Contents

1 Details of cell lines and RNA-Seq procedures

1.1 Cell lines used

In the analyses presented in the main paper, we used data generated by sequencing RNA from 69 lymphoblastoid cell lines obtained from Corriell. The cell lines were derived from Yoruban individuals from Nigeria by the International HapMap Project, and have been extensively genotyped (Frazer et al., 2007). Many of the individuals in the HapMap are parts of trios; we used only the parents in these families. The full list of individuals is in Supplementary Table 1; we included 54 individuals from HapMap "plate 1" and 15 from HapMap "plate 2".

1.2 RNA Sequencing

Sequencing libraries for the Illumina GA2 platform were created from the polyadenylated fraction of RNA from each cell line. Total RNA was extracted using an RNeasy Mini Kit (Qiagen) and assessed using an Agilent Bioanalyzer. mRNA was then isolated with Dyna1 oligo-dT beads (Invitrogen) from 10 μg of total RNA. The mRNA was randomly fragmented using the RNA fragmentation kit from Ambion. First-strand cDNA synthesis was performed using random primers and SuperScriptII reverse-transcriptase (Invitrogen). This was followed by second-strand cDNA synthesis using DNA Polymerase I and RNase H (Invitrogen). The Illumina adaptor was ligated to the ends of the double-stranded cDNA fragments and a 200 bp size-selection of the final product was performed by gel-excision, following the Illumina-recommended protocol. 200 bp cDNA template molecules with the adaptor attached were enriched by PCR to create the final library.

Each library that was prepared was sequenced twice, once at the Yale sequencing center using 35 bp reads, and once at the Argonne sequencing center using 46 bp reads. In the course of examining variability between libraries, multiple libraries were prepared and sequenced for a subset of cell lines; we found that it increased power to include these libraries by averaging expression levels across libraries of the same cell line. Image analysis and base calling were done with the Illumina pipeline version 1.3.2.

1.2.1 Mapping reads to the genome

The genome sequence used for read mapping was a slightly modified version of hg18 – we converted the pseudo-autosomal region of the Y chromosome to all Ns and removed the files labeled as "random", as these often contain known duplicated regions. We then mapped all reads to this genome sequence using MAQ v0.6.8 (Li et al., 2008) using the default parameters (the default settings allow two mismatches in the first 24 bases of a read). All reads that failed to map were then mapped to a database that we constructed of every possible exon-exon junction between Ensembl exons. This database was created by concatenating, for each gene, the last 50 bases of every exon with the first 50 bases of every other exon. Reads mapping to this database were assigned to the exon in which the first base of the read fell. Reads that did not map to either the genome or the exon-exon junction database were examined for evidence of having originated in the poly-A tail. We extracted all reads that either started or ended with a run of at least four As or Ts, trimmed off those terminal As or Ts, and attempted to map these trimmed reads back to the whole genome.

To obtain a rough estimate of how well we could expected mapping to perform, we simulated 35bp sequencing reads tiling the genome on both the plus and minus strands, and then mapped these reads back to the genome sequence described above. This mapping used BWA (Li and Durbin, 2009) instead of MAQ for reasons of computational speed. For each base, we calculated its "mappability" as the number of reads that correctly mapped back to the base divided by the total number of simulated reads covering the base (in this case 70). We compared the true length of mRNAs to the "mappable" length (Supplementary Figure 11A), defined as the sum of the mappability of all bases in the gene (the length of a gene was defined as the union of all transcripts associated with the gene). For most genes, the mappable length of the gene is near or equivalent to the length of the gene; however, for about 1% of genes we failed to map reads efficiently, if at all (Supplementary Figure 11B).

We note that SNPs in the genome may lead to biases in mapping, in that reads that do not carry the allele present in the reference genome sequence may not map correctly. For this reason, we experimented with mapping against a genome in which all variable positions had been masked (converted to a base not known to segregate in humans). Perhaps surprisingly, this method actually *increases* the fraction of SNP sites that show biased mapping (Degner et al., 2009). For this reason, we used the unmasked reference sequence, however, in analyses of allele-specific expression, we screened for areas of mapping biases using simulations (see section 8).

1.2.2 Quality control

The number of reads generated per lane varied from 1.2M to 11.7M, with a median of 8.4M. Of these, a median of 69% of reads mapped uniquely to the genome (range: 43-72%). Of these unique matches, a median of 86% (range: 64-91%) mapped within exons. A median of 6% of reads per lane (range: 3%-9%) mapped to exon-exon junctions. See Supplementary Table 1 for all numbers.

We compared SNPs identified in the RNA-Seq data to those from the HapMap. To do this, we generated SNP calls for all lanes using MAQ (using the default parameters). We then extracted all sites that were both typed in the HapMap and called as heterozygous in the RNA-Seq data, and checked to see if the SNP was also called as a heterozygote in the HapMap. For all samples, there was excellent concordance (>95%) between these genotype calls.

2 Identifying unannotated transcription

See Supplementary Figure 1 for a flow chart of the step for gene annotation described below. We calculated, for each base in the genome, the average rate across lanes at which it was covered by mapped sequencing reads in our data, and divided the genome into contiguous regions with evidence of expression. As a first approximation, we defined the "expression level" of a region as the maximum per base coverage of bases in the region. In general, the higher the expression level of a region in the data, the more likely it was to overlap a known exon (we define "known" as being present in the RefSeq, Ensembl, UCSC, or Vega gene databases). We chose a threshold of an average expression level of 5×10^{-8} (or 0.05 reads/million) to consider a region expressed, and merged together regions separated by less than 15 bases.

At this threshold, there are 191,982 regions of putative transcription in the human LCLs, 141,164 (74%) of which overlap annotated exons. Because of results described below (see the section "Comparison to chimpanzee") that those regions of putative transcription that overlap "most conserved" elements are observed in chimpanzee at a higher rate than those that do not, we focused on the 4,031 putative novel regions of expression that overlap these elements and are aligned between human and chimp, rhesus, mouse, rat, and dog in the 28-way vertebrate alignment from UCSC (see section on dN/dS analysis below). These putative novel exons have a median length of 452 bases, and range from 65 to 10,057 bases in length. Overall, these regions have lower rates of transcription than previously annotated exons. 36% of these regions fall in introns of known genes, while the remainder are intergenic.

We examined a number of other characteristics of these regions. First, we examined how often these regions overlap previously characterized non-coding RNAs (Khalil et al., 2009). Only 1.3% of them overlap these regions. Second, we used splice junctions identified *ab initio* from the RNA-Seq data (Section 4) to determine whether these putative exons show evidence of being part of spliced transcripts. 24.6% of these regions show evidence of being spliced (compared to 9.1% of transcribed regions that do not show evidence of conservation); of these regions, 74% show evidence of splicing to known transcripts. Two examples are shown in Supplementary Figure 6. The remaining 259 spliced regions appear to be part of previously unannotated genes.

The results of the dN/dS test (see below) suggest that the majority of transcribed regions, even those that are spliced to known genes, do not code for protein. We examined this possibility in more detail. There are 696 unannotated regions that show evidence of being spliced to known genes, 45 of which show evidence of having protein-coding function from our analysis of evolutionary conservation. We reasoned that any novel exon with evidence of being spliced between two known protein-coding exons should also be protein coding. Of the 696 unannotated regions spliced to known genes, 99 fit this profile. This still leaves the majority of unannotated transcribed regions with no evidence of having protein-coding potential. Given this, we suggest that many of them form parts of unannotated UTRs.

2.1 dN/dS analysis

To determine whether the novel, conserved transcribed regions represent unannotated protein-coding exons, we applied a test based on the ratio of non-synonymous to synonymous substitutions (dN/dS). For each putative novel exon, we first extracted the region from the 28-way alignment from UCSC (Miller et al., 2007). We observed that a number of regions called as "conserved" were present in human and distant species sequenced at low coverage (e.g., medaka), while not being present in deeply sequenced genomes like mouse or rat; manual inspection led us to conclude that many of these conserved regions are in fact due to misalignment of non-orthologous DNA, often due to pseudogenes (not shown). For this reason, we extracted only those regions that were aligned between human and chimp, rhesus, mouse, rat, and dog (as these are the most complete genomes), though we included all species for analysis of these regions. For each alignment, for each of all six possible reading frames (three on the (+) strand and three on the (-) strand), we calculated a likelihood ratio for a model under which dN/dS is estimated versus a model in which dN/dS is 1 using PAML (Yang, 2007). To format alignments for PAML, gaps in the human sequence were removed, gaps in non-human sequences were converted to Ns, and each of the six possible frames was trimmed to be a multiple of three bases.

To estimate a null distribution for calculation of the false discovery rate, we sampled 100,000 times from the distribution of the maximum of six χ^2_1 variables. This distribution was simulated using R. At a false discovery rate of 1% (corresponding to a likelihood ratio of 17.2), there are 115 transcribed regions with patterns of conservation consistent with protein-coding function. This set of 115 includes includes only 7 of the 99 regions classified as being likely protein-coding above.

2.1.1 Power estimation

This dN/dS test has been shown to be a powerful test for protein-coding potential in the fly phylogeny (Lin et al., 2008). To test our power in the mammalian phylogeny, we randomly sampled 5,000 conserved transcribed regions which overlap protein-coding exons annotated in Ensembl and performed the same procedure as above. At the same likelihood ratio threshold used to call novel regions as protein-coding (17.2), 4026 (81%) of the known protein-coding exons are significant. This suggests that we have good power in this analysis, and that the majority of conserved, unannotated regions that we identify as transcribed are not protein-coding.

2.2 Comparison to chimpanzee

We used an RNA-Seq dataset generated by sequencing the polyadenylated fraction of RNA from five chimpanzee LCLs to analyze the expression of putative novel exons identified in the human LCLs (A. Pai and Y. Gilad, unpublished data). For each human exon (novel or in Ensembl), we converted the coordinates from hg18 to PanTro2 using the LiftOver utility from the UCSC Genome Browser. We then counted the reads in each exon in the chimpanzee cell lines. Putative novel exons which overlap a “most conserved” region replicated at a higher rate than putative novel exons which do not overlap such a region, supporting our decision to focus only on those putative novel exons which are conserved (Supplementary Figure 5).

We compared how often Ensembl exons and putative novel exons with similar expression levels (in humans) are expressed in the chimpanzees. The rationale is that, if novel exons are due to spurious transcription, they should be observed in chimp at a lower rate than known, annotated exons. If, however, the novel exons are largely from true transcripts, they should be observed in chimp at approximately the same rate as annotated exons. This latter situation appears to be the case (Supplementary Figure 7).

