Research

In silico discovery of gene-coding variants in murine quantitative trait loci using strain-specific genome sequence databases

Kriste E Marshall*, Elizabeth L Godden*, Fan Yang*, Sonya Burgers*, Kari J Buck† and James M Sikela*

Addresses: *Department of Pharmacology and Human Medical Genetics Program, University of Colorado Health Sciences Center, Denver CO 80262, USA. †Department of Behavioral Neuroscience and Portland Alcohol Research Center, Oregon Health & Science University and VA Medical Center, Portland, OR 97201, USA.

Correspondence: James M Sikela. E-mail: James.Sikela@UCHSC.edu

Published: 27 November 2002

Genome ***Biology*** 2002, **3(12)**:research0078.1–0078.9

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0078

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 20 September 2002
Revised: 17 October 2002
Accepted: 22 October 2002

Abstract

Background: The identification of genes underlying complex traits has been aided by quantitative trait locus (QTL) mapping approaches, which in turn have benefited from advances in mammalian genome research. Most recently, whole-genome draft sequences and assemblies have been generated for mouse strains that have been used for a large fraction of QTL mapping studies. Here we show how such strain-specific mouse genome sequence databases can be used as part of a high-throughput pipeline for the *in silico* discovery of gene-coding variations within murine QTLs. As a test of this approach we focused on two QTLs on mouse chromosomes 1 and 13 that are involved in physical dependence on alcohol.

Results: Interstrain alignment of sequences derived from the relevant mouse strain genome sequence databases for 199 QTL-localized genes spanning 210,020 base-pairs of coding sequence identified 21 genes with different coding sequences for the progenitor strains. Several of these genes, including four that exhibit strong phenotypic links to chronic alcohol withdrawal, are promising candidates to underlie these QTLs.

Conclusions: This approach has wide general utility, and should be applicable to any of the several hundred mouse QTLs, encompassing over 60 different complex traits, that have been identified using strains for which relatively complete genome sequences are available.

Background

The discovery of genes underlying multigenic diseases and traits is one of the most important challenges currently facing genetic researchers. This effort has been aided by quantitative trait locus (QTL) mapping methods, which have now been applied to numerous complex phenotypes in a range of species, including many behavioral phenotypes of high interest. A QTL is a chromosomal region that contains a gene or genes that influence a quantitative trait. The power of this approach was first demonstrated in plants [1] and later in yeast [2], flies [3], livestock [4,5], rodents [6-9] and humans [10-12].

Historically, a typical approach to going from QTL to gene has been to select one or a few of the best biological candidate genes from within the QTL interval and search for sequence differences that predict differential expression and/or structure of the gene product. An alternative strategy

is to carry out comparative sequencing of large numbers of potential candidate genes located within the QTL interval, which is feasible given the automated sequencing methods now available [13]. However, these approaches are limited because the gene underlying a QTL may not be recognized as a good candidate gene if little is known about the gene's function and/or if a QTL region is large, in which case sequencing every gene within the QTL requires considerable time, cost and labor to complete.

The recent development of murine whole-genome draft sequences [14] should speed the process of identifying disease genes underlying QTLs because several of the strains used for genome sequencing are the same as those that have been used to develop the majority of mouse QTLs so far identified. The public mouse genome sequencing effort used C57BL/6J (B6) mice, while the private effort by Celera has sequenced the mouse genome in four strains: DBA/2J (D2), A/J, 129X1/SvJ, and 129S1/SvImJ. Combinations of these strains, for example B6xD2 and B6xA/J, have been used for the identification of over 250 murine QTLs. Because of these resources, we reasoned that sequence variations in QTL genes might now be identified simply by direct *in silico* alignment of the sequences obtained directly from the databases of the two relevant strains, obviating much of the need for *de novo* sequencing. We show that, once several factors were addressed, such as sequencing error detection and discrimination between closely related genes, a high-throughput pipeline could be developed that allows large numbers of gene-coding regions from QTL intervals to be rapidly compared *in silico* and interstrain allelic sequence differences quickly identified and targeted for further hypothesis-driven analyses. We also believe that this strategy has considerable general utility, and should be applicable to any QTL as long as the strains used are those for which relatively complete genome sequences are available.

While our work was in progress, Celera incorporated the public B6 sequence into its mouse genome assembly and developed a searchable database of mouse single nucleotide polymorphisms (SNPs) [15]. Comparison of the gene variants identified by our approach to the SNPs identified by a search of Celera's SNP database indicates that the latter significantly underestimates the number of SNPs present and that the method described here provides the most complete and accurate compilation of QTL gene variants.

We tested this approach on two QTLs involved in physical dependence on ethanol; these are located on murine chromosomes 1 and 13 and were identified using two mapping populations derived from the B6 and D2 progenitor strains. The well documented difference in susceptibility to withdrawal after chronic ethanol exposure between the B6 and D2 inbred mouse strains provided an excellent starting point for dissecting genetic influences involved in physical dependence on ethanol and to study how common allele variants influence genetic predisposition to physical dependence on ethanol. Genome-wide evaluations of chronic ethanol withdrawal convulsions identified QTLs on murine chromosomes 1, 13, 19 and 4 [16]. The QTLs on chromosome 1 and chromosome 13 (a sex-limited QTL) were detected in two mapping populations, that is, the BXD recombinant inbred strains and a B6D2 F2 intercross [16,17].

