Identification and quantification of disease-related gene clusters

Gábor Firneisz1, Idit Zehavi2,3,4, Csaba Vermes1, Anita Hanyecz1, Joshua A. Frieman2,3,4 and Tibor T. Glant1,*

1Section of Biochemistry and Molecular Biology, Departments of Biochemistry, Orthopedic Surgery and Internal Medicine, Rush University at Rush-Presbyterian-St Luke’s Medical Center Chicago, IL 60612, USA; 2Department of Astronomy and Astrophysics, 3Center for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA; 4NASA/Fermilab Astrophysics Center, Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA

Received on July 16, 2002; revised on February 8, 2003; accepted on April 16, 2003

ABSTRACT

Motivation: DNA microarray technology and the completion of human and mouse genome sequencing programs are now offering new avenues for the investigation of complex genetic diseases. In particular, this makes possible the study of the spatial distribution of disease-related genes within the genome. We report on the first systematic search for clustering of genes associated with a polygenic autoimmune disease.

Results: Using a set of cDNA microarray chip experiments in two mouse models of rheumatoid arthritis, we have identified ∼200 genes based on their expression in inflamed joints and mapped them into the genome. We compute the spatial autocorrelation function of the selected genes and find that they tend to cluster over scales of a few megabase pairs. We then identify significant gene clusters using a friends-of-friends algorithm. This approach should aid in discovering functionally related gene clusters in the mammalian genome.

Contact: tglant@rush.edu

INTRODUCTION

The completion of the human genome sequencing program and subsequent annotation of previously unidentified genes have opened a new epoch in biology and revealed that the human genome harbors approximately 35,000 genes coding for proteins (Venter et al., 2001; Lander et al., 2001). Although the existence of clusters of functionally related genes in the genome, for example the major histocompatibility genes (MHC) (Trowsdale, 2001; Flajnik and Kasahara, 2001) or immunoglobulins (Celera Discovery System, http://www.celera.com), is well known, the identification and significance of such clusters have not been studied with rigorous statistical techniques.

METHODS

Microarray experiments

To investigate the distribution of genes associated with rheumatoid arthritis, an autoimmune inflammatory joint disease characterized by genetic complexity, we used five independent cDNA microarray experiments. Specifically, we screened two genetically distinct mouse models of rheumatoid arthritis induced by systemic immunization of type II collagen (Courtenay et al., 1980), or cartilage proteoglycan (Glant et al., 1987). In each experiment, pooled and reverse-transcribed mRNA samples from inflamed paws of 3-4 arthritic mice, collected within 2-10 days after the onset of arthritis, were labeled with Cy5 fluorescent marker and co-hybridized with Cy3-labeled corresponding samples of non-immunized control mice. The DNA microarray chips contained a total of 9500 known genes and enhanced sequence tag (EST) clones (Mouse Gem 1 Screen; Genome System Inc.), the highest gene number available when the experiments were initiated. The microarray results were analyzed with the GemTools™ (v2.4) software (Incyte Genomics), using

*To whom correspondence should be addressed.
1.5-fold differential expression threshold level. A total of 221 cDNA clones were selected by the following criterion: representing genes were up- or down-regulated in at least four of the five independent hybridization experiments.

**Gene mapping**

Based on sequence identity, 217 of the 221 cDNA clones were successfully mapped into the genome using the Celera Genomics database system (Celera Discovery System, http://www.celera.com). Chromosomal localizations of 125 genes were also confirmed using public gene databases (http://www.ncbi.nlm.nih.gov; http://www.jax.org; http://www.gsc.riken.go.jp, http://www.ensembl.org) and in four cases using the T31 mouse–hamster hybridization radiation hybrid panel (Research Genetics). In several cases, there were two different EST clones representing the same gene, reducing the available sample to a total of 203 mapped genes (hereafter, called selected genes). For this analysis, each selected gene was treated as a point-like object and assigned a spatial position along its chromosome (in megabase pair units; Mb) corresponding to the beginning of the gene. Since the genes are small (typically 10–100 kb) compared to the scale of clustering we find below (of order 1–few Mb), this idealization does not affect our results. In some cases the genes were localized in the genome as duplicate or multiple copies in consecutive order, in which case we included only the most proximal localization in the subsequent analysis. We have verified that this also has no significant effect on the statistical results below.