2.3 Comparison to other tissues

We downloaded the RNA-Seq dataset generated by sequencing RNA from multiple human tissues by Wang et al. (2008) and remapped all sequencing reads to our modified version of hg18 (described above) using MAQ. We then counted the number of reads from each tissue that mapped within each novel and Ensembl exon. We compared, for the data from each tissue, the fraction of novel versus Ensembl exons in which we observed sequencing reads as for chimpanzee (Supplementary Figures 7). For nearly all tissues, the fraction of Ensembl exons observed is significantly higher than the number of novel exons observed, across expression levels. There are two exceptions to this: lymph node and breast tissue. It seems reasonable to expect that lymph node and lymphoblastoid cell lines would show similar expression levels, as they are similar tissue types. It is less clear why breast tissue would also show this pattern, although we note that Wang et al. (2008) observed that breast and lymph node also show similar patterns of alternative splicing. Overall, the observation that unannotated exons in these cells are lowly expressed and more tissue-specific than annotated exons is consistent with previous work annotating exons using sequence conservation alone (Siepel et al., 2007).

3 Identifying potential novel polyadenylation sites

As described above, our mapping strategy allowed us to find putative novel polyadenylation sites. We first identified all sequencing reads that did not initially map to the genome and either began or ended with a run of at least four As or T. We then trimmed off the run of As or Ts and remapped the reads to the genome using MAQ. At those reads that then mapped uniquely to the genome, we inferred the precise base where cleavage occurred. To filter out cleavage sites possibly due to sequencing errors, we removed putative polyadenylation sites where the downstream genomic regions contained at least three As or Ts, reasoning that a sequencing error at the non-A or T site might lead to mis-mismapping and spurious calling of a poly-A site. After filtering out these sites, and pooling across all individuals, we identified 39,729 putative poly-A sites. To examine our error rate in this analysis, we used the fact that the majority of polyadenylated mRNAs contain an AATAAA motif approximately 15-30 bases upstream of the cleavage site. We estimated the frequency of this motif for sites supported by only a single read versus sites seen more than once; while sites supported by a single read show a significant enrichment for this hexamer, sites supported by multiple reads show a much stronger enrichment (Supplementary Figure 8A). For this reason, we focused our attention on sites supported by at least two reads.

We examined the enrichment of single base variants of the consensus hexamer upstream of these sites, as some of these may play the same role as the consensus hexamer (Tian et al., 2005). Only the ATTAAA hexamer showed any enrichment upstream of these putative cleavage sites (Supplementary Figure 8B).

3.1 Estimation of the false discovery rate

We used the spatial distribution of single base-pair variants of AATAAA to estimate the false discovery rate in our analysis. For each hexamer (excluding ATTAAA), we calculated the fraction of sites containing the hexamer in the 15-30 bases upstream of the predicted polyadenylation site (Supplementary Figure 8B). This is an approximate null model for the distribution of the AATAAA hexamer. We then calculated this same fraction for the AATAAA hexamer. Of all predicted polyadenylation sites over 500 bases from a known polyadenylation site (known sites are from the Ensembl, RefSeq, Vega, and UCSC databases), 10.8% have an AATAAA hexamer 15-30 bases upstream. The average fraction for all other single base variants of AATAAA is 1.4%. This gives an FDR of 12.9% for the class of predicted poly-A sites that lie at least 500 bases from a known site and contain a match to the AATAAA hexamer upstream of the site.

4 *Ab initio* prediction of splice junctions

In the analysis of unannotated transcription and in Figures 1A and 3A in the main text, we analyze splice junctions identified from the RNA-Seq data that are not present in current databases. In this section, we describe the identification and properties of these junctions. We note that our approach for *ab initio* identification of splice junctions is similar to previous approaches (Trapnell et al., 2009; Yassour et al., 2009), but tailored specifically to our data.

4.1 Gapped alignment procedure

To identify splice junctions without reference to a set of known exons, we used only those Illumina lanes sequenced at the Argonne sequencing center, as these reads are 46 bases long (as opposed to 35 at the Yale sequencing center). We initially removed all sequencing reads that mapped (either uniquely or non-uniquely) to the reference genome (note we do not filter based on matches to known exon-exon junctions). This left us with 47 million sequencing reads. We then attempted to map both the first 20 bases and the last 20 bases of each of these reads to the genome independently using MAQ. We identified reads where at least one end mapped uniquely to the genome. If both ends mapped uniquely, we filtered for those pairs in which both mapped to the same strand of the same chromosome with a minimum distance between them of 70 bases, and in the "correct" order (i.e., for reads mapping to the plus strand, the position of the starting 20-mer of the read must be before the position of the ending 20-mer). We then extended the alignment between these two seeds by exhaustively searching the possible extensions and choosing the one with the least number of mismatches.

For reads where only one of the two ends mapped uniquely, we extended the alignment on the mapped end as long as the reference and the sequencing read were perfect matches. This left us with an unmapped end with a length between six (chosen as the minimum length) and 26 bases. We then searched within 20kb of the mapped end for a unique perfect match to this remaining sequence. Together, this left us with 20 million reads spanning putative splice junctions.

4.2 Identification of junction boundaries and analysis of specificity

Each individual read is generally compatible with more than one possible junction boundary due to the sequence organization of the splice junctions. We grouped individual sequence reads together if there was any overlap in the set of splice junctions compatible with each. We then grouped together junctions that had both the 5' and 3' ends located within two bases of each other. After this procedure, we arrived at a set of 408,826 predicted splice junctions.

To evaluate whether these predicted splice junctions represent the outcomes of true splicing reactions, we evaluated how often our predicted junctions contained the consensus GT-AG splice site in the intronic two bases. We also compared these predicted junctions to those present in current gene annotations. In Supplementary Figure 10, we show the fraction of predicted junctions that match the consensus GT-AG splice site as a function of the number of reads supporting the junction. The majority of junctions seen more than once match the consensus, approaching nearly 95% of those seen more than 20 times. As a control, we also evaluated how often these junctions match the pair GT-TC (note that the strand of the 3' splice site is switched in this control). Almost no junctions match this pair of dinucleotides (Supplementary Figure 10). Further, the more reads that support the existence of a junction, the more likely it is to be present in current databases of gene models. We infer the presence of a large number of unannotated splice junctions–though (for example) 70% of the 33,923 splice junctions supported by two sequencing reads match the consensus splice site, only 23% of these are annotated.

5 HapMap genotypes and imputation

For association mapping, we used the HapMap combined Phase 2 and 3 genotypes from release 27. As some SNPs were typed in one panel but not in the other, we performed genotype imputation using the fastPHASE model (Scheet and Stephens, 2006) implemented in bimbam (Guan and Stephens, 2008). The EM algorithm was run 5 times with 20 steps per run (the bimbam options are -e 5 -s 20). All association analyses use the posterior mean genotype, as recommended by Guan and Stephens (2008).

For analyses of allele-specific expression, we used the phased haplotypes from Phase 2 of the HapMap (release 22). These individuals were all phased in trios, so the phasing is highly accurate.

6 Association mapping

In this section, we describe the normalization and correction procedures used for the eQTL and sQTL analyses, and give additional summaries of the results not presented in the main text. See Supplementary Figure 2 for a flow chart of the steps described in detail below.

6.1 eQTL mapping

As noted in the main text, we performed a number of correction and normalization steps on the RNA-Seq data to gain power for the mapping part of the paper. Below, we describe these steps in full. We started by counting, for each exon in each lane, the number of sequencing reads mapped to that exon. Let this count be x_{ij}, where i indexes exon and j indexes lane.

6.1.1 Correction for GC content

We noticed that different lanes (even of the same individual) often showed preferential sequencing of genes or exons at different levels of GC content (Supplementary Figure 12). To correct for this, we performed the following procedure:

1. Assign all exons to 200 approximately equally-sized bins based on GC content. Let s_{lj} be the number of reads in bin l from lane j.

2. For each bin, for each lane, calculate the $\log_2$ relative enrichment, f_{lj}, of reads in each GC bin: $f_{lj} = \log\left(\frac{s_{lj}/\sum_j s_{lj}}{\sum_l s_{lj}/\sum_l \sum_j s_{lj}}\right)$

3. For each lane, fit a spline to the plot of f_{lj} against the mean GC content for the bin (Supplementary Figure 12). We used the R function "smooth.spline" with a smoothing parameter of 1.

4. Now, estimate the over/under-representation of each individual exon in each lane from the spline (we used the "predict" function in R); let $\hat{g_{ij}}$ be the predicted log over/under-representation of exon i in lane j.

5. Set $x_{ij}^{new} = x_{ij} 2^{-\hat{g_{ij}}}$

After this step, we computed gene-level expression by summing the corrected exon-level values in each gene and dividing by the total (corrected) number of reads in the lane, as follows. If we let y_{jk} be the "expression level" of gene k in lane j, and Z be the set of indices of all exons in gene k, we can write this as:

$$y_{jk} = \frac{\sum_{i \in Z} x_{ij}}{\sum_i x_{ij}}.$$

Here, y_{jk} is our estimate of the fraction of all reads in lane j from gene k.

6.1.2 Correction for center and concentration effects

As each sample was sequenced at two different centers, we included an explicit correction for differences between the centers. We also sequenced some cell lines at a concentration of 2.5 pM, then switched to 3.5 pM later in the experiment. We included an explicit correction for the differences between these sequencing concentrations as well. In both cases we calculated the median expression level for each gene in both groups, then adjusted the expression level in one of the groups. Let μ_{k1} be the median expression level of gene k in group 1 (one of the sequencing centers or concentrations), and μ_{k2} be the corresponding expression level in group 2. Recall that y_{jk} is the adjusted expression level of gene k in lane j. We adjusted these expression levels such that (if lane j is in group 1):

$$y'_{jk} = y_{jk}\frac{\mu_{k2}}{\mu_{k1}}$$

After this step, we averaged expression levels across lanes of the same individual. Let z_{kl} be the expression level of gene k in individual l, and S be the set of indices of all lanes measuring the expression of individual l. Then:

$$z_{kl} = \frac{\sum_{j\in S} y_{jk}}{|S|}$$

After averaging expression levels across lanes of the same individual, we removed all genes with a median expression level of zero.

6.1.3 Quantile normalization

One of the assumptions of the linear regression model used for identification of eQTLs is that the expression values follow a Gaussian distribution within genotype classes. This assumption is violated by outliers or non-normality (for whatever reason) of expression levels inferred from the sequencing data. While one approach would be to examine the robustness of each individual regression to dropping outliers or different transformations of the data, this is infeasible for the millions of regressions performed in this study. We thus guaranteed that the overall distribution of expression levels for each gene is normal by transforming the ranks of the expression values for each gene to their respective quantiles of a $N(0,1)$ distribution (using the “qqnorm” function in R). Ties (in practice only due to estimated expression levels of zero in some genes for some individuals) were broken randomly.

