BMC
Medical Genomics

RESEARCH ARTICLE Open Access

Co-regulatory expression quantitative trait loci mapping: method and application to endometrial cancer

Kenneth S Kompass, John S Witte*

Abstract

Background: Expression quantitative trait loci (eQTL) studies have helped identify the genetic determinants of gene expression. Understanding the potential interacting mechanisms underlying such findings, however, is challenging.

Methods: We describe a method to identify the *trans*-acting drivers of multiple gene co-expression, which reflects the action of regulatory molecules. This method-termed *co-regulatory expression quantitative trait locus* (creQTL) *mapping*-allows for evaluation of a more focused set of phenotypes within a clear biological context than conventional eQTL mapping.

Results: Applying this method to a study of endometrial cancer revealed regulatory mechanisms supported by the literature: a creQTL between a locus upstream of STARD13/DLC2 and a group of seven IFNβ-induced genes. This suggests that the Rho-GTPase encoded by STARD13 regulates IFNβ-induced genes and the DNA damage response.

Conclusions: Because of the importance of IFNβ in cancer, our results suggest that creQTL may provide a finer picture of gene regulation and may reveal additional molecular targets for intervention. An open source R implementation of the method is available at http://sites.google.com/site/kenkompass/.

Background

Expression Quantitative Trait Locus (eQTL) mapping searches across the genome for markers associated with individual transcripts to identify loci containing regulatory elements [1,2]. Although *cis* regulators are easily interpreted, assigning function to *trans* regulators is more difficult. At the transcriptional level, genes in *trans* are often co-regulated by transcription factors binding the same regulatory elements in the noncoding sequences of multiple genes [3]. When looking genome-wide, however, genes across many ontologies acting upstream of transcription factors participate in the co-regulation of genes [4,5]. Because of the difficulty in predicting the *trans*-regulators, the majority of transcriptional regulatory proteins remain unknown.

Identification of genetic components of gene co-regulation is important because there is compelling evidence that aberrant gene co-regulation participates in human disease [6,7]. Investigations into the genetic basis of gene co-regulation have used Bayesian networks to identify *cis*- and *trans*-acting factors controlling modules of co-regulated genes [8-10] or gene clusters. A key result here was the identification of associations that would have been missed when genes were tested individually, as in traditional eQTL. Biologically, this is very compelling because genes typically do not perform their functions in isolation but rather in coordinated groups. The existing Bayesian methods have focused primarily on the identification of yeast regulatory programs where other sources of information, such as sequence conservation, transcription factor binding site (TFBS) data, or protein interaction data are readily available and serve as prior information [8,10]. Extension of these methods to human genetics with data from HapMap subjects has shown that sequence conservation and *cis*-regulatory information were the most useful prior data [8].

Studies in other human cohorts and mice have used directed Bayesian networks or undirected weighted gene

* Correspondence: jwitte@ucsf.edu
Department of Epidemiology and Biostatistics and Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA

coexpression networks to incorporate existing marker and phenotype data into models that have made biologically validated predictions [11-16]. These network-based methods and their alternatives (e.g. [17-21]) show great promise but often have high computational complexity, making them most practical for smaller datasets with limited numbers of traits. Furthermore, prior information is generally not readily available in humans. For example, most TF binding sites remain unknown and even within the same tissue, the vast majority of TFBS appear divergent between human and mouse [22]. This suggests that relying on sequence conservation or the presence of conserved TF binding motifs may miss some key associations and that agnostic, complementary methods should be developed.

Many algorithms that search for the determinants of gene co-regulation assign each gene to a single cluster (e.g. [4,9,23]), which is limiting, because genes can belong to different clusters under different biological conditions [24]. More recent network approaches overcome this problem by examining differential canonical correlation between multiple states, such as healthy and diseased, or with a reference [11]; these approaches, however, may rely on methods that are not robust to non-normal data to find correlated genes. This may be a problem for gene microarray expression data, which is often not normally distributed. Alternatively, robust, "mega-clustering" methods have been developed to provide improved estimates of co-regulation for microarray data [25]. One such algorithm-the 'Gene Recommender'-has successfully predicted previously unknown interactions that were verified experimentally in a multicellular organism [26]. A key property of the Gene Recommender is the categorization of genes into clusters under different conditions (i.e., allowing for "biclusters") where different samples' contributions to the given cluster may vary. Since inexpensive genotyping platforms can presently interrogate >1 million SNPs and we are rapidly shifting into the era of whole genome sequencing, existing genetics and systems biology methods would be nicely complemented by computationally feasible, agnostic approaches to the detection of *trans*-acting factors that regulate groups of genes.

Therefore, here we extend the Gene Recommender with an approach that can systematically identify *trans* loci controlling gene co-regulation. We broadly refer to the identification of gene co-expression trait loci as 'co-regulatory expression quantitative trait loci' (creQTL) mapping. Unlike *trans*-eQTL analysis, our method does not consider individual transcripts but rather focuses on multiple co-regulated transcripts. We provide a genome-wide implementation of the Gene Recommender and a statistical framework for association testing. The key steps of the creQTL approach are: gene clustering; calculation of each sample's similarity to each cluster; and statistical testing of how well genotype explains the similarity. We applied creQTL to a study of germline variants and tumor expression in endometrial cancer [27] and identified many loci significantly associated with gene co-regulation. These loci were commonly in noncoding regions closely in *cis* with genes encoding proteins required for transcription, signaling, cell adhesion, and development. Our results suggest that associating genetic variants with co-regulation via creQTL mapping provides an efficient and agnostic avenue for detecting biological factors important in the coordinate regulation of groups of genes.

Methods

creQTL Approach

To identify *trans* regulatory associations, creQTL mapping tests for the association between genetic variants spanning the genome and clusters of co-regulated genes. The variant data can be obtained, for example, from single nucleotide polymorphism (SNP) arrays. Gene expression data can be obtained from standard gene expression arrays or other high-throughput methods for mRNA quantification.

Gene clustering is based on a recursive, heuristic implementation of the Gene Recommender (GR) [26], automated for genome-wide application. GR is a rank-based biclustering algorithm with an existing, open source R implementation [28]. Here we outline relevant portions of the GR. First, gene expression data are normalized to a uniform distribution with mean zero and variance 1/3. A suggested gene list, based on prior knowledge, of at least two putative co-regulated genes is input to GR. For each putative cluster input, each sample's relative score, the "$Z_E(j)$", is computed, as in Equation 1.

$$Z_{E(j)} = \sqrt{k_j}\,\frac{\overline{Y}_{Q,j}}{\sqrt{\hat{V}_{Q,j}+\dfrac{1}{3p^2}}} \tag{1}$$

where k_j is the number of gene expression values in sample *j*, *p* is the number of samples, and $\overline{Y}_{Q,j}$ is the median gene cluster expression value in sample *j*. Note that the original implementation of Gene Recommender used the mean gene cluster expression value in the numerator of $Z_E(j)$ [26]; later versions of the algorithm used the median [28]. Here we used the latest version. $\hat{V}_{Q,j}$ is the sample variance, given by

$$\hat{V}_{Q,j} = \frac{k_j - 1}{k_j}\,\mathrm{var}(Y_{Q,j}) \tag{2}$$

where $\mathrm{var}(Y_{Q,\ j})$ is the variance of the genes within the gene cluster in sample *j* (both equations, ref. [26]). $Z_E(j)$ has an approximate Student's t null distribution and larger values indicate tighter co-regulation of clustered genes within sample *j*. Once Z_E(j) has been computed, an incremental approach is used to compute the correlation coefficient, s.g.i., with a scoring function that minimizes the number of nonquery genes scoring higher than those in the query. The final s.g.i., which is proportional to the Euclidean distance, is then based on the most informative experiments.

It is not computationally feasible to compute all possible gene clusters from the dataset with GR, so heuristics are necessary to find tightly co-regulated gene clusters. Our extensions permit genome-wide use of the GR by automatically selecting putative co-regulated genes, running the algorithm, and then recursively modifying the query using leave-one-out cross validation (LOOCV) as a scoring function until the cluster converges at a point where all query genes contribute approximately equally to the cluster. The procedure is as follows. First, using each gene as a seed for a potential cluster, initial predictions of co-regulated genes are made using Spearman's rank correlation. If the number of highly correlated genes is less than 5 or more than 20, the putative cluster is expanded or trimmed to the most correlated 5 or 20 genes, respectively. Next, the Gene Recommender algorithm is run with this initial gene cluster, and then rerun using the top hits (by s.g.i, the Gene Recommender normalized correlation metric) from the initial run. If the seed gene scores highly after the second run (i.e., within the top 50 genes most correlated with the putative cluster) LOOCV is used to trim the cluster to only the highest scoring hits (regardless of gene set size). Once a tightly regulated cluster has been found, the next-highest scoring genes are added incrementally while stringently keeping LOOCV scores low. Ultimately, all genes within a predicted cluster will have an approximately equivalent contribution to the cluster, as determined by LOOCV.