**Statistical clustering**

The autocorrelation function $\xi$ is a useful statistic for characterizing the clustering of objects, e.g. the distribution of galaxies in the Universe (Peebles, 1980; Zehavi et al., 2002). The probability that an object, in our case a selected gene, is found in the small genome interval $dr$ is $dP = n dr$, where $n$ is the mean number density of objects. The autocorrelation function $\xi(r)$ is defined by the joint probability of finding objects in both intervals $dr_1$ and $dr_2$ separated by a distance $r_{12}$, $dP = n^2 [1 + \xi(r_{12})] dr_1 dr_2$. For a uniform random Poisson distribution, the probabilities of finding objects in $dr_1$ and $dr_2$ are independent, so the joint probability is just the product $dP = n^2 dr_1 dr_2$; that is, $\xi = 0$ for uncorrelated objects. If the object positions are correlated, then $\xi > 0$; if they are anticorrelated, $-1 < \xi < 0$. In other words, $\xi$ measures the excess probability over random of finding a pair of objects with a given separation.

Operationally, $\xi(r)$ can be estimated by measuring the excess number of pairs of genes with separation $r$ over that expected for a random Poisson distribution, $\xi(r) = [N_{DD}(r)/N_{DR}(r)](n_R/n_D) - 1$ (Davis and Peebles, 1983). Here, $N_{DD}(r)$ is the number of data–data pairs with separation in the interval $r - \Delta r, r + \Delta r$; $N_{DR}$ is the number of data–random pairs in the same interval, obtained by cross-correlating the data with an artificially created random distribution, and $n_D, n_R$ are the number densities of points in the data and random distributions. The use of $N_{DR}$ instead of $N_R$ in the denominator generally gives a more accurate estimate in the presence of sample boundaries; other estimators can be used, but the differences between them are small in this case. To minimize errors arising from the finite random sample size, we used a random sample with 100 times as many points as the number of selected genes displaying differential expression, comparable to the total number of genes in the mouse genome.

Since the mean density of objects enters the estimate of the correlation function, we must choose between several possibilities for estimating the density of selected genes. The simplest choice, that adopted in this analysis, is to assume that the density of selected genes is constant across the genome, independent of chromosome number. We have checked the consistency of this assumption by comparing the actual number of selected genes on each chromosome to that expected from their genome-average number density of 0.078 transcripts/Mb. The actual number of selected genes is within one standard deviation (18) of the expected number for 65% of the chromosomes and within 28 for 95% of them, consistent with Poisson fluctuations around a constant density. There is one outlier chromosome, number 3, for which the number of selected genes is more than 36 higher than expected from the genome-mean.

An alternative choice would be to estimate the density of selected genes, and then the correlation function, separately for each chromosome; we found that this does not change our results for $\xi$ qualitatively. We also examined the possibility of scaling the mean density of selected genes on each chromosome by the most recent (preliminary) mouse chromosome-specific gene density (Celera Discovery System). While the latter does display significant inter-chromosome variability, it appears to be smaller than that of the human genome (Venter et al., 2001), and this choice would not qualitatively alter the results on arthritis/autoimmune gene clustering.

To interpret the clustering of the arthritis-selected genes, we also compute the correlation function of the full mouse genome. In addition, since the arthritis-selected sample is drawn from the 9500 genes contained in the cDNA microarray, we must assess the degree to which the gene chip represents a fair (unbiased) sample of the mouse genome. To do this, we compute the correlation function for six samples of 203 genes each drawn randomly from the microarray, for which we have obtained gene localizations.

