nature genetics



# Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'

Leonid Bystrykh[1], Ellen Weersing[1], Bert Dontje[1], Sue Sutton[2], Mathew T Pletcher[2], Tim Wiltshire[2], Andrew I Su[2], Edo Vellenga[3], Jintao Wang[4,5], Kenneth F Manly[4,5], Lu Lu[5], Elissa J Chesler[5], Rudi Alberts[6], Ritsert C Jansen[6], Robert W Williams[5], Michael P Cooke[2] & Gerald de Haan[1]

**We combined large-scale mRNA expression analysis and gene mapping to identify genes and loci that control hematopoietic stem cell (HSC) function. We measured mRNA expression levels in purified HSCs isolated from a panel of densely genotyped recombinant inbred mouse strains. We mapped quantitative trait loci (QTLs) associated with variation in expression of thousands of transcripts. By comparing the physical transcript position with the location of the controlling QTL, we identified polymorphic *cis*-acting stem cell genes. We also identified multiple *trans*-acting control loci that modify expression of large numbers of genes. These groups of coregulated transcripts identify pathways that specify variation in stem cells. We illustrate this concept with the identification of candidate genes involved with HSC turnover. We compared expression QTLs in HSCs and brain from the same mice and identified both shared and tissue-specific QTLs. Our data are accessible through WebQTL, a web-based interface that allows custom genetic linkage analysis and identification of coregulated transcripts.**

The developmental potential of stem cells is tightly regulated by genetic and epigenetic factors that collectively define a stem cell–specific transcriptome. Irrespective of the tissue from which stem cells are isolated, they are typically defined by their extensive proliferative capacity, enabling rapid production of a large number of fully differentiated daughter cells. To ensure maintenance of their compartment, stem cells must undergo self-renewing divisions[1]. To identify key stem cell genes that specify this poorly understood process of self-renewal, several groups have embarked on genome-wide gene expression studies, comparing embryonic, neural and hematopoietic stem cells[2,3]. Although unique stem cell transcripts have been identified by each group, the overlap between the various data sets is limited[4]. Therefore, the remaining challenge is to delineate those unique transcriptional circuits in stem cells that collectively result in appropriate transitions in gene expression patterns and that distinguish stem cells from nonstem cell progeny.

In previous studies, we used a genetic approach to identify loci associated with variation in attributes of HSC populations[5,6]. We showed that HSCs isolated from the bone marrow of DBA/2 (D2) mice had higher turnover rates than those isolated from C57BL/6 (B6) mice. The variation in the percentage of cells in S phase is a cell-autonomous trait and is largely independent of cellular microenvironment, indicating that it originates from distinct gene expression patterns in HSCs themselves[7,8]. Using a large panel of BXD recombinant inbred (RI) strains of mice generated by crossing strains B6 and D2, we defined a QTL on chromosome 11 called stem cell proliferation-2 (*Scp2*) that modulates the percentage of cells in S phase[6]. The same locus was associated with the difference in mean mouse lifespan between these two strains[6], suggesting that increased stem cell turnover is one of the factors that underlie the aging process. The relevance of this 10-cM region in isolation was confirmed in an extensive analysis of backcrossed mice and, ultimately, in a congenic mouse model[9]. In humans, the corresponding region maps to 5q31.1. Deletions in this region are associated with myelodysplastic syndrome and acute myeloid leukemia[10,11], confirming the presence of unknown essential genes in this region that regulate stem cell behavior.

To identify candidate genes, we have now used a 'genetical genomics' approach. Genetical genomics entails an analysis of high-throughput transcript expression patterns in a pedigree of genetically distinct subjects in which variable levels of gene expression segregate. The concept of this technique was first suggested by one of us[12,13] and was recently shown to dissect transcriptional regulation successfully in fruit flies and yeast[14–16]. Here we used this new approach to identify variation in gene expression patterns in HSCs isolated from fully homozygous BXD RI strains of mice. In an accompanying paper, Chesler and colleagues dissected variation in expression profiles in forebrain of the same strains of mice[17]. One of the advantages of this approach is that for any given transcript, on average, half of all samples will carry the B6 allele whereas the other half will carry the D2 allele. Therefore, there is an inherently large number of replicate

[1]Department of Stem Cell Biology, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, the Netherlands. [2]Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. [3]Department of Hematology, Academic Hospital Groningen, Groningen, the Netherlands. [4]Molecular and Cellular Biology Department, Roswell Park Cancer Institute, Buffalo, New York, USA. [5]University of Tennessee Health Science Center, Memphis, Tennessee 38163, USA. [6]Groningen Bioinformatics Centre, University of Groningen, Groningen, the Netherlands. Correspondence should be addressed to G.d.H. (g.de.haan@med.rug.nl).

Published online 13 February 2005; doi:10.1038/ng1497

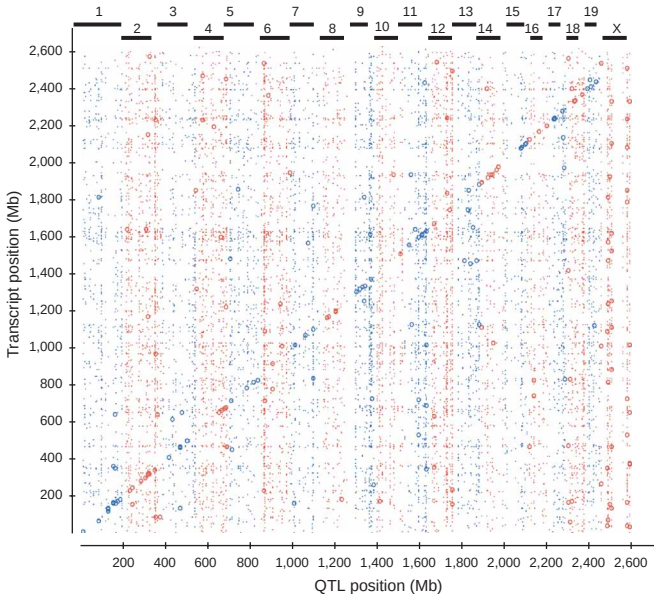


**Figure 1** Mapping QTLs that modulate gene expression in HSCs. The variation in transcript levels across 30 BXD HSC samples was correlated with the presence of B6 or D2 alleles at 779 loci throughout the genome. Each dot in the figure represents a single transcript. The physical position of each transcript is indicated on the *y* axis, and the position of the locus that is most strongly associated with variation of the corresponding transcript levels is shown on the *x* axis. Transcripts on the diagonal are *cis*-regulated (*i.e.*, modulated by a QTL in close proximity to the gene; **Table 1** and **Supplementary Table 2** online). To represent the data graphically, the entire mouse genome was aligned, resulting in a total genome size of ~2,600 Mb. Actual chromosomal positions are indicated at the top and highlighted by alternating red and blue coloring. Large circles represent transcripts with significant genome-wide linkage statistics ($P < 0.05$).

transcript-specific tests. Together with the fact that replicate sampling can be done easily using isogenic RI strains, this large number of tests increases the statistical power of this type of array experiments substantially[13]. Finally, by using a fixed reference population of RI strains, we can explore gene pleiotropy and tissue-specific expression patterns, in this case, by comparing HSCs to a population of forebrain neurons and glial cells.

