

Genetics of gene expression in
humans:
Data structures, algorithms, and
inference

ISMB 2011 Tutorial

VJ Carey <stvjc@channing.harvard.edu>

Mapping determinants of human gene expression by regional and genome-wide association

Vivian G. Cheung^{1,2,3}, Richard S. Spielman², Kathryn G. Ewens², Teresa M. Weber^{2,3}, Michael Morley³ & Joshua T. Burdick³

To study the genetic basis of natural variation in gene expression, we previously carried out genome-wide linkage analysis and mapped the determinants of ~1,000 expression phenotypes¹. In the present study, we carried out association analysis with dense sets of single-nucleotide polymorphism (SNP) markers from the International HapMap Project². For 374 phenotypes, the association study was performed with markers only from regions with strong linkage evidence; these regions all mapped close to the expressed gene. For a subset of 27 phenotypes, analysis of genome-wide association was performed with >770,000 markers. The

present at the marker. In contrast, allelic association with a linked marker requires correlation with a particular SNP allele; that is, linkage disequilibrium. Even if there are several different alleles at the determinant ('allelic heterogeneity'), linkage can be detected. But if there is allelic heterogeneity, it is less likely that there will be detectable association. Therefore, it was not obvious that evidence for linkage would predict evidence for association. So, for a set of phenotypes with *cis* linkage, we performed association analysis with SNPs within the target genes and within 50 kilobases (kb) of the 5' and 3' ends, and compared results with those from the previous

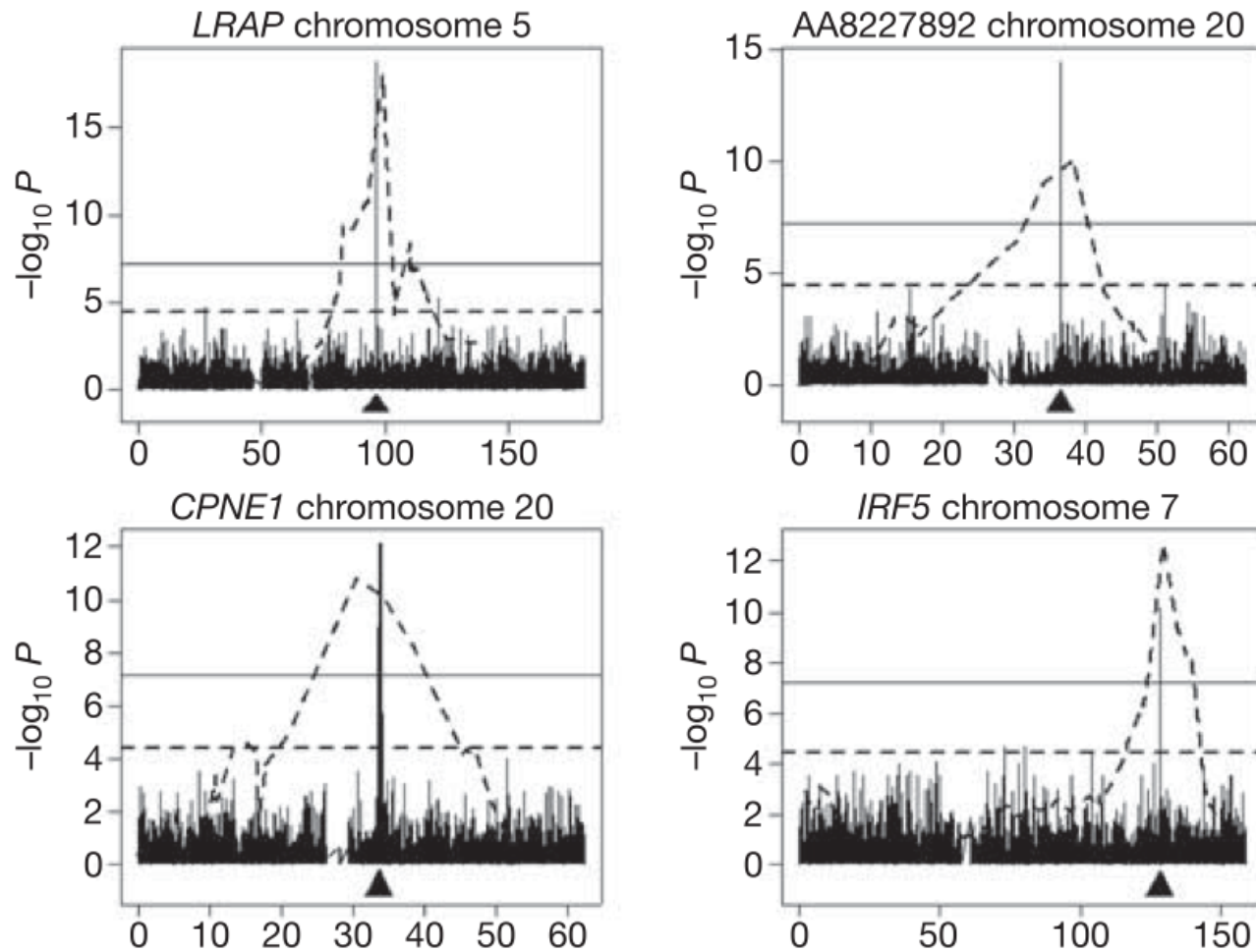


Figure 2 | Results of genome-wide linkage analysis (dotted line) superimposed on those from genome-wide association (bars) for the chromosome where the expressed gene is located. The location of the expressed gene is indicated by an arrow. The dotted horizontal line is for data from linkage scans and corresponds to $t = 4, P = 3.7 \times 10^{-5}$. The solid horizontal line is for data from GWA and corresponds to $P = 0.05$ after Šidák correction. The x axis indicates chromosome location in megabases.

The influence of genetic variation on gene expression

Rohan B.H. Williams,^{1,2,3,4} Eva K.F. Chan,^{1,3,5} Mark J. Cowley,^{1,4} and Peter F.R. Little^{1,6}

¹*School of Biotechnology and Biomolecular Sciences, University of New South Wales, Randwick, NSW 2052, Australia;*

²*Ramaciotti Centre for Gene Function Analysis, University of New South Wales, Randwick, NSW 2052, Australia*

The view that changes to the control of gene expression rather than alterations to protein sequence are central to the evolution of organisms has become something of a truism in molecular biology. In reality, the direct evidence for this is limited, and only recently have we had the ability to look more globally at how genetic variation influences gene expression, focusing upon inter-individual variation in gene expression and using microarrays to test for differences in mRNA levels. Here, we review the scope of these experimental analyses, what they are designed to tell us about genetic variation, and what are their limitations from both a technical and a conceptual viewpoint. We conclude that while we are starting to understand the impact of this class of genetic variation upon steady-state mRNA levels, we are still far from identifying the potential phenotypic and evolutionary outcomes.

The conceptual framework

specific DNA probes, and we have not attempted to extend our review into this area.

How DNA variants might affect variation in mRNA abundance/function (RBH Williams+ 2007)

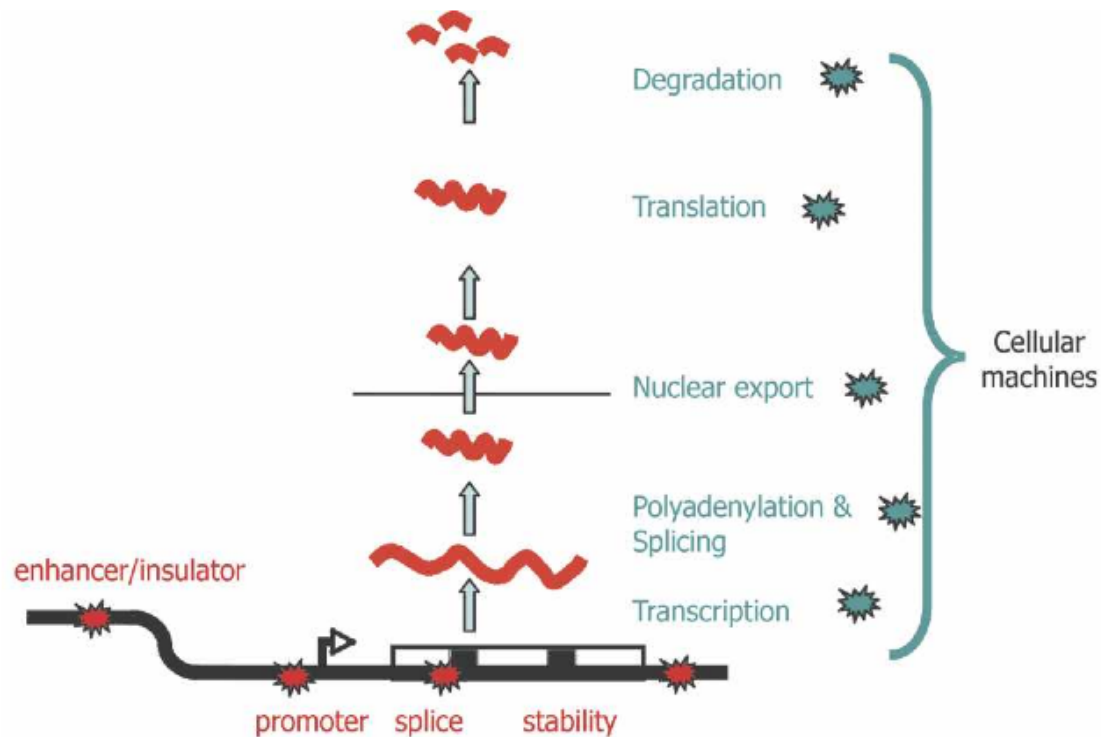


Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in *cis* (red stars), or there may be variations in the molecular machinery that interact with *cis*-regulatory sequences and so act genetically in *trans* (blue stars).