Once gene clusters were assigned, for creQTL association testing, we computed a modified version of Z_E(j), called $Z2_E$(j), to describe how tightly regulated the gene cluster was within a given sample. The modified version replaces the numerator of Z_E(j) with the mean of all genes' contributions to the cluster and adds a small positive constant, *s0*, to the denominator in order to moderate variance and avoid extremely large values of $Z2_E$(j). *s0* was chosen by minimizing the coefficient of variation of the denominator of Z_E(j) across moving windows of data, akin to the strategy of Tusher et al. [29] for the moderation of t statistics. The modified statistic for association testing is

$$Z2_{E(j)} = \sqrt{k_j}\,\frac{mean(\bar{Y}_{Q,j}Y_{Q,j})}{\sqrt{\hat{V}_{Q,j}+\frac{1}{3p^2}}+s0} \qquad (3)$$

In place of the median expression value in the numerator of the original Z_E(j), we multiplied the sample's median gene expression value with the expression values for each individual gene within the cluster, and computed the mean of the resulting vector. Although this strategy will produce a higher $Z2_E$(j) in the case of a cluster with strongly driven outliers, it will also produce a lower $Z2_E$ (j) in the presence of weakly driven outliers, even when the cluster is otherwise strongly driven. Thus, a larger priority was given to penalizing strongly driven clusters with weakly driven outliers than the opposite. Overall, this produced smaller values of $Z2_E$(j) and gave a slightly greater dispersion of clusters with outliers when compared to taking the median gene expression value (not shown). Our software implementation provides a choice of either calculation for the numerator.

The association between genotypes and the $Z2_E$(j) statistic was then evaluated with Bartlett's k-sample variance test [30], using the 'bartlett.test' function in R to find significant differences in the variance of $Z2_E$(j) across each genotype using additive coding (i.e. AA, AB, BB). Resulting p-values were then adjusted for FDR using the 'p.adjust' function in R and the method of Benjamini and Hochberg [31].

Application of creQTL to Endometrial Cancer

We applied creQTL mapping to a study of 52 endometrial tumor samples with genotype and expression data available (NCBI GEO 14860; [27]). We downloaded the raw genotype and normalized microarray data from GSE14860. Genotype data were from Affymetrix 100K SNP chips, and calls were made with the 'crlmm' function from the 'oligo' library (v1.10.0) [32] for Bioconductor (v2.5; [33]) and R (v2.10.0; [34]). Prior to association testing, SNP markers were prefiltered to satisfy the following stringent conditions for each marker: a call confidence value of at least 95% in 51/52 samples, no significant deviations from Hardy-Weinberg equilibrium at p = 0.05 (with 'snpMatrix' v1.8.0 [32]), and at least two different genotypes observed with at least 5 samples for each genotype. This was analogous to requiring a SNP minor allele frequency of 5% if two genotypes were observed, and 14% if all three genotypes were seen. To enrich the microarray data for abundant and strongly-hybridized transcripts, we calculated the variance of each probe and excluded the lowest 50% from our analyses. Microarray data were ranked and normalized to mean 0 and variance 1/3 across each gene, the default for the Gene Recommender algorithm. SNP r^2 values were calculated with 'snpMatrix.' Because all 52 samples were from a single county in Norway [27], we did not perform any adjustments for population stratification.

For comparison with the creQTL results, we also undertook an eQTL analysis of the EC dataset. Here, the microarray data were normalized to have a mean of zero and variance equal to one within each individual sample. *cis* eQTL analysis evaluated all SNPs within 5 MB of a gene using an analysis of variance (ANOVA) of expression level on genotype and including covariates identified with surrogate variable analysis [35]. *trans* eQTL was done similarly, focusing on the remaining SNPs > 5 MB from any protein-coding gene. For *cis* and *trans* eQTL analyses, p-values were adjusted separately for FDR [31].

Finally, we evaluated whether associated genes were overrepresented in broad gene ontology categories for creQTL. We organized all Gene Ontology [36] annotations into Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) subcategories (Affymetrix, Santa Clara CA). SNPs were assigned to genes, as for *cis*-eQTL, using the minimum q-value for association across all tested SNPs within a 5 MB distance cutoff (to generate per-gene q-values). For all GO subcategories with at least 10 genes, the two-sample Kolmogorov-Smirnov test (one-tailed) was used to test for enrichment of significant genes in each Gene Ontology category. Specifically, to compute exact p-values for each category, we compared the observed q-value for a given gene set to those calculated from 10,000 same-sized sets of randomly drawn q-values. Because of the large number of categories, many of which contained overlapping genes, resulting p-values (for each GO category) were then adjusted for FDR [31].

Results

The key steps in our creQTL approach and application to identify loci controlling co-regulation of genes are outlined in Figure 1. Because of the overall unreliability of low-expression measurements in gene microarray expression data and our interest only in genes with dynamic expression patterns, we prefiltered those data to exclude the bottom 50% of probes by variance, leaving 8,543 for inclusion in our analysis of gene clustering and QTL testing. 68,523 SNPs from the Affymetrix 100K chip met cutoffs for inclusion in our analysis.

From the expression data, we identified 282 gene clusters with at least 3 co-regulated genes. The distribution of gene cluster sizes is shown in Figure 2. Many genes present within the same clusters had overlapping, known biological functions, and genes within individual clusters often were syntenic, such as the IFNβ-induced genes *IFI44, IFI44L*, and *IFIT1*, *-2*, and *-3*, on chromosomal regions 1p31.1 and 10q23-26, contained within a cluster significantly associated with a SNP/locus upstream of *RFC3* and *STARD13*.

We found 453 associations between genotype and $Z2_E$ (j) with q-values < = 0.005 (Additional File 1 contains annotated associations) after 19,323,486 variance tests of 282 gene clusters and 68,523 SNPs. Note that a FDR q-value of 0.005 corresponded to an unadjusted p-value of approximately 1 × 10^{-7}. Table 1 and Figure 3 highlight two of the most significant creQTLs that were within 5 MB of a protein-coding gene. As shown in Figure 3, for statistically significant creQTLs, the SNP genotypes exhibit varying contributions to the clusters. Moreover, samples that contribute most to a given cluster have more extreme, more tightly-clustered expression values; samples that contribute little to a given cluster have expression values (and contributions) closer to zero with greater variability. Additional File 2 shows the q-value density for associations as a function of gene cluster size. As can be seen in the figure, there was no relationship between q-values and gene cluster size.

For the association between rs9315220 and the cluster noted above, several other nearby SNPs (rs9315219, rs9315215, rs7981602, rs4943110) were also significantly associated with this cluster (see Additional File 1), reflecting the high linkage disequilibrium (LD) among them (the pairwise r^2 across all five SNPs ranged from 0.92 to 0.96 in these 52 samples). The smallest q-value was nearest to the *STARD13* locus (q-value = 1.13 × 10^{-4}). Another strong association (q = 5.40 × 10^{-4}) shown in Figure 2 was between the intronic SNP rs2296697 in *LPHN2* (latrophilin 2) and a cluster of genes including *C21orf58*, *CPNE6*, *SOX11*, *MMP8*, *LRTM1*, and one unannotated gene. The genes within this cluster have established roles in Ca^{2+}-dependent signaling and cancer [37-41] (**Discussion**). Interestingly, a conventional *trans*-eQTL analysis between the SNPs and individual expression values of these co-regulated genes did not detect any associations (see Table 1); these results were typical of other clusters (data not shown).