Results

A flow chart showing the steps involved in the *in silico* comparison of the coding region of genes within the two alcohol withdrawal QTLs is shown in Figure 1. The National Center for Biotechnology Information (NCBI)/Mouse Genome Database (MGD) human/mouse homology map and the Celera mouse genome assembly were searched to obtain a list of genes and predicted genes within each QTL. The selected genes included plausible candidate genes that have been implicated in functions relevant to alcohol action, as well as genes whose functions are not well understood or were not previously suspected to be related to alcohol action. For the QTL on chromosome 1, a list of 121 genes was

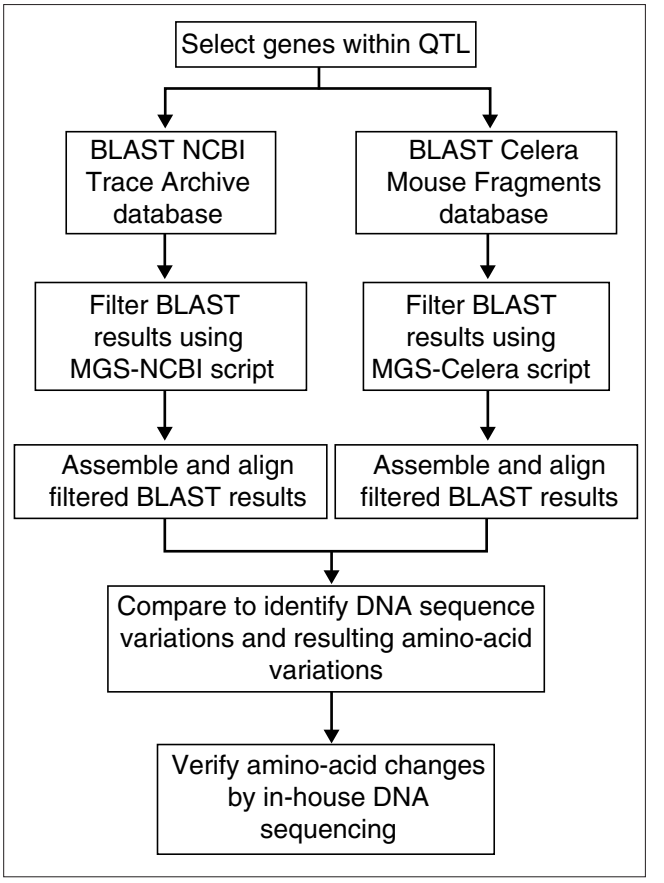

Figure 1
Strategy for *in silico* identification of coding-sequence variations within QTLs.

selected, most of which are within 4 million bases (Mb) upstream or 4 Mb downstream of the peak log of odds (LOD) score (approximately 18-24 centimorgans (cM) from the centromere). For chromosome 13, a list of 78 genes was selected, most of which are within 2 Mb upstream or 2 Mb downstream of the peak LOD score (approximately 37 cM from the centromere). The complete coding region of each gene, as available, was retrieved either from Celera or NCBI to be used as the query for subsequent strain-specific BLAST searches [18]. B6 sequences were retrieved from the mmtrace database at NCBI using a MegaBLAST search, and D2 sequences were retrieved from Celera by BLAST search of the All Mouse Fragments (masked) database. All data were retrieved from Celera and NCBI between August 2001 and April 2002. Two in-house programs were used to parse the BLAST results and remove any hits with low nucleotide-similarity scores or an insignificant e-value. The filtered BLAST results from the B6 and D2 strains were then aligned using AutoAssembler and allelic sequence variations were detected.

On proximal chromosome 1, 121 genes covering 125,385 base-pairs (bp) were selected and the sequences aligned, and for chromosome 13, 78 genes covering 84,635 bp were selected and aligned. Altogether, the alignment encompassed 199 genes spanning 210,020 bp. Because of some gaps that remain in the available genome sequences for the B6 and D2 strains, complete coding regions were sometimes not available and in such cases the coverage was therefore incomplete. Our calculations indicate that the *in silico* coverage of the coding region of the 199 genes for the B6 and D2 strains was 86% and 69%, respectively. The lower percent coverage of the D2 genome is likely to be due to the fact that it is only one out of four mouse strains used to generate Celera's 5.2x coverage of the mouse genome. When the B6 and D2 data are combined, 61% of the length of the coding regions was covered by both genome sequences (Table 1).

Previous efforts to identify human SNPs by database mining found it important to filter results in order to eliminate the false detection of SNPs [19]. With this in mind, we took the following steps to ensure the quality of SNPs detected in our data. First, to avoid the inclusion of sequences from additional closely related paralogous genes, we used a BLAST hit cutoff of 98%, so that only reads at least 98% similar to the query sequence were selected. Furthermore, reads that potentially affected amino acid sequence were used for a BLAST search against the mouse genome and removed from the B6/D2 alignment if the best BLAST hit was on a sequence scaffold located in another part of the genome. Second, to avoid the inclusion of changes due to poor-quality sequence, Phred [20] scores of individual B6 reads were checked at changed nucleotide positions and poor-quality changes (Phred score ≤ 10) eliminated. This step was not done with D2 sequences because individual sequence traces and Phred scores were not available from Celera. Third, by in-house sequencing of 44 gene coding variations found in the *in silico* alignments of three genes from B6, D2, A/J, 129X1/SvJ, and 129S1/SvImJ DNA, we verified that Celera had not mislabeled the strain from which sequence reads were derived and showed that no mix-up or cross-contamination of samples had occurred. Even after the above steps were taken, there were still instances where some reads had one sequence and other reads had a slightly different sequence for the same strain. In-house sequencing of B6 and D2 DNA for a subset of interstrain frame shifts (10% of the total) showed that all the apparent differences were false. In-house sequencing for a subset of intrastrain single base-pair substitutions (21% of the total) was also carried out. This analysis showed that when two different intrastrain sequences (for example, in the D2 strain) were detected, the correct sequence was always identical to the other strain (for example, B6), so that no interstrain variant appeared at that position. On the basis of this in-house sequence analysis we chose to operationally classify all frameshifts as false and interstrain substitutions present in 50%, or fewer than 50%, of the reads for one strain as false.