## RESULTS

### Transcript QTLs in HSCs

We used highly purified Lin$^-$ Sca-1$^+$ c-kit$^+$ cells, containing all HSCs and a subset of more committed progenitors, from the bone marrow of female mice of 30 BXD strains. We deposited a limited number of purified single cells in microtiter plates using *in vitro* long-term bone marrow cultures to verify functional activity of each sample (**Supplementary Table 1** online). We isolated 16,000–118,000 stem cells from three mice per strain and isolated total RNA from ~10,000 cells, amplified using a linear amplification protocol and hybridized to Affymetrix U74Av2 oligonucleotide arrays.

We then compared the strain distribution pattern of each individual transcript with the genetic distribution of B6 and D2 alleles at 779 markers mapping throughout the genome using WebQTL (see URL

**Table 1 HSC transcripts showing strongest evidence of *cis* regulation**

| Gene | Probe set | Name | Transcript position (Mb) | QTL marker[a] | QTL chromosome | Marker position (Mb) | LRS[b] | Genome-wide *P* value[c] |
|---|---|---|---|---|---|---|---|---|
| *Srp9* | 101579_at | Signal recognition particle 9 kDa | 183 | *D1Mit426* | 1 | 181 | 40.954 | 0.00000 |
| *Ctse* | 104696_at | Cathepsin E | 132 | *D1Mit218* | 1 | 128 | 85.621 | 0.00000 |
| *Creg1* | 160502_at | Cellular repressor of E1A-stimulated genes | 166 | *D1Mit145* | 1 | 168 | 35.823 | 0.00000 |
| *Cd1d2* | 101896_at | CD1d2 antigen | 466 | *D3Mit155* | 3 | 467 | 53.468 | 0.00000 |
| *F2r* | 95474_at | Coagulation factor II (thrombin) receptor | 1.854 | *D13Mit145* | 13 | 1.854 | 45.042 | 0.00000 |
| *Cst3* | 99586_at | Cystatin 3 | 347 | *D2Mit423* | 2 | 347 | 42.038 | 0.00001 |
| *Ctsc* | 161251_f_at | Cathepsin C | 1.074 | *D7Mit350* | 7 | 1.070 | 47.264 | 0.00001 |
| *Runx1* | 92399_at | Runt related transcription factor 1 | 2.196 | *D16Mit86* | 16 | 2.196 | 30.157 | 0.00001 |
| *Cnih* | 97528_at | Cornichon homolog (*Drosophila*) | 1.918 | *D14Mit121* | 14 | 1.920 | 33.537 | 0.00002 |
| *Fli1* | 94698_at | Friend leukemia integration 1 | 1.296 | *D9Mit297* | 9 | 1.298 | 37.242 | 0.00003 |
| *Dctn6* | 160327_at | Dynactin 6 | 1.166 | *D8Mit294* | 8 | 1.172 | 34.101 | 0.00006 |
| *Ptprv* | 92662_g_at | Protein tyrosine phosphatase, receptor type, V | 135 | *D1Mit218* | 1 | 128 | 34.349 | 0.00008 |
| *Flot1* | 95095_at | Flotillin 1 | 2.237 | *D17Mit175* | 17 | 2.233 | 31.288 | 0.00008 |
| *Ccr2*[d] | 93397_at | Chemokine (C-C) receptor 2 | 1.389 | *D9Rp2* | 9 | 1.387 | 34.321 | 0.00019 |
| *Gcet2* | 101147_at | Germinal center expressed transcript | 2.148 | S16Gnf042.995 | 16 | 2.148 | 29.911 | 0.00029 |
| *Scoc* | 95467_at | Short coiled coil protein | 1.216 | *D8Mit75* | 8 | 1.215 | 30.694 | 0.00048 |
| *Il3ra* | 92955_at | Interleukin 3 receptor α | 1.889 | *D14Mit99* | 14 | 1.892 | 20.336 | 0.00054 |
| *Cd59a* | 101516_at | CD59a antigen | 302 | *D2Mit43* | 2 | 302 | 30.006 | 0.00057 |
| *Birc1f*[d] | 160605_s_at | Neuronal apoptosis inhibitory protein 6 | 1.214 | *D8Mit75* | 8 | 1.215 | 21.858 | 0.00120 |
| *Hs1bp1* | 96578_r_at | HS1 binding protein | 470 | S03Gnf106.500 | 3 | 486 | 18.817 | 0.00220 |
| *Gfer* | 160269_at | Growth factor, erv1 (*S. cerevisiae*)-like | 2.226 | S17Gnf021.275 | 17 | 2.225 | 23.695 | 0.00300 |
| *F11r* | 103816_at | F11 receptor | 172 | *D1Mit113* | 1 | 173 | 20.841 | 0.00467 |
| *Hars* | 92580_at | Histidyl tRNA synthetase | 2.335 | *D18Mit94* | 18 | 2.336 | 27.513 | 0.00600 |
| *Fgf3*[d] | 92957_at | Fibroblast growth factor 3 | 1.132 | *D7Mit259* | 7 | 1.131 | 18.869 | 0.01200 |

[a]Marker most strongly associated with variation in transcript expression. [b]Calculation of strength of the linkage association. [c]Significance of linkage, calculated using permutation test. [d]These transcripts are preferentially or differentially expressed in Lin$^-$ Sca-1$^+$ c-kit$^+$ Rho$^{low}$ cells[2]. A complete list of all *cis*-regulated stem cell genes is given in **Supplementary Table 2** online.