The strongest regulatory loci for creQTL (q = 5.50 × 10^{-4}) are shown in Table 2. We calculated pairwise r^2 between all 453 SNPs from creQTLs with q < 0.005 and eliminated those with r^2 values above 0.5, leaving 338 creQTLs. Of the 338 SNPs remaining, 190 were in noncoding regions; 145 were intronic; two were in coding sequences; and one was in the 3' untranslated region. Interestingly, many of the associated creQTLs contained SNPs in noncoding sequences >500 kb from the nearest gene. After assigning the 68,523 SNPs to 8,398 protein-coding genes in *cis* (**Methods**), we plotted the resulting 8,398 q-values versus *cis*-location relative to the nearest gene in Figure 4. Consistent with the pattern for the highest-scoring hits at q < 0.005, associations appeared most enriched in the intronic regions, and within the non-intronic, noncoding regions, significant associations

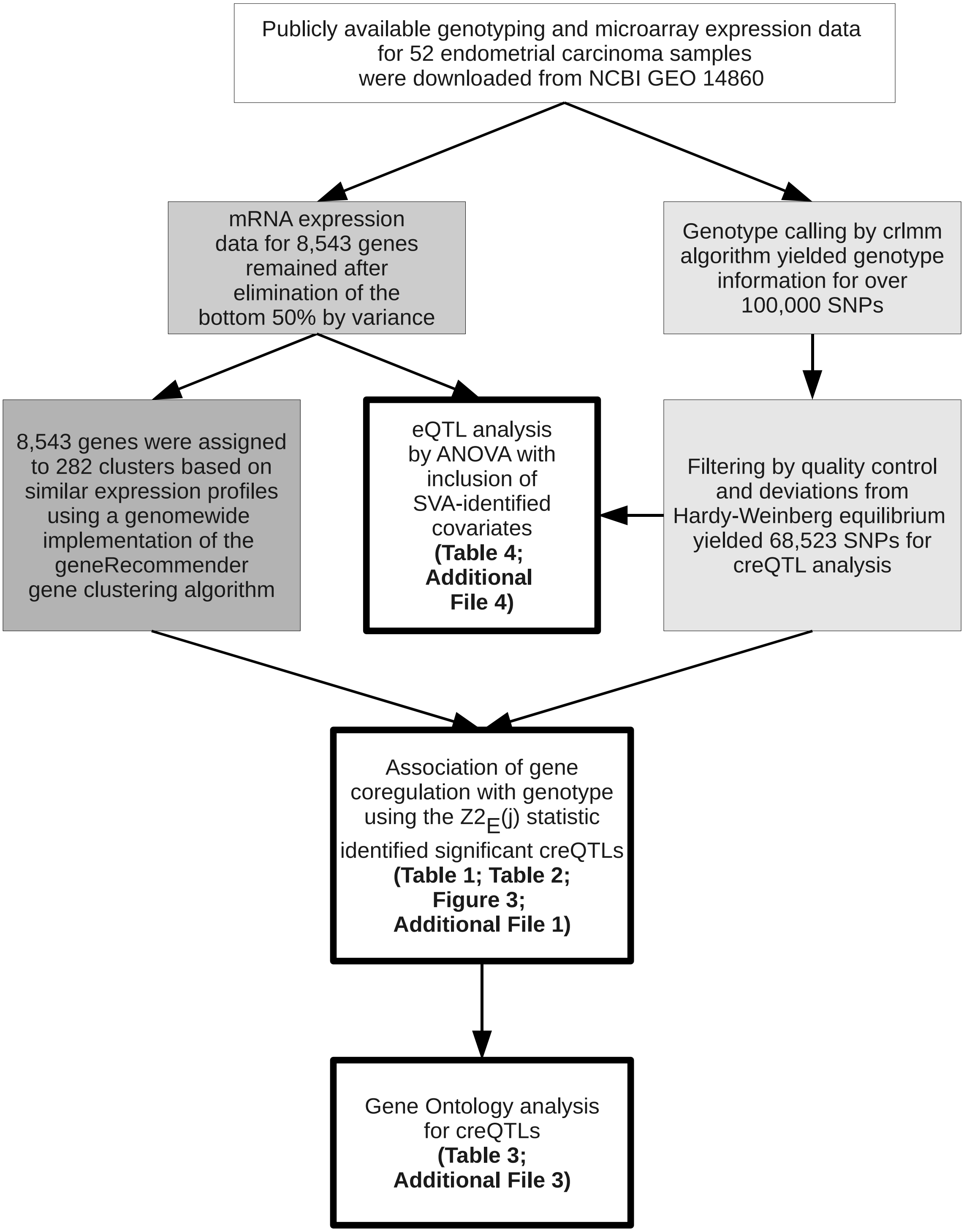
Publicly available genotyping and microarray expression data for 52 endometrial carcinoma samples were downloaded from NCBI GEO 14860
mRNA expression data for 8,543 genes remained after elimination of the bottom 50% by variance
Genotype calling by crlmm algorithm yielded genotype information for over 100,000 SNPs
8,543 genes were assigned to 282 clusters based on similar expression profiles using a genomewide implementation of the geneRecommender gene clustering algorithm
eQTL analysis by ANOVA with inclusion of SVA-identified covariates (Table 4; Additional File 4)
Filtering by quality control and deviations from Hardy-Weinberg equilibrium yielded 68,523 SNPs for creQTL analysis
Association of gene coregulation with genotype using the Z2E(j) statistic identified significant creQTLs (Table 1; Table 2; Figure 3; Additional File 1)
Gene Ontology analysis for creQTLs (Table 3; Additional File 3)

Figure 1 Overview of analysis.

appeared most in regions >10 KB away from the coding sequence.

We further analyzed the creQTL results with Gene Ontology in order to determine if any broad categories of genes participated in the regulation of gene clusters. There were 648, 176, and 310 Gene Ontology Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) annotations, respectively, for gene set enrichment testing that were represented on the SNP arrays after assigning the 68,523 SNPs to 8,398 protein-coding genes

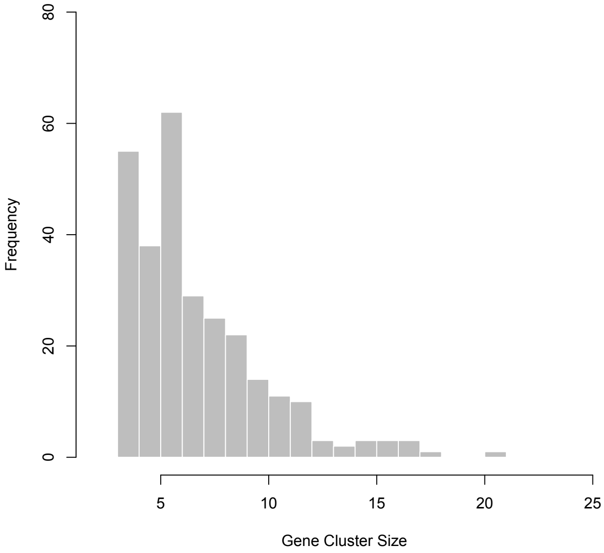

Figure 2 Histogram of gene cluster sizes identified from the EC gene expression data.

in *cis*, using an arbitrary minimum cutoff of 10 genes for a category to be tested. Top hits are shown in Table 3 at a more relaxed q = 0.05. The most significant ontologies were related to cell adhesion, potassium- and calcium ion transport, binding, channel activity, development, and the nervous system, particularly development and plasticity. For Biological Process and Molecular Function, as expected, some of the most enriched categories were related to transcription factors and other aspects of transcriptional regulation. For Molecular Function, categories '0003700' (transcription factor activity) and '0003714' (transcription corepressor activity), with 497 and 74 genes, respectively, were just beyond the FDR significance cutoff with adjusted p-values of 0.056 and 0.078 (Additional File 3). For Cellular Component, nearly all significant categories were directly related to various aspects or cell-type specific (e.g. neuronal) specializations of the cellular membrane (9 of 13 hits at q = 0.05).

eQTL analysis identified no significant *cis*-eQTLs (q = 0.05); the smallest FDR-adjusted significance level for *cis*-eQTLs was q = 0.098 (not shown). There were 63 significant *trans*-eQTLs (q = 0.05; Additional File 4), and the most significant are shown in Table 4. A raw p-value of approximately 6×10^{-9} corresponded to q = 0.05. Even with a high 50% FDR cutoff, there were only 1,293 *cis*-eQTLs (of 1,468,327), or 0.09%, and 290,625 (of 507,010,655), or 0.06% *trans*-eQTLs. Although marginally significant (q = 0.049), the association between *ZNF639* and *TCF12* is included in Table 4 due to association of 3q26 with survival in this dataset [27] (**Discussion**). We did not perform gene ontology analysis of the eQTL experiment since there were no clear *cis*-eQTLs detected.

Discussion

creQTL provides an agnostic framework for predicting *trans* regulators of clusters of genes. It does not require any one particular biclustering method or statistical test and can be extended to any species for which genetic markers (including those derived from linkage) and gene expression data can be obtained; for example, creQTL mapping in mice would allow the study of gene regulation in many disease models and in normal biological processes such as development, which are highly relevant to human disease. Like eQTLs, creQTLs may be incorporated into additional analyses. As recent studies have used eQTL data to assemble genetic loci into directed, Bayesian networks to predict causal relationships (e.g. [16,42]), future studies could assemble creQTLs into directed or undirected gene networks. Because creQTL mapping groups co-regulated genes into a smaller number of modules, network construction using creQTL information in place of eQTL information should be computationally more efficient.