The number of nonsynonymous changes was totaled from all 199 genes within the chronic alcohol withdrawal QTLs on proximal chromosome 1 and mid-chromosome 13 for which allelic sequences were aligned, together covering 7 cM and 12 Mb of the mouse genome and representing 210,019 bp of coding region. In total, 21 of the 199 genes surveyed from the chromosome 1 and 13 QTL regions showed nonsynonymous

Table 1

Summary of genes surveyed within chronic alcohol withdrawal QTLs on proximal chromosome 1 and chromosome 13

	Number of genes	Base-pairs	% Coverage*			Number of amino-acid changes†	Number of genes with at least one amino-acid change†
			B6	D2	Both		
Chromosome 1	121	125,385	84 (85)	77 (80)	67 (70)	8 (8)	6 (6)
Chromosome 13	78	84,635	89 (90)	57 (59)	52 (55)	36 (45)	15 (19)
Total	199	210,020	86 (87)	69 (71)	61 (64)	44 (53)	21 (25)

*The first number listed indicates *in silico* percent coverage; numbers in parentheses indicate percent coverage after in-house sequencing was done to fill in gaps in the *in silico* alignment. †The first number indicates changes where multiple reads were available in each strain or the change has been verified by in-house sequencing; the number in parentheses includes changes where only one read was available for either B6 or D2.

changes in their coding regions between B6 and D2 animals (Table 1). A search of Celera's SNP database found approximately half of the nonsynonymous B6/D2 sequence changes which we identified by multiple alignment of BLAST hits. Of the nonsynonymous changes that we identified by multiple alignment of BLAST hits and verified by in-house sequencing, fewer than half are listed in the SNP database.

On chromosome 1, where 121 genes were compared, eight amino-acid changes were identified in six genes (Table 2). Of these, two are known genes and four are based only on computational predictions by Celera. One of the two known genes for which allelic variation was identified within its coding region sequence was *Slc5a7*. This encodes a high-affinity choline transporter expressed in the brain and spinal cord and may be considered a plausible candidate gene for underlying the QTL (see Discussion).

On chromosome 13, when looking only at changes that have multiple read coverage or were verified by in-house sequencing, 36 amino-acid changes were detected within 15 genes (Table 2). Nine additional changes in four other genes within the chromosome 13 QTL were also identified, but these were all of limited reliability in that only a single sequence read was available for either the B6 or D2 allele. Of the *in silico* predicted changes based on single-read coverage that were sequenced in house, approximately half were proven false. The proteins showing allelic differences with multiple read coverage span a range of functional classes and include a zinc metalloprotease, 60S ribosomal protein, hyaluronic acid binding protein (gene *Habp4*), developmental related protein (gene *Ptch*), transport system kinase and several genes for which no functional assignments have yet been made. As discussed below, three genes for which allelic variations were detected, *Srd5a1*, *Nrif-2*, and *Hsd17b3* are thought to be particularly promising biological candidates to underlie the chromosome 13 QTL.

Discussion

While other studies have mined EST databases to detect SNPs [19], the work presented here is the first demonstration of the use of strain-specific genome sequence databases to discover gene-coding variants for murine QTLs. Whereas previous studies have focused on detection of SNPs in expressed sequence tags (ESTs) that map to locations throughout the genome for use as genetic markers, the present work searched strain-specific genomic sequence databases for coding-region SNPs underlying disease-related QTLs. In this first application of the strategy, we surveyed 199 QTL-localized genes covering approximately 7 cM and 12 Mb of the mouse genome and identified 21 genes that had altered coding regions between the B6 and D2 progenitor strains.

At least one of the allelically variant genes - *Slc5a7*, detected within the proximal chromosome 1 QTL interval - is a promising biological candidate gene. *Slc5a7* encodes a high-affinity choline transporter expressed in cholinergic neurons in the brain and spinal cord. Prolonged ethanol exposure decreases high-affinity choline uptake in rat cerebral cortex [21], and acute ethanol exposure decreases choline transport in erythrocytes [22]. There was a single amino-acid change of arginine (B6) to histidine (D2) at position 38, which is located in the first cytoplasmic loop of the transporter, with the arginine also being conserved in rat and human. However, while synteny data between human and mouse suggest that the gene is within the QTL, data from Celera apparently places it on another chromosome. Therefore the question of whether *Slc5a7* remains a good candidate for the proximal 1 QTL must await resolution of this issue.