6.1.4 Unidentified confounders

The measured expression levels of sets of genes in some individuals may be correlated for a number of both technical and biological reasons. This should not generate false positives in our study design, but may reduce power. To increase power, we removed these correlations using principal components analysis. This approach was motivated by the success of similar approaches in the analysis of expression microarray data (Kang et al., 2008; Leek and Storey, 2007). We calculated the principal components of the $N \times M$ matrix of N (quantile-normalized) expression values by M individuals. For each individual, we extracted their loadings on the first 16 principal components (PCs) using the R function “prcomp”. For each gene, we performed a linear regression of expression value on the first 16 PCs, and replaced the expression value of individual l at gene k with their residual in that regression. That is,

$$z_{kl}^{new} = z_{kl} - \sum_{n=1}^{16} \hat{\beta_{nk}} x_{nl}$$

where x_{nl} is the loading for individual l on PC n and $\hat{\beta_{nk}}$ is the regression coefficient from the regression of the expression level of gene k on PC n. We chose to remove 16 PCs because this empirically gave the largest number of eQTLs in downstream analysis. As these residuals may be slightly non-normal, we performed a second round of quantile normalization after this step. These expression values z_{kl}^{new} were used as input to the regressions performed in the eQTL study as described below. Overall, we noticed the largest increases in power in the entire procedure came from the quantile normalization and the PC correction.

6.2 sQTL mapping

A number of studies have used exon microarrays or quantitative PCR to identify SNPs influencing the relative expression of different isoforms of a gene (Fraser and Xie, 2009; Heinzen et al., 2008; Hull et al., 2007; Kwan et al., 2008; Zhang et al., 2009b). Here we describe our approach using RNA-Seq. The correction and normalization steps done to the data before performing the sQTL study were similar to those steps described above for the eQTL study. The differences are described below:

6.2.1 Correction for GC content

After the GC correction, we did not convert the exon-level counts to gene-level counts or correct for difference in read depth across lanes. Instead, we converted the counts from each exon to the fraction of reads in that exon out of all reads in the gene. In the notation from the section on GC content above, let y_{ij} be the "exon expression level" of exon i in lane j, and Z be the set of indices of all exons in gene k, we can write this as:

$$y_{ij} = \frac{x_{ij}}{\sum_{i \in Z} x_{ij}}$$

The corrections for center and concentration effects, averaging across lanes of the same individual, and quantile normalization were performed on these values as described above.

6.2.2 Correction for unidentified confounders

The PC correction in the sQTL analysis was similar to that for the eQTL analysis, but we removed the effect of eight, rather than 16, PCs (this choice was made because removing the effects of 8 PCs gave us the largest number of sQTLs). Similarly to the above procedure, we calculated the principal components of the $N \times M$ matrix of N (quantile-normalized) *exon* expression values by M individuals. For each individual, we extracted their loadings on the first 8 principal components

(PCs) using the R function "prcomp". For each gene, we performed a linear regression of exon expression value on the first 8 PCs, and replaced the expression value of individual l at exon i with their residual in that regression. If, analogously to the notation above, z_{il} is the uncorrected exon expression level for exon i in individual l,

$$z_{il}^{new} = z_{il} - \sum_{n=1}^{8} \hat{\beta_{ni}} x_{nl}$$

where x_{nl} is the loading for individual l on PC n and $\hat{\beta_{ni}}$ is the regression coefficient from the regression of the exon expression level of gene i on PC n. As above, we performed a second round of quantile normalization after this step. These expression values z_{il}^{new} were used as input to the regressions performed in the sQTL study described below.

6.3 Linear regression and estimation of FDR

For "local" association studies (focused on SNPs falling in a candidate region including the gene and 200kb on either side of the gene), least-squares linear regressions between expression levels (either the gene or exon expression levels described above) and genotypes at SNPs within 200 kb of the gene were performed in R. To estimate the false discovery rate, for each gene we permuted the phenotypes three times, recalculated the linear regressions, and recorded the minimum P-value for the gene for each permutation. This set of minimum P-values forms the empirical null distribution for the P-values. We then compared the true distribution of the minimum P-values to this null distribution to estimate the FDR. That is, we found the P-value z such that $\frac{P(p_0<z)}{P(p_1<z)} = x$, where x is the desired FDR, p_0 is a P-value from the null distribution, p_1 is a P-value from the true distribution, $P(p_0 < z)$ is the fraction of minimum P-values from the permutations that fall below the P-value threshold, and $P(p_1 < z)$ is the corresponding fraction in the non-permuted data.

For genome-wide association studies, we used bimbam version 0.99 (Guan and Stephens, 2008) to perform Bayesian linear regression (Servin and Stephens, 2007) between the normalized gene expression levels described above and all SNPs genome-wide.

6.4 Summary of QTL mapping results

We first summarize the results from the genome-wide association studies using bimbam. At a significance threshold of a log10(Bayes Factor) of 6, there are 76 genes with an eQTL; of all significant SNPs, 92% are within 200kb of the respective gene. As different genes have different numbers of significantly associated SNPs, we also considered the top Bayes Factor for each gene. Of these, 71% are within 200kb of the respective gene and 79% are on the same chromosome. If we limit ourselves to a more stringent threshold of a log10(BF) of 8, there are 23 eQTLs, all but two of which are on the some chromosome as the affected gene. Inspection of the most significant distant eQTL revealed that it is also a local eQTL for a nearby gene; we suspect this is a false distant eQTL due to sequencing reads originating from a small section of a nearby gene mapping far from the gene. Overall, then, the overwhelming majority of strong associations are "local" associations.

We now summarize the results of the "local" eQTL and sQTL studies. Linear regressions and permutations were performed as described above. As reported in the main text, at an FDR of 10%, we identified 929 and 187 genes with eQTLs and sQTLs, respectively. Numbers for different FDR thresholds are in Supplementary Table 2.

6.5 Distribution of eQTLs around the gene

We examined the distribution of eQTLs with respect to distance from the transcription start and end sites, using the Bayesian hierarchical model developed by Veyrieras et al. (2008) (Supplementary Figure 13). Previous eQTL studies using Illumina microarrays reported peaks of eQTL density near both the transcription start and end sites (Cheung et al., 2005; Veyrieras et al., 2008). In our study using RNA-Seq, the peak of eQTL density near the 3' end of genes is much attenuated (Supplementary Figure 13). This does not appear to be due to measurements of expression levels by micoarrays being confounded by SNPs in array probes, as the 3' peak in the array study remains even after removing all probes containing SNPs in these individuals, as determined by data from the 1000 Genomes Project (not shown). We now believe that much of the 3' peak in eQTL density in the Illumina array data is due to SNP effects on splicing of the exon containing the probe. We will provide more analysis of this issue in a future publication.

6.6 Relationship between read depth and power

After identifying eQTLs and sQTLs, we examined the rate at which we detected QTLs at genes/exons in different bins of expression level (Supplementary Figure 15). For overall gene expression levels, we found that the rate of eQTLs approached a plateau at a mean expression rate of approximately 1 read per million (Supplementary Figure 15A), corresponding to approximately 10 reads/individual in our data. About 65% of expressed genes (defined as genes with a median number of reads per individual greater than one) are covered at this level or higher in our data, suggesting that deeper coverage of individuals would result in greater power for about 35% of genes. This is necessarily a very rough approximation, as it is unknown if eQTL rates vary across expression levels for biological, as well as statistical, reasons.

For splicing QTLs, the rate of QTLs approaches a maximum only at very high levels of expression (Supplementary Figure 15B)–upwards of 1 read per 10,000, an expression level achieved by only around 0.5% of all exons in our data. This implies greater power to detect sQTLs for longer exons (which have more reads). An additional consideration is that we take the most significant exon in a gene as the target of the sQTL, which may further increase the skew in our results towards highly expressed exons. However, these results suggest that deeper sequencing of individuals will increase power to detect sQTLs at nearly all genes.

6.7 Comparison of eQTLs across populations

We obtained P-values from the top 500 SNP-gene associations (i.e., the lowest P-values) from an RNA-Seq eQTL study performed in LCLs derived from individuals of Northern European descent (S. Montgomery et al., sumbitted), and calculated the P-values for these same SNP-gene pairs in our data. After excluding monomorphic SNPs in the Nigerian population, there are 460 SNP-gene associations which we could compare. Among the 460 associations identified in the European populations, there is a clear enrichment for significant P-values in the Nigerian population (Supplementary Figure 16). For example, of the top 50 associations identified in the European population, 20 have $p < 0.01$ in our data, a 40-fold enrichment. Overall, of the top 460 associations identified in the European population, 51 have $P < 0.01$ in our data, an 11-fold enrichment over background. We note that the data collection and data analysis strategies likely differed between our eQTL study and the eQTL study performed in the European population, so direct comparison of P-values, as done here, is only a very rough approximation of the amount of overlap between the two studies.

7 Comparison of RNA-Seq to genome-wide expression microarrays

We compared the expression levels inferred from RNA-Seq in this study to those inferred from a data set generated using genome-wide Affymetrix exon microarrays (Huang et al., 2007). We performed similar comparisons to the data generated for the same individuals using Illumina microarrays (Stranger et al., 2007) (which have in general a single probe per gene), though we focus on the exon microarrays due the fact that they, like RNA-Seq, attempt to measure expression using information from the entire transcript.

7.1 Pre-processing of exon array data

We downloaded the exon array data from GEO and extracted the data for the 53 individuals in common between this study and the study of Huang et al. (2007). We remapped the probe sequences to hg18 using MAQ, and excluded those probes overlapping a SNP in the June 2009 release of the 1000 Genomes Project. Expression levels from all probes that mapped to the exons of a given Ensembl gene were combined to obtain a gene-level expression measurement using RMA (Irizarry et al., 2003).

7.2 Correlation between gene expression assayed with two technologies

To compare these exon array measurements to those from the RNA-Seq, for each individual we defined the RNA-Seq expression level of each gene as the fraction of reads mapping to exons of the gene (using the union of all annotated transcripts) divided by the mappable length of the gene. Spearman correlations between the exon array and RNA-Seq measurements of expression levels were similar across individuals (Supplementary Figure 3).