In our application to endometrial cancer, there was an abundance of creQTLs in noncoding regions and introns, and relatively very few in coding regions. Polymorphisms in coding regions, although observed less frequently, would likely generate more pronounced effects, and for signaling molecules, these could result in substantial downstream effects important for carcinogenesis due to a change in function of the protein. However, these alleles, which are likely to be under strong negative selection and very rare, would not be covered by the genotyping platform used in the present study. Unlike the results of eQTL studies, our strongest hits were not typically located in *cis* regions very close to the coding sequences of individual genes, but rather in more distant, noncoding regions located over 10 KB from the nearest gene. This suggests that regions that regulate clusters of genes, unlike those regulating individual genes, are located farther from coding sequences and may possibly function as general enhancers. However, without careful experimentation, such as promoter reporter gene assays that would specifically test the enhancer activity of these sequences, it is difficult to predict their actual roles. Future studies could address the relationship between location, allele frequency, and other empirical aspects of creQTLs in larger cohorts and healthy tissue. Collapsing SNPs into a dominant model would permit testing of less common SNPs, as requiring five observations in each genotype essentially restricted our analysis to more common variants.

Our results for creQTL from the Gene Ontology enrichment analysis indicated a regulatory hierarchy, consistent with data from wet-lab studies, where signal transduction molecules receive signals at the plasma membrane and transduce them through various intracellular intermediates

Table 1 Two representative, significant creQTL associations

creQTL 1

SNP Annotation

Locus	dbSNP	Location (bp)	Position	Gene Symbol	Gene Title
chr13q12.3-q13	rs9315220	304602	upstream	*RFC3*	replication factor C (activator 1) 3, 38 kDa
chr13q12-q13	rs9315220	227702	upstream	*STARD13*	StAR-related lipid transfer (START) domain containing 13

Gene Cluster Annotation

Locus	Entrez ID	q_{creQTL}	p_{eQTL}	Gene Symbol	Gene Title
chr1p31.1	10561	1.13×10^{-4}	0.21	*IFI44*	interferon-induced protein 44
chr1p31.1	10964		0.28	*IFI44L*	interferon-induced protein 44-like
chr10q23-q25	3433		0.21	*IFIT2*	interferon-induced protein with tetratricopeptide repeats 2
chr10q24	3437		0.49	*IFIT3*	interferon-induced protein with tetratricopeptide repeats 3
chr10q25-q26	3434		0.19	*IFIT1*	interferon-induced protein with tetratricopeptide repeats 1
chr12q24.1	4938		0.52	*OAS1*	2',5'-oligoadenylate synthetase 1, 40/46 kDa
chr21q22.3	4599		0.56	*MX1*	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)

creQTL 2

SNP Annotation

Locus	dbSNP	Location (bp)	Position	Gene Symbol	Gene Title
chr1p31.1	rs2296697	0	intron	*LPHN2*	latrophilin 2

Gene Cluster Annotation

Locus	Entrez ID	q_{creQTL}	p_{eQTL}	Gene Symbol	Gene Title
chr2p25	6664	5.40×10^{-4}	0.92	*SOX11*	SRY (sex determining region Y)-box 11
chr3p14.3	57408		0.38	*LRTM1*	leucine-rich repeats and transmembrane domains 1
chr11q22.3	4317		0.10	*MMP8*	matrix metallopeptidase 8 (neutrophil collagenase)
chr14q11.2	9362		0.38	*CPNE6*	copine VI (neuronal)
chr21q22.3	54058		0.64	*C21orf58*	chromosome 21 open reading frame 58
NA	NA		0.35	NA	NA

For each of the two creQTLs shown in this table (labeled 'creQTL 1' and 'creQTL 2'), the associated SNP (labeled 'SNP Annotation') is first shown with annotation for the nearest coding genes in *cis*. Beneath the SNP annotation is annotation for the significantly associated, co-regulated gene cluster (labeled 'Gene Cluster Annotation'). These two significant associations correspond to the two loci shown in Figure 3. 'Position' indicates the relative location of the SNP relative to the nearest coding gene; 'Location (bp)' indicates the physical distance, in base pairs, between the SNP and the nearest coding genes; p_{eQTL} indicates unadjusted significance of standard eQTL analysis.

to the nucleus to activate specific transcription factors, other DNA-binding proteins, and other transcriptional regulatory proteins. The relative abundances of DNA-binding proteins at the promoter region may control gene expression [43]. Because of the enrichment of cell adhesion and signaling ontologies at the cell junction and membrane at the highest significance levels, the major steps of gene regulation in endometrial cancer may occur not within the nucleus but rather at the cellular membrane. In addition, GO enrichment in genes required for telomere maintenance suggests additional target genes and loci for further investigation, as telomerase has long been considered an attractive target for therapy in endometrial carcinoma [44] and more recently, a mouse model of endometrial cancer showed that telomere length affected initiation of type II carcinogenesis [45].

Gene Ontology analysis is not without its own pitfalls. Besides the obvious effects of overlapping annotations for the majority of genes and annotation bias, the categories themselves often do not describe real biological phenomena. Given that transcription factors are very cell-type specific, a category that includes all of them is not particularly useful. In light of this, the q-values for transcription factor-related categories are likely to be overly conservative, and would probably be smaller if those TFs not expressed in this tissue were excluded. Future studies could address this problem using array and wet lab expression data.

For the association shown in Figure 2, the nearest coding sequence to rs9315220 is *STARD13*, a.k.a. *DLC2* (deleted in liver cancer 2). All genes of this cluster but one (OAS1), including *MX1, IFI44, IFIT2, IFIT3, IFI44L*, and *IFIT1*, are known to be induced by IFNβ in humans [46]. IFNβ, besides its role in the treatment of multiple sclerosis (MS), is anti-tumorigenic [47-50]. Similarly, many genes from this cluster also have established or

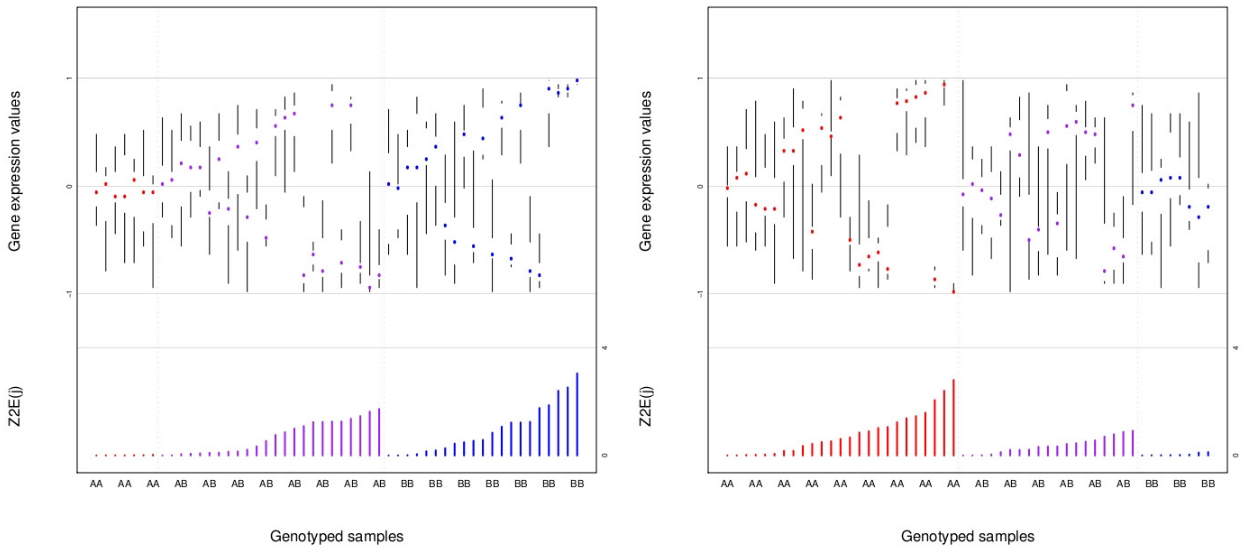

Figure 3 creQTL1 (left panel) and creQTL2 (right panel) from Table 1, with SNP IDs rs9315220 and rs2296697, respectively. For each association, co-regulation of a given cluster and a SNP is shown. The 52 endometrial carcinoma samples are color coded by genotype on the x-axis (upper panel), with every third sample labeled beneath the axis. On the y-axis (upper panel), normalized gene cluster expression values are plotted. The median, 25th/75th quantiles, minimum, and maximum gene values across all genes in the cluster are shown for each sample. The bottom panel of each figure shows the $Z2_E(j)$ statistic for each sample and the given cluster.