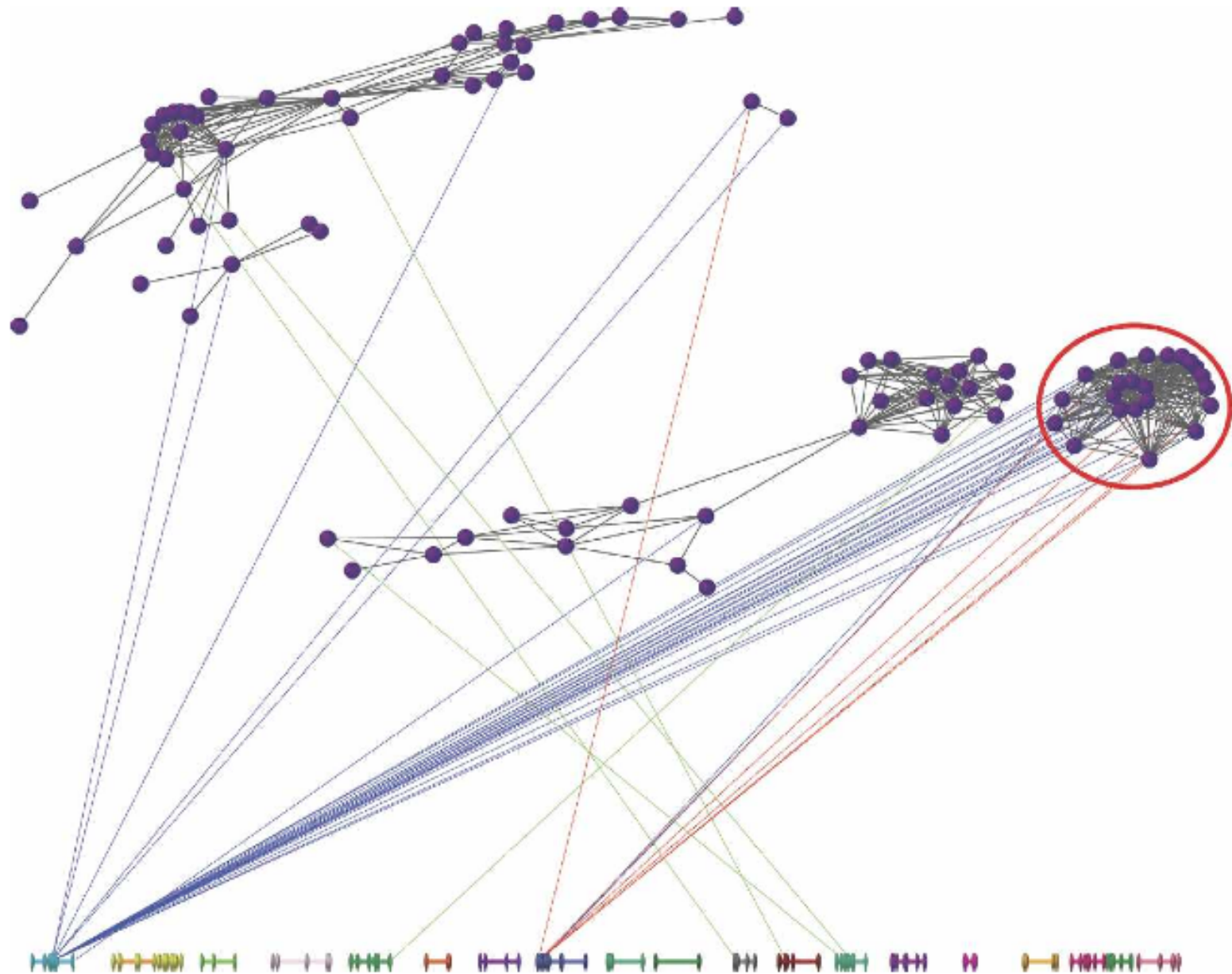


Figure 2. Regulon analysis of genes. Following sparse latent factor analysis, each gene is represented as a purple vertex connected to other genes by a gray line if the posterior probability of being correlated across the three tissues is ≥ 0.90 . Linkage is drawn to the chromosomes below (1–19, X, left to right) if $P < 10^{-4}$: line color indicates the relevant tissue (blue, brain; green, kidney; red, liver). Note

Guest Editorial

Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes

'In Nature's infinite book of secrecy, a little I can read.'

Antony and Cleopatra [Act I, Scene 2], William Shakespeare

Pathological mutations occurring within the extended consensus sequences of exon–intron splice junctions account for ~10 per cent of all inherited lesions logged in The *Human Gene Mutation Database* (HGMD[®]; <http://www.hgmd.org>)¹ and are frequently encountered in mutation screening studies.² Mutations residing in other intronic locations (including the canonical branch-point sequence,³ 5'-YURAY-3'), however, may

variants will have been seriously under-ascertained to date. Although most of these variants are single nucleotide polymorphisms (SNPs), others may be of the insertion/deletion type.⁸ With the advent of genome-wide association studies (GWAS), an increasing number of potentially functional intronic variants are being identified.⁹ In the majority of cases, however, it is unclear whether such variants are of direct functional significance, as opposed to simply being in linkage disequilibrium with another (as yet unidentified) functional SNP in the vicinity.¹⁰ Even when GWAS studies deem a newly

Table 1. Selected examples of *in vitro* characterised human functional intronic polymorphisms located more than ~30 bp from the nearest splice site

Gene	Disease/phenotype	Chromosomal location	Polymorphism, intronic location and dbSNP number	Consequences for gene expression or mRNA splicing	Reference
<i>AGTR2</i>	Predisposition to congenital anomalies of the kidney and urinary tract	Xq22-q23	IVS1, AS, A > G, -29 (rs1403543)	SNP occurs within branchpoint motif and alters splicing efficiency	Nishimura <i>et al.</i> (1999) ^a
<i>BANK1</i>	Susceptibility to systemic lupus erythematosus	4q23	IVS1, AS, T > C, -43 (rs17266594)	SNP occurs within branchpoint motif and risk allele alters expression of alternative transcripts	Kozyrev <i>et al.</i> (2008) ^b
<i>CD244</i>	Susceptibility to rheumatoid arthritis	1q23.1	IVS3, AS, T > C, -164 (rs6682654)	Risk allele associated with increased transcriptional activity	Suzuki <i>et al.</i> (2008) ^c
<i>CD244</i>	Susceptibility to rheumatoid arthritis	1q23.1	IVS5, DS, G > A, +526 (rs3766379)	Risk allele associated with increased transcriptional activity	Suzuki <i>et al.</i> (2008) ^c
<i>COL1A1</i>	Reduced bone density/osteoporosis	17q21.33	IVS1, AS, G > T, -440 (rs1800012)	SNP occurs within Sp1-binding site; risk allele alters Sp1 binding and transcriptional activity	Mann <i>et al.</i> (2001) ^d

Summary

- DNA variants affecting mRNA abundance observable using standard microarray platforms
- “Regulon” models for coexpression networks have been proposed
- Numerous ‘functional’ polymorphisms connected with disease, some mechanistic explanations
- How can mechanics be further elaborated?
 - Structural localization
 - Details of mRNA processing; alternate splicing
 - Identifying and understanding allelic imbalance

High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation

Jean-Baptiste Veyrieras^{1*}, Sridhar Kudaravalli¹, Su Yeon Kim², Emmanouil T. Dermitzakis³, Yoav Gilad^{1*}, Matthew Stephens^{1,2*}, Jonathan K. Pritchard^{1,4*}

1 Department of Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, **2** Department of Statistics, The University of Chicago, Chicago, Illinois, United States of America, **3** Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, **4** Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America

Abstract

Recent studies of the HapMap lymphoblastoid cell lines have identified large numbers of quantitative trait loci for gene expression (eQTLs). Reanalyzing these data using a novel Bayesian hierarchical model, we were able to create a surprisingly high-resolution map of the typical locations of sites that affect mRNA levels in *cis*. Strikingly, we found a strong enrichment of eQTLs in the 250 bp just upstream of the transcription end site (TES), in addition to an enrichment around the transcription start site (TSS). Most eQTLs lie either within genes or close to genes; for example, we estimate that only 5% of eQTLs lie more than 20 kb upstream of the TSS. After controlling for position effects, SNPs in exons are ~2-fold more likely than SNPs in introns to be eQTLs. Our results suggest an important role for mRNA stability in determining steady-state mRNA levels, and highlight the potential of eQTL mapping as a high-resolution tool for studying the determinants of gene regulation.

Citation: Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet* 4(10): e1000214. doi:10.1371/journal.pgen.1000214