A perhaps more important comparison is how well the expression levels inferred from the microarray and RNA-Seq data correlate *within* genes, across individuals. For each gene, we calculated Spearman correlations between the expression levels determined by the RNA-Seq data and by the array data. These correlations were smaller, with a median around 0.12 (Supplementary Figure 4A). The highest correlations were found at genes with intermediate expression levels in the RNA-Seq data (Supplementary Figure 4B), consistent with the possibility that arrays measure expression levels less well at the low and high ends of the expression range.

7.3 Comparison between eQTLs identified using two technologies

We also compared the eQTLs identified using exon arrays to those identified using RNA-Seq in the same set set of 53 individuals measured in both studies. For each gene, we quantile-normalized the expression levels as measured by the exon array, and performed association studies for these genes (all association studies used SNPs within 200kb of the gene). By permutation (as described in Section 6.3), we identified 138 genes with eQTLs using expression levels from the microarray data. In Supplementary Figure 14, we plot the effect size in both the array and sequencing association studies for these 138 gene-SNP pairs. As noted in the main text, 93% of the associations are in the same direction, and 70% are significant at a p-value threshold of 0.05.

8 Identification of allele-specific expression

RNA-Seq data can be used to identify regions of allele-specific expression (Degner et al., 2009; Heap et al., 2009; Lee et al., 2009; Zhang et al., 2009a). For these analyses, we used only those individuals and SNPs from HapMap release 22, as they were phased in trios and thus the phasing is highly accurate. For each gene with an eQTL (at an FDR of 10%), we identified all individuals heterozygous for the most significant SNP. Then, in these individuals we identified all sequencing reads overlapping all heterozygous exonic SNPs in the gene, and assigned each read to either the high-expression haplotype or the low-expression haplotype using the phased data. If the most significant SNP was in release 27 of the HapMap (used for the eQTL mapping) but not in release 22, we used the best proxy from release 22 to define the high- and low-expression haplotypes (if there was no SNP with a proxy with $r^2 > 0.8$, we excluded the gene).

Some SNPs have a “mapping bias”, in that sequencing reads derived from one of the alleles map preferentially to the genome (Degner et al., 2009). These SNPs will cause errors in the estimation of allele-specific expression. To identify potentially biased reads, for each exonic SNP, we simulated all possible reads (of both 35 and 46 base pairs in length) that could overlap the SNP from either allele. We then mapped these reads back to the genome using MAQ. Any SNP that showed a mapping bias in favor of either allele (in that an unequal number of reads from each allele successfully mapped back to the genome in these simulations) was excluded from analysis.

8.1 Beta-binomial model

For each gene with an eQTL, we used a beta-binomial model to estimate the fraction of reads coming from the (+) haplotype in individuals heterozygous for the most significant SNP. We took this approach rather than using, for example, a binomial model, to allow for over-dispersion in the data due to heterogeneity in the fraction of reads from the (+) haplotype across individuals (due to, for example, differences in genetic background).

After removing potentially biased SNPs, we called a gene "informative" if at least two individuals had at least five reads that could be assigned to either the low- or high-expression haplotype. Let X be the number of reads coming from the high-expressing haplotype, and y be the total number of reads that could be assigned to individual haplotypes. For each individual, then,

$$X \sim BeBi(y, \alpha, \beta) \tag{1}$$

The probability density for the beta-binomial distribution is

$$p(x|y, \alpha, \beta) = \frac{B(\alpha + x, \beta + y - x)}{B(\alpha, \beta)} \tag{2}$$

where $B(a, b)$ is the beta function. The mean of the beta-binomial is $\mu = \frac{\alpha}{\alpha+\beta}$. The overall log-likelihood of the data, assuming each individual is independent, is:

$$\log(P(D)) = \sum_i \log\left[\frac{B(\alpha + x_i, \beta + y_i - x_i)}{B(\alpha, \beta)}\right] \tag{3}$$

where i indexes individual.
We then considered the following hypotheses:

1. H0: $\alpha = \beta$ (i.e., $\mu = 0.5$)
2. H1: $\alpha \neq \beta$

Parameters were estimated using a maximum likelihood approach. As these models are nested, $-2[l(\hat{H}0) - l(\hat{H}1)] \sim \chi^2_1$, where $l(\hat{H}0)$ and $l(\hat{H}1)$ denote the likelihoods calculated at the MLEs for each model. All maximizations of the likelihoods were performed using the "optim" or "optimize" functions in R. In Figure 2B in the main text, we plot the histogram of $\hat{\mu}$ estimated under H1 for the set of 244 informative genes.

8.2 Comparison to eQTL effect size

From the beta-binomial model above, we have an estimate, for each eQTL, of the fraction of reads coming from the (+) haplotype in individuals heterozygous for the eQTL. We then wanted to compare this fraction to our expectation given the magnitude of the effect size of the QTL. To do this, we defined the expression level of a gene in a lane as the number of reads falling into exons of that gene divided by the total number of exonic reads in the lane, then averaged this fraction across lanes sequencing RNA from the same cell line. We did this, rather than use the normalized and corrected expression levels used in the initial eQTL study, because it is unclear how the estimates of allele-specific expression relate to expression levels after transformation. For each gene, we performed a linear regression of the uncorrected expression level of the gene, Y_i (the number of reads in the gene divided by the number of reads in the individual), on the genotype of the most significant SNP, g_i:

$$E[Y_i] = a + Bg_i \tag{4}$$

where i indexes individual. This gives an estimated allelic effect $\hat{B}$ and the estimated expression level in the homozygote, $\hat{a}$. The expected fraction of reads coming from the high expressing haplotype is then $\frac{\hat{a}+2\hat{B}}{2(\hat{a}+\hat{B})}$. In Figure 2C in the main text, we plot this estimate of the allelic effect against the estimate using allele-specific expression for each of the 222 informative genes.

9 Details of the hierarchical model

We used a modification of the hierarchical model presented in Veyrieras et al. (2008) to estimate the effects of SNP and exon annotations on the probability of a gene to have a splicing QTL. This model has several components. First, there is the calculation of the Bayes factor, defined as $\frac{P^1_{ijk}}{P^0_{ij}}$, where i indexes gene, j indexes exon, and k indexes SNP. That is, it is the probability of the data under a model where SNP k influences the inclusion level of exon j in gene i, divided by the probability under a model where it does not. We use priors from Servin and Stephens (2007) that allow us to compute the Bayes factor analytically. We note that we must specify a parameter that corresponds to the variance of the expected effect size (this is σ_a^2 in the notation of Servin and Stephens (2007)); we computed the Bayes factor under six different values of σ_a^2 (0.05, 0.1, 0.2, 0.4, 0.8, 1.6), and averaged these Bayes factors for each SNP, as recommended by Servin and Stephens (2007). (Note that the level of exon inclusion has been quantile normalized, and so a normal prior on the effect size is reasonable). In all cases, the prior on dominance (σ_d^2) was 0; this was motivated by the observation that nearly all eQTLs act in a strictly additive fashion (Veyrieras et al., 2008). These Bayes factors end up giving nearly identical ranks to SNPs as using a P-value from a standard linear regression (not shown).

The rest of the model requires specification of the prior probability that a given exon has a QTL, and the prior probability that a given SNP influences exon inclusion. Below, we define these and lay out the full hierarchical model.

9.1 Modeling the probability that an exon has a QTL

In the model, a gene either has a single SNP that affects the inclusion of a single exon, or no such SNP. Conditional on gene i having one such SNP, let π_{ij} be the prior probability that the exon affected is exon j. We want to allow this probability to depend on any of L annotations (for example, whether the exon in question it the first or last exon). We define this as:

$$\pi_{ij} = \frac{e^{x_{ij}}}{\sum_j e^{x_{ij}}} \tag{5}$$

where

$$x_{ij} = \sum_l \lambda_l I_{jl} \tag{6}$$

where λ_l is the effect of annotation l and I_{jl} is an indicator variable denoting whether exon j falls in annotation l.

9.2 Modeling the probability that a SNP causes a QTL

As above, we place a prior on the probability that SNP k influences the inclusion level of exon j, and to allow this probability to depend on any of M SNP-level annotations (note that there are both SNP-level and exon-level annotations, in contrast to the model in Veyrieras et al. (2008) where only SNPs have annotations). As above,

$$\pi_{ijk} = \frac{e^{x_{ijk}}}{\sum_k e^{x_{ijk}}} \tag{7}$$

where

$$x_{ijk} = \sum_m \lambda_m I_{jkm} \tag{8}$$

where λ_m is the effect of SNP annotation m and I_{jkm} is an indicator variable denoting whether SNP k falls in annotation m for exon j. Note that the SNP annotation is allowed to vary depending on the exon, such that a SNP that is annotated as a splice site SNP for one exon will be an intronic SNP for other exons.

9.3 Modification of the hierarchical model to include exon effects

Let P_i be the likelihood in this model of the data at gene i, P_i^0 be the likelihood of the data at gene i under the null hypothesis that no SNP affects the inclusion of any exon of the gene, P_i^1 be the likelihood of the data at gene i under the alternative hypothesis that the inclusion of exactly one exon of the gene is influenced by exactly one SNP, Π_0 be the prior probability of the null model, and Π_1 be the prior probability of the alternative.

$$P_i = \Pi_0 P_i^0 + \Pi_1 P_i^1 \tag{9}$$

In this case,

$$P_i^0 = \prod_{j=1}^{n_i} P_{ij}^0 \tag{10}$$

where j indexes exon and n_i is the number of exons in gene i. Additionally,

$$P_i^1 = \sum_{j=1}^{n_i} \left[\pi_{ij} P_{ij}^1 \prod_{k \neq j} (1 - \pi_{ik}) P_{ik}^0 \right] \tag{11}$$

That is, conditional on there being a single exon affected, the likelihood is the sum over the likelihood over all configurations of single affected exons. Putting all this together,

$$\frac{P_i}{P_i^0} = \Pi_0 + \Pi_1 \frac{\sum_{j=1}^{n_i} \left[\pi_{ij} P_{ij}^1 \prod_{k \neq j} (1 - \pi_{ik}) P_{ik}^0 \right]}{P_i^0} \tag{12}$$

$$\frac{P_i}{P_i^0} = \Pi_0 + \Pi_1 \sum_{j=1}^{n_i} \left[\pi_{ij} \frac{P_{ij}^1}{P_{ij}^0} \prod_{k \neq j} (1 - \pi_{ik}) \right] \tag{13}$$

$\frac{P_{ij}^1}{P_{ij}^0}$ is the probability of the observed data in exon j under the model allowing one QTL, divided by the probability of the observed data in exon j under the model allowing no QTLs, and,

$$P_{ij}^1 = \sum_{k=1}^{m_i} \pi_{ijk} P_{ijk}^1 \tag{14}$$

where k indexes SNPs, and m_i is the number of SNPs in the region. Then, defining $BF_{ijk} = \frac{P_{ijk}^1}{P_{ij}^0}$, the full likelihood is

$$P_i = P_i^0 \left[\Pi_0 + \Pi_1 \sum_{j=1}^{n_i} \left(\pi_{ij} \prod_{k \neq j} (1 - \pi_{ik}) \sum_{k=1}^{m_i} \pi_{ijk} BF_{ijk} \right) \right] \tag{15}$$

We maximized the log-likelihood using the Nelder-Mead algorithm as implemented in the GNU Scientific Library (Gough, 2003). Maximization is over Π_1 and the set of all λ. Initial estimates of the parameters were set to 0.