putative roles in cancer; *IFI44* was part of a small set of genes independently validated for recurrence status in non-small cell lung carcinoma [37], is anti-proliferative in melanoma cell lines [38], and is upregulated in squamous cell carcinoma [39]. *IFIT2* inhibits migration and proliferation in squamous cell carcinoma cultures [41]. *IFIT1* was one of a small group of genes that predict carcinogenesis after training on DNA-damage response [40], while *STARD13* is upregulated in human lymphocytes following gamma-irradiation [51], suggesting that

Table 2 Top scoring creQTLs at a significance threshold of q = 5.50 × 10^{-4}

Locus	dbSNP	Location (bp)	Position	Gene Symbol	Gene Title	q_{creQTL}
chr6q16	rs959186	125994	downstream	*RNGTT*	RNA guanylyltransferase and 5'-phosphatase	1.13 × 10^{-4}
chr13q12.3-q13	rs9315215	348455	upstream	*RFC3*	replication factor C (activator 1) 3, 38 kDa	1.13 × 10^{-4}
chr13q12-q13	rs9315215	183849	upstream	*STARD13*	StAR-related lipid transfer (START) domain containing 13	1.13 × 10^{-4}
chrXq21.33	rs4969656	695921	upstream	*DIAPH2*	diaphanous homolog 2 (Drosophila)	1.93 × 10^{-4}
chrXq21.3-q22	rs4969656	2315179	upstream	*NAP1L3*	nucleosome assembly protein 1-like 3	1.93 × 10^{-4}
chr3q13.1	rs4856121	2246645	upstream	*ALCAM*	activated leukocyte cell adhesion molecule	1.93 × 10^{-4}
chr5q31-q32	rs319217	0	intron	*PPP2R2B*	protein phosphatase 2 (formerly 2A), regulatory subunit B, beta isoform	1.13 × 10^{-4}
chr1p31.1	rs2296697	0	intron	*LPHN2*	latrophilin 2	5.40 × 10^{-4}
chr2p21	rs222471	31099	downstream	*KCNG3*	potassium voltage-gated channel, subfamily G, member 3	4.91 × 10^{-4}
chr2p21	rs222471	49701	upstream	*COX7A2L*	cytochrome c oxidase subunit VIIa polypeptide 2 like	4.91 × 10^{-4}
chr4q21	rs1992489	59023	upstream	*CXCL13*	chemokine (C-X-C motif) ligand 13	5.19 × 10^{-4}
chr4q21.1	rs1992489	282671	downstream	*CCNG2*	cyclin G2	5.19 × 10^{-4}
chr14q13-q21	rs1951319	680177	downstream	*RPL10L*	ribosomal protein L10-like	5.48 × 10^{-4}
chr14q21.2	rs1951319	717439	upstream	*C14orf106*	chromosome 14 open reading frame 106	5.48 × 10^{-4}
chr8p23.2	rs1714757	0	intron	*CSMD1*	CUB and Sushi multiple domains 1	1.47 × 10^{-4}
chr16p13.12	rs1159167	147599	upstream	*ERCC4*	excision repair cross-complementing rodent repair deficiency, complementation group 4	4.91 × 10^{-4}
chr16p13.12	rs1159167	968670	upstream	*CPPED1*	calcineurin-like phosphoesterase domain containing 1	4.91 × 10^{-4}
chr22q12.1	rs10483151	0	intron	*TTC28*	tetratricopeptide repeat domain 28	1.81 × 10^{-4}

Associated gene clusters have been omitted for brevity and are listed in Additional File 1. 'Position' and 'Location (bp)' as for Table 1; q_{creQTL}' indicates FDR-adjusted significance.

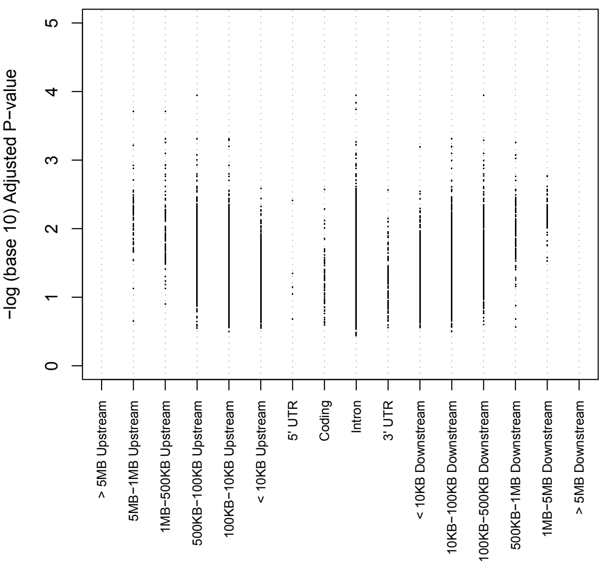

Figure 4 Genewise genomic locations of SNP q-values. After assigning 68,523 SNPs to 8,398 genes based on the minimum association q-value within 5 MB of the nearest gene's coding sequence, q-values were binned into categories based on their relative position to the nearest gene. Note that the q-values do not reach the bottom of the plot as they were chosen as the *minimum* value within 5 MB of the nearest gene.

IFIT1 may be a downstream target of *STARD13* following DNA damage.

STARD13/DLC2 encodes a Rho GTPase activating protein and was identified from a region of chromosome 13 exhibiting a loss of heterozygosity in hepatocellular carcinoma [52]. It is a tumor suppressor that antagonizes Rho expression [53], and its downregulation or deletion has been reported in multiple cancers [54,55]. Supporting its role as a regulator of IFNβ-induced genes, an intronic SNP in *STARD13* was amongst eighteen that were replicated in a study of responders to IFNβ therapy in MS patients [56]. Further, SNPs in *OAS1* have been associated with MS susceptibility [57]. Therefore, this particular gene cluster and its predicted regulatory partner, *STARD13*, may be interacting determinants of the response to various negative stimuli. This also proposes the hypothesis that the role of *STARD13* as a tumor suppressor is from its induction in response to DNA damage, and that deletion of this region in at least ten different cancers is pathogenic possibly due to hyper-activation of Rho, which promotes migration, proliferation, and invasion [58]. In this case, beyond improvement of the total lack of significant association from the individual eQTLs, creQTL proposed a more complete picture of the biological pathway.

We detected a strong association of rs2296697, within an intron of *LPHN2* (latrophilin 2), with a cluster of genes including *C21orf58, CPNE6, SOX11, MMP8*, and *LRTM1*. *LPHN2* belongs to the latrophilin subfamily of G-protein coupled receptors, with known roles in cell adhesion and signal transduction [59]. Cell adhesion was among the most enriched categories from Gene Ontology analysis of creQTL. *CPNE6* (copine VI), is a member of the ubiquitous copine family, which bind phospholipid in Ca^{2+}-dependent manner and have roles in cellular division and growth [60]. Although very little is known regarding *CPNE6*, other members of the copine family have been shown to promote tumor cell migration [61] and repress *NF-KB* transcription [62]. *SOX11* is a TF with important roles in brain development that is strongly upregulated in lymphoma [63] and malignant glioma, where it may affect tumorigenesis [64,65]. *SOX11* may also be a prognostic marker for recurrence-free survival in another uterine neoplasia, epithelial ovarian cancer [66]. *MMP8* (human neutrophil collagenase) is a member of the matrix metallopeptidase family. Recent work has shown that a SNP in *MMP8* may be a predictor of lung cancer risk [67], and that higher plasma levels of *MMP8* may protect against lymph node metastasis in breast cancer [68]. Consistent with these findings, somatic mutations of *MMP8* were common in melanomas, and mutated *MMP8* failed to inhibit tumor formation *in vivo* [69], suggesting widespread roles for *MMP8* in cancer progression. Three members of this gene cluster have no known roles, yet the functions of the other cluster members suggest they could contribute to cancer progression. In this case, the association of this cluster of genes with *LPHN2* has suggested that *LPHN2* may be a regulatory point for the modulation of genes with important roles in cancer, including *MMP8* and *SOX11*.

eQTL analysis was not done in the original paper presenting this dataset [27], and the general lack of findings from the eQTL analysis may be from a lack of power, given that there were only 52 samples. This is not surprising: an eQTL study of 60 samples from the CEU cohort in the HapMap data identified only 10 *cis*-eQTLs and 94 in *trans* [70], although there may have been some technical issues here [71].