**Table 2 HSC transcripts showing strongest evidence of *trans* regulation**

| Gene | Probe set | Name | Transcript chromosome | Transcript position (Mb) | QTL marker[a] | QTL chromosome | Marker position (Mb) | LRS[b] | Genome-wide *P* value[c] |
|---|---|---|---|---|---|---|---|---|---|
| AI594671 | 96499_at | EST AI594671 | 11 | 1.563 | *D7Mit301* | 7 | 1.078 | 58.284 | 0.00000 |
| *G22p1* | 103036_at | Thyroid autoantigen, 70kD | 15 | 2.162 | *D15Mit71* | 15 | 2.077 | 50.193 | 0.00000 |
| AA415817 | 94312_at | KIAA0251 | 3 | 469 | *D16Mit88* | 16 | 2.115 | 44.269 | 0.00000 |
| *Fmod* | 161373_r_at | Fibromodulin | 1 | 134 | X.057.845 | X | 2.500 | 24.914 | 0.00000 |
| *1810037I17Rik* | 161955_f_at | Reverse transcriptase | Unknown | | *D3Mit347* | 3 | 501 | 49.477 | 0.00001 |
| *Ceacam2* | 101907_s_at | CEA-rel cell adhesion molecule 2 | 7 | 1.014 | *D6Mit149* | 6 | 952 | 32.253 | 0.00001 |
| *Asb3* | 161466_r_at | Ankyrin repeat and SOCS box-containing | 11 | 1553 | *D11Mit19* | 11 | 1.548 | 42.345 | 0.00001 |
| *Proc* | 161656_r_at | Protein C | 18 | 2.330 | *DXMit25* | X | 2.507 | 27.568 | 0.00001 |
| *Mela* | 97282_at | Melanoma antigen, 80 kDa | 8 | 1.257 | *D9Mit263* | 9 | 1.340 | 41.358 | 0.00003 |
| *Psmb5-ps* | 101741_at | Proteasome subunit | 11 | 1.587 | *D14Mit140* | 14 | 1.923 | 40.412 | 0.00004 |
| *1110015E22Rik*[d] | 104217_at | Hypothetical protein MGC4171 | 7 | 1.113 | X.057.845 | X | 2.500 | 28.138 | 0.00006 |
| AA638002 | 96755_at | EST AA638002 | 18 | 2.333 | *DXMit25* | X | 2.507 | 23.065 | 0.00007 |
| *Mbd3* | 101385_at | Methyl-CpG binding domain protein 3 | 10 | 1.471 | *D4Mit111* | 4 | 593 | 28.739 | 0.00012 |
| *Psmd9* | 97929_r_at | Proteasome 26S subunit, non-ATPase, 9 | 5 | 814 | DXNds3 | X | 2.539 | 27.111 | 0.00015 |
| *Cnot7* | 161123_i_at | CCR4-NOT transcription complex | 8 | 1.173 | S02Gnf118.650 | 2 | 319 | 30.133 | 0.00020 |
| AA673511 | 95612_at | CS box-containing WD protein (WSB-2) | 5 | 808 | S18Gnf008.065 | 18 | 2.308 | 24.017 | 0.00029 |
| *Pmm2* | 101949_at | Phosphomannomutase 2 | 16 | 2.110 | *D19Mit19* | 19 | 2.429 | 23.993 | 0.00041 |
| *Lmna* | 98060_at | Lamin A | 3 | 468 | *DXMit223* | X | 2.597 | 31.761 | 0.00041 |
| *2600013G09Rik* | 102117_at | RAB, member of RAS oncogene family | 15 | 2.154 | *D15Mit239* | 15 | 2.075 | 27.364 | 0.00044 |
| *C81072* | 96489_at | EST C81072 | 3 | 455 | *D9Mit91* | 9 | 1.301 | 25.902 | 0.00048 |
| *Traf6* | 98874_at | Tnf receptor-associated factor 6 | 2 | 300 | *D4Mit17* | 4 | 601 | 23.029 | 0.00055 |
| *Trim21* | 92942_at | Tripartite motif protein 21 | 6 | 909 | *Mod2* | 7 | 1.076 | 29.514 | 0.00065 |
| *Hsp60* | 93277_at | Heat shock protein, 60 kDa/chaperonin | 1 | 55 | *D2Msw142* | 2 | 339 | 25.609 | 0.00075 |

[a]Marker most strongly associated with variation in transcript expression. [b]Calculation of strength of the linkage association. [c]Significance of linkage, calculated using permutation test. [d]This transcript is preferentially or differentially expressed in $Lin^-$ $Sca\text{-}1^+$ $c\text{-}kit^+$ $Rho^{low}$ cells[2]. A complete list of all *trans*-regulated stem cell genes is given in **Supplementary Table 3** online.



below). This genetic linkage analysis resulted in the assignment of genetic loci and intervals that are most strongly linked to the variation in gene expression of each individual transcript. As the physical position of almost all transcripts is known, we were able to produce a two-dimensional scatter plot in which, for each transcript, the *x* axis indicates the position of the best controlling locus (QTL) and the *y* axis identifies the physical chromosomal position of the corresponding gene (**Fig. 1**). Two patterns became immediately apparent. First, 478 transcripts were associated by a QTL mapping within 20 Mb of the gene itself. We refer to these as *cis*-acting QTLs. Typically, the likelihood ratio statistic (LRS) value, indicating the strength of association of the controlling locus with expression levels, was high for these *cis*-acting QTLs. Association statistics for 162 of the 478 *cis*-acting transcripts (34%) passed thresholds for significant genome-wide linkage. If we assume a total mouse genome size of 2,600 Mb and evaluate 12,422 transcripts, the null expectations are that at least 5% of 12,422 tested probe sets should meet or exceed the statistical criterion across the entire genome and that of these 621 false positives, ~0.8% or only 5 spurious QTLs would fall within 20 Mb of the parent gene. Most of these *cis*-regulated genes contain polymorphisms in regulatory elements that affect expression levels in B6 and D2 stem cells. A small subset of the oligonucleotides on the U74Av2 array (~0,3%) have a sequence that overlaps with one or more of the ~1.2 million SNPs that distinguish B6 and D2 (ref. 17, original SNP data from Celera Genomics). Most of these SNP-bearing probes do not map as *cis*-acting QTLs. Several hematopoietic genes are polymorphic and differentially expressed in B6 and D2 HSCs, including *Gpi1* (ref. 18), *H2-D1* and *Fli1* (ref. 19). These transcripts were

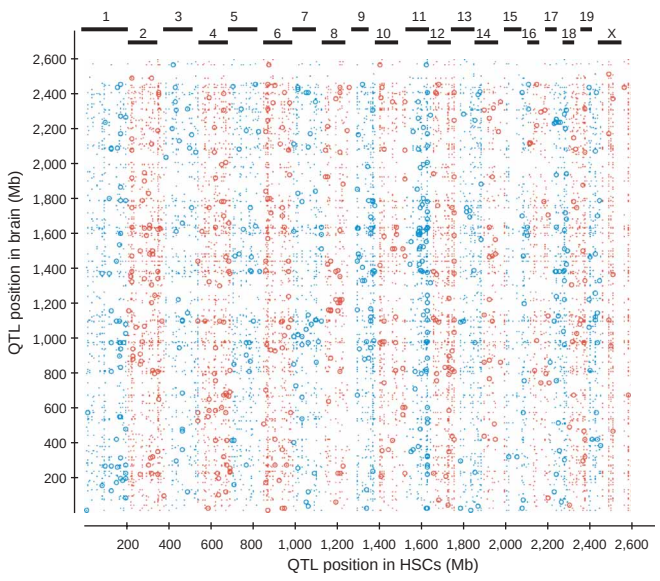


**Figure 2** Comparison of brain and HSC QTLs. For each transcript on the Affymetrix array, the locations of modifying QTL in brain and HSCs were compared. Brain data were taken from ref. 17. Transcripts positioned on the diagonal are controlled by the same QTL in both tissues (*i.e.*, are stable) but are not necessarily *cis*-acting (all transcripts significantly modulated by stable QTLs are listed in **Supplementary Tables 4** and **5** online). Chromosomal positions are indicated at the top and highlighted by alternating red and blue coloring. Large circles represent transcripts that are *cis*-regulated in HSCs.