**Correlation function results**

Figure 1 shows the resulting autocorrelation function of the various gene samples discussed above. The filled dots and dashed curve shows the correlation function for the 203 selected genes associated with autoimmune arthritis. The errorbars shown on this estimate of $\xi$ are 16 errors obtained from Poisson statistics, i.e. from the assumption that the selected genes
Disease-related gene clusters

Fig. 1. The correlation functions of three different gene samples, the mouse genome itself, random samples of the cDNA-gene chip and the arthritis-related genes, as a function of separation measured in megabase pairs (Mb). Round shaped symbols indicate the correlation function of the differentially expressed arthritis-related genes in arthritic mice, with errorbars obtained from Poisson statistics from the number of gene pairs in each bin. Triangle signs show the correlation function measured for the full mouse genome. Square symbols and errorbars indicate the average and scatter between six samples chosen at random from the microarray gene sample. This is a sensible choice when testing the null hypothesis that the genes are randomly distributed. The selected genes exhibit significant positive correlations for separations smaller than a few Mb, that is, a strong tendency for the selected genes to be clustered on these scales. No significant correlations on larger separations are apparent.

The triangle curve shows the correlation function for the full mouse genome (due to the large number of genes, the Poisson errors in this case are negligible). It is interesting that the mouse genome is not a random Poisson distribution but is itself significantly clustered on similar scales, up to a separation of \( \sim 4 \) Mb. Filled squares represent \( \xi \) measured for the ‘random’ microarray samples, where the dots indicate the average over the six samples and the error bars indicate the scatter between the samples. The microarray genes appear not to be a ‘fair sample’ of the full genome, since they are more strongly clustered than the genome. The correlation function of the disease-related genes lies systematically above that of the microarray samples, but only at the 1\( \delta \) level; we cannot confidently rule out that the arthritis-selected genes are only as clustered as a random subset of the mouse genome. We conclude that the arthritis genes are certainly significantly clustered but not necessarily more clustered than the genome itself.

Gene cluster finding

Having established statistically significant clustering of arthritis-associated genes, the next step is to identify individual gene clusters. Figure 2 shows the transcript density of the entire mouse genome calculated chromosome by chromosome using 1 Mb windows. There are clearly non-random over-and under-dense regions of the genome; this structure is reflected in the non-zero autocorrelation function
of the genome measured above. We generally expect arthritis-related gene clusters to be found in gene-dense regions, and indeed we will see this is the case. As an example, the second peak of the transcript density on chromosome 17 coincides with the MHC gene cluster in the mouse genome, where we also see an arthritis-related cluster. To systematically identify arthritis gene clusters, we use a friends-of-friends algorithm, a standard technique for identifying clusters of galaxies in astronomical surveys (e.g. Coles and Lucchin, 1995). To belong to a cluster, a gene must be within a distance $l$, the linking length, of at least one other member of the cluster.

**Detected gene clusters**

Figure 3 shows the spatial distribution of the 203 selected genes in the mouse genome and the clusters we identify for two choices of linking length, $l = 1.4$ Mb (top panel) and $l = 3.4$ Mb (bottom panel). These choices for $l$ are both comparable to the scale at which $\xi \sim 1$, which demarcates strong from weak correlations. To focus on the prominent clusters, we consider only those with $N \geq 3$ members for $l = 1.4$ Mb and $N \geq 4$ members for $l = 3.4$ Mb. These thresholds in $N$ are chosen such that, for a random Poisson sample of 203 genes with the same gene density as the arthritis gene sample, we would find 1.7 clusters on average with $N = 3$ members for $l = 1.4$ Mb and the same expected number with $N = 4$ members for $l = 3.4$ Mb. In general, for a random Poisson sample of $N_G$ genes with mean density $n$ and linking length $l$, we expect $N_M = \langle N_G \rangle \exp(-2nl)[1 - \exp(-nl)]^{M-1}$ clusters with $M$ members to be found on average (Sheth, 2001).