Editor: Greg Gibson, The University of Queensland, Australia

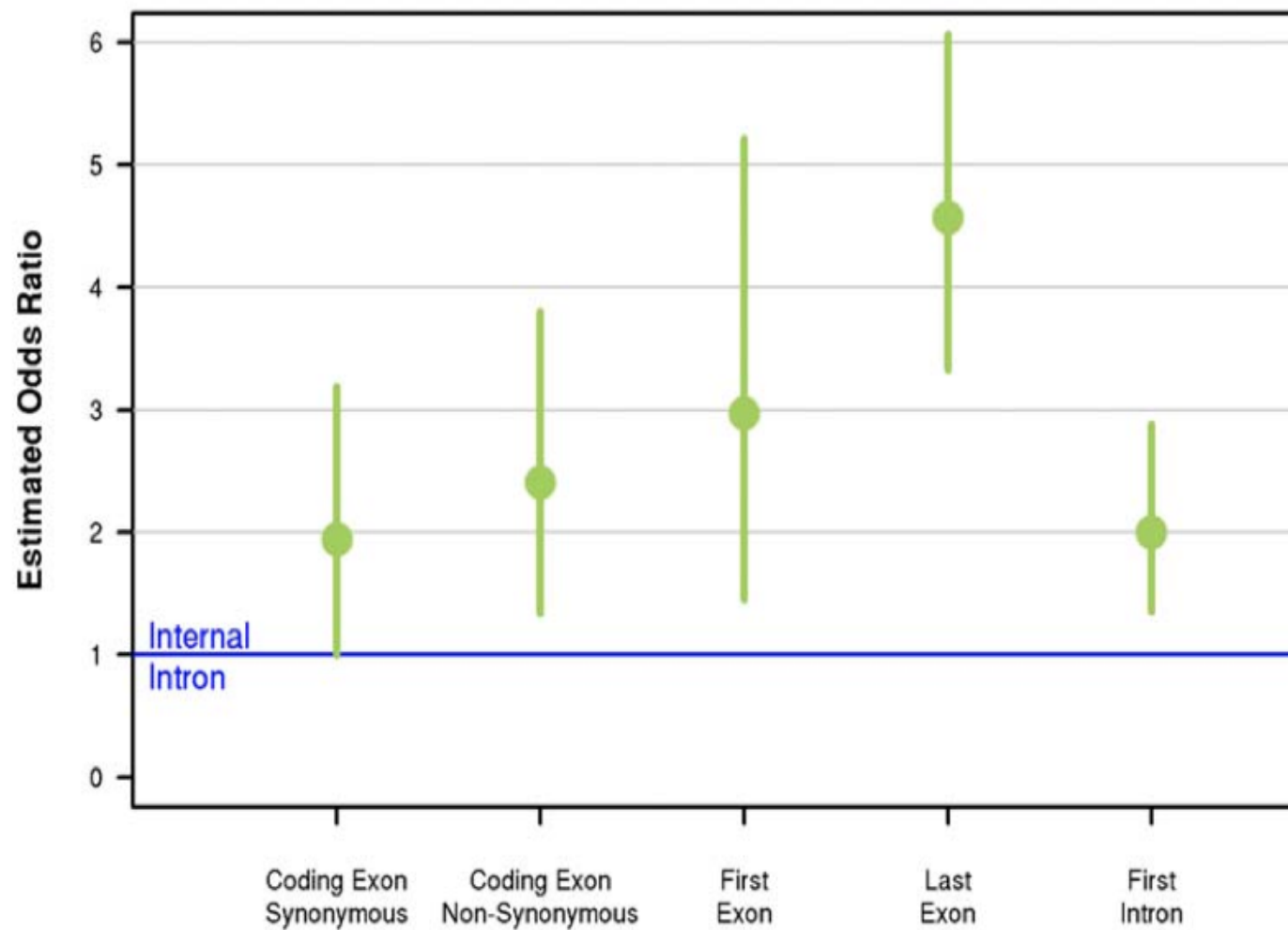


Figure 5. Expression-QTNs are under-represented in coding sequence introns, even after controlling for position effects. The figure shows the odds ratios for the probability that a SNP in a particular part of the gene (e.g., coding exon) is inferred to be an eQTN, relative to the probability for a SNP in an “internal” intron (i.e., an intron within the coding sequence). The odds ratios are estimated using the hierarchical model with internal introns fixed at a value of 1, and control for SNP position using the TSS+TES model. The vertical bars show 95% confidence intervals. doi:10.1371/journal.pgen.1000214.g005

Functional integration of transcriptional and RNA processing machineries

Shatakshi Pandit, Dong Wang and Xiang-Dong Fu

Cotranscriptional RNA processing not only permits temporal RNA processing before the completion of transcription but also allows sequential recognition of RNA processing signals on nascent transcripts threading out from the elongating RNA polymerase II (RNAPII) complex. Rapid progress in recent years has established multiple contacts that physically connect the transcription and RNA processing machineries, which centers on the C-terminal domain (CTD) of the largest subunit of RNAPII. Although cotranscriptional RNA processing has been substantiated, the evidence for 'reciprocal' coupling starts to emerge, which emphasizes functional integration of transcription and RNA processing machineries in a mutually beneficial manner for efficient and regulated gene expression.

Address

Department of Cellular and Molecular Medicine, University of California

fashion along the gene [3]. As illustrated in Figure 1, this dynamic change in CTD phosphorylation suggests that the RNAPII complex is rearranging its content during transcription to allow sequential action of distinct machineries for cotranscriptional RNA processing. However, sequential action does not necessarily mean sequential recruitment of RNA processing factors to the RNAPII complex because many 'downstream' factors seem to be recruited at the very beginning of transcription [8,9].

Although most studies focus on understanding how RNA processing takes advantage of the transcriptional machinery to execute cotranscriptional processing for efficient gene expression, increasing evidence suggests that transcription may also benefit from and/or depend on specific

How DNA variants might affect variation in mRNA abundance/function (RBH Williams+ 2007)

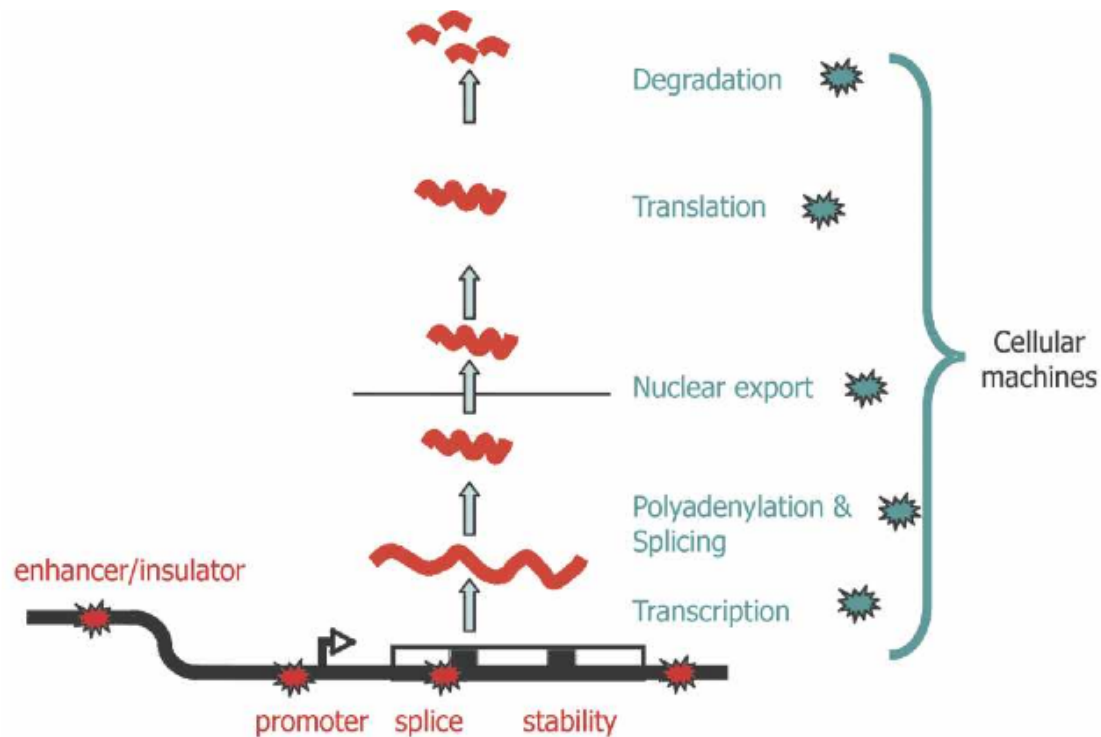
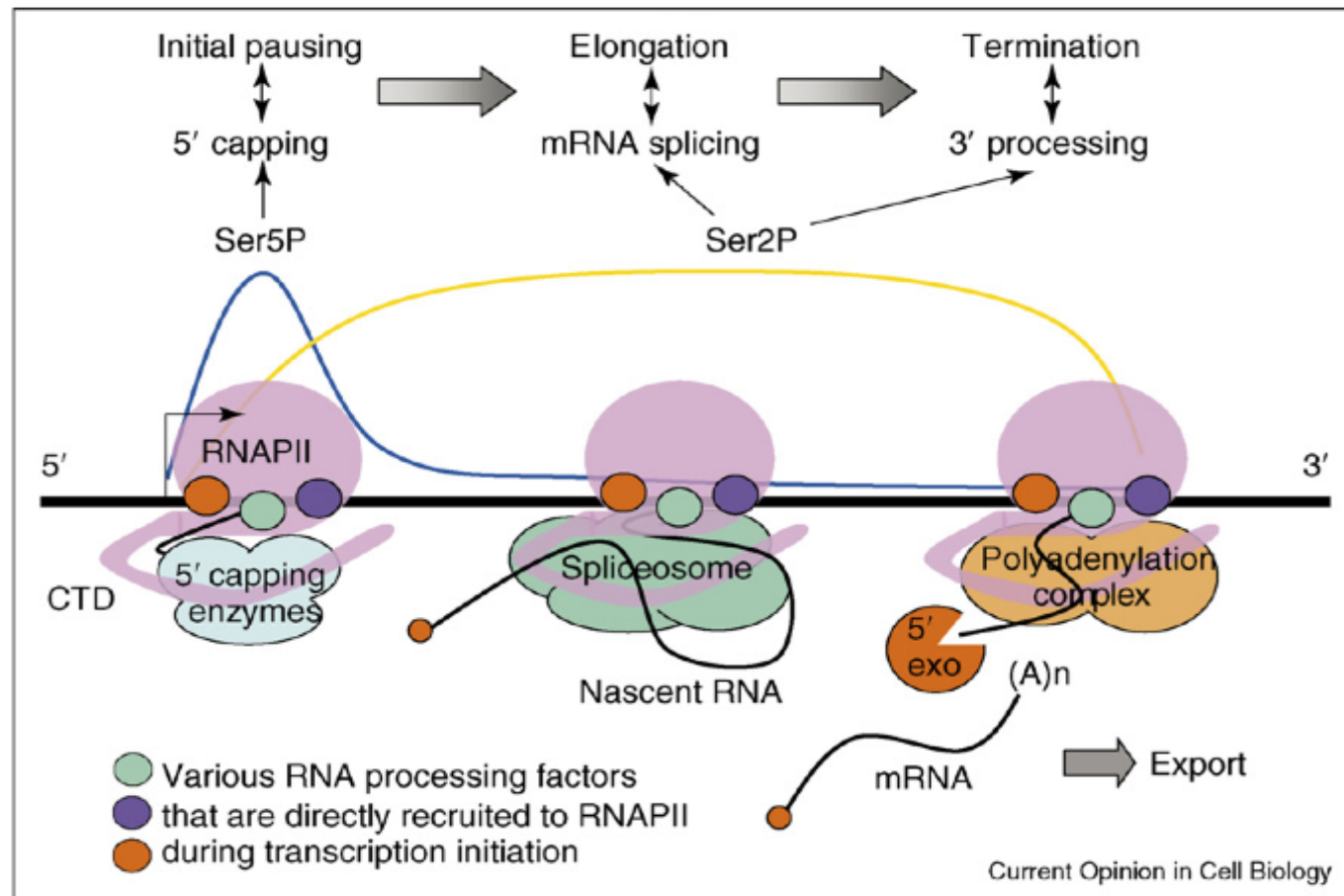


Figure 1. Plausible sites of action for genetic determinants of mRNA levels. Genetic variations influencing gene expression may reside within the regulatory sequences, promoters, enhancers, splice sites, and secondary structure motifs of the target gene and so be genetically in *cis* (red stars), or there may be variations in the molecular machinery that interact with *cis*-regulatory sequences and so act genetically in *trans* (blue stars).