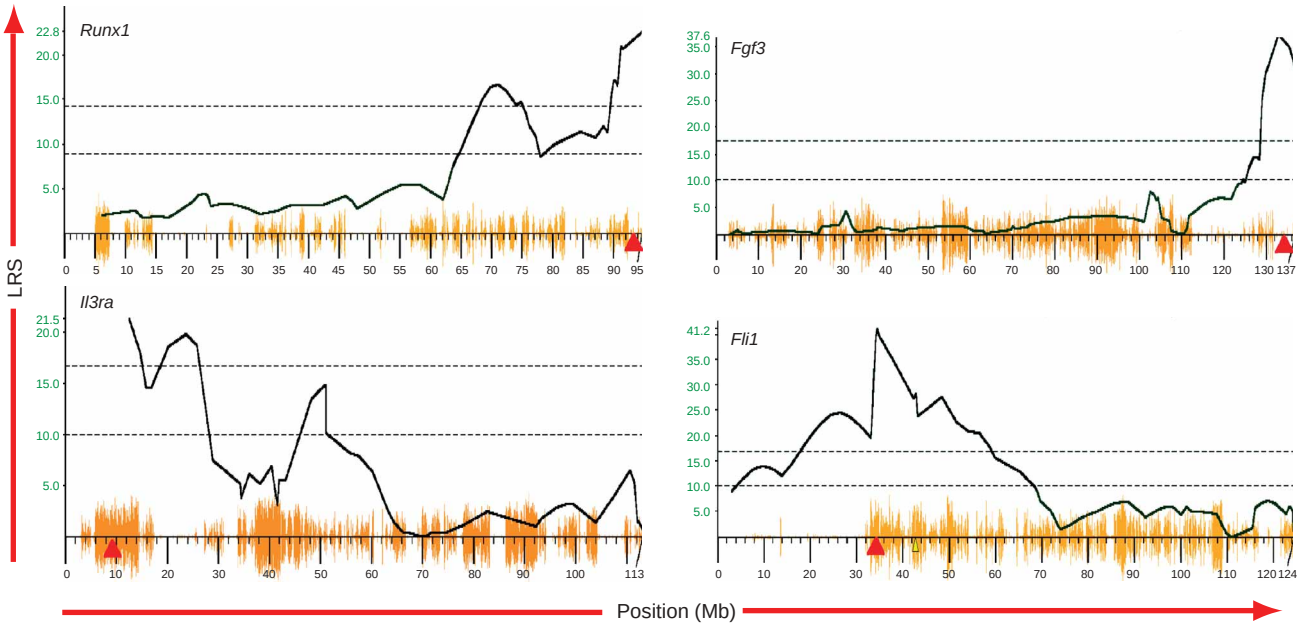


**Figure 3** Linkage analysis of four strongly *cis*-regulated stem cell transcripts showing genome-wide significant linkage to an interval mapping in close proximity to the gene (gene position is indicated by red triangle). The two dotted lines in each graph indicate suggestive (lower) and significant (upper) genome-wide linkage. The yellow seismogram reflects SNP density across each chromosome. SNP analysis comparing B6 and D2 alleles detected the presence of multiple polymorphisms in each gene (**Table 3**).



strongly *cis*-regulated. Thus, our genetical genomics approach immediately identified large numbers of genes carrying allelic polymorphisms. The strongest *cis*-acting genes, some of which have a critical role in HSC function, are listed in **Table 1**. A complete list of all 162 significant *cis*-regulated HSC transcripts is provided in **Supplementary Table 2** online.

Notably, we identified multiple QTLs that modulate expression levels of a large number of transcripts mapping throughout the genome. These controlling loci, which we refer to as *trans*-acting QTLs, are identified as vertical bands (**Fig. 1**). Horizontal bands result from local variation in gene density and incomplete representation of transcripts on the array. Although, in general, linkage statistics for *cis*-regulated transcripts were higher than those for *trans*-regulated transcripts, some *trans*-regulated genes showed essentially mendelian inheritance patterns (**Table 2**). Among the strongest *trans*-regulated transcripts, six were regulated by loci on the X chromosome. We detected 136 transcripts that were significantly linked (genome-wide linkage $P < 0.005$) to a single marker. Genomic distribution of all significant *trans*-acting QTLs is listed in **Supplementary Table 3** online.

## Comparing brain and stem cell QTLs

An advantage of the RI panel is that mice can be repeatedly phenotyped, and gene expression levels in distinct tissues can be compared easily *in silico*. From parallel studies[17], we have detailed information on gene expression levels in forebrain of the same panel of RI mice, enabling us to assess whether genes were regulated by the same QTLs in HSCs and brain (**Fig. 2**). We found that 297 genes were associated with the same regulatory QTL (within 20 Mb) in both HSCs and brain. Of these genes, only 75 were *cis*-regulated in HSCs (**Supplementary Table 4** online). Therefore, 222 *trans*-regulated transcripts were stable (*i.e.*, their QTL location was identical in both HSC and in brain; **Supplementary Table 5** online).

## Using WebQTL to detect gene networks

The concept of genetical genomics, though intuitively straightforward, has been tested only twice in a mammalian system[20,21]. Therefore, very little is known of the molecular nature of *cis*-acting and, even more so, *trans*-acting QTLs. In yeast, *trans*-acting QTLs do not map specifically to transcription factors but rather are broadly dispersed across distinct classes of genes. But the extensive coverage of the yeast genome and its lower molecular complexity allowed researchers to conclude that clustered genes with known and similar function very often mapped to the same QTL[16].