However, in the present study, creQTL appeared to identify interactions that were supported by recent work in the literature. A very compelling benefit of creQTL over *trans*-eQTL is greater computational efficiency and a reduced multiple testing burden, as there was an approximate 25-fold reduction in the number of statistical tests. By organizing genes into co-regulated clusters prior to statistical testing, the problem becomes more computationally feasible and the resulting output more biologically interpretable. The identification of more significant hits after adjustment was likely not simply due to a lighter multiple testing penalty; *cis*-eQTL required 1,468,327 tests, far less than the 19,323,486 tests for

Table 3 Gene Ontology enrichment analysis for creQTL at q = 0.05

Gene Ontology Biological Process

Category ID	Category Size	Category Title	q_{KS}
7214	18	gamma-aminobutyric acid signaling pathway	< 0.008
7612	21	learning	< 0.008
7411	62	axon guidance	< 0.008
7156	69	homophilic cell adhesion	< 0.008
7417	80	central nervous system development	< 0.008
7155	320	cell adhesion	< 0.008
45944	203	positive regulation of transcription from RNA polymerase II promoter	0.008
7399	225	nervous system development	0.008
45595	12	regulation of cell differentiation	0.014
48754	19	branching morphogenesis of a tube	0.018
30900	42	forebrain development	0.018
45165	29	cell fate commitment	0.027
43065	69	positive regulation of apoptosis	0.04
7224	16	smoothened signaling pathway	0.043
7420	74	brain development	0.043
30326	35	embryonic limb morphogenesis	0.043
7169	56	transmembrane receptor protein tyrosine kinase signaling pathway	0.043
6813	94	potassium ion transport	0.043

Gene Ontology Cellular Component

Category ID	Category Size	Category Title	q_{KS}
5913	18	cell-cell adherens junction	< 0.004
45202	156	synapse	< 0.004
30054	231	cell junction	< 0.004
14069	49	postsynaptic density	0.004
45211	79	postsynaptic membrane	0.004
30424	83	axon	0.006
781	13	chromosome, telomeric region	0.013
5912	27	adherens junction	0.013
34707	36	chloride channel complex	0.016
5923	40	tight junction	0.016
5886	1504	plasma membrane	0.037
42734	21	presynaptic membrane	0.038
5578	174	proteinaceous extracellular matrix	0.038

Gene Ontology Molecular Function

Category ID	Category Size	Category Title	q_{KS}
8066	11	glutamate receptor activity	< 0.008
5216	187	ion channel activity	< 0.008
5509	491	calcium ion binding	< 0.008
4993	11	serotonin receptor activity	< 0.008
4970	12	ionotropic glutamate receptor activity	0.009
5234	12	extracellular-glutamate-gated ion channel activity	0.009
43565	273	sequence-specific DNA binding	0.009
5244	85	voltage-gated ion channel activity	0.016
4890	17	GABA-A receptor activity	0.017
5246	11	calcium channel regulator activity	0.019
31404	49	chloride ion binding	0.023
30594	18	neurotransmitter receptor activity	0.028
30165	32	PDZ domain binding	0.029

Table 3 Gene Ontology enrichment analysis for creQTL at q = 0.05 *(Continued)*

5001	13	transmembrane receptor protein tyrosine phosphatase activity	0.029
5267	53	potassium channel activity	0.029
5230	24	extracellular ligand-gated ion channel activity	0.037
5254	48	chloride channel activity	0.041
8146	26	sulfotransferase activity	0.041
16455	11	RNA polymerase II transcription mediator activity	0.042

Results are presented for Biological Process, Cellular Component, and Molecular Function, respectively. Category IDs are presented minus leading zeroes. ‘q_{KS}’ indicates FDR-adjusted significance of the category.

creQTL, and produced no statistically significant results even at the more liberal q-value cutoff of 0.05. Overall, these results suggest that individual SNP-gene interactions are more difficult to detect (as in eQTL) when compared to relative changes in the expression levels of tightly co-regulated genes, in agreement with previous work in yeast [9]. creQTL may be better-suited to the identification of the strongest *trans* drivers of gene expression because it better explains the data; i.e., genes do not work individually to exert their biological effects, but rather in tightly coordinated groups. However, our results regarding the location of these drivers of co-regulation relative to coding sequences suggests that they might be disjoint from *cis* drivers of individual gene expression.

Some *trans*-eQTL hits were significant and may merit further investigation based on previously established biological roles for associated loci, indicating complementary utility of eQTL and that the two methods may be most useful when applied side-by-side. Copy number analysis of this EC dataset by the original authors revealed two regions of gain that were predictive of survival - 3q26.32 and 12p12.1 [27]. Deletions were not considered. Because *PIK3CA* is located in this region, the authors proposed a role for the PI3-kinase pathway, and found supporting evidence for this theory based on indirect, bioinformatic analysis of existing, *in vitro* data. Located 200 kb from *PIK3CA* is *ZNF639*, a.k.a. *ZASC1*, which is often contained within the same amplicon in cancer [72,73]. *ZNF639* was originally identified in squamous cell carcinoma as a Kruppel-like transcription factor with mRNA expression levels prognostic for survival and metastasis [72,73]. *trans*-eQTL identified a SNP less than 35 kb upstream of *ZNF639* (SNP_A-1686963) driving expression of the *TCF12* gene at 15q21. *TCF12*, a.k.a. *HTF4* or *ME1*, is a basic helix-loop-helix TF with key roles in development [74]. In a mouse experiment, *TCF12* was associated with obesity [14], a key prognostic factor for endometrial cancer [75]. Because the stoichiometric concentration of TFs within the nucleus may control gene expression [43], this association proposes a link between obesity and *TCF12* through *ZNF639*. Because *ZNF639* regulates anchoring of E-cadherin to the cytoskeleton through alpha N-catenin [76], given the

Table 4 The most significant associations from *trans*-eQTL analysis at q = 0.05

SNP Locus	dbSNP	Location (bp)	Position	SNP Gene Symbol	mRNA Locus	Entrez ID	mRNA Gene Symbol	q_{eQTL}
chrXp22.11	rs4828879	23770	upstream	*PRDX4*	chr11p15.5	7140	*TNNT3*	0.003
chr7q32.3	rs277491	0	intron	*PLXNA4*	chr3p26-p25	7862	*BRPF1*	0.003
chr4q32.3	rs10517754	0	intron	*FSTL5*	chr22q12.1	25770	*C22orf31*	0.005
chr7q34	rs2363830	0	intron	*DENND2A*	chr11q23-q24	56	*ACRV1*	0.016
chr3q22.1	rs938243	0	intron	*CPNE4*	chr16q23.3	93517	*SDR42E1*	0.018
chr2q13	rs10496425	0	intron	*CCDC138*	chr7q11.23	7461	*CLIP2*	0.019
chr2q13	rs7591305	0	intron	*CCDC138*	chr7q11.23	7461	*CLIP2*	0.019
chr10q21.1	rs10509024	0	intron	*PCDH15*	chrXp11.23	778	*CACNA1F*	0.029
chr3p24.3	rs964910	561192	upstream	*SGOL1*	chr1q44	114548	*NLRP3*	0.029
chr21q21.1	rs2826728	0	intron	*NCAM2*	chr11p15.5	7140	*TNNT3*	0.029
chr21q21.1	rs2155798	0	intron	*NCAM2*	chr11p15.5	7140	*TNNT3*	0.029
chr1q43	rs2790645	135107	upstream	*CHRM3*	chr5q31-q32	5521	*PPP2R2B*	0.029
chr11q14.1	rs62388	0	intron	*DLG2*	chr7q34	154790	*CLEC2L*	0.029
chr3q26.33	rs9290675	34917	upstream	*ZNF639*	chr15q21	6938	*TCF12*	0.049

’Position’ indicates SNP genomic location; ‘ q_{eQTL}’ indicates FDR-adjusted significance. Annotation is shown for SNPs in the first 5 columns, then for mRNA expression; e.g. for the first (non-header) row of this table, the mRNA expression levels for gene *TNNT3* were significantly associated with genotype at rs4828879, which is 23,770 base pairs upstream of the *PRDX4* locus.

importance of cell adhesion proteins in cancer [77], future studies of endometrial cancer might focus on *ZNF639*.

Conclusions

With the advent of low-cost SNP and expression arrays, human genetics has become a common tool for the dissection of gene regulation. Genetic approaches to the study of transcriptional regulation are compelling because they can detect regulatory molecules at all stages of gene regulation. Our approach expands upon previous methods and uses genetic variation to help identify transcriptional regulatory mechanisms, providing a biologically intuitive approach for detecting potential links between genotype and gene co-regulation.

Additional material

Additional file 1: Significant creQTLs. All significant creQTLs at q = 0.005 with annotation.