9.4 Annotations considered in the model

There are two types of annotation considered in the model, as noted above: exon annotation and SNP annotations. For exon annotations, we classified exons as being the first exon, the last exon, or an interior exon. Exons that could fall into either annotation (due to known alternative transcripts) were preferentially assigned to be first or last. We first fit a model with no SNP annotations. The last exon has a log odds ratio of 1.05 [0.35, 1.65] (compared to interior exons) while first and interior exons are indistinguishable. Part of the effect of last exons may by due to their larger exon size (and thus increased power). We note, however, that both Fraser and Xie (2009) and Kwan et al. (2008) and also found that the inclusion of last exons was more likely to be affected by polymorphic variation; further work is needed to fully characterize this pattern.

We considered several different SNP annotations. We first considered SNPs in the exon being tested for alternative splicing, SNPs in other exons, intronic (non-splice site) SNPs, intergenic SNPs, and SNPs within the canonical splice sites (the two bases intronic of either splice site). The splice sites are significantly enriched for sQTLs (Supplementary Figure 19). We tested whether including SNPs near the canonical splice sites improved the fit of the model. We defined SNPs in the 5' splice site as all SNPs falling two bases exonic to six bases intronic of the 3' end of an exon, and SNPs in the 3' splice site as all SNPs falling one base exonic to 20 bases intronic of the 5' end of an exon; these are approximately the binding sites for the U1 snRNP and U2AF splice factors (Watson et al., 2008). Using this definition improved the optimized log-likelihood of the model by 8.8 units without adding additional parameters. We then considered allowing different parameters for the 5' and 3' splice sites; this did not improve the model ($P = 0.41$, Supplementary Figure 19). We also considered splitting SNPs in the tested exon into those which disrupt splicing enhancers (as defined by Fairbrother et al. (2002)) and those which do not. The odds ratio for SNPs in the exon which disrupt splicing enhancers to affect exon inclusion is higher than that of SNPs in the exon which do not (a log-odds ratio of 3.9 versus 3.4), however, allowing this additional parameter did not significantly improve the model ($P = 0.56$).

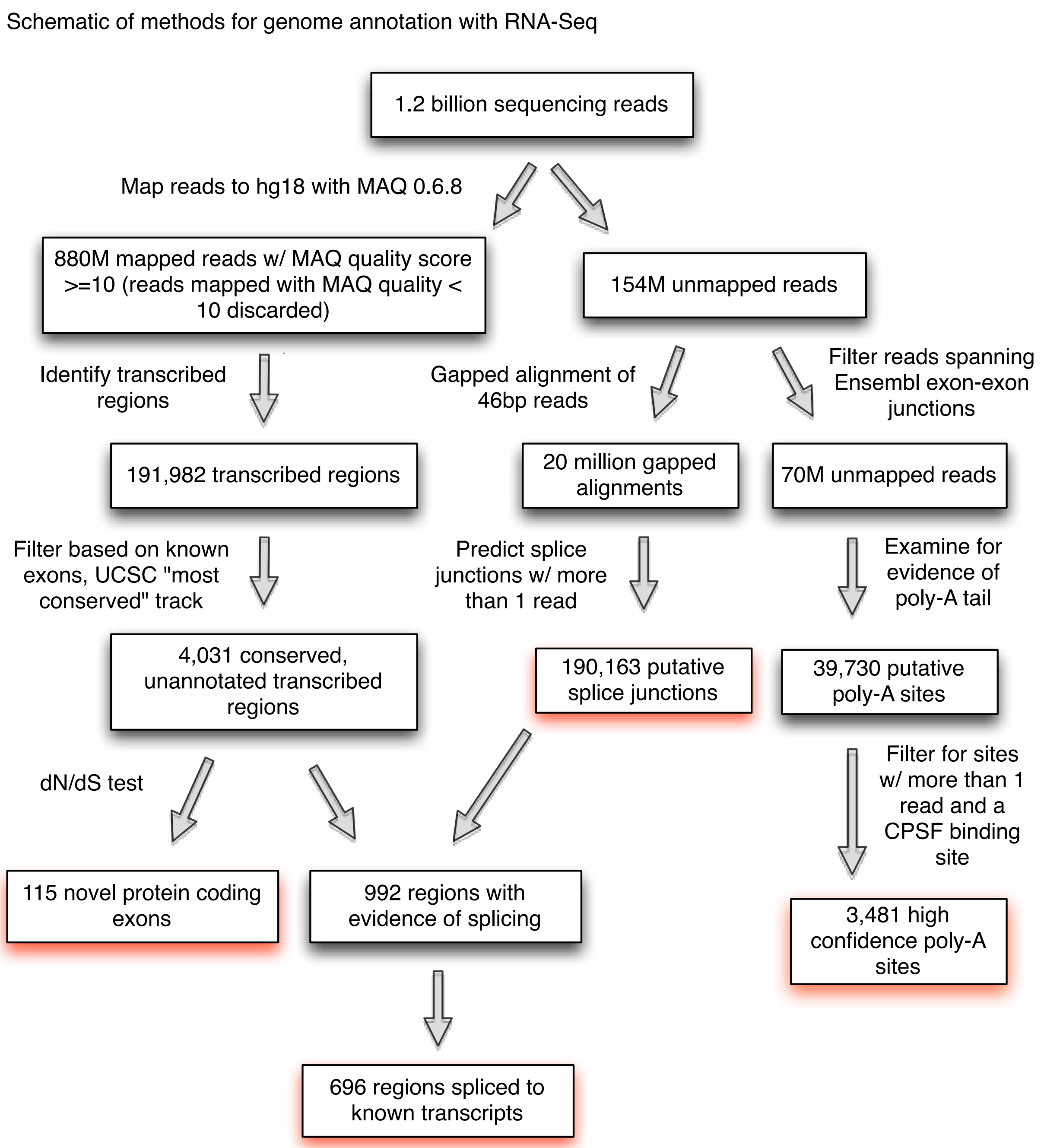
Schematic of methods for genome annotation with RNA-Seq
1.2 billion sequencing reads
Map reads to hg18 with MAQ 0.6.8
880M mapped reads w/ MAQ quality score >=10 (reads mapped with MAQ quality < 10 discarded)
154M unmapped reads
Identify transcribed regions
Gapped alignment of 46bp reads
Filter reads spanning Ensembl exon-exon junctions
191,982 transcribed regions
20 million gapped alignments
70M unmapped reads
Filter based on known exons, UCSC "most conserved" track
Predict splice junctions w/ more than 1 read
Examine for evidence of poly-A tail
4,031 conserved, unannotated transcribed regions
190,163 putative splice junctions
39,730 putative poly-A sites
dN/dS test
Filter for sites w/ more than 1 read and a CPSF binding site
115 novel protein coding exons
992 regions with evidence of splicing
3,481 high confidence poly-A sites
696 regions spliced to known transcripts

Figure 1: **Flow chart of the methods used for annotating gene models.**

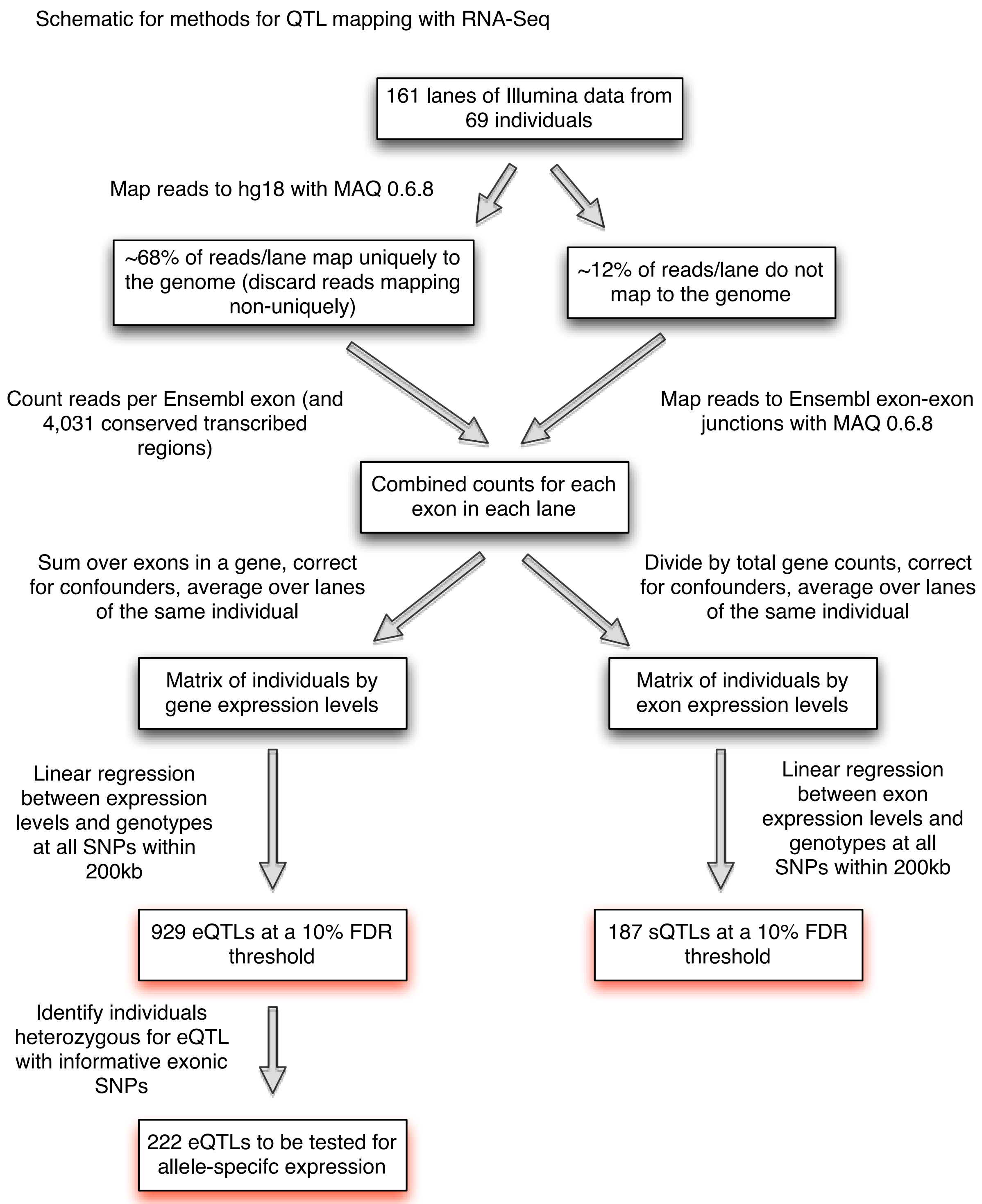
Schematic for methods for QTL mapping with RNA-Seq
161 lanes of Illumina data from 69 individuals
Map reads to hg18 with MAQ 0.6.8
~68% of reads/lane map uniquely to the genome (discard reads mapping non-uniquely)
~12% of reads/lane do not map to the genome
Count reads per Ensembl exon (and 4,031 conserved transcribed regions)
Map reads to Ensembl exon-exon junctions with MAQ 0.6.8
Combined counts for each exon in each lane
Sum over exons in a gene, correct for confounders, average over lanes of the same individual
Divide by total gene counts, correct for confounders, average over lanes of the same individual
Matrix of individuals by gene expression levels
Matrix of individuals by exon expression levels
Linear regression between expression levels and genotypes at all SNPs within 200kb
Linear regression between exon expression levels and genotypes at all SNPs within 200kb
929 eQTLs at a 10% FDR threshold
187 sQTLs at a 10% FDR threshold
Identify individuals heterozygous for eQTL with informative exonic SNPs
222 eQTLs to be tested for allele-specifc expression