Similarly, we propose that collections of coregulated transcripts, identified by vertical *trans*-acting bands (**Fig. 1**), consist largely of downstream targets of polymorphic genes. To substantiate this proposal and to document the ability of our approach to identify target genes, we selected four strongly *cis*-regulated transcripts with known function and searched for coregulated genes using WebQTL's correlation search (**Fig. 3**). *Runx1*, a transcription factor that has an essential role in normal blood cell development, was highly *cis*-regulated. By searching for transcripts that had similar strain distribution patterns as *Runx1*, we identified *Tcrb* and *Csf1r*, which are well-known downstream targets of this transcription factor (**Table 3**). We also found that several other receptors, most notably those binding activin A and ephrin B3, varied with *Runx1* levels. Similarly, we identified *Mapk1*, *Ccnd3* and *Rac1* as putative downstream targets of *Il3ra*. We found *Bmp8a*, *Efnb3*, *Pbx1* and *Mapk6* to be downstream of *Fgf3*, and we identified multiple well-known proto-oncogenes as new putative targets of *Fli1* (**Table 3**).

## Identification of *Scp2* candidate genes

Using a similar approach, we searched for candidate genes involved in variation in HSC turnover. We recently mapped this trait to a 10-cM region on chromosome 11 between markers *D11Mit279* and *D11Mit41* (ref. 9). Here, we first identified all transcripts on the

**Table 3 Identification of putative targets of four *cis*-regulated HSC transcripts**

| | *Trans*-regulated | | |
|---|---|---|---|
| *Cis*-regulated | Affymetrix ID | Description | Interaction status |
| *Runx1* (92399_at; chromosome 16; 2 3′ UTR SNPs, 74 intronic SNPs) | 103617_at | Decay accelerating factor 1 | Unknown |
| | 93208_at | TCR-beta chain | PMID 11564801 |
| | 98317_at | Paired mesoderm homeobox 2b | Unknown |
| | 162175_at | Defender against cell death 1 | Unknown |
| | 95808_g_at | CSF1-r | PMID 10891464 |
| | 99323_at | IL12-R | Unknown |
| | 100448_at | Activin A receptor | Unknown |
| | 98726_at | Progesteron receptor | Unknown |
| | 93469_at | Eph receptor B3 | Unknown |
| | 161713_f_at | Prostaglandin F receptor | Unknown |
| *Il3ra* (92955_at; chromosome 14; 1 silent mutation, 18 intronic SNPs) | 160834_at | CDK4-binding protein | PMID 7862452 |
| | 101650_at | Protocadherin 6 | Unknown |
| | 93252_at | Map kinase 1 | PMID 10362354 |
| | 101122_at | Eph receptor A6 | Unknown |
| | 104568_at | Mixed lineage leukemia | Unknown |
| | 160545_at | Cyclin D3 | PMID 8415743 |
| | 103001_at | Vegf-b | PMID 11157721 |
| | 103038_at | Guanylate cyclase activator | Unknown |
| | 101555_at | Rac1 | PMID 12384416 |
| | 161456_f_at | GATA1 | PMID 8265595 |
| *Fgf3* (92957_at; chromosome 7; 4 intronic SNPs) | 100707_at | Plenty of SH3 domains | PMID 9811447 |
| | 92982_at | Bmp8a | PMID 11493538 |
| | 102829_s_at | Map kinase kinase 6 | PMID 11802165 |
| | 101657_at | Bmp8b | PMID 11493538 |
| | 103075_at | POU domain TF | Unknown |
| | 94160_at | Ephrin B3 | PMID 10611251 |
| | 98407_at | Ephrin B1 | PMID 10611251 |
| | 102257_at | Pbx/knotted homeobox | PMID 12431378 |
| *Fli1* (94698_at; chromosome 9; 2 silent mutations, 249 intronic SNPs) | 92951_at | Hox D11 | Unknown |
| | 160687_r_at | Activator of S-phase | Unknown |
| | 102265_at | Myf6 | Unknown |
| | 102873_at | AbcB3 | Unknown |
| | 103530_at | Fanconi anemia Compl. G | Unknown |
| | 93231_at | Hic1 | Unknown |
| | 98500_at | IL-1 receptor like 1 | Unknown |
| | 95296_r_at | Flt3 | Unknown |
| | 96941_at | Ras oncogene family-like 4 | Unknown |
| | 98731_at | Ras-related GTP binding | Unknown |

WebQTL was used to identify coregulated and *trans*-regulated targets of four *cis*-regulated polymorphic transcripts: *Runx1*, *Il3ra*, *Fgf3* and *Fli1*. The interaction status refers to whether or not data are available in PubMed that support potential interaction (identified by PubMed identification number, PMID). If no hit was retrieved in PubMed, interaction status was considered unknown.



Affymetrix array that mapped to the critical interval and then used the variation in gene expression levels across the 30 BXD strains to assess which of these transcripts was *cis*-regulated. Acknowledging that we have evaluated expression data for only ~25% of all genes in the mouse genome, we identified eight *cis*-acting genes that map to the critical interval (**Fig. 4**). Three of these are also *cis*-regulated in brain, one is *trans*-regulated in brain, and the other four are HSC-specific. Notably, we had previously identified three of these genes using a subtractive hybridization approach[9]. The eight *cis*-acting candidate genes can be divided in two clusters. The first cluster contains three very strong *cis*-regulated transcripts (*Kif1c*, *Psmb6* and *6330403K07Rik*, an unknown Riken gene); the second cluster (*Lig3*, *Ccl9*, *Ggnbp2*, *Mpo* and *Dlc2*) maps ~14 Mb telomeric. Haplotype analysis[22] showed that the entire *Scp2* interval is polymorphic between B6 and D2 (**Fig. 4**). We searched for mutations in transcribed sequences for these eight genes by comparing B6 and D2 genomes *in silico* by exploiting public and Celera databases. Polymorphisms were abundant in all eight genes. We sequenced *6330403K07Rik* and *Mpo* B6 and D2 alleles and confirmed sequence variations in both the coding and promoter sequences (**Supplementary Table 6** online).

The phenotype of interest (HSC turnover) is complex in itself and can be caused by mutations in a wide variety of genes or even clusters of genes. This renders our model system substantially more complex than the yeast model previously described[16]. Their study showed, however, that highly coregulated and *trans*-regulated transcripts can uncover the function of the underlying QTL gene. Therefore, we assessed which transcripts were highly correlated with each of the eight *cis*-acting candidates (**Table 4**). Although these transcripts themselves may be located anywhere in the genome, their expression levels are significantly associated by QTLs in the *Scp2* interval ($P < 0.05$).