Dissecting transcription and pre-mRNA processing: initiation, cotranscription (splicing, capping,...), post-tx

Figure 1



Coupling between transcription and pre-mRNA processing. The RNA polymerase II (RNAPII) is modified on its CTD with Ser5 phosphorylation predominately at the beginning of the gene (blue line) and Ser2 phosphorylation in the middle and end of the gene (yellow line). 5'-Capping enzymes are recruited through direct interactions with Ser5 phosphorylated CTD to catalyze the cotranscriptional capping reaction. Various RNA processing factors are recruited during the elongation phase of transcription, most of which in a CTD Ser2 phosphorylation-dependent manner, to facilitate cotranscriptional splicing. The 3'-end formation is functionally tied to transcription termination. Importantly, increasing evidence now suggests that the transcription and RNA processing machineries are functionally integrated in a reciprocal fashion such that individual cotranscriptional

The study of eQTL variations by RNA-seq: from SNPs to phenotypes

Jacek Majewski and Tomi Pastinen

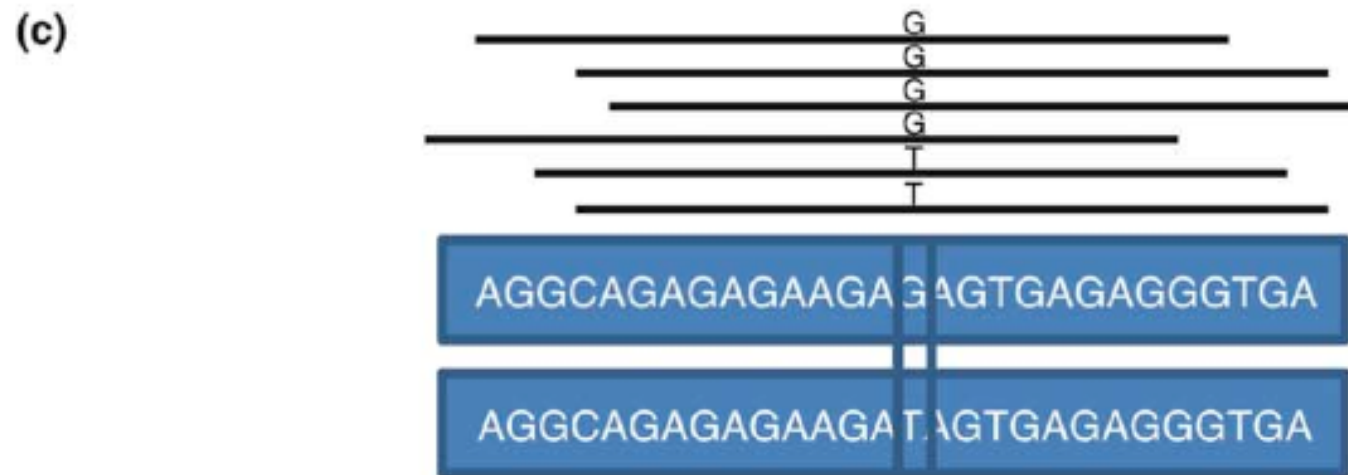
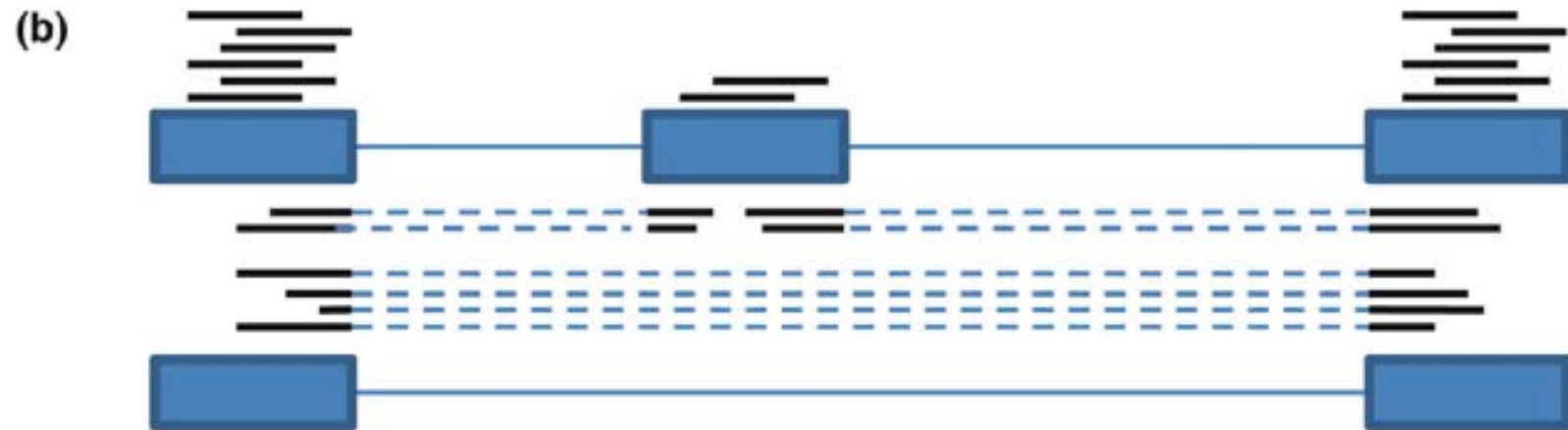
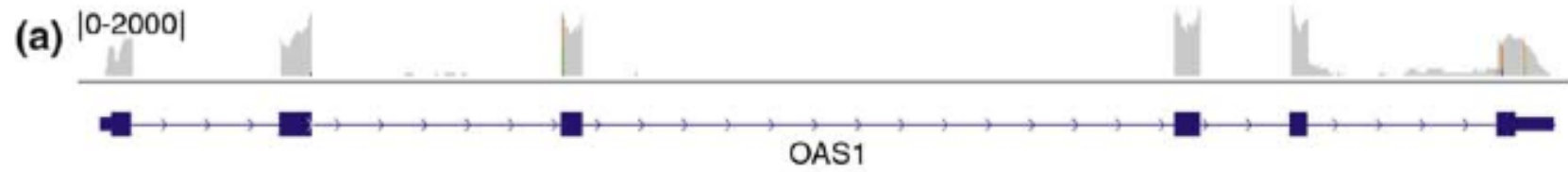
Department of Human Genetics, McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield Avenue, Rm 7210, Montreal, Quebec, H3A 1A4, Canada

Common DNA variants alter the expression levels and patterns of many human genes. Loci responsible for this genetic control are known as expression quantitative trait loci (eQTLs). The resulting variation of gene expression across individuals has been postulated to be a determinant of phenotypic variation and susceptibility to complex disease. In the past, the application of expression microarray and genetic variation data to study populations enabled the rapid identification of eQTLs in model organisms and humans. Now, a new technology promises to revolutionize the field. Massively parallel RNA sequencing (RNA-seq) provides unprecedented resolution, allowing us to accurately monitor not only the expression output of each genomic locus but also reconstruct and quantify alternatively spliced transcripts. RNA-seq also provides new insights into the regulatory mechanisms underlying eQTLs. Here, we discuss the

more widespread importance of noncoding or regulatory DNA alterations in disease as implied by GWAS now calls for approaches to characterize such a variation and its links to disease phenotypes.

Genome-wide identification of loci controlling gene expression

The parallel assessment of thousands of transcripts using DNA microarrays is clearly one of the revolutionary technologies that launched the 'genomic' era. The genome-wide association of genetic and transcriptome variations was first achieved in yeast [6], where expression traits of the progeny were shown to be largely correlated with the genetic contribution of parental genotypes. The excitement of observing thousands of quantitative traits, or eQTLs, in a technically straightforward experiment quickly spread to studies in more complex genomes [7] including



Fine-Scale Variation and Genetic Determinants of Alternative Splicing across Individuals

Jasmin Coulombe-Huntington^{1,2}, Kevin C. L. Lam², Christel Dias², Jacek Majewski^{1,2*}

1 Department of Human Genetics, McGill University, Montreal, Québec, Canada, **2** McGill University and Génome Québec Innovation Centre, Montréal, Québec, Canada

Abstract

Recently, thanks to the increasing throughput of new technologies, we have begun to explore the full extent of alternative pre-mRNA splicing (AS) in the human transcriptome. This is unveiling a vast layer of complexity in isoform-level expression differences between individuals. We used previously published splicing sensitive microarray data from lymphoblastoid cell lines to conduct an in-depth analysis on splicing efficiency of known and predicted exons. By combining publicly available AS annotation with a novel algorithm designed to search for AS, we show that many real AS events can be detected within the usually unexploited, speculative majority of the array and at significance levels much below standard multiple-testing thresholds, demonstrating that the extent of cis-regulated differential splicing between individuals is potentially far greater than previously reported. Specifically, many genes show subtle but significant genetically controlled differences in splice-site usage. PCR validation shows that 42 out of 58 (72%) candidate gene regions undergo detectable AS, amounting to the largest scale validation of isoform eQTLs to date. Targeted sequencing revealed a likely causative SNP in most validated cases. In all 17 incidences where a SNP affected a splice-site region, *in silico* splice-site strength modeling correctly predicted the direction of the micro-array and PCR results. In 13 other cases, we identified likely causative SNPs disrupting predicted splicing enhancers. Using *Fst* and REHH analysis, we uncovered significant evidence that 2 putative causative SNPs have undergone recent positive selection. We verified the effect of five SNPs using *in vivo* minigene assays. This study shows that

Schemata for SNP-associated splicing events (Coulombe-Huntington 2009)

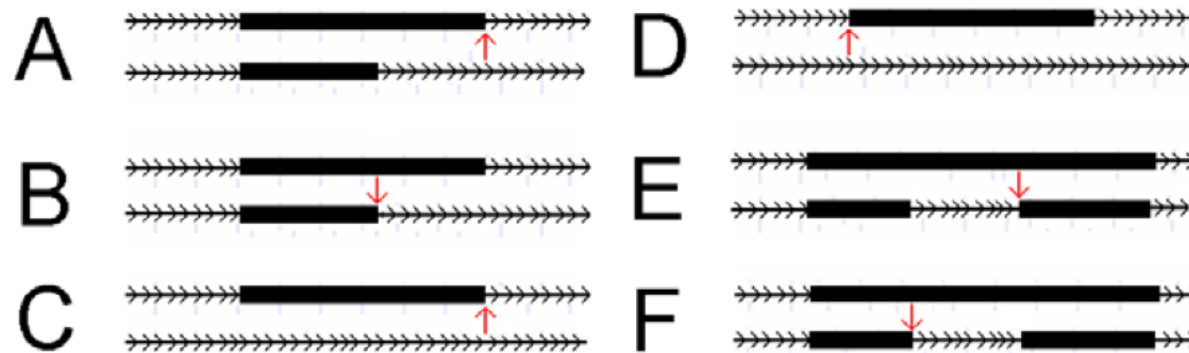


Figure 2. AS type and affected splice-site for SNPs identified in Table 2 and Table 3. The arrow indicates the splice-site affected by the polymorphism. The genes are read from left to right, as indicated by the intersecting arrow heads. The type of AS event and which splice-site is affected is essential to understanding the relation between the probeset expression change and the theoretical efficiency of splicing. In (A,C,D), the correlation should be positive since the use of the splice-site produces a longer transcript, while in (B,E,F), an inverse relation is expected since the use of the splice-site produces a shorter transcript. doi:10.1371/journal.pgen.1000766.g002

Table 2. SNPs affecting splice-sites.

