.

21. Chesler,E.J., Miller,D.R., Branstetter,L.R. *et al.* (2008) The collaborative cross at oak ridge national laboratory: developing a powerful resource for systems genetics. *Mamm. Genome*, **19**, 382–389.

22. Morahan,G., Balmer,L. and Monley,D. (2008) Establishment of ''The Gene Mine'': a resource for rapid identification of complex trait genes. *Mamm. Genome*, **19**, 390–393.

23. Collaborative Cross Consortium. (2012) The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics*, **190**, 389–401.

24. Keane,T.M., Goodstadt,L., Danecek,P. *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294.

25. Wang,J., Williams,R.W. and Manly,K.F. (2003) WebQTL: web-based complex trait analysis. *Neuroinformatics*, **1**, 299–308.

26. Zou,W., Aylor,D.L. and Zeng,Z.B. (2007) eQTL Viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics*, **8**, 7.

27. Yang,T.P., Beazley,C., Montgomery,S.B. *et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.

28. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

29. Churchill,G.A., Gatti,D.M., Munger,S.C. *et al.* (2012) The diversity outbred mouse population. *Mamm. Genome*, **23**, 713–718.

30. Morahan,G., Peeva,V., Munish,M. *et al.* (2008) Systems genetics can provide new insights in to immune regulation and autoimmunity. *J. Autoimmun.*, **31**, 233–236.

31. Beck,B.D., Park,S.J., Lee,Y.J. *et al.* (2008) Human Pso4 is a metnase (SETMAR)-binding partner that regulates metnase function in DNA repair. *J. Biol. Chem.*, **283**, 9023–9030.

32. David,C.J. and Manley,J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.*, **24**, 2343–2364.

33. Rosenfeld,J.A., Wang,Z., Schones,D.E. *et al.* (2009) Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, **10**, 143.

34. Hublitz,P., Albert,M. and Peters,A.H. (2009) Mechanisms of transcriptional repression by histone lysine methylation. *Int. J. Dev. Biol.*, **53**, 335.