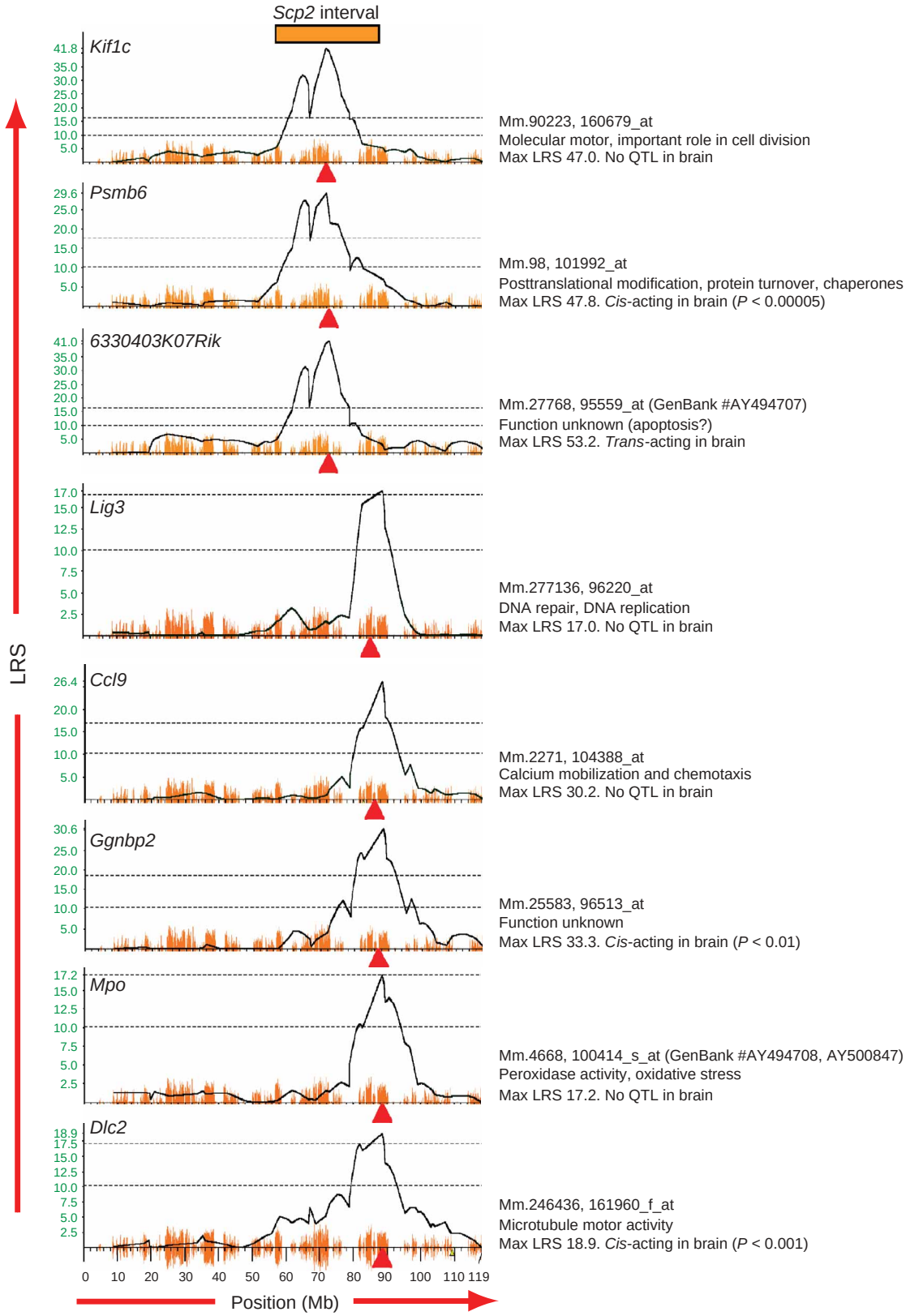


**Figure 4** Candidate genes affecting HSC proliferation. Eight *cis*-acting transcripts were identified that physically map to the *Scp2* locus, which was identified previously[9]. Graphs for each of these eight transcripts show linkage statistics on chromosome 11. The two dotted lines in each graph indicate suggestive (lower) and significant (upper) genome-wide linkage. The yellow seismogram reflects SNP density across chromosome 11 comparing B6 and D2 alleles. The physical position of the gene encoding each transcript is indicated by the triangle below each *x* axis. Peak LRS scores, and additional information on these genes, are shown next to each linkage graph.

Because the eight primary transcripts in each of the two *cis*-acting clusters are highly linked, we are not formally able to assign each specific *trans*-regulated transcript exclusively to an individual *cis*-acting candidate. Functional annotation showed clustering of transcripts with overlapping or interacting function. For example, *Dlc2*, which is associated with microtubule motor activity, was highly correlated with *Myog*, *Mdfi* and *Myl4*. In addition, this transcript was correlated with two seemingly unrelated seven-transmembrane receptors. Also, differences in *Mpo* expression were correlated with *Txnip*, which, like *Mpo*, is involved in oxidative stress. *6330403K07Rik*, which shows homology with a rat Ced-4-like apoptosis protein, is associated with several extracellular matrix molecules (*Pcsk4*, *Sparc* and *Col4a2*).

We cannot exclude the possibility that, as we have suggested before[9], a combination of the genes that we identified act in concert to confer the cell cyclus trait. We provide a preliminary list of candidate genes that is subject to more rigorous biological confirmation. It is notable, however, that we found several transcripts that interact directly with the DNA replication and repair machinery. These genes include a *cis*-regulated ligase *Lig3*; two *trans*-regulated helicases, *Cetn1* and *Dhx40*; the ribonuclease *Dnase1l2*; the polymerase *Pold4*; and *Tep1*, a telomerase-associated protein (**Table 4**). Mutation analysis detected the presence of a single base-pair frameshift insertion in the coding sequence of the B6 allele of *Lig3* (**Supplementary Table 6** online). The established role in the aging process of enzymes involved in DNA repair[23] and our observation that stem cell turnover and organismal aging are genetically linked[6] provide a conceptual framework that could integrate our findings.