Gene	SNP ID/new SNP	AS Type ¹	Splice-site sequence ²	Maximum Entropy Score ³	Probeset Expression ⁴
C8orf59	new SNP	A	aagGTaaaa	8.38	138
			aagGAaaaa	0.19	12
DMKN	rs4254439	C	cggGTgagc ⁵	8.18	117
			aggGTgagc	7.75	11
ERAP2	rs2248374	B ⁶	atgGTaagg ⁵	9.33	69
			atgGTgagg	7.61	297
MGC16169	rs12639869	C	aagGTatgt ⁵	9.79	225
			aatGTatgt	5.87	26
PLD2	rs3764897	A	cagGTagag ⁵	7.10	140
			cggGTagag	2.04	43
SH3YL1	rs62114506	C	atgGTaagt ⁵	11.01	118
			atgGTaact	6.06	22
TMEM77	rs3762374	C	gttGTgagt ⁵	6.59	2552
			gttGTgaat	−4.72	394
ZNF419	rs11672136	D	ccatAGgtt ⁵	8.87	56
			ccaaAGgtt	6.65	13

Summary

- Structural localization of eQTL depends on finding them and connecting them with relevant annotation – we will do that
- Alternative splicing analysis will not be covered in this tutorial
- Data from Stranger et al (2007) Cheung et al (2010) will be primary resources for assessing cis- and trans-associated eQTL

Polymorphic *Cis*- and *Trans*-Regulation of Human Gene Expression

Vivian G. Cheung^{1,2,3,4*}, Renuka R. Nayak⁵, Isabel Xiaorong Wang¹, Susannah Elwyn⁴, Sarah M. Cousins⁴, Michael Morley⁴, Richard S. Spielman^{3†}

1 Howard Hughes Medical Institute, Philadelphia, Pennsylvania, United States of America, **2** Department of Pediatrics, University of Pennsylvania, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **3** Department of Genetics, University of Pennsylvania, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **4** University of Pennsylvania, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Medical Scientist Training Program, University of Pennsylvania, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America

Abstract

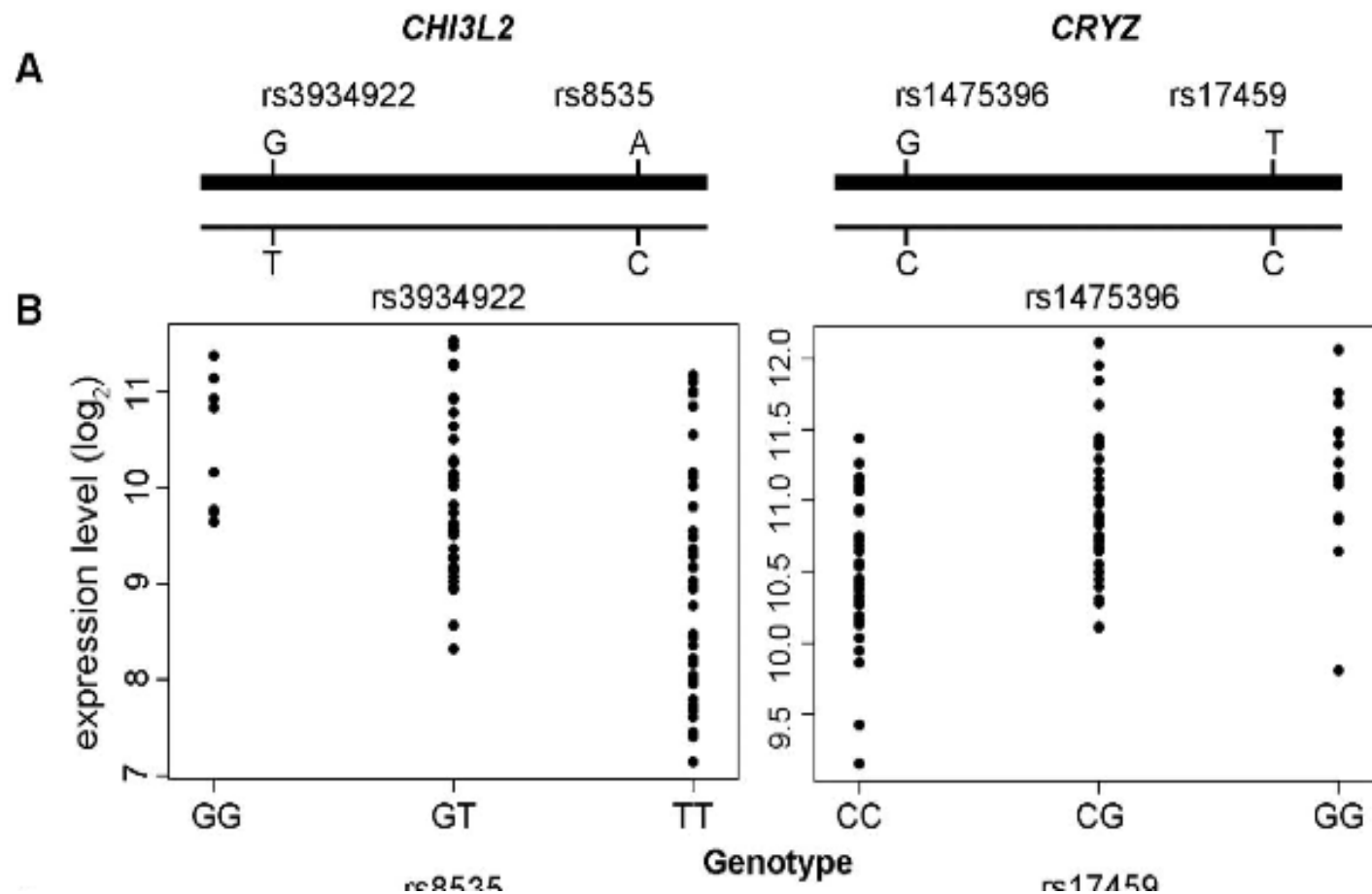
Expression levels of human genes vary extensively among individuals. This variation facilitates analyses of expression levels as quantitative phenotypes in genetic studies where the entire genome can be scanned for regulators without prior knowledge of the regulatory mechanisms, thus enabling the identification of unknown regulatory relationships. Here, we carried out such genetic analyses with a large sample size and identified *cis*- and *trans*-acting polymorphic regulators for about 1,000 human genes. We validated the *cis*-acting regulators by demonstrating differential allelic expression with sequencing of transcriptomes (RNA-Seq) and the *trans*-regulators by gene knockdown, metabolic assays, and chromosome conformation capture analysis. The majority of the regulators act in *trans* to the target (regulated) genes. Most of these *trans*-regulators were not known to play a role in gene expression regulation. The identification of these regulators enabled the characterization of polymorphic regulation of human gene expression at a resolution that was unattainable in the past.

Citation: Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, et al. (2010) Polymorphic *Cis*- and *Trans*-Regulation of Human Gene Expression. PLoS Biol 8(9): e1000480. doi:10.1371/journal.pbio.1000480

Academic Editor: Jonathan Flint, The Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom

The digital nature of the sequence data allows us to use the heterozygous genotypes in each transcript to determine whether two allelic forms of a transcript are expressed in equal abundance [31–33]. Among the 107 expression phenotypes with proximal linkage peaks, 67 have at least one SNP where there are 2 individuals who are heterozygous at that SNP (see Methods). We examined these heterozygous samples for evidence of DAE. For many of these genes, we have data for multiple SNPs from an average of 7.2 individuals (median = 6). Among the 67 genes, 43 genes (64%) showed significant evidence ($p < 0.01$, chi-square test) of departure from equal expression of the two allelic forms of the genes. For the 273 exonic SNPs in these 43 genes, we calculated an “allelic expression ratio” $a/(a+b)$, where a and b are the numbers of sequence reads for the two alleles. Figure S1 shows

Cheung et al 2010: RNA-seq for differential allelic expression



Relative frequency of A- vs C-bearing forms of the CHI3L2 transcript (left panel)