Three genes for which allelic variations were detected - *Srd5a1*, *Nrif-2* and *Hsd17b3* - appear to be particularly promising biological candidates to underlie the chromosome 13 QTL. The most compelling of these is *Srd5a1*, which encodes a steroid 5α-reductase 1 and is expressed in brain and is consequently termed a neurosteroid. 5α-reduced metabolites of the neurosteroids are thought to be involved in myelination [23], and one of the key features of alcoholism is the loss of cerebral white matter [24]. Steroid 5α-reductase is also required for the reduction of progesterone to 5α-dihydroprogesterone, which is further metabolized to 3α,5α-progesterone (allopregnanolone), which has neuroactivity at $GABA_A$ receptors [25]. Some symptoms of ethanol withdrawal appear to be produced through neuroadaptive changes in GABA-mediated neurotransmission [26]. It is known that many neuroactive steroids have anticonvulsant activity [27], and during ethanol withdrawal the B6 and D2 mice have differential sensitivity to the anticonvulsant effects of allopregnanolone [28]. We identified three amino acid changes within this gene between the B6 and D2 strains, one of which occurs at position 7 with arginine (B6) being replaced by cysteine (D2). Unlike arginine, cysteine can participate in disulfide bridge formation and thus has the potential to produce a marked difference in the functional activity of 5α-reductase 1 in B6 compared to D2 animals.

Another candidate gene is *Nrif-2*, a Krüppel-type zinc finger protein thought to function as a transcriptional repressor [29]. One of the identified amino-acid changes replaces the arginine at position 593 in the D2 strain with a glutamine residue in B6 mice, and this occurs within a known binding site for the intracellular domain of the neurotrophin receptor p75NTR [30]. The *Nrif-2* gene is expressed in brain and several other tissues, but expression is highest in testis. Interestingly, the chromosome 13 QTL is sex-limited and affects severity of physical dependence on alcohol only in males [16]. In males the D2 allele was associated with more severe withdrawal responses than the B6 allele, whereas the opposite direction of effect was found in female mice. Because the behavioral trait is sex-limited, this suggests that the actions of the underlying gene are sex-limited. Such a phenotype might

Table 2

QTL genes with nonsynonymous changes

(a) Altered genes within the chronic alcohol withdrawal QTL on proximal chromosome 1

	Gene		*In silico* percent coverage*			
Symbol	Gene ID	Location	B6	D2	Both	Amino-acid variations
	mCG55019	30,510,903 - 30,520,321	100	100	100	M72V†
	mCG52566	31,637,483 - 31,666,730	95	95	95	S6A
	mCG49360	32,140,873 - 32,141,648	100	87	87	P74L†, C123R†
	mCG66997	36,796,676 - 36,805,521	100	100	100	stop61Y
	mCG4911	39,571,491 - 39,572,134	54	100	54	R74H, E79D
Slc5a7	NM_022025	Uncertain	100	99 (100)	99 (100)	R38H†

(b) Altered genes within the sex-limited chronic alcohol withdrawal QTL on chromosome 13

	Gene		*In silico* percent coverage*			
Symbol	Gene ID	Location	B6	D2	Both	Amino-acid variations
	mCG22018	54,248,540 - 54,266,542	100	86	86	I155F
Ptch	mCG21490	54,702,052 - 54,733,739	79	68	68	T1267N
	mCG67381	54,910,868 - 54,921,322	95	95	95	N48D
	mCG49128	55,141,914 - 55,156 036	82	100	82	S38G, H96R
Hsd17b3	NM_008291	55,258,374 - 55,287,747	69 (100)	61 (100)	45 (100)	C95R†
	mCG67239	55,349,545 - 55,350,013	100	100	100	T47M, P124R‡
Habp4	NM_019986	55,358,811 - 55,383,494	89 (96)	90 (96)	84 (96)	A65- (in/del)†, P187L†, A266T†
	mCG65010	55,397,330 - 55,399,305	84	100	84	S48N†, R76W†
	mCG51118	55,652,165 - 55,652,438	31	100	31	F6L‡
	mCG50935	55,841,328 - 55,842,516	100	100	100	A210V
	mCG50411	56,005,687 - 56,054,181	100	66	66	G200E†
	mCG55714	56,067,926 - 56,074,495	96	90	90	S99P‡, L361S‡, Q557R‡, A851V‡
	mCG59053	56,106,754 - 56,107,033	84	100	84	S64G‡
	mCG59052	56,122,823 - 56,169,063	93	86	79	W53R, M62V, T171, V197A, T198P, A324S, D645N, R728G, W846R†
Nrif-2	AJ319726	56,376,102 - 56,425,620	96 (100)	78 (100)	74 (100)	A491T†, Q593R†, I679V†
Ptdss1	mCG67987	57,190,261 - 57,251,587	100	89	89	G322C†
	mCG67995	57,426,346 - 57,427,146	100	83	83	G4E
	mCG67997	57,732,668 - 57,740,675	53 (100)	100	53 (100)	K24T†, L30R†, S36P†, I48S†, K62R†, V89E†, N94T†
Srd5a1	mCG7698	58,982,812 - 59,009,801	78 (100)	78 (100)	78 (100)	R7C†, V142I†, E176D†

Gene IDs beginning with the letters mCG are Celera gene ID numbers; all other gene ID numbers are GenBank accession numbers for transcript sequences. Although genes are identified by the Celera genomic (mCG) ID, transcript sequence was used as query for the BLAST searches for each gene. All data were retrieved from Celera and NCBI between August 2001 and April 2002. Only changes where the reading frame is known and sequence is available in both B6 and D2 are displayed. The B6 amino acid is given first, then the position, followed by the D2 amino acid. *Numbers in parentheses indicate percent coverage after in-house sequencing. †These changes were verified by in-house sequencing. ‡These changes are supported by only a single read in either B6 or D2 and have not been verified by in-house sequencing.

be expected if the causal variant gene was expressed and/or active in gonadal tissue, as is the case with *Nrif-2*.