Additional file 2: creQTL q-values by gene cluster size. Plot of q-values for all associations, grouped by gene cluster size. The right panel is a zoomed view of the left panel, highlighting the rejection region. For both panels, the legend at upper left indicates the number of genes in the cluster.

Additional file 3: creQTL Gene Ontology. All results of Gene Ontology testing for creQTL.

Additional file 4: *trans*-eQTL. Results of *trans*-eQTL analysis at q = 0.05, annotated.

List of Abbreviations

ANOVA: analysis of variance; BP: gene ontology biological process; CC: gene ontology cellular component; creQTL: co-regulatory expression quantitative trait locus; EC: endometrial cancer; EDF: empirical distribution-free; eQTL: expression quantitative trait locus; FDR: false discovery rate; GO: gene ontology; KS: Kolmogorov-Smirnov test; LD: linkage disequilibrium; LOOCV: leave-one-out cross-validation; MF: gene ontology molecular function; mRNA: messenger ribonucleic acid; MS: multiple sclerosis; SNP: single nucleotide polymorphism; TF: transcription factor; TFBS: transcription factor binding site.

Acknowledgements

We appreciate helpful feedback from Joel Mefford, Jun Song, Sung Kim, Jill Hardin, Mike Baldwin, and Niall Cardin. This work was supported by grant CA88164 from the National Institutes of Health. The funding body had no role in the design, collection, analysis, interpretation of data, writing, or decision to submit the manuscript for publication. The authors have no competing interests, financial or otherwise.

Authors' contributions

Conceived and designed the experiments: KSK and JSW. Analyzed the data: KSK. Wrote software: KSK. Wrote the paper: KSK and JSW. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 23 July 2010 Accepted: 12 January 2011
Published: 12 January 2011

References

1. Brem RB, Yvert G, Clinton R, Kruglyak L: Genetic dissection of transcriptional regulation in budding yeast. *Science* 2002, **296(5568)**:752.
2. Jansen RC, Nap JP: Genetical genomics: the added value from segregation. *TRENDS in Genetics* 2001, **17(7)**:388-391.
3. Wang T, Stormo GD: Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 2003, **19(18)**:2369.
4. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics* 2003, **35(1)**:57-64.
5. Troyanskaya O, Dolinski K, Owen A, Altman R, Botstein D: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences* 2003, **100(14)**:8348.
6. Kostka D, Spang R: Finding disease specific alterations in the co-expression of genes. *Bioinformatics* 2004, **20(Suppl 1)**:i194.
7. Choi JK, Yu U, Yoo OJ, Kim S: Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* 2005, **21(24)**:4348.
8. Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D: Learning a prior on regulatory potential from eQTL data. *PLoS Genetics* 2009, **5(1)**:e1000358.
9. Lee SI, Pe'Er D, Dudley AM, Church GM, Koller D: Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proceedings of the National Academy of Sciences* 2006, **103(38)**:14062.
10. Gat-Viks I, Meller R, Kupiec M, Shamir R: Understanding Gene Sequence Variation in the Context of Transcriptional Regulation in Yeast. *PLoS Genet* 2010, **6(1)**:e1000800.
11. Zhang B, Horvath S: A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 2005, **4(1)**:1128.
12. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S: Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome* 2007, **18(6)**:463-472.
13. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S: Genetics of gene expression and its effect on disease. *Nature* 2008, **452(7186)**:423-428.
14. Yang X, Deignan JL, Qi H, Zhu J, Qian S, Zhong J, Torosyan G, Majid S, Falkard B, Kleinhanz RR: Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nature Genetics* 2009, **41(4)**:415-423.
15. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK: Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008, **452(7186)**:429-435.
16. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C: An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* 2005, **37(7)**:710-717.
17. Kim S, Xing EP: Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *PLoS Genet* 2009, **5(8)**:e1000587.
18. Kim S, Sohn KA, Xing EP: A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* 2009, **25(12)**:i204.
19. Liu B, De La Fuente A, Hoeschele I: Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* 2008, **178(3)**:1763.
20. Huang Y, Wuchty S, Ferdig MT, Przytycka TM: Graph theoretical approach to study eQTL: a case study of Plasmodium falciparum. *Bioinformatics* 2009, **25(12)**:i15.
21. Biswas S, Storey J, Akey J: Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics* 2008, **9(1)**:244.
22. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics* 2007, **39(6)**:730-732.
23. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* 2003, **34(2)**:166-176.

