Genetical genomics aims at identifying quantitative trait loci (QTLs) for molecular traits such as gene expression or protein levels (eQTL and pQTL, respectively). One of the central concepts in genetical genomics is the existence of hotspots [1], where a single polymorphism leads to widespread downstream changes in the expression of distant genes, which are all mapping to the same genomic locus. Several groups have hypothesized that many genetic polymorphisms—e.g., in major regulators or transcription factors—would lead to large and consistent biological effects that would be visible as eQTL hotspots.

Rather surprisingly, however, there have been only very few verified hotspots in published genetical genomics studies to date. In contrast to local eQTLs, which coincide with the position of the gene and are presumably acting in cis—e.g., by polymorphisms in the promoter region—distant eQTLs have been found to be more elusive. They seem to show smaller effect sizes and are less consistent, perhaps due to the indirect regulation mechanism, resulting in lower statistical power to detect them and, consequently, an inability to reliably delimit hotspots [2]. While there are typically hundreds to thousands of strong local eQTLs per study, the number of associated hotspots is much lower. For example, a recent very large association study in about 1,000 humans did not find a single significant hotspot [3]. Other studies have reported up to about 30 hotspots, far less than the number of significant local eQTLs (Table 1). The molecular basis is known for less than a handful of cases. An example is the Arabidopsis erecta locus, which leads to a drastic phenotypic change in the plant and has broad pleiotropic effects on many molecular [and morphological] traits [4].

Recently, Wu et al. [5] reported the large-scale identification of hotspots. They studied gene expression in adipose tissue of 28 inbred mouse strains and performed eQTL analysis by genome-wide association analysis. The paper reports the identification of over 1,600 candidate hotspots, each with a minimum hotspot size of 50 target genes. Furthermore, they demonstrated that these hotspots are biologically coherent by showing that in about 25% of cases, the hotspot targets are enriched for functional gene sets derived from Gene Ontology, the KEGG pathways database, and the Ingenuity Pathways Knowledge Base. These findings suggested that genetic polymorphisms can indeed lead to large and consistent biological effects that are visible as eQTL hotspots.

However, the authors chose a relatively permissive threshold of \( p = 0.003 \) for QTL detection, uncorrected for multiple testing. In total, 886,440 eQTLs were identified at this threshold, i.e., 134 per gene. A permutation test (C. Wu and A. I. Su, unpublished data) shows that this results in a false discovery rate of 64%, largely resulting from multiple testing across 157,000 SNPs and 6,601 probe sets. This relatively permissive threshold was chosen because the focus of the analysis was on patterns of eQTL hotspots and not on individual eQTL associations. Analysis of eQTL patterns is relatively robust to individual false positives, and a permissive threshold allows for relatively greater sensitivity in detecting signal [6]. The authors observed an enrichment of specific biological functions among the genes in the reported hotspots. The study also reported that enriched categories tended to match the annotation of candidate regulators. Moreover, one predicted regulator was experimentally validated. In sum, these data seem to support the hypothesis that hotspots are downstream of a common master regulator linked to the eQTL.

However, we suggest here that these observations may also be explained by clusters of genes with highly correlated expression. If one gene shows a spurious eQTL, many correlated genes will show the same spurious eQTL, in particular if the false discovery rate for individual eQTLs is very high [2, 7–9]. There are many nongenetic mechanisms that can create strongly correlated clusters of functionally related genes. On the one hand, such clusters may be a result of a concerted response to some uncontrolled environmental factor. On the other hand, dissected tissue samples can contain slightly varying fractions of individual cell types, leading to cell-type–specific gene clusters, which vary in a correlated manner. The resulting correlation patterns represent potentially confounding effects, both for the correct determination of a significance threshold and for the biological interpretation of the resulting hotspots.

Consequently, a key consideration in eQTL analysis is in the effective design of a permutation strategy to assess statistical significance. The approach used in [5] permuted the observed eQTLs among genes (Figure 1B). However, this approach has the disadvantage of ignoring the expression correlation between genes so that their spurious eQTLs no longer cluster along the genome. This permutation strategy leads to a potentially severe underestimate of the null distribution of the size of hotspots, when there are correlated clusters as described above.

An alternative strategy would have been to permute the strain labels as shown in Figure 1A, maintaining the correlation of the expression traits while destroying any


Editor: Gary A. Churchill, The Jackson Laboratory, United States of America
Published: October 24, 2008

Copyright: © 2008 Breitling et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

* E-mail: asu@gnf.org (AIS); z.c.jansen@rug.nl (RCJ)
Of course, this does not imply that all hotspots are necessarily false positives. As described above, about 5% of the co-mapping clusters in [5] are not only functionally coherent but also map to a locus that contains a gene of the same functional class. This number is not statistically significant, but it is still suggestive of an enrichment of functional associations ($p<0.16$, false discovery rate = 67%; C. Wu and A. I. Su, unpublished data). Some of these prioritized hotspots could correspond to true hotspots, and indeed one of them has been verified experimentally: cyclin H was validated as a new upstream regulator of cellular oxidative phosphorylation, as well as a transcriptional regulator of genes composing a hotspot [5].

Other studies, which used much stricter thresholds for defining their hotspots, also demonstrated the potential of interpreting putative hotspots by a closer study of the associated genetic locus [11,12]. An example is the recent work of Zhu et al. [12]; by combining eQTL information, transcription factor binding sites, and protein–protein interaction data in a Bayesian network approach, they were able to predict causal regulators for nine out of the 13 hotspots (69%) originally reported in [13]. With integrated methods like these, it should be possible to identify those hotspots that are more than just clusters of co-expressed genes. As a result, the number of identified, functionally relevant hotspots could ultimately increase beyond the small numbers reported in Table 1. This would create new opportunities for gene regulatory network reconstruction.

In any case, for the time being it seems that distant eQTLs and their hotspots are still scarce and hard to find, and that those that are reported should be interpreted...
with caution. This rarity of convincing hotspots in genetic genomics studies is intriguing. It could be due to the limited power of the initial studies, but it could also have a more profound reason. For example, it might well be that biological systems are so robust against subtle genetic perturbations that the majority of heritable gene expression variation is effectively “buffered” and does not lead to downstream effects on other genes, protein, metabolites, or phenotypes [14–17]. Experimental evidence for phenotypic buffering of protein coding polymorphisms is well established [18,19]. In fact, it has been shown that phenotypic buffering is a general property of complex gene-regulatory networks [20]. Also, if small heritable changes in transcript levels were transmitted unbuffered throughout the system, there would be a grave danger that genetic recombination would lead to unhealthy combinations of alleles and, consequently, to systems failure. Hotspots with large pleiotropic effects are thus more likely to be removed by purifying selection. If, as thus expected, common alleles are predominantly buffered by the robust properties of the system and hence largely inconsequential for the rest of the molecules in the system, this will have profound consequences for the design and interpretation of genetic genomics studies of complex diseases. Most importantly, it could turn out that even so-called common diseases—like diabetes, asthma, or rheumatoid arthritis—are not necessarily the result of common, small-effect variants in a large number of genes, but are rather caused by changes at a few crucial fragile points of the system (hotspots), which cause large, system-wide disturbances [21,22]. Future studies in genetic genomics should aim at further elucidating the striking rarity of eQTL hotspots.

References