A third altered candidate gene within the chromosome 13 QTL is *Hsd17b3*. This gene encodes a 17β-hydroxysteroid dehydrogenase expressed primarily in testis but also in several other tissues, including brain [30]. *Hsd17b3* catalyzes the interconversion of androstenedione and testosterone, favoring the reduction reaction [31]. Ethanol exposure has been shown to inhibit the reduction of

androstenedione to testosterone in rat Leydig cells *in vitro* [32]. The amino acid change within this gene replaces a cysteine (B6) at residue 95 with an arginine (D2).

While these data are consistent with the variant genes identified on chromosomes 1 and/or 13 being the gene that underlies the respective QTL, it is also possible that the polymorphisms identified in these genes are merely linked to the polymorphisms that actually underlie these two QTLs. Although *Slc5a7*, *Nrif-2*, *Srd5a1*, and *Hsd17b3* are promising candidates to underlie these two QTLs, a significant number of altered genes were also identified that cannot formally be ruled out at this time, either in place of, or in addition to, these three genes. Additional fine mapping of these QTLs is needed to eliminate as many false candidate genes as possible from these QTL intervals. In addition, definitive confirmation that *Slc5a7*, *Nrif-2*, *Srd5a1* and *Hsd17b3* (or other promising candidate genes if they arise) are involved in predisposition to physical dependence on ethanol will probably require verification using transgenics (for example, allele substitution or bacterial artificial chromosome transgenics).

Another feature of the approach described here is that it not only rapidly identifies gene variants within QTL intervals but, because it will find many genes that are unchanged between strains, can also quickly eliminate large numbers of gene-coding regions as possible candidates underlying the QTL. Finally, it should be pointed out that even when a synonymous sequence difference is found between QTL strains, it can still be used as a new, easily scored marker to further narrow the QTL interval using fine-mapping resources such as interval-specific congenic recombinant mice [33]. In this regard we identified many such synonymous changes between B6 and D2 that fall within the proximal chromosome 1 or mid-chromosome 13 QTLs, and therefore can serve as new informative markers for fine mapping of these QTLs.

The approach described here has considerable general applicability. Many other murine QTLs, encompassing a wide range of complex phenotypes of interest, have been identified using mapping populations derived from two of the five mouse strains for which whole-genome draft sequences are now available, and therefore lend themselves to this *in silico* approach. When we carried out a PubMed literature database search, 276 QTLs, corresponding to 57 different phenotypes, were identified that utilized two of these five mouse strains (Table 3). The QTL intervals that were used in the present study were rather large, and many other QTLs that use strains amenable to an *in silico* approach are considerably smaller.

While the present study focused on QTL gene-coding region comparisons because of their potential functional importance [5], it should be pointed out that gene-regulatory regions, which are also of considerable functional importance and have been shown to underlie some QTLs [34,35], can be studied in the same way. For example, the approach described here can be combined with data from high-density DNA microarray studies of the B6 and D2 strains (J.M.S., unpublished work). In such a scenario, interstrain comparison of the expression levels of tens of thousands of genes would be carried out, the genes that are differentially expressed between strains identified, and their chromosome location determined from the genome databases. Those differentially expressed genes that map to a QTL region would be selected and the upstream regulatory regions aligned *in silico* to search for sequence differences that affect predicted transcriptional binding sites. Such binding-site differences, when present, would be promising candidates to underlie the QTL in which the gene resides. Although less is known about sequences affecting gene regulation than about those affecting protein sequence, there has been considerable progress in the past few years in our understanding of the sequences involved in controlling gene regulation [36,37].

It is also possible that sequence changes in the noncoding regions of mRNAs can have functional effects. For example, sequence in the 5′ untranslated region (UTR) can affect mRNA translation, and sequence in the 3′ UTR can affect mRNA stability. The 3′ UTR region can also contain enhancer elements that affect gene regulation and expression. Whereas less is known about the functional relevance of sequence changes in noncoding parts of mRNAs than those in the protein-coding region, they can nevertheless be functionally important and, as such, could be included in surveys using the *in silico* approach described here.

While Celera's SNP database has the ability to quickly identify interstrain SNPs in QTL genes, the information available from the SNP database is not as complete as the data generated by our *in silico* approach. Specifically, Celera uses a different strategy for assigning SNPs in the mouse genome and the SNP database makes very conservative SNP predictions; fewer than half of the SNPs we identified *in silico* and verified by in-house sequencing are in the SNP database. The SNP database also misclassifies some SNPs as missense mutations when the SNPs are actually in the 5′ or 3′ UTR (J.M.S., unpublished data). And finally, the Celera SNP database does not indicate which regions of a gene's coding sequence are covered by reads in each strain, and therefore does not distinguish between completely sequenced regions in which no variants have been detected and regions where no variants are found because sequence coverage for one of the strains is absent. Such a limitation provides an incomplete picture of the interstrain variants that exist in a given QTL, which in turn could result in passing over the critical SNP underlying the QTL.