**Table 4** ***Trans*-modulated transcripts controlled by QTLs in the *Scp2* interval**

| *Cis* transcript | Linked *trans* transcript | *P* value | Function |
|---|---|---|---|
| *Kif1c* | *Hspcb* | $3.25 \times 10^{-7}$ | Heat shock protein |
| | AV046379 | $3.25 \times 10^{-7}$ | Unknown |
| | *Atp7b* | $5.18 \times 10^{-5}$ | Cu-transport |
| | *Nkx2-6* | $7.38 \times 10^{-5}$ | Homeobox containing transcription factor |
| *Psmb6* | *Fmo1* | $9.83 \times 10^{-8}$ | Flavocontaining monooxygenase |
| | *Cetn1* | $2.31 \times 10^{-5}$ | Helicase activity, chromosome partitioning |
| | *Hspc150* | $2.40 \times 10^{-4}$ | Heat shock protein |
| | *Lamb* | $3.30 \times 10^{-4}$ | Extracellular matrix |
| *6330403K07Rik* | *Lif* | $1.29 \times 10^{-5}$ | Leukemia inhibitory factor, cytokine |
| | *Pcsk4* | $5.46 \times 10^{-5}$ | Serine protease |
| | *Sparc* | $1.50 \times 10^{-4}$ | Extracellular matrix, osteonectin |
| | *Col4a2* | $2.23 \times 10^{-5}$ | Extracellular matrix, procollagen |
| | *4733401H14Rik* | $8.70 \times 10^{-5}$ | Deoxyribonuclease 1-like 2 |
| *Lig3* | *Tep1* | $2.89 \times 10^{-5}$ | Telomerase associated protein-1 |
| | *Akr1c13* | $3.99 \times 10^{-5}$ | Aldo-keto reductase family member 13 |
| *Ccl9* | *Sftpc* | $2.06 \times 10^{-5}$ | Surfactant protein |
| *Mpo* | *Pold4* | $1.09 \times 10^{-5}$ | DNA polymerase |
| | *Rga* | $1.35 \times 10^{-5}$ | Rag1 gene activated |
| | *Fusip1* | $3.65 \times 10^{-5}$ | Mitosis |
| | *Txnip* | $6.25 \times 10^{-5}$ | Thioredoxin interacting, oxidative stress |
| | *Psmd3* | $6.25 \times 10^{-5}$ | Proteasome subunit |
| | *Ctsg* | $6.36 \times 10^{-5}$ | Proteolysis |
| | *Pbx1* | $8.56 \times 10^{-5}$ | Pre B cell leukemia transcription factor |
| *Dlc2* | *Fpr-rs2* | $5.97 \times 10^{-7}$ | Seven-transmembrane receptor |
| | *Sema5b* | $6.14 \times 10^{-6}$ | Seven-transmembrane receptor |
| | *Myog* | $7.76 \times 10^{-6}$ | HLH transcription factor |
| | *Sca2* | $1.01 \times 10^{-5}$ | Protein binding |
| | *Kpn* | $1.40 \times 10^{-5}$ | Protein transport |
| | *Dhx40* | $2.40 \times 10^{-5}$ | Helicase |
| | *Mdfi* | $4.36 \times 10^{-5}$ | Inhibition of myoD |
| | *Myl4* | $5.18 \times 10^{-5}$ | Cell division and partitioning |

WebQTL was used to identify transcripts that are highly correlated to one or more of the *cis* candidates on chromosome 11. Colors indicate genes with overlapping or interacting function (red, protein trafficking/degradation; blue, cell cycling; green, extracellular matrix; orange, DNA repair; black, other).

## DISCUSSION

Together with recent reports using similar approaches[14–16,20], our results document the power of genetical genomics to dissect complex traits. Molecular networks associated with phenotypic differences immediately become accessible as collections of coregulated genes controlled by a single locus, and key candidate genes within such a locus can be identified by their physical position. The HSC data set, the brain data set and the BXD genotypes were collectively deposited in a database, accessible through WebQTL. This analysis engine allows custom searches to identify new gene expression pathways and is valuable to the research community. Coregulated stem cell genes can easily be retrieved. Also included in WebQTL are phenotypes of previously published BXD traits, which now can be correlated *in silico* with the HSC and brain expression patterns. Forty-six additional BXD strains were recently added to this RI family[24]. Adding data from these mice will further improve the power and precision of QTLs in this cross between two sequenced strains. Additional cell type– and tissue-specific *cis*- or *trans*-regulation patterns can easily be incorporated in the WebQTL database. The advent of DNA chips that contain much larger samples of transcripts, and related efforts in the field of proteomics[21], will make this approach even more comprehensive and powerful. We expect that this approach will also be relevant for the identification of human complex and quantitative traits.

## METHODS

**Stem cell purification.** We purchased BXD RI mice from the Jackson Laboratory and housed them in clean conventional conditions in the Central Animal Facility of the University of Groningen, the Netherlands. We used female mice between 3 and 6 months of age. We flushed bone marrow cells from the femurs and tibias of three mice and pooled them. After standard erythrocyte lysis, we stained nucleated cells with a panel of biotinylated lineage-specific antibodies (Mouse Lineage Panel, containing antibodies to CD3e, -CD45R/B220, CD11b (Mac-1), TER119 (Ly-76) and Gr-1 (Ly-6G); Pharmingen), fluorescein isothiocyanate–conjugated antibody to Sca-1 and allophycocyanin-conjugated antibody to c-kit (Pharmingen). We washed cells twice and incubated them for 30 min with streptavidin-phycoerythrin (Pharmingen). After two washes, we resuspended cells in phosphate-buffered saline with 0.2% bovine serum albumin and purified them using a MoFlo flowcytometer. We defined the lineage-depleted bone marrow cell population as the 5% of cells showing the least phycoerythrin-fluorescence intensity. Stem cell yield across all BXD samples varied from 16,000 to 118,000 $Lin^-$ $Sca\text{-}1^+$ $c\text{-}kit^+$ cells. We tested a small aliquot of each sample of purified cells functionally for stem cell activity by directly depositing single cells in a cobblestone area forming cell assay using the automated cell deposition unit

(**Supplementary Table 1** online). We immediately collected the remainder of the cells in RNA lysis buffer. All animal experiments were approved by the Groningen University Animal Care Committee.

**Cobblestone area forming cell assays.** We carried out the cobblestone area forming cell assay as described[5]. We seeded cells of the stromal cell line FBMD-1 in 96-well plates (Costar) in Dulbecco's modified Eagle medium containing L-glutamine (GIBCO-BRL, Life Technologies), 5% horse serum, 15% fetal bovine serum (sera from GIBCO-BRL), $10^{-4}$ mol $l^{-1}$ β-mercaptoethanol, $10^{-5}$ mol $l^{-1}$ hydrocortisone (Sigma), 80 U $ml^{-1}$ penicillin, 80 μg $ml^{-1}$ streptomycin (both from GIBCO-BRL) and 25 mmol $l^{-1}$ $NaHCO_3$. We incubated plates at 33 °C in 5% $CO_2$ and used them 10–14 d later. We seeded sorted HSCs onto these preestablished stromal layers as single cells (one cell per well). At this time, we switched the medium from 5% horse serum and 15% fetal bovine serum to 20% horse serum. We evaluated all wells weekly for 5 weeks for the presence or absence of cobblestone areas, defined as colonies of at least five small nonrefractile cells growing beneath the stromal layer.

**RNA isolation and labeling.** We isolated total RNA derived from pooled HSC samples from three mice using StrataPrep Total RNA Microprep kit (Stratagene) as described by the manufacturer. We dissolved RNA pellets in 500 μl of absolute ethanol and sent them on dry ice by courier to GNF.

We quantified total RNA using RiboGreen, split it into two equal aliquots of ~10 ng, representing RNA from ~10,000 cells, and labeled it using three rounds of RNA amplification, exactly as described previously[25]. We used two microarrays per strain (three mice × two arrays). We fractionated labeled cRNA and hybridized it to the U74Av2 microarray from Affymetrix in accordance with the manufacturer's protocol. We scanned arrays and analyzed images as previously described using MAS 5.0 software. To generate .TXT files, we analyzed .CEL files using MAS 5.0 with the global value of each array scaled to 200 units.