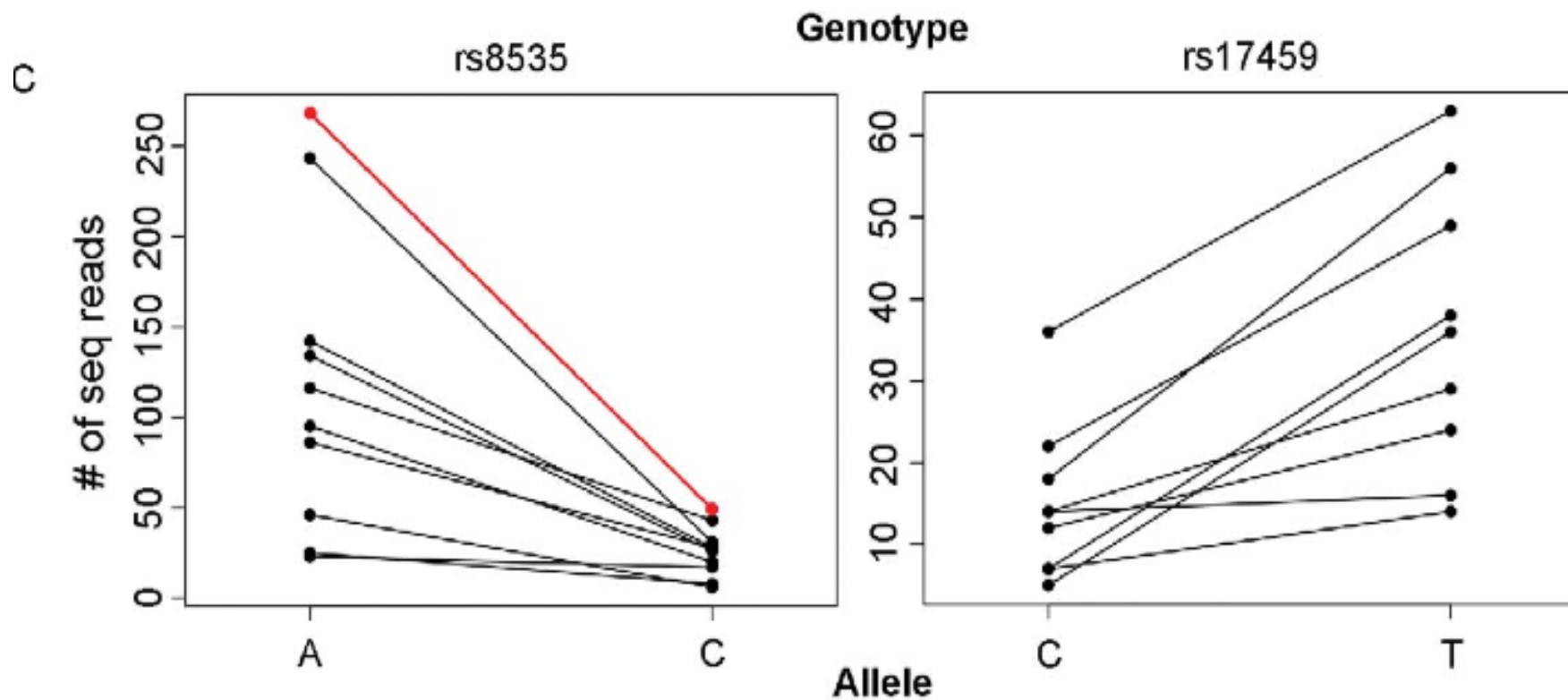
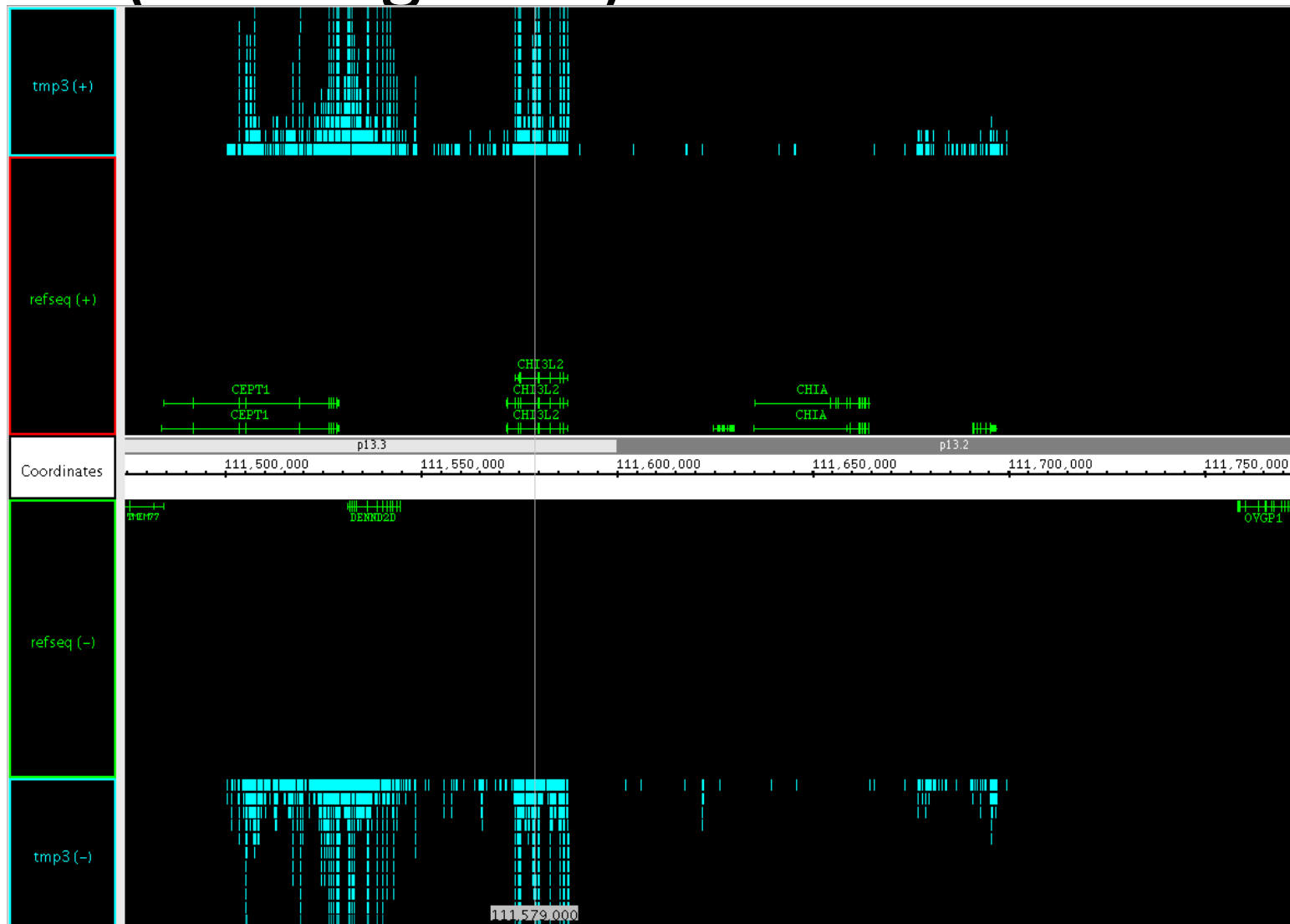


Figure 2. Allelic expression from RNA-Seq confirms prediction by association analysis. Graphical presentations of two genes showing differential allelic expression. The thick lines represent the higher expressing allelic forms of *CHI3L2* and *CRYZ* (A). Regression of ex

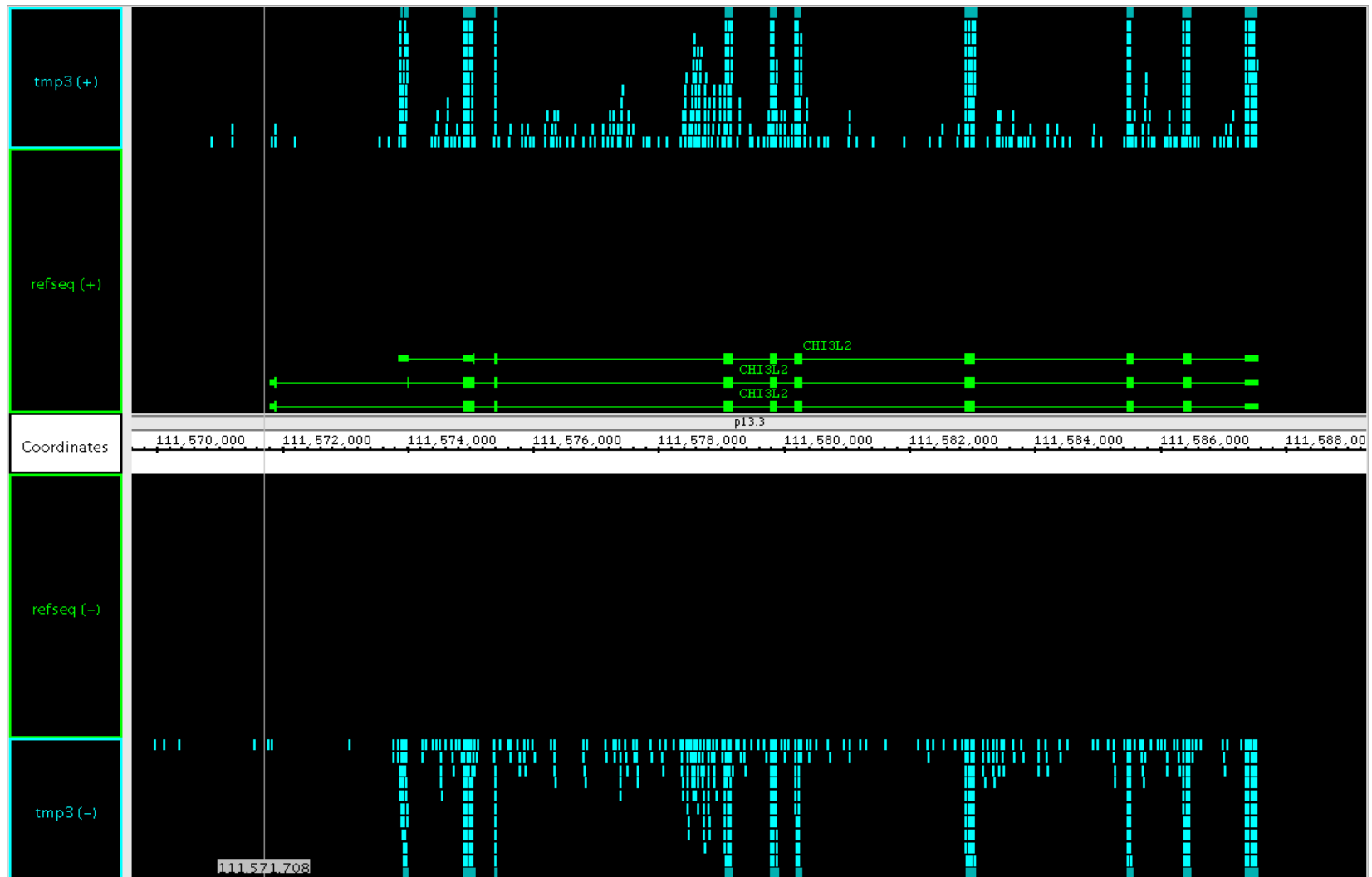
Summary

- SNP associated with
 - overall differential expression
 - alternative splicing
 - allelic imbalance
- Not reviewed: Specific to context: tissue, disease, developmental stage ...
- Tutorial: what is a reasonable computational environment to foster progress with these investigations?