Figure 2: **Flow chart of the methods used for eQTL and sQTL mapping.**

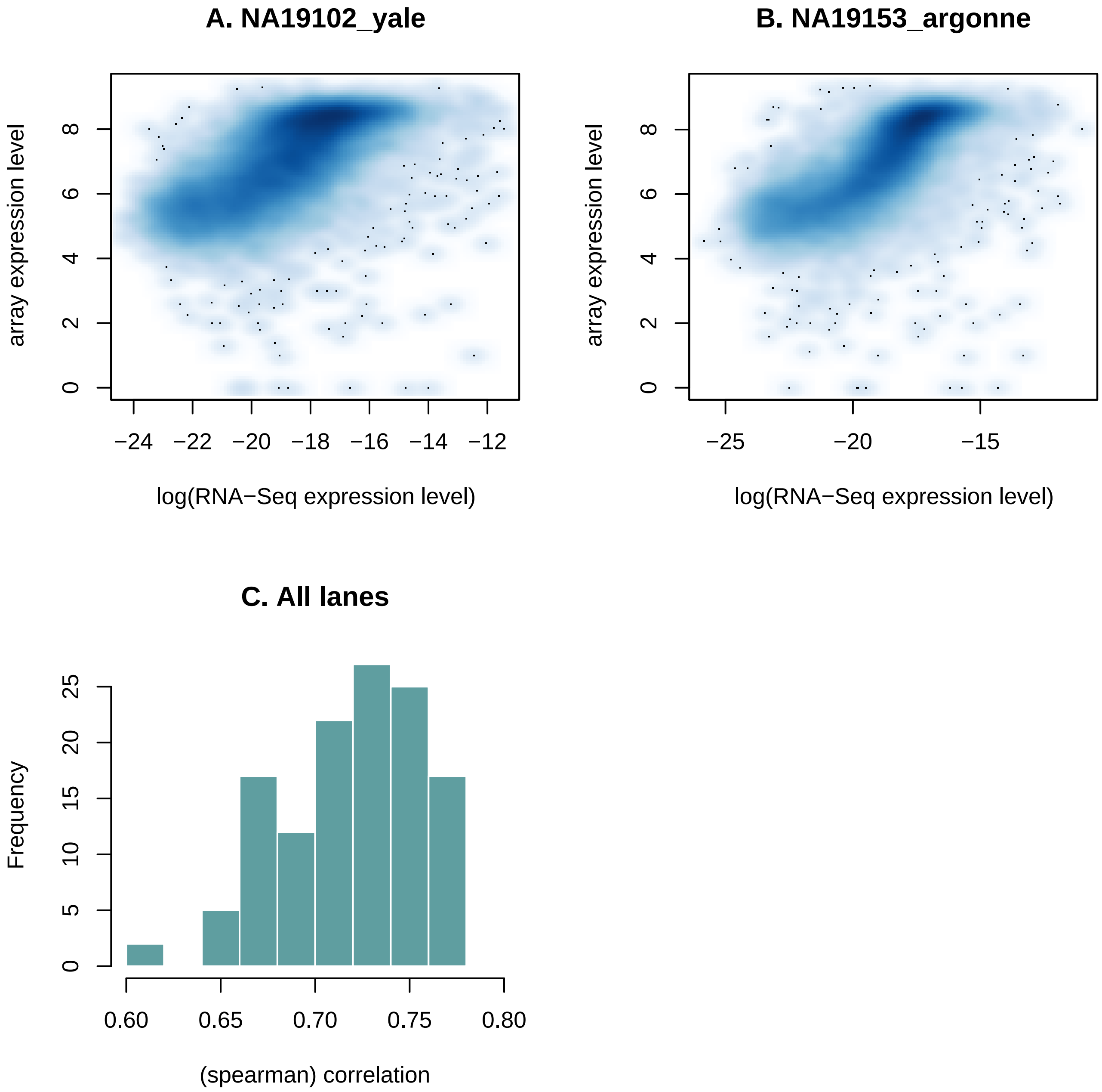

Figure 3: **Correlation of exon array and sequencing measurements of expression levels within individuals**. We compared the expression levels estimated by the two technologies for the subset of 53 individuals for whom we have estimates of expression levels from both. Here, the expression level of a gene in a lane is the number of reads mapping to exons of the gene divided by the total number of exonic reads in the lane and divided again by the mappable length of the gene; we averaged this quantity across lanes to get a mean expression level in the RNA-Seq data for each gene. **A, B. Correlation between exon array and sequencing estimates of expression for two lanes**. The expression estimates derived from the array (log2 intensities) are plotted against the (natural log of the) expression measurements derived from the sequencing. In A is the lane with the lowest correlation, and in B is the lane with the highest. **C. Similar correlations between the array and sequencing expression levels for all lanes**. Plotted is a histogram of the Spearman correlations for all lanes.

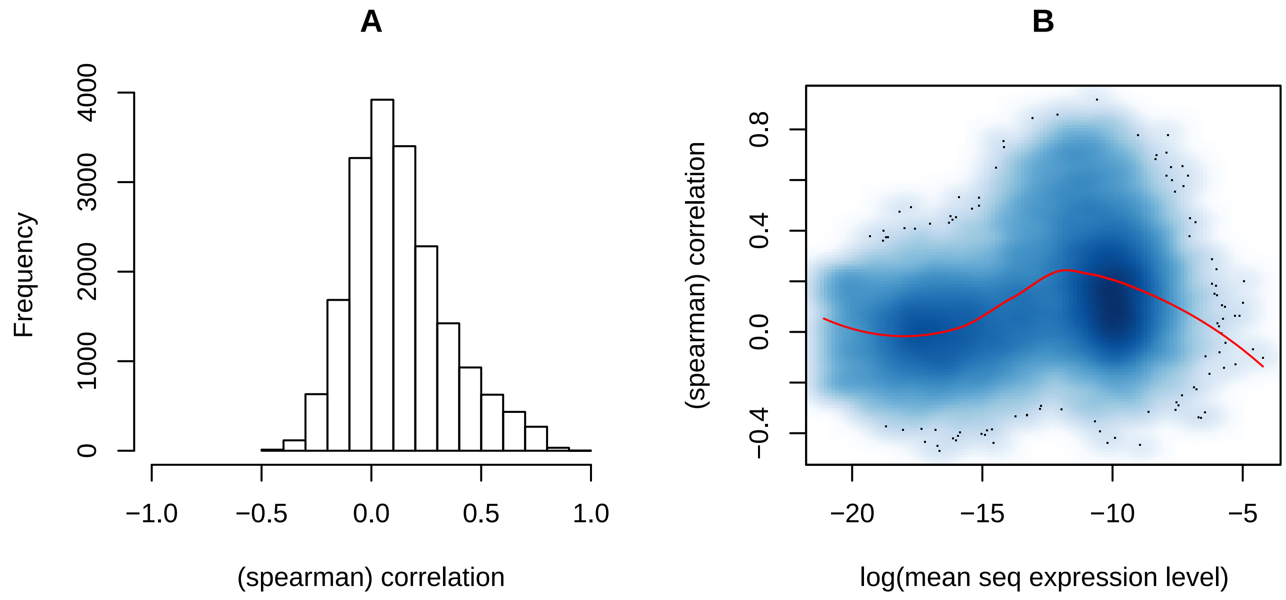

Figure 4: **Correlation of exon array and sequencing measurements of expression levels across individuals. A. Histogram of all correlations.** For each gene, we calculated the Spearman correlation between the expression levels derived from the array and sequencing data. Here, we estimated the expression level of each gene from the RNA-Seq data as the number of reads mapped to exons of the gene divided by the number of reads mapped to all exons, divided again by the "mappable" length of the gene. Plotted is the histogram for all genes. **B. Increased concordance between expression levels estimated using the two platforms at the middle of the expression range**. We plot the correlation between expression levels assayed by arrays and sequencing as a function of the natural log of the expression level of the gene (in the sequencing data). The correlations are highest towards the middle of the expression range, and close to zero at low expression levels.