**Data acquisition and normalization used for WebQTL: probe (cell) level data from the .CEL file.** The .CEL values produced by MAS 5.0 are the 75% quantiles from a set of 36 pixel values per cell (the pixel with the twelfth highest value represents the whole cell). Step 1: We added an offset of 1.0 to the .CEL expression values for each cell to ensure that all values could be logged without generating negative values. Step 2: We took the $\log_2$ of each cell. Step 3: We computed the Z score for each cell. Step 4: We multiplied each Z score by 2. Step 5: We added 8 to the value of each Z score. The consequence of this simple set of transformations is to produce a set of Z scores with a mean of 8, a variance of 4 and a standard deviation of 2. The advantage of this modified Z score is that a twofold difference in expression level corresponds to a difference of approximately one unit. Step 6: We computed the arithmetic mean of the values for the set of microarrays for each of the individual strains.

**Probe set data from the .TXT file.** We generated the .TXT files using MAS 5.0. We applied the same steps described above to these values. Every microarray data set therefore has a mean expression of 8 with a standard deviation of 2. A one-unit difference represents a roughly twofold difference in expression level. Expression levels below 5 are usually close to background noise levels.

**Mapping.** We carried out linkage mapping for 12,422 transcript expression traits using strain averages of probe set expression levels obtained using RMA or MAS 5.0. We carried out QTL mapping using a custom program, QTL Reaper, that does simple regression implemented in Python and C. Permutation tests (up to $10^6$ permutations) established empirical *P* values. Significant and suggestive linkage refer to the conventional criteria for QTL mapping[26] (1,000 permutations with *P* values of 0.05 and 0.63).

There are several hematopoietic databases available in WebQTL. Our data presented here are based on the GNF Hematopoietic U74Av2 Cells September 2003 database. Genome scans for all traits can be replicated and recomputed using a variety of transforms and analytic methods in WebQTL.

**URLs.** WebQTL is available at http://www.webqtl.org/. The GNF SNP database is available at http://www.gnf.org/SNP/.

**GenBank accession numbers.** D2 *6330403K07Rik* allele, AY494707; D2 *Mpo* allele, AY494708 and AY500847.

*Note: Supplementary information is available on the Nature Genetics website.*

**ACKNOWLEDGMENTS**

We thank G. Mesander and H. Moes for flow cytometry support and O. Sibon and R. van Os for critically reading the manuscript. This work was supported by grants from the Royal Netherlands Academy of Sciences, a Genomics Fellowship from the Netherlands Organization for Scientific Research, the Dutch Cancer Society and the National Heart, Lung, and Blood Institute (to G.d.H.) and by the National Institute of Mental Health, National Institute on Drug Abuse, the National Institute on Alcohol Abuse and Alcoholism and the National Science Foundation (to R.W.W.).

**COMPETING INTERESTS STATEMENT**

The authors declare that they have no competing financial interests.

Received 24 September; accepted 29 November 2004
Published online at http://www.nature.com/naturegenetics/

1. Potten, C.S. & Loeffler, M. Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt. *Development* **110**, 1001–1020 (1990).
2. Ivanova, N.B. *et al.* A stem cell molecular signature. *Science* **298**, 601–604 (2002).
3. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C. & Melton, D.A. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597–600 (2002).
4. Fortunel, N.O. *et al.* Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* **302**, 393 (2003).
5. de Haan, G. & Van Zant, G. Intrinsic and extrinsic control of hemopoietic stem cell numbers: mapping of a stem cell gene. *J. Exp. Med.* **186**, 529–536 (1997).
6. De Haan, G. & Van Zant, G. Genetic analysis of hemopoietic cell cycling in mice suggests its involvement in organismal life span. *FASEB J.* **13**, 707–713 (1999).
7. Muller-Sieburg, C.E., Cho, R.H., Sieburg, H.B., Kupriyanov, S. & Riblet, R. Genetic control of hematopoietic stem cell frequency in mice is mostly cell autonomous. *Blood* **95**, 2446–2448 (2000).
8. Kamminga, L.M. *et al.* Autonomous behavior of hematopoietic stem cells. *Exp. Hematol.* **28**, 1451–1459 (2000).
9. De Haan, G. *et al.* A genetic and genomic analysis identifies a cluster of genes associated with hematopoietic cell turnover. *Blood* **100**, 2056–2062 (2002).
10. Boultwood, J., Lewis, S. & Wainscoat, J.S. The 5q-syndrome. *Blood* **84**, 3253–3260 (1994).
11. Lai, F. *et al.* Transcript map and comparative analysis of the 1.5-Mb commonly deleted segment of human 5q31 in malignant myeloid diseases with a del(5q). *Genomics* **71**, 235–245 (2001).
12. Jansen, R.C. & Nap, J. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
13. Jansen, R.C. Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.* **4**, 145–151 (2003).
14. Wayne, M.L. & McIntyre, L.M. Combining mapping and arraying: An approach to candidate gene identification. *Proc. Natl. Acad. Sci. USA* **99**, 14903–14906 (2002).
15. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
16. Yvert, G. *et al.* Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.* **35**, 57–64 (2003).
17. Chesler, E.J. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* advance online publication, 13 February 2005 (doi:10.1038/ng1518).
18. Pearce, S.R., Morgan, M.J., Ball, S., Peters, J. & Faik, P. Sequence characterization of alleles Gpi1-Sa and Gpi1-Sb at the glucose phosphate isomerase structural locus. *Mamm. Genome* **6**, 537–539 (1995).
19. Ben-David, Y., Giddens, E.B. & Bernstein, A. Identification and mapping of a common proviral integration site Fli-1 in erythroleukemia cells induced by Friend murine leukemia virus. *Proc. Natl. Acad. Sci. USA* **87**, 1332–1336 (1990).
20. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
21. Klose, J. *et al.* Genetic analysis of the mouse brain proteome. *Nat. Genet.* **30**, 385–393 (2002).
22. Wiltshire, T. *et al.* Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci. USA* **100**, 3380–3385 (2003).
23. Hasty, P., Campisi, J., Hoeijmakers, J., van Steeg, H. & Vijg, J. Aging and genome maintenance: lessons from the mouse? *Science* **299**, 1355–1359 (2003).
24. Peirce, J.L., Lu, L., Gu, J., Silver, L.M. & Williams, R.W. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.* **5**, 7 (2004).
25. Scherer, A. *et al.* Optimized protocol for linear RNA amplification and application to gene expression profiling of human renal biopsies. *Biotechniques* **34**, 546–550, 552–554, 556 (2003).
26. Lander, E.S. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).