Although the strategy presented here provides a significant new tool in going from QTL to gene, a number of challenges still remain on the way to realizing this goal. For

Table 3

Phenotypes for which QTLs have been generated using strains that have whole genome draft sequence available

	Phenotype	Number of QTLs (B6xD2)
1	Alcohol drinking - preference	7
2	Alcohol drinking - acceptance	1
3	Alcohol conditioned taste aversion	3
4	Cocaine seizures	3
5	Morphine preference	1
6	Morphine analgesia	2
7	Screen test sensitivity	8
8	Ethanol-induced locomotor activity	3
9	Basal activity	1
11	Acute alcohol withdrawal	5
12	Acute pentobarbital withdrawal	3
13	Chronic alcohol withdrawal	3
14	LORR duration	4*
15	LORR Bec	14*
16	Free-choice ethanol consumption	6*
17	Behavioral effect of stress	7*
18	Porphyria	3
19	Liver injury and porphyria	3
20	Lymphocyte proportion B220 (%)	2
21	Lymphocyte proportion CD4 (%)	1†
22	Lymphocyte proportion CD8 (%)	1
23	Hepatic lipid peroxidation potential	1
24	Peak bone mass	16*
25	Variation in TRBV4 expansion	2
26	Maximal electroshock seizure threshold	4
27	Variation in cerebellar size	4
28	Variation in IGL volume	1
29	Susceptibility to mycobacterium tuberculosis	2, 1†
30	Acute behavioral sensitivity to paraoxon	6*
31	Circadian period of locomotor activity	1†
32	Short-term intake of saccharin	1
33	Short-term intake of sucrose	1
34	Short-term intake of quinine	1
35	Testicular weight	1
36	Bone density	6
37	Contextual fear conditioning	5
38	Hypothermic sensitivity to quinpirole	1
39	Tolerance to quinpirole	1†
40	Sensitivity and tolerance to quinpirole-induced hypothermia	1†
41	Baseline locomoter activity	3†
42	Locomotor sensitivity to quinpirole	1†
43	Hypnotic sensitivity to ethanol	1*, 1†
44	Body weight	1
45	Tail length	1
46	Methamphetamine responses	
	Chewing	5
	Climbing	14
	Home cage locomotor activity	20
	Body temperature change	22
47	Pcp-induced behavior	3*
48	Amp-induced behavior	6*
49	Antinociceptive responsiveness to N_2O	7
50	Corticosterone response to ethanol	5
51	High-affinity choline uptake	1†
	Phenotype	**Number of QTLs (B6xA/J)**
52	Cocaine-induced locomotor activity	2
53	Seizure susceptibility (beta-CCM administration)	3
54	Habituated open field behavior	2, 5†
55	Hormone-induced ovulation rate	2, 5†
56	Light-to-dark transition behavior	3, 5†
57	Center time behavior	5†
58	Initial ambulation in the open field	1, 4†
59	Vertical rearings	1, 2†
60	Trypanosomosis resistance	1
61	Nickel-induced acute lung injury	1, 2†
62	PKC-alpha content	1
63	PKC activity	1
	Total number of QTLs	276

*QTLs identified as suggestive in the original publications. †QTLs identified as provisional in the original publications.

example, functional confirmation that an interstrain gene alteration underlies a QTL, while feasible, is not trivial. As more QTL-localized genes are identified that have potentially important sequence changes, as described here for example, it becomes more difficult to test them all functionally. A related ongoing development that should help this situation by reducing the number of potential candidate genes that remain within the QTL interval is reduction of the interval by fine mapping using specialized congenic strains or recombinant progeny testing [33,38]. A number of interval-specific congenic strains have recently been developed for a number of QTLs and, although straightforward, fine mapping is still a large-scale enterprise because of the need to generate and test large numbers of animals to identify and behaviorally assess informative recombinants, and because of the need for markers to detect such recombination events.

While such challenges remain, the new genomics tools and resources that are now available or are becoming available are clearly providing a major impetus to this field. The power of high-density microarray studies to identify QTL-related genes and pathways, draft genomic sequences for the primary mouse strains used for QTL analysis, and evolving bioinformatics tools and databases are all making major impacts on this field, with more improvements to be expected. For example, after this work was completed, we used the *in silico* approach on a QTL identified using B6 and A/J strains and found that coverage was considerably higher than reported in this study. We found that in an analysis of over 30 QTL genes the *in silico* coverage for B6 and A/J was 98% and 91%, respectively, with a total coverage for both of 90% (J.M.S., unpublished data). Finally, the anticipated completion of the B6 genome sequence and deeper sequence coverage of the D2 and A/J genomes, in particular, if carried out would remove the last remaining obstacles to virtually complete *in silico* discovery of gene variants for the several hundred QTLs identified using these strains.

Materials and methods

Identification of genes within QTL regions

A list of known genes within the 1.0 LOD confidence interval for each QTL was found using the NCBI vs MGD human/mouse homology maps at NCBI [39]. This list was screened for candidate genes on the basis of gene expression within the brain and functional relevance to the QTL phenotype. GenBank accession numbers for the complete coding region sequence of each candidate gene were found on LocusLink. Coding region sequences were retrieved from GenBank and used as the query for BLAST searches.

A more complete list of genes and gene predictions within the QTL was found on the Celera Discovery System [40]. The genetic marker with the highest LOD score for a QTL was localized on Celera's physical map and a list of genes and predicted genes within an interval on either side of this marker was generated and screened for candidate genes. Coding region sequences of the selected genes were retrieved either from Celera or NCBI and used as the query for BLAST searches.

BLAST searches

A MegaBLAST search of the mmtrace database on NCBI [41] was used to retrieve B6 sequences from the public database. Low complexity was filtered, percent identity was set to 98%, and the number of alignments to return was set to 100. Results were saved as a text file. A BLASTN search of the All Mouse Fragments (masked) database on Celera [42] was used to search for D2 sequences. The e-value was set to 0.1, and the number of hits to return was set to 250. The complete entries view of the results was exported to a file which was given a .txt extension.