For $l = 1.4$ Mb, we find eight clusters comprising 14% of the selected genes; six of them are gene triplets ($N = 3$), and two larger ones contain $N = 5$ members each. In the random Poisson sample, we would expect to find 0.02 clusters with $N = 5$, and the probability of finding two of them is 0.0002.

One of the larger clusters appears on chromosome 3 (see Fig. 4), located roughly at 90 Mb, and is in close proximity to another triplet-gene cluster. These clusters on chromosome 3 overlap with one of the major susceptibility loci identified in both rheumatoid arthritis and insulin-dependent diabetes mellitus (Marrack et al., 2001; Jawheer et al., 2001). The other $N = 5$ cluster appears on chromosome 17 (at $\sim 35$ Mb) and appears to correspond to the known MHC cluster.

Using the larger linking length ($l = 3.4$ Mb), we identify four clusters with $N \geq 4$ members, comprising 12% of the selected genes. Two small clusters are found on chromosomes 2 and 11. The cluster found on chromosome 17 now includes six members; the probability of such a cluster occurring in the random Poisson sample is 10%. The prominent cluster on chromosome 3 now comprises 11 genes and is extremely unlikely (less than 0.002%) to be found in the Poisson sample, even if we were to estimate the probability using the larger density of selected genes on chromosome 3. This cluster exhibits a hierarchical structure in the sense that it includes two smaller clusters identified with the smaller linking length.

Figure 4 provides a detailed picture of the genes in this cluster. Finally, we note that comparison of Figures 2 and 3 shows that the arthritis clusters generally are found in gene-dense regions of the mouse genome, as expected. We performed the same cluster-finding analysis on the six sets of random samples drawn from the microarray genes, using the larger linking length and again identifying clusters with $N \geq 4$ members . The small cluster on chromosome 2 is not found in any of the random sets, while the other small cluster on chromosome 11 appears in 2 of the 6. The prominent cluster on chromosome 3 is found in half of these cases, but with fewer members than in the arthritis-selected gene sample, while the chromosome 17 cluster is identified in only one of these samples, with $N = 4$ members. We have done similar tests with six samples of 203 genes drawn at random from the full mouse genome. In this case, the chromosome 3 cluster is not identified in any of the random samples, and the cluster on chromosome 17 is detected once. We reiterate that these 12 random samples are not random Poisson samples, since
we have shown that both the microarray and the full genome display positive correlations; consequently, we expect to find more clusters in these random samples than in a corresponding Poisson sample.

The fact that clusters are occasionally found in a random Poisson distribution, and that they are found more often in random subsets of the mouse genome and microarray genes, does not vitiate the reality of the arthritis-related clusters we have identified: while some of these clusters may have formed ‘accidentally’, they are nevertheless groupings of genes associated with a given disease and therefore of potential clinical interest.

**DISCUSSION**

In addition to the cluster on chromosome 3 and the MHC on chromosome 17, it is interesting to speculate whether the other clusters comprising up- or down-regulated genes may also play significant roles in regard to the inflammatory joint diseases. As DNA chips covering the whole genome become available, we may also find additional disease-related gene clusters in inflammatory joint diseases or other polygenic autoimmune disorders. We believe that the algorithm employed here should provide a useful technique for identifying and cataloging disease-related gene clusters. One can envision employing it in future screening studies to create disease-specific gene-cluster databases, which might help in understanding the underlying genetic complexity of autoimmune diseases (Wanstrat and Wakeland, 2001).

**ACKNOWLEDGEMENTS**

We thank R.K. Sheth, V.A. Adarichev and K. Miecz for useful discussions. This work has been supported in part by the NIH (AR40310 and AR47652), Musculoskeletal Research Foundation and J.O. Galante endowment chair (TTG), by the DOE and NASA grant NAG 5-10842 at Fermilab, and by NSF grant PHY-0079251 and the Center for Cosmological Physics at Chicago (IZ and JAF).

**REFERENCES**