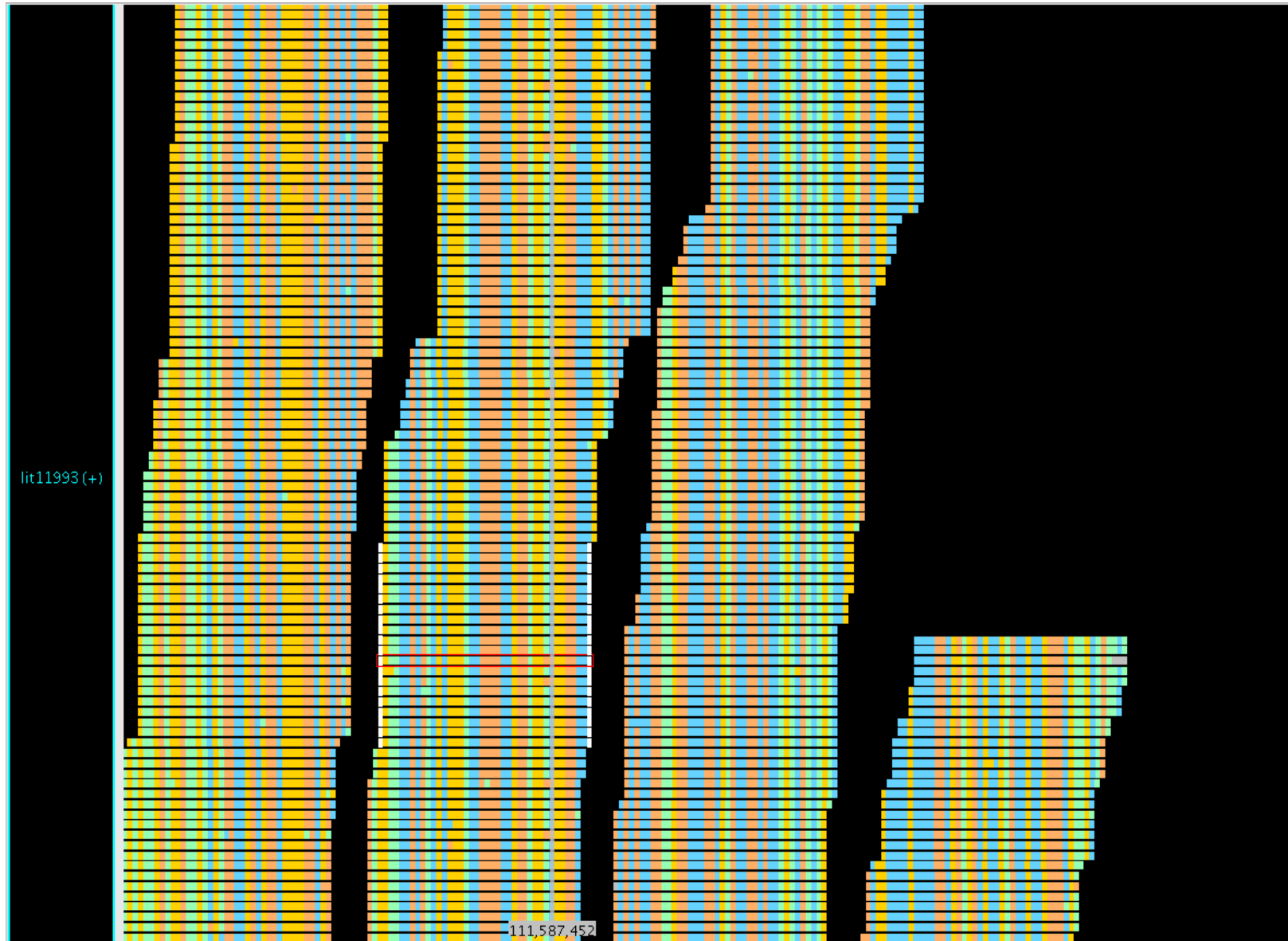
A collection of reads from GSE16921 (Cheung et al) around CHI3L2



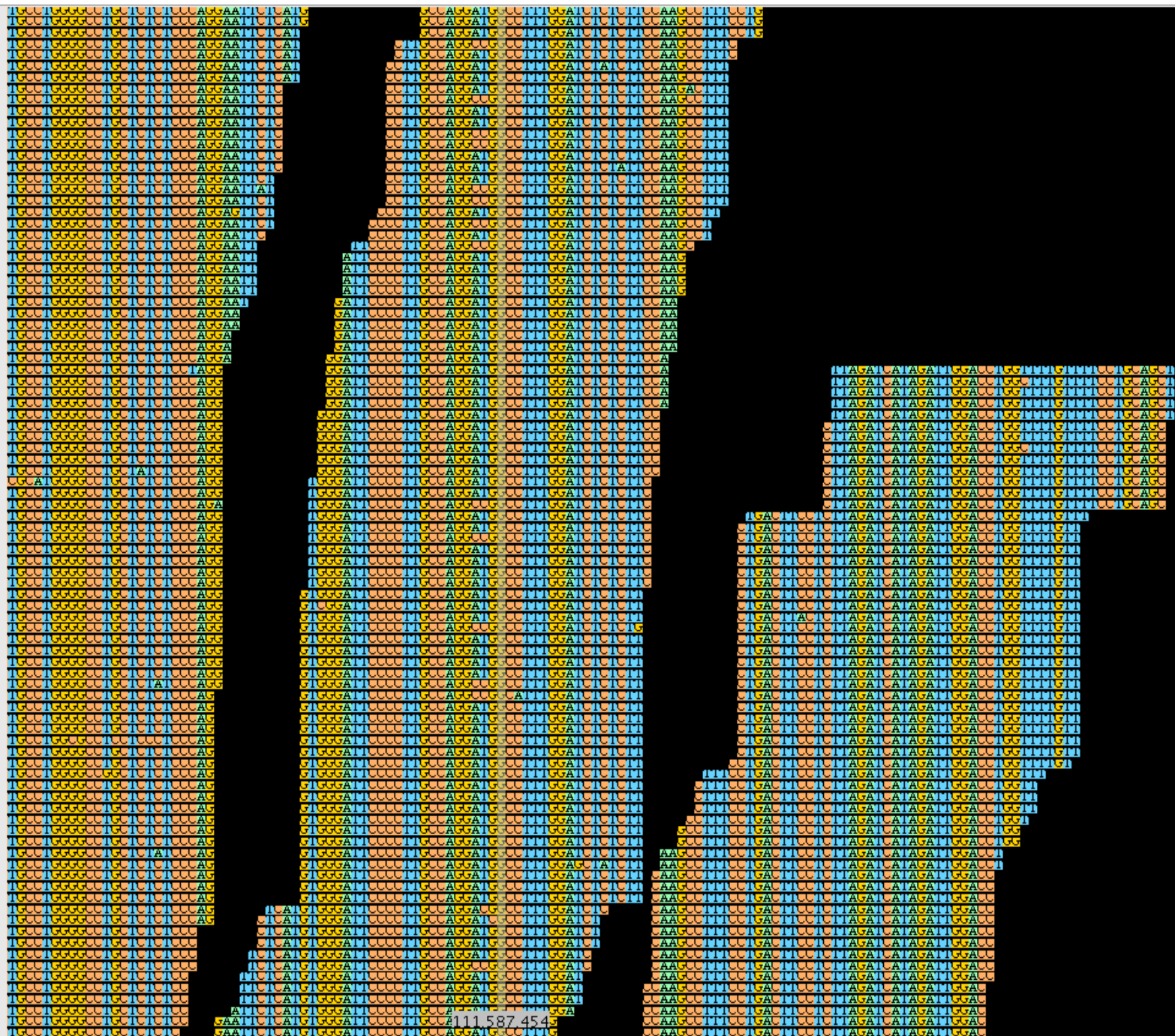
Read densities compatible with gene model (roughly)



Seeking the gold in a band of blue



lit11993(+)



111,587,454

Data structures, algorithms, and inference for genetics of gene expression: moving targets

- ca. 2008:
 - SNP chips with 1 million loci (+ CNV)
 - Expression arrays with 50K (u133+2.0) to many more (exon arrays) expression features
- Current (ambitions)
 - Genome-wide or exome-wide DNA sequencing
 - RNA-seq
 - Platforms, protocols, required depth?

Perspectives of the tutorial

- Platforms for array-based methods for transcript profiling and genotyping are reasonably stable; archives of these should be easily harvested by any interested computational biologist
- Concepts that work in this domain need some extension for sequence-based context, but the basic principles should carry over
- Despite 10+ years of transcript profiling with microarrays, **frameworks for establishing optimal methods**, common in biostatistical applications, have not been established

Basic computational and interpretive methods required

Array context

- Preprocessing
 - Quality assessment, background correction, image registration....
- Normalization
 - ‘Removal’ of non-biologic sources of variation
 - Establishment of between-sample comparability
- Inference

Sequencing context

- Preprocessing
 - Image analysis/base-calling ... basically proprietary
 - Quality assessment, filtering
- Normalization
 - ‘Removal’ of non-biologic sources of variation
 - Establishment of between-sample comparability
- Inference

Details of computations to be reviewed in tutorial

- Container design
 - Link experiment metadata, assay results, sample characteristics for reliable filtering, reshaping, and analysis
 - Allow a unified view even if the resources are decomposed for computational efficiency
- Decomposable workflow
 - Permit concurrent execution of embarrassingly parallelizable tasks
- Avoid large memory footprints whenever possible

Contents

1	Basic concepts with array-based data: cis-associated variants	2
1.1	Functional relations between DNA variants and mRNA abundance	2
1.2	Direct computation to search for eQTL	2
1.2.1	Exercises 1	5
1.3	Transcriptome-wide searches for eQTL	6
1.3.1	Managing millions of test results; resolving focused queries	6
1.3.2	Surveying transcriptome-wide test collections	7
1.3.3	Assessing false discovery rates using statistics computed after per- mutation	8
1.3.4	Locations and contexts: the eQTL landscape of a chromosome . .	11
1.3.5	High-level tools for locus annotation: ChIPpeakAnno	16
2	Imputation using 1000 genomes genotypes	19
3	Identifying and reducing expression heterogeneity for enhanced eQTL discovery	23
3.1	Unsupervised approach: PCA for covariates	23
3.2	Supervised approach: surrogate variable analysis	25
4	Investigating trans associations	27
5	Leveraging RNA-seq: details of transcriptomic diversity	31
5.1	Some key observations and their approximate reproduction	31
5.2	Surveying a read set for transcript variants	34

Some abstractions for array-based resources

- Enumerations
 - Genes (or expression probes) $1, \dots, G$
 - Samples $1, \dots, N$ (may be clustered into families)
 - Sample characteristics $1, \dots, R$
 - SNPs on chromosome c , $1, \dots, S_c$
- ExpressionSet X unites
 - Assay data: $\text{exprs}(X)$ is $G \times N$
 - Sample level data: $\text{pData}(X)$ is $N \times R$
 - MIAME: $\text{experimentData}(X)$