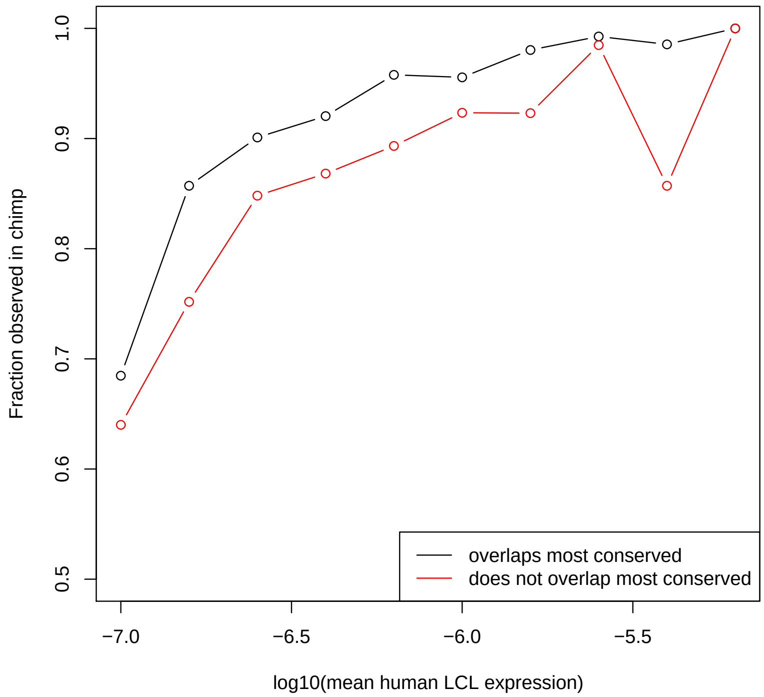

Figure 5**: Putative exons conserved at the DNA level are more likely than unconserved exons to be transcribed in chimpanzee**. We split unannotated transcribed regions into those that overlap most conserved regions (as defined by phastCons on a 28-way alignment of vertebrates) and those that do not. We then tested each region for evidence of expression in chimpanzee. For exons expressed at different expression levels, we plot the fraction observed in chimpanzee. In red are exons which do not overlap conserved regions, and in black are those which do. The expression level of a transcribed region in a lane is defined as the number of reads mapping to the region divided by the total number of reads mapping to exons from the lane. We then averaged this fraction to get the mean expression level of the region. Across all expression levels, exons that overlap most conservered regions replicate at a higher rate than those that do not.

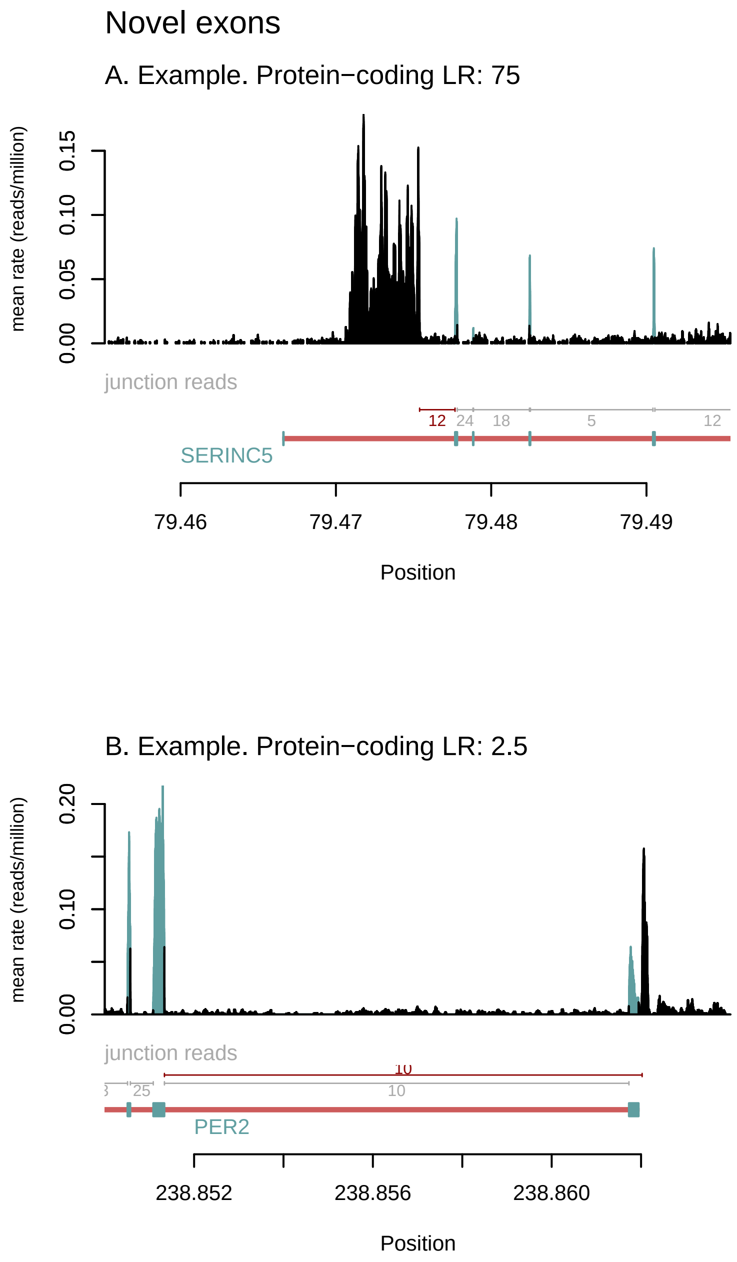

Figure 6: **Examples of novel exons with evidence of splicing to known genes**. See the legend to Figure 1 in the main text for full description of the panels. In the top panel, we show an alternative protein-coding last exon in *SERINC5*, and in the bottom panel an alternative (presumably non-coding) first exon in *PER2*. In both panels, the annotated genes are transcribed off the (-) strand. The likelihood ratios in the description refer to a test of the dN/dS ratio in the region defined by the black peak of expression.

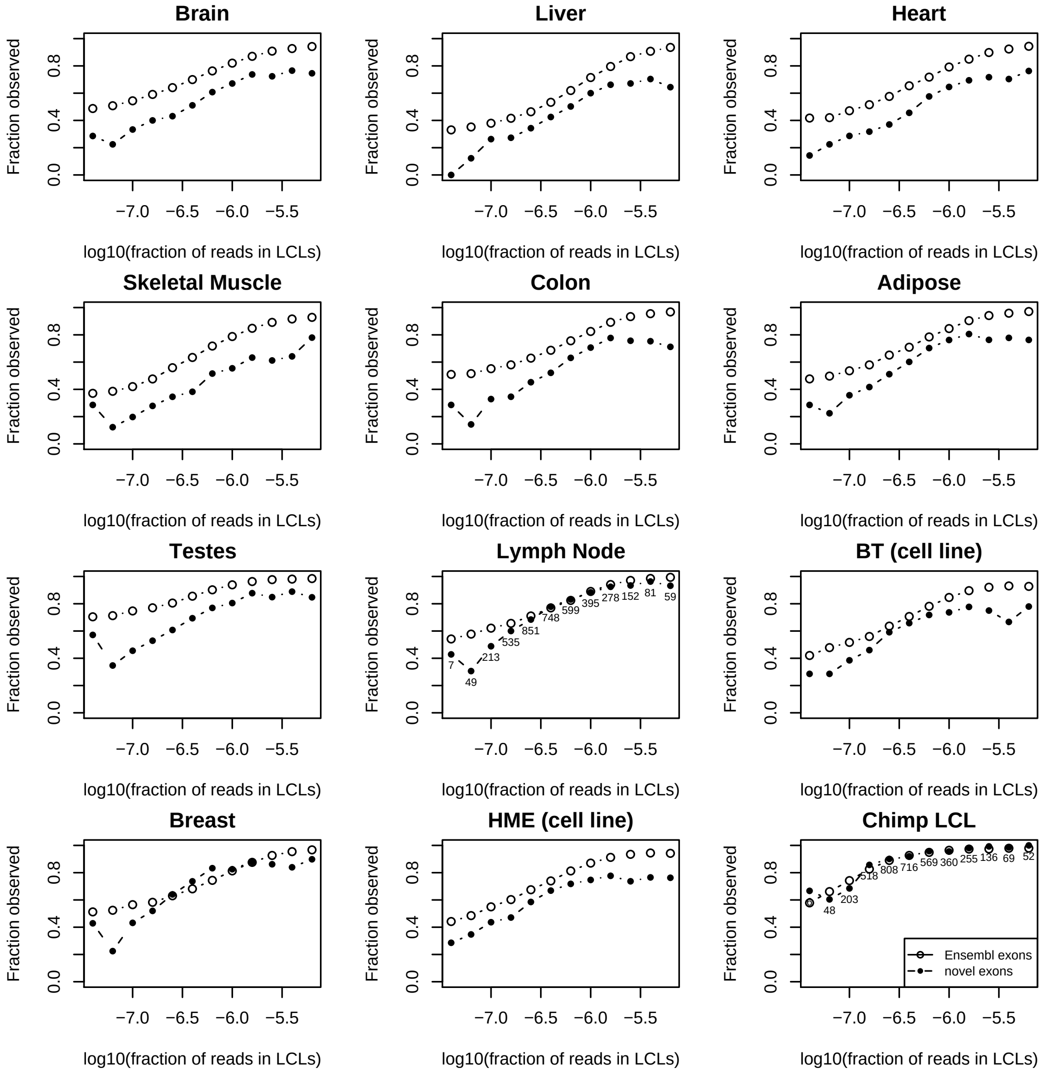

Figure 7: **Previously unannotated exons are somewhat more tissue-specific than annotated exons.**. We used the RNA-Seq dataset from Wang et al. (2008) to estimate the expression levels of all unannotated and Ensembl exons in a number of tissues. We divided the exons according to expression level in the human LCLs, and plot, for each expression level, the fraction of Ensembl or unannotated exons that are observed in each tissue. Above each plot is the cell type. The numbers are the number of unannotated exons in each data point (these numbers are the same within all human tissue. For Ensembl exons, there are thousands to tens of thousands of observations underlying each data point).