JAVA scripts to process BLAST results

A script, named Mouse Gene Selection NCBI (MGS_NCBI), was written in JAVA and used to parse BLAST results from NCBI. This script filters the BLAST results and only selects hits which have percent identity greater than or equal to 98% and e- value, which is a measure of the statistical significance of BLAST hits, less than or equal to 0.0001. The e-value is the number of hits with the same degree of similarity one would expect to find by chance if there were no true matches in the database. Another script, Mouse Gene Selection Celera (MGS_Celera), was written in JAVA and used to parse BLAST results from Celera. This script filters the BLAST results and only selects hits with strain name DBA/2J, e-value less than or equal to 0.0001 and percent identity greater than or equal to 98%. After filtering the BLAST results, the script saves each alignment as an individual text file. When one hit aligns with more than one region of the query sequence, the entire hit is discarded if any of the alignments does not meet the above criteria. As both scripts are written in JAVA, they can be run on Unix, Macintosh or PC platforms.

Alignment of filtered BLAST results

Filtered BLAST results were aligned using AutoAssembler 2.1. For each gene, one alignment file was made of the filtered B6 BLAST results and a separate alignment file was made of the filtered D2 BLAST results. The B6 alignment file was then compared to the D2 alignment file to identify sequence changes. When sequence variations were found within a strain they were regarded as true only if more than half the reads for a strain contained the change. This criterion was checked for validity by in-house sequencing.

In-house sequencing

In-house sequencing was used to test *in silico* sequence predictions and also occasionally to fill in gaps that remained in certain genes. C57BL/6J and DBA/2J mice were obtained from the Jackson Laboratory. Animals were euthanized and the brains dissected. Whole-brain RNA was isolated using an RNeasy Maxi Kit (Qiagen). cDNA was synthesized by RT-PCR using SUPERSCRIPT First Strand Synthesis System for RT-PCR (Invitrogen). C57BL/6J, DBA/2J, A/J, 129X1/SvJ, and 129S1/SvImJ genomic DNAs were obtained from the Jackson Laboratory. Primer3 software [43] was used to design primers to amplify either cDNA or genomic DNA. Template DNA was amplified by PCR in the Perkin Elmer GeneAmp PCR system 9700. A typical reaction cycle was denaturation for 3 min at 94°C followed by 35 cycles of denaturation for 15 s at 94°C, annealing for 1 min 15 s at 54°C, and extension for 1 min 15 s at 72°C with a final extension step at 72°C for 10 min. Products were run on a 1.7% agarose gel, stained with ethidium bromide, and purified using QIAquick Gel Extraction kits (Qiagen). PCR products were reamplified when necessary.

PCR products were sequenced using ABI Big Dye Terminator v.2 chemistry and cycle sequencing and sequence traces

read on an ABI PRISM 3100 Gene Analyzer. When in-house sequencing was done to fill gaps from the *in silico* alignment, a minimum of one forward read and one reverse read was run for each sample of PCR products. When-in house sequencing was done to verify nonsynonymous changes found in the *in silico* alignment, a minimum of two forward reads and two reverse reads was run for each sample of PCR products. Sequencing data for each strain were analyzed with the Consed software suite [44]. Sequencing was repeated for regions of poor quality (that is, Phred score < 40). High-quality sequence was exported and B6 sequence was compared to D2 sequence using AutoAssembler to identify changes between the two strains.