24. Cheng Y, Church GM: **Biclustering of expression data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:93-103.
25. Turner HL, Bailey TC, Krzanowski WJ, Hemingway CA: **Biclustering models for structured microarray data.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005, **2(4)**:316-329.
26. Owen AB, Stuart J, Mach K, Villeneuve AM, Kim S: **A gene recommender algorithm to identify coexpressed genes in C. elegans.** *Genome Research* 2003, **13(8)**:1828.
27. Salvesen HB, Carter SL, Mannelqvist M, Dutt A, Getz G, Stefansson IM, Raeder MB, Sos ML, Engelsen IB, Trovik J: **Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation.** *Proceedings of the National Academy of Sciences* 2009, **106(12)**:4834.
28. **Gene Recommender.** [http://www.bioconductor.org/help/bioc-views/release/bioc/html/geneRecommender.html].
29. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences* 2001, **98(9)**:5116-5121.
30. Bartlett MS: **Properties of sufficiency and statistical tests.** *Proceedings of the Royal Society of London Series A* 1937, **160**:268-282.
31. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57(1)**:289-300.
32. Carvalho B, Bengtsson H, Speed TP, Irizarry RA: **Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.** *Biostatistics* 2007, **8(2)**:485.
33. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10)**:R80.
34. Ihaka R, Gentleman R: **R: A language for data analysis and graphics.** *Journal of Computational and Graphical Statistics* 1996, **5(3)**:299-314.
35. Leek JT, Storey JD: **Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis.** *PLoS Genet* 2007, **3(9)**:e161.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25(1)**:25-29.
37. Lee ES, Son DS, Kim SH, Lee J, Jo J, Han J, Kim H, Lee HJ, Choi HY, Jung Y: **Prediction of Recurrence-Free Survival in Postoperative NonñSmall Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression.** *Clinical Cancer Research* 2008, **14(22)**:7397.
38. Hallen LC, Burki Y, Ebeling M, Broger C, Siegrist F, Oroszlan-Szovik K, Bohrmann B, Certa U, Foser S: **Antiproliferative activity of the human IFN--inducible protein IFI44.** *Journal of Interferon and Cytokine Research* 2007, **27(8)**:675-680.
39. Ye H, Yu T, Temam S, Ziober BL, Wang J, Schwartz JL, Mao L, Wong DT, Zhou X: **Transcriptomic dissection of tongue squamous cell carcinoma.** *BMC Genomics* 2008, **9(1)**:69.
40. Weichselbaum RR, Ishwaran H, Yoon T, Nuyten DSA, Baker SW, Khodarev N, Su AW, Shaikh AY, Roach P, Kreike B: **An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer.** *Proceedings of the National Academy of Sciences* 2008, **105(47)**:18490.
41. Lai KC, Chang KW, Liu CJ, Kao SY, Lee TC: **IFN-induced protein with tetratricopeptide repeats 2 inhibits migration activity and increases survival of oral squamous cell carcinoma.** *Molecular Cancer Research* 2008, **6(9)**:1431.
42. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C: **Mapping the genetic architecture of gene expression in human liver.** *PLoS Biol* 2008, **6(5)**:e107.
43. MacArthur S, Li XY, Li J, Brown J, Chu HC, Zeng L, Grondona B, Hechmer A, Simirenko L, Keränen S: **Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions.** *Genome Biology* 2009, **10(7)**:R80.
44. Yokoyama Y, Wan X, Shinohara A, Takahashi Y, Tamaya T: **Hammerhead ribozymes to modulate telomerase activity of endometrial carcinoma cells.** *Human cell: official journal of Human Cell Research Society* 2001, **14(3)**:223.
45. Akbay EA, Contreras CM, Perera SA, Sullivan JP, Broaddus RR, Schorge JO, Ashfaq R, Saboorian H, Wong KK, Castrillon DH: **Differential roles of telomere attrition in type I and II endometrial carcinogenesis.** *American Journal of Pathology* 2008, **173(2)**:536.
46. Serrano-Fernandez P, Muller S, Goertsches R, Fiedler H, Koczan D, Thiesen HJ, Zettl UK: **Time course transcriptomics of IFNB1b drug therapy in multiple sclerosis.** *Autoimmunity* 2010, 33-2702.
47. Ogasawara SYH, Momosaki S, Akiba J, Nishida N, Kojiro S, Moriya F, Ishizaki H, Kuratomi K, Kojiro M: **Growth inhibitory effects of IFN-beta on human liver cancer cells in vitro and in vivo.** *J Interferon Cytokine Res* 2007, **27(6)**:507-516.
48. Recchia F, Frati L, Rea S, Torchio P, Sica G: **Minimal residual disease in metastatic breast cancer: treatment with IFN-beta, retinoids, and tamoxifen.** *J Interferon Cytokine Res* 1998, **18(1)**:41-47.
49. Flörcken A, Denecke T, Kretzschmar A, Gollasch H, Reich G, Westermann J: **Long-Lasting Remission of Pulmonary Metastases of Renal Cell Cancer under IFN- Therapy in a Patient with Multiple Sclerosis.** *Onkologie* 2006, **29(8-9)**:382-384.
50. Murase T, Hotta T, Saito H, Ohno R: **Effect of recombinant human tumor necrosis factor on the colony growth of human leukemia progenitor cells and normal hematopoietic progenitor cells.** *Blood* 1987, **69(2)**:467.
51. Mitsuhashi M, Peel D, Ziogas A, Anton-Culver H: **Enhanced expression of Radiation-induced Leukocyte CDKN1A mRNA in Multiple primary Breast cancer patients: potential new Marker of cancer susceptibility.** *Biomark Insights* 2009, **4**:201-209.
52. Ching YP, Wong CM, Chan SF, Leung THY, Ng DCH, Jin DY, Ng IO: **Deleted in liver cancer (DLC) 2 encodes a RhoGAP protein with growth suppressor function and is underexpressed in hepatocellular carcinoma.** *Journal of Biological Chemistry* 2003, **278(12)**:10824.
53. Leung THY, Ching YP, Yam JWP, Wong CM, Yau TO, Jin DY, Ng IOL: **Deleted in liver cancer 2 (DLC2) suppresses cell transformation by means of inhibition of RhoA activity.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(42)**:15207.
54. Ullmannova V, Popescu NC: **Expression profile of the tumor suppressor genes DLC-1 and DLC-2 in solid tumors.** *International Journal of Oncology* 2006, **29(5)**:1127.
55. de Tayrac M, Etcheverry A, Aubry M, Saokali S, Hamlat A, Quillien V, Le Treut A, Galibert MD, Mosser J: **Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression.** *Genes, Chromosomes and Cancer* 2008, **48(1)**:55-68.
56. Comabella M, Craig DW, Morcillo-Suarez C, Rio J, Navarro A, Fernandez M, Martin R, Montalban X: **Genome-wide Scan of 500 000 Single-Nucleotide Polymorphisms Among Responders and Nonresponders to Interferon Beta Therapy in Multiple Sclerosis.** *Archives of Neurology* 2009, **66(8)**:972.
57. Fedetz M, Matesanz F, Caro-Maldonado A, Fernandez O, Tamayo JA, Guerrero M: **OAS1 gene haplotype confers susceptibility to multiple sclerosis.** *Tissue Antigens* 2006, **68(5)**:446-449.
58. Karlsson R, Pedersen ED, Wang Z, Brakebusch C: **Rho GTPase function in tumorigenesis.** *Biochimica et Biophysica Acta Reviews on Cancer* 2009, **1796(2)**:91-98.
59. Matsushita H, Lelianova VG, Ushkaryov YA: **The latrophilin family: multiply spliced G protein-coupled receptors with differential tissue distribution.** *FEBS letters* 1999, **443(3)**:348-352.
60. Tomsig JL, Creutz CE: **Copines: a ubiquitous family of Ca 2+-dependent phospholipid-binding proteins.** *Cellular and Molecular Life Sciences* 2002, **59(9)**:1467-1477.
61. Heinrich C, Keller C, Boulay A, Vecchi M, Bianchi M, Sack R, Lienhard S, Duss S, Hofsteenge J, Hynes NE: **Copine-III interacts with ErbB2 and promotes tumor cell migration.** *Oncogene* 2009, **29(11)**:1598-1610.
62. Ramsey CS, Yeung F, Stoddard PB, Li D, Creutz CE, Mayo MW: **Copine-I represses NF- B transcription by endoproteolysis of p65.** *Oncogene* 2008, **27(25)**:3516-3526.
63. Dictor M, Ek S, Sundberg M, Warenholt J, Gyorgy C, Sernbo S, Gustavsson E, Abu-Alsoud W, Wadstrom T, Borrebaeck C: **Strong lymphoid nuclear expression of SOX11 transcription factor defines lymphoblastic neoplasms, mantle cell lymphoma and Burkitt's lymphoma.** *Haematologica* 2009, **94(11)**:1563.
64. Hide T, Takezaki T, Nakatani Y, Nakamura H, Kuratsu J, Kondo T: **Sox11 Prevents Tumorigenesis of Glioma-Initiating Cells by Inducing Neuronal Differentiation.** *Cancer Research* 2009, **69(20)**:7953.
65. Weigle B, Ebner R, Temme A, Schwind S, Schmitz M, Kiessling A, Rieger MA, Schackert G, Schackert HK, Rieber EP: **Highly specific overexpression of the transcription factor SOX11 in human malignant gliomas.** *Oncology Reports* 2005, **13(1)**:139.

66. Brennan DJ, Ek S, Doyle E, Drew T, Foley M, Flannelly G, O'Connor DP, Gallagher WM, Kilpinen S, Kallioniemi OP: **The transcription factor Sox11 is a prognostic factor for improved recurrence-free survival in epithelial ovarian cancer.** *European Journal of Cancer* 2009, **45(8)**:1510-1517.
67. Gonzalez-Arriaga P, Lopez-Cima MF, Fernandez-Somoano A, Pascual T, Marron MG, Puente XS, Tardon A: **Polymorphism+ 17 C/G in matrix metalloprotease MMP 8 decreases lung cancer risk.** *BMC Cancer* 2008, **8(1)**:378.
68. Decock J, Hendrickx W, Vanleeuw U, Van Belle V, Van Huffel S, Christiaens MR, Ye S, Paridaens R: **Plasma MMP 1 and MMP 8 expression in breast cancer: Protective role of MMP 8 against lymph node metastasis.** *BMC Cancer* 2008, **8(1)**:77.
69. Palavalli LH, Prickett TD, Wunderlich JR, Wei X, Burrell AS, Porter-Gill P, Davis S, Wang C, Cronin JC, Agrawal NS: **Analysis of the matrix metalloproteinase family reveals that MMP8 is often mutated in melanoma.** *Nature Genetics* 2009, **41(5)**:518-520.
70. Spielman R, Bastone L, Burdick J, Morley M, Ewens W, Cheung V: **Common genetic variants account for differences in gene expression among ethnic groups.** *Nature Genetics* 2007, **39(2)**:226-231.
71. Akey J, Biswas S, Leek J, Storey J: **On the design and analysis of gene expression studies in human populations.** *Nature Genetics* 2007, **39(7)**:807-808.
72. Imoto I, Yuki Y, Sonoda I, Ito T, Shimada Y, Imamura M, Inazawa J: **Identification of ZASC1 encoding a Kruppel-like zinc finger protein as a novel target for 3q26 amplification in esophageal squamous cell carcinomas.** *Cancer Research* 2003, **63(18)**:5691.
73. Lin SC, Liu CJ, Ko SY, Chang HC, Liu TY, Chang KW: **Copy number amplification of 3q26-27 oncogenes in microdissected oral squamous cell carcinoma and oral brushed samples from areca chewers.** *The Journal of Pathology* 2005, **206(4)**:417-422.
74. Uittenbogaard M, Chiaramello A: **Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis.** *Gene Expression Patterns* 2002, **1(2)**:115-121.
75. Fader AN, Arriba LN, Frasure HE, von Gruenigen VE: **Endometrial cancer and obesity: epidemiology, biomarkers, prevention and survivorship.** *Gynecologic Oncology* 2009, **114(1)**:121-127.
76. Bogaerts S, Vanlandschoot A, van Hengel J, van Roy F: **Nuclear translocation of [alpha] N-catenin by the novel zinc finger transcriptional repressor ZASC1.** *Experimental Cell Research* 2005, **311(1)**:1-13.
77. Makrilia N, Kollias A, Manolopoulos L, Syrigos K: **Cell adhesion molecules: role and clinical significance in cancer.** *Cancer Investigation* 2009, **27(10)**:1023.

Pre-publication history

The pre-publication history for this paper can be accessed here: http://www.biomedcentral.com/1755-8794/4/6/prepub

doi:10.1186/1755-8794-4-6
Cite this article as: Kompass and Witte: **Co-regulatory expression quantitative trait loci mapping: method and application to endometrial cancer.** *BMC Medical Genomics* 2011 **4**:6.