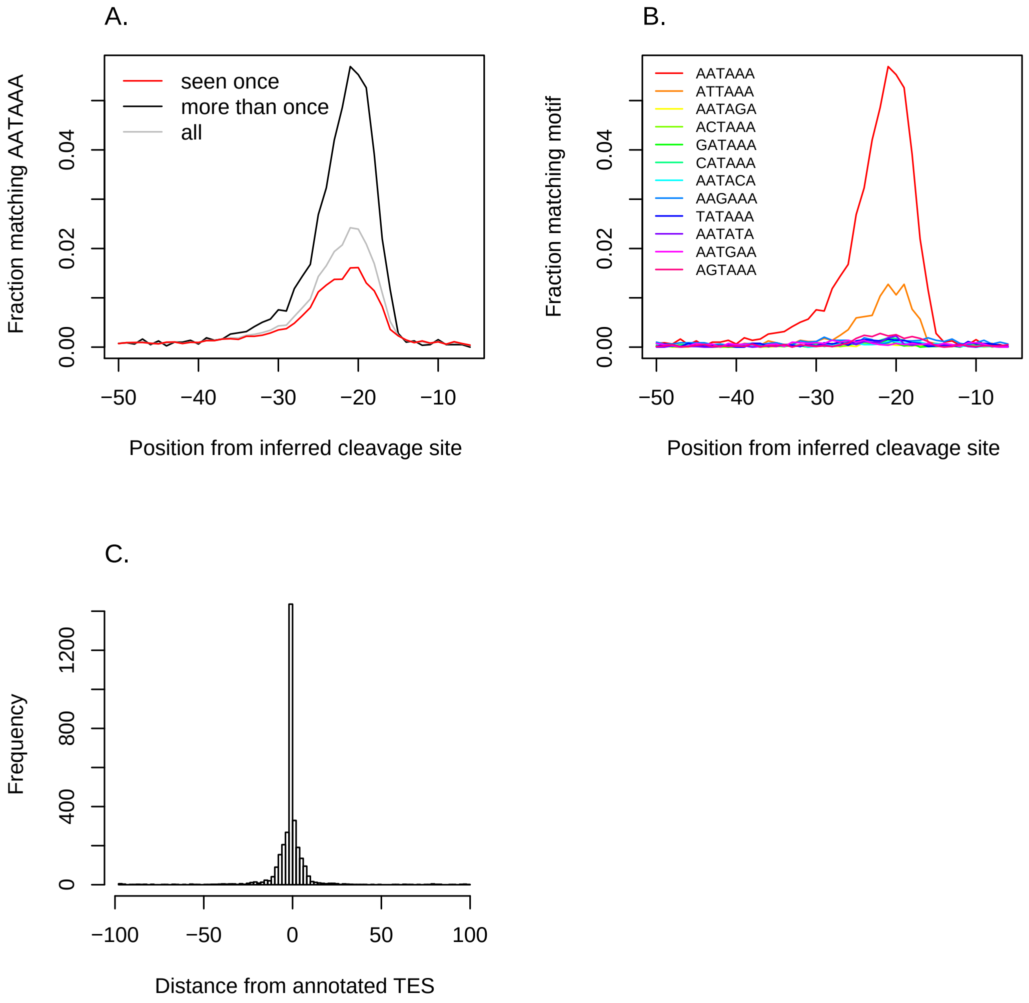

Figure 8: **Motif enrichment around predicted polyadenylation sites. A. Sites supported by more than a single sequencing read show a stronger enrichment of the consensus binding site.** We split putative polyadenylation sites into those supported by only one read in the data and those observed more than once. We extracted the 50 bases upstream of each site and plot, for each base, the fraction of sites which match AATAAA. **B. Single base variants of AATAAA show little enrichment upstream of predicted polyadenylation sites.** For each site supported by more than one sequencing read, we extracted the 50 bases upstream of each site. We plot the fraction of sites matching each hexamer. Only ATTAAA shows a significant enrichment downstream of our predicted sites. **C. Predicted sites cluster close to known cleavage sites.** For each site supported by more than one sequencing read and containing an AATAAA hexamer in the 15-30 bases upstream, we calculated the distance to the nearest known cleavage site in the Ensembl, UCSC, RefSeq and Vega databases. We plot a histogram of these distances (removing all distances greater than 100 bases). There is a clear peak near known cleavage sites.

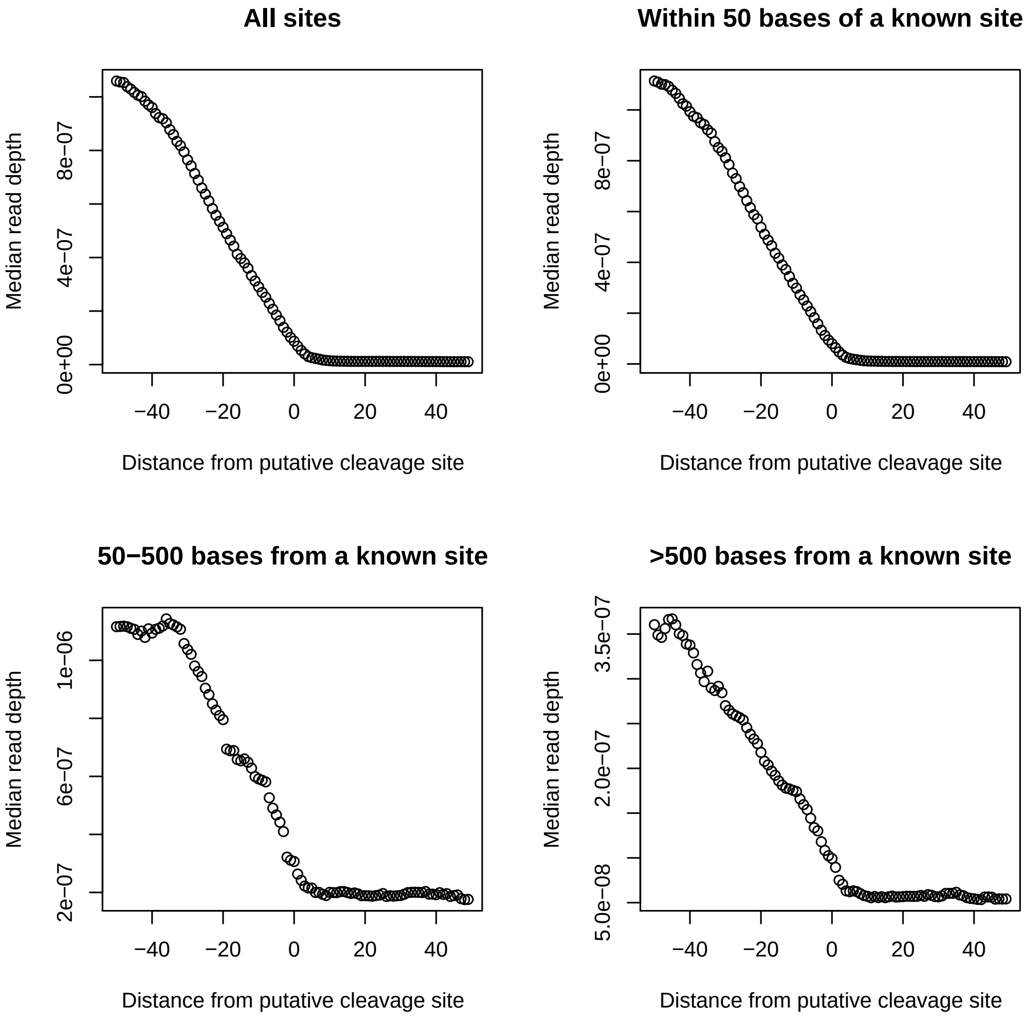

Figure 9: **Enrichment of transcription upstream of putative poly-A cleavage sites.** We divided putative cleavage sites according to their distances from known cleavage sites, and calculated the median RNA-Seq read depth in the 50 bases upstream and downstream of the site. The read depth of a base in a lane is defined as the fraction of all reads in the lane that cover the base. Plotted are these median read depths for sites at different distances from known sites. In all cases, the upstream bases show dramatically more evidence for transcription than the downstream bases, as is expected if the predicted cleavage sites indeed represent the ends of mRNAs.

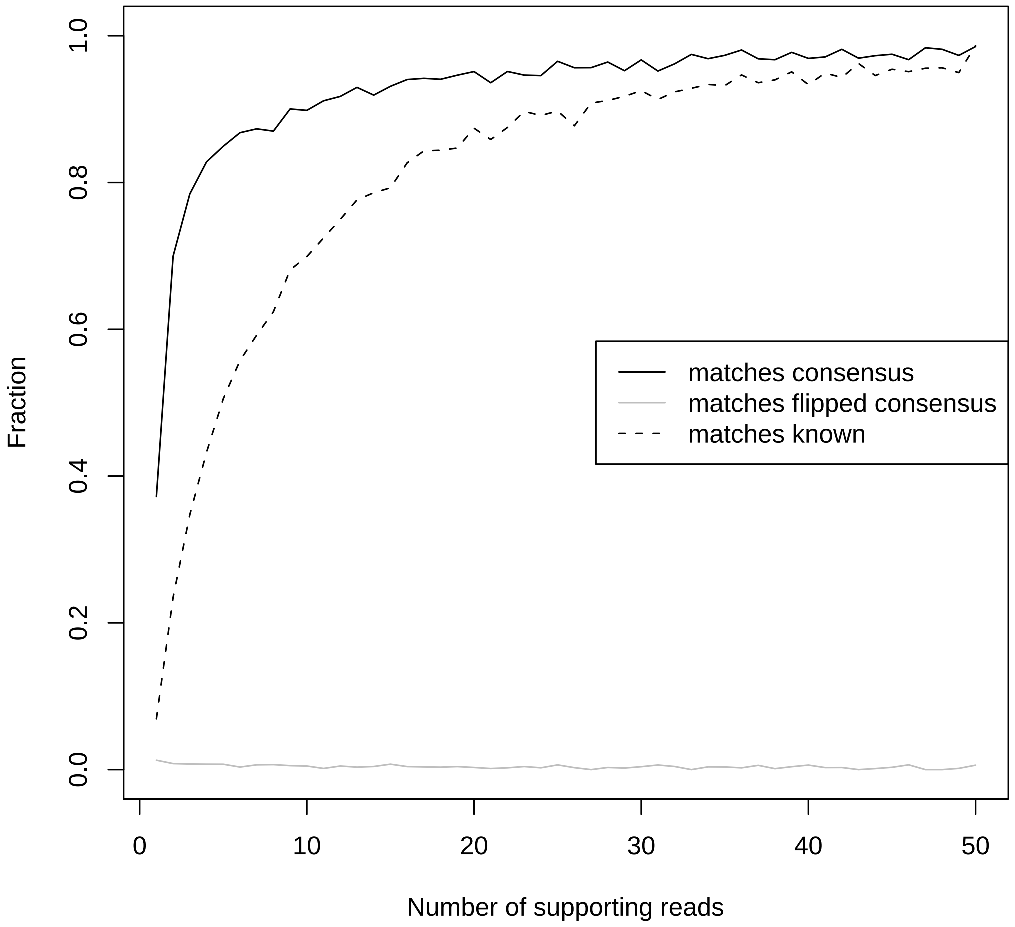

Figure 10: **Validation of exon-exon junctions identified *ab initio* with RNA-Seq.** We divided putative exon-exon junctions identified in the RNA-Seq data according to the number of reads supporting the junction. For each class, we calculated the fraction of junctions that overlap known junctions (from the Ensembl, Refseq, UCSC, and Vega annotations), as well as the fraction consistent with a GT-AG consensus splice site and the fraction consistent with a control "flipped" splice site of GT-TC. Plotted are these fractions. We conclude that the majority of these inferred exon-exon junctions are truly the result of splicing reactions.