References

1. Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Lincoln SEL, Tanksley SD: **Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments.** *Genetics* 1991, **127:**181-197.
2. Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW: **Dissecting the architecture of a quantitative trait locus in yeast.** *Nature* 2002, **416:**326-330.
3. Mackay TF: **Quantitative trait loci in Drosophila.** *Nat Rev Genet* 2001, **2:**11-20.
4. Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, Andersson-Eklund L, Edfors-Lilja I, Fredholm M, Hansson I, *et al.*: **Genetic mapping of quantitative trait loci for growth and fatness in pigs.** *Science* 1994, **263:**1771-1774.
5. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, *et al.*: **Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition.** *Genome Res* 2002, **12:**222-231.
6. Crabbe JC, Phillips TJ, Buck KJ, Cunningham CL, Belknap JK: **Identifying genes for alcohol and drug sensitivity: recent progress and future directions.** *Trends Neurosci* 1999, **22:**173-9.
7. Kreutz R, Hubner N, James MR, Bihoreau MT, Gauguier D, Lathrop GM, Lindpaintner K: **Dissection of a quantitative trait locus for genetic hypertension on rat chromosome 10.** *Proc Natl Acad Sci USA* 1995, **92:**8778-8782.
8. Garrett MR, Rapp JP: **Multiple blood pressure QTL on rat Chromosome 2 defined by congenic Dahl rats.** *Mammalian Genome* 2002, **13:**41-44.
9. Flint J, Mott R: **Finding the molecular basis of quantitative traits: successes and pitfalls.** *Nat Rev Genet* 2001, **2:**437-445.
10. Zhu X, McKenzie CA, Forrester T, Nickerson DA, Broeckel U, Schunkert H, Doering A, Jacob HJ, Cooper RS, Rieder MJ: **Localization of a small genomic region associated with elevated ACE.** *Am J Hum Genet* 2000, **67:**1144-1153.
11. Knoblauch H, Muller-Myhsok B, Busjahn A, Ben Avi L, Bahring S, Baron H, Heath SC, Uhlmann R, Faulhaber HD, Shpitzen S, *et al.*: **A cholesterol-lowering gene maps to chromosome 13q.** *Am J Hum Genet* 2000, **66:**157-166.
12. Zimatkin SM, Deitrich RA: **Aldehyde dehydrogenase activities in the brains of rats and mice genetically selected for different sensitivity to alcohol.** *Alcohol Clin Exp Res* 1995, **19:**1300-1306.
13. Ehringer MA, Thompson J, Conroy O, Xu Y, Yang F, Canniff J, Beeson M, Gordon L, Bennett B, Johnson TE, *et al.*: **High-throughput sequence identification of gene coding variants within alcohol-related QTLs.** *Mammalian Genome* 2001, **12:**657-663.
14. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, *et al.*: **A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome.** *Science* 2002, **296:**1661-1671.
15. **Celera Mouse Reference SNP Database** [http://www.celera.com/genomics/academic/home.cfm?ppage=cds&cpage=mousesnps
16. Buck KJ, Rademacher BS, Metten P, Crabbe JC: **Mapping murine loci for physical dependence on ethanol.** *Psychopharmacology* 2002, **160:**398-407.
17. Crabbe JC: **Provisional mapping of quantitative trait loci for chronic ethanol withdrawal severity in BXD recombinant inbred mice.** *J Pharmacol Exp Ther* 1998, **286:**263-271.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
19. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M: **Mining SNPs from EST databases.** *Genome Res* 1999, **9:**167-74.
20. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.
21. Floyd EA, Young-Seigler AC, Ford BD, Reasor JD, Moore EL, Townsel JG, Rucker HK: **Chronic ethanol ingestion produces cholinergic hypofunction in rat brain.** *Alcohol* 1997, **14:**93-98.
22. Deves R, Krupka RM: **Inhibition of choline transport in erythrocytes by n-alkanols.** *Biochim Biophys Acta* 1990, **1030:**32-40.
23. Celotti F, Melcangi RC, Martini L: **The 5 alpha-reductase in the brain: molecular aspects and relation to brain function.** *Front Neuroendocrinol* 1992, **13:**163-215.
24. Kril JJ, Halliday GM, Svoboda MD, Cartwright H: **The cerebral cortex is damaged in chronic alcoholics.** *Neuroscience* 1997, **79:**983-998.
25. Lambert JJ, Belelli D, Hill-Venning C, Peters JA: **Neurosteroids and $GABA_A$ receptor function.** *Trends Pharmacol Sci* 1995, **16:**295-303.
26. Buck KJ: **New insight into the mechanisms of ethanol effects on $GABA_A$ receptor function and expression, and their relevance to behavior.** *Alcohol Clin Exp Res* 1996, **20:**198A-202A.
27. Paul SM, Purdy RH: **Neuroactive steroids.** *FASEB J* 1992, **6:**2311-2322.
28. Finn DA, Gallaher EJ, Crabbe JC: **Differential change in neuroactive steroid sensitivity during ethanol withdrawal.** *J Pharmacol Exp Ther* 2000, **292:** 394-405.
29. Benzel I, Barde YA, Casademunt E: **Strain-specific complementation between NRIF1 and NRIF2, two zinc finger proteins sharing structural and biochemical properties.** *Gene* 2001, **281:**19-30.
30. Sha JA, Dudley K, Rajapaksha WR, O'Shaughnessy PJ: **Sequence of mouse 17beta-hydroxysteroid dehydrogenase type 3 cDNA and tissue distribution of the type 1 and type 3 isoform mRNAs.** *J Steroid Biochem Mol Biol* 1997, **60:**19-24.
31. Luu-The V, Zhang Y, Poirier D, Labrie F: **Characteristics of human types 1, 2 and 3 17 beta-hydroxysteroid dehydrogenase activities: oxidation/reduction and inhibition.** *J Steroid Biochem Mol Biol* 1995, **55:**581-587.
32. Widenius TV, Orava MM, Vihko RK, Ylikahri RH, Eriksson CJ: **Inhibition of testosterone biosynthesis by ethanol: multiple sites and mechanisms in dispersed Leydig cells.** *J Steroid Biochem* 1987, **28:**185-188.
33. Fehr C, Shirley RL, Belknap JK, Crabbe JC, Buck K: **Congenic mapping of alcohol and pentobarbitol withdrawal liability quantitative trait loci to a <1 cM region of mouse chromosome 4: identification of Mpdz as a candidate gene.** *J Neurosci* 2002, **22:**3730-3738.
34. Lyman RF, Lai C, Mackay TF: **Linkage disequilibrium mapping of molecular polymorphisms at the scabrous locus associated with naturally occurring variation in bristle number in *Drosophila melanogaster*.** *Genet Res* 1999, **74:**303-311.
35. Wang RL, Stec A, Hey J, Lukens L, Doebley J: **The limits of selection during maize domestication.** *Nature* 1999, **398:**236-239.
36. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Res* 2000, **28:**316-319.
37. Workman CT, Stormo GD: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity.** *Pac Symp Biocomput* 2000, 467-478.
38. Darvasi A: **Interval-specific congenic strains (ISCS): an experimental design for mapping a QTL into a 1-centimorgan interval.** *Mammalian Genome* 1997, **8:**163-167.
39. **NCBI Human/Mouse Homology Maps** [http://www.ncbi.nlm.nih.gov/Homology]
40. **Celera Discovery System BioMolecule Library** [http://cds.celera.com/biolib/cdsTopLibrary.jsp]
41. **NCBI Trace Archive Database MegaBLAST Search** [http://www.ncbi.nlm.nih.gov/blast/mmtrace.html]
42. **Celera Discovery System Sequence Analysis** [http://cds.celera.com/servlet/com.celera.web.cds.servlets.PortalProxy]
43. **Primer3** [http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi]
44. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8:**195-202.