Containers for integrative genomics experiments

- A `SnpMatrix` for a SNP panel from chromosome c is a container for observed or imputed genotypes, using 1 byte per locus, as a matrix of dimensions $N \times S_c$
- An `smList` collects `SnpMatrix` instances for a collection of chromosomes
- An `smlSet` X combines an `ExpressionSet` with an `smList`
 - As before `exprs(X)` is $G \times N$
 - `smList(X)[[c]]` is $N \times S_c$
 - `pData()` and `experimentData()` function as before

CEPH CEU GENEVAR+HapMap ph 2

```
> library(GGtools)
> library(GGdata)
> c17 = getSS("GGdata", "17", renameChrs="chr17")
> class(c17) # smlSet links SnpMatrix instances and expression data
```

```
[1] "smlSet"
attr(,"package")
[1] "GGBase"
```

```
> c17
```

```
SnpMatrix-based genotype set:
number of samples: 90
number of chromosomes present: 1
annotation: illuminaHumanv1.db
Expression data dims: 47293 x 90
Phenodata: An object of class "AnnotatedDataFrame"
  sampleNames: NA06985 NA06991 ... NA12892 (90 total)
  varLabels: famid persid ... male (7 total)
  varMetadata: labelDescription
```

Various views of chr17 SNP

```
> dim(smList(c17)[["chr17"]])
[1] 90 89701
> as(smList(c17)[["chr17"]][1:3,1:5], "matrix")
      rs6565733 rs1106175 rs17054921 rs8064924 rs8070440
NA06985      03      02      03      03      03
NA06991      03      01      03      03      03
NA06993      03      01      03      03      03
> as(smList(c17)[["chr17"]][1:3,1:5], "numeric")
      rs6565733 rs1106175 rs17054921 rs8064924 rs8070440
NA06985        2        1        2        2        2
NA06991        2        0        2        2        2
NA06993        2        0        2        2        2
> as(smList(c17)[["chr17"]][1:3,1:5], "character")
      [,1] [,2] [,3] [,4] [,5]
[1,] "B/B" "A/B" "B/B" "B/B" "B/B"
[2,] "B/B" "A/A" "B/B" "B/B" "B/B"
[3,] "B/B" "A/A" "B/B" "B/B" "B/B"
> □
```

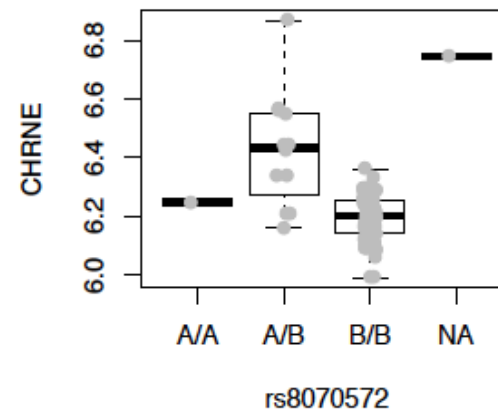
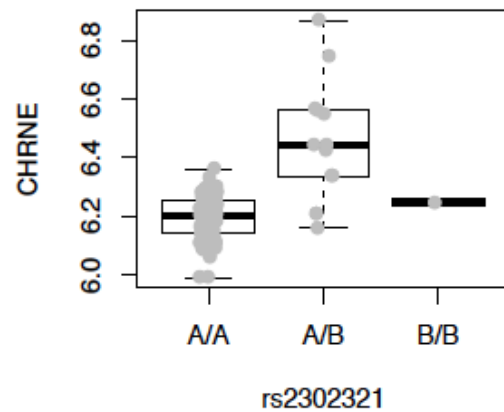
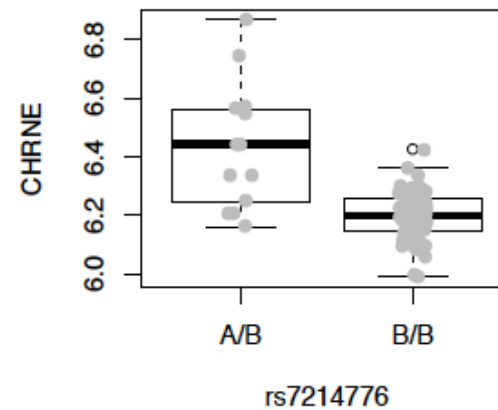
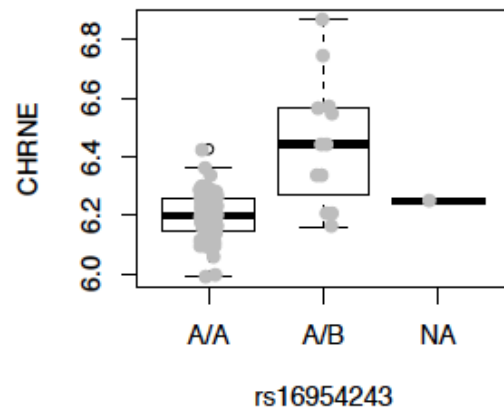
Compute and manage an eQTL screen for one gene

```
> unix.time(t1 <- gwSnpTests(genesym("CHRNE")~male, c17, chrnum("chr17")))
  user  system elapsed
 0.576   0.024   0.602
> length(p.value(t1@.Data[[1]]))
[1] 89701
> topSnps(t1,n=5)
               p.val
rs16954243 2.925728e-09
rs7214776  7.564315e-09
rs8081611  7.564315e-09
rs2302321  4.838786e-08
rs8070572  2.505861e-07
> t1@.Data[[1]][rownames(.Last.value),]
      Chi.squared Df      p.value
rs16954243    35.23268 1 2.925728e-09
rs7214776     33.38401 1 7.564315e-09
rs8081611     33.38401 1 7.564315e-09
rs2302321     29.78032 1 4.838786e-08
rs8070572     26.59736 1 2.505861e-07
> □
```

```

> par(mfrow=c(2,2))
> plot_EvG(genesym("CHRNE"), rsid("rs16954243"), c17)
> plot_EvG(genesym("CHRNE"), rsid("rs7214776"), c17)
> plot_EvG(genesym("CHRNE"), rsid("rs2302321"), c17)
> plot_EvG(genesym("CHRNE"), rsid("rs8070572"), c17)
> par(mfrow=c(1,1))

```



Summary on container

- Unification of array-generated expression and genotype data is reasonably straightforward with R packages
- One can handle large numbers of samples with 50k probes and 1 million SNPs without special structures
- With large SNP panels, loading and unloading chromosome-specific images is more sensible for interactive work; `externalize()/getSS()` in GGtools

Summary on single-gene test

- D. Clayton's snpMatrix package includes numerous facilities for import and fast analysis of large genotype panels
- Byte-encoding of genotypes or genotype probabilities (for imputation) saves space
- gwSnpTests connects this to the eQTL context with flexible specification of association model

Surveying large sets of genes for eQTL

```
> library(ggtut)
> f1 = observed17ceu()
> f1
```

```
eqtlTools results manager, computed Fri May  6 16:05:50 2011
```

```
gene annotation: illuminaHumanv1.db
```

```
There are 1 chromosomes analyzed.
```

```
some genes (out of 498): GI_10190685-S GI_10835020-S ... hmm23927-S hmm5188-S
```

```
some snps (out of 60967): rs6565733 rs1106175 ... rs7502145 rs4986109
```

```
> f1@call
```

```
eqtlTests(smlSet = c17, rhs = ~male, targdir = "c17c", geneApply = mclapply,
          genegran = 1)
```

The object `f1` holds results of 30361566 tests for expression-genotype association. Note

Compare sizes and locations of peaks to results when expression permuted against genotype



Additional topics in the array context

- SNP imputation using regression or population haplotype models
- Expression heterogeneity reduction using surrogate variable analysis
- Details of searching for trans-associated eQTL

The sequencing context

- We are concerned mainly with processing and interpreting RNA-seq data
- We assume filtering and alignment are done well
- We use the BAM format to manage all short reads, one BAM file per sample
- Idioms comparable to the X[G,S] filtering specification for ExpressionSet instances are available with BamViews in Rsamtools

A basic product of interest, for a specified SNP

	NA07055	NA06985	NA06993	NA06994	NA07000	NA07022	1
A	0	66	0	86	51	0	
C	53	0	193	30	1	39	
G	0	0	0	1	0	0	
T	0	0	0	0	0	0	
	NA11004	NA11829	NA11830	NA11831	NA11832	NA11839	1
A	144	0	0	22	33	119	
C	27	7	52	8	1	47	
G	0	0	0	0	0	0	
T	0	0	0	0	0	0	
	NA11993	NA12003	NA12004	NA12005	NA12006	NA12043	1
A	203	0	111	0	47	0	
C	43	2	0	266	0	51	
G	0	0	0	0	0	0	
T	0	0	0	0	0	0	
	NA12044	NA12045	NA12046	NA12047	NA12048	NA12049	1

Conclusions

- Comprehensive eQTL search – a readily solved problem with 50k probes vs 10 million SNP
- Interpreting associations found in such searches is challenging; scalable and linkable access to result sets is essential
- Working with RNA-seq data to dissect mechanisms by which DNA variants influence mRNA abundance is in early stages, but feasible with tools described here

References

- Cheung, V. G., R. R. Nayak, et al. (2010). "Polymorphic cis- and trans-regulation of human gene expression." PLoS Biol **8**(9).
- Cheung, V. G., R. S. Spielman, et al. (2005). "Mapping determinants of human gene expression by regional and genome-wide association." Nature **437**(7063): 1365-9.
- Cooper, D. N. (2010). "Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes." Hum Genomics **4**(5): 284-8.
- Coulombe-Huntington, J., K. C. L. Lam, et al. (2009). "Fine-scale variation and genetic determinants of alternative splicing across individuals." PLoS Genet **5**(12): e1000766.
- Majewski, J. and T. Pastinen (2011). "The study of eQTL variations by RNA-seq: from SNPs to phenotypes." Trends Genet **27**(2): 72-9.
- Pandit, S., D. Wang, et al. (2008). "Functional integration of transcriptional and RNA processing machineries." Curr Opin Cell Biol **20**(3): 260-5.
- Stranger, B. E., A. C. Nica, et al. (2007). "Population genomics of human gene expression." Nat Genet **39**(10): 1217-1224.
- Veyrieras, J.-B., S. Kudaravalli, et al. (2008). "High-resolution mapping of expression-QTLs yields insight into human gene regulation." PLoS Genet **4**(10): e1000214.
- Williams, R. B. H., E. K. F. Chan, et al. (2007). "The influence of genetic variation on gene expression." Genome Research **17**(12): 1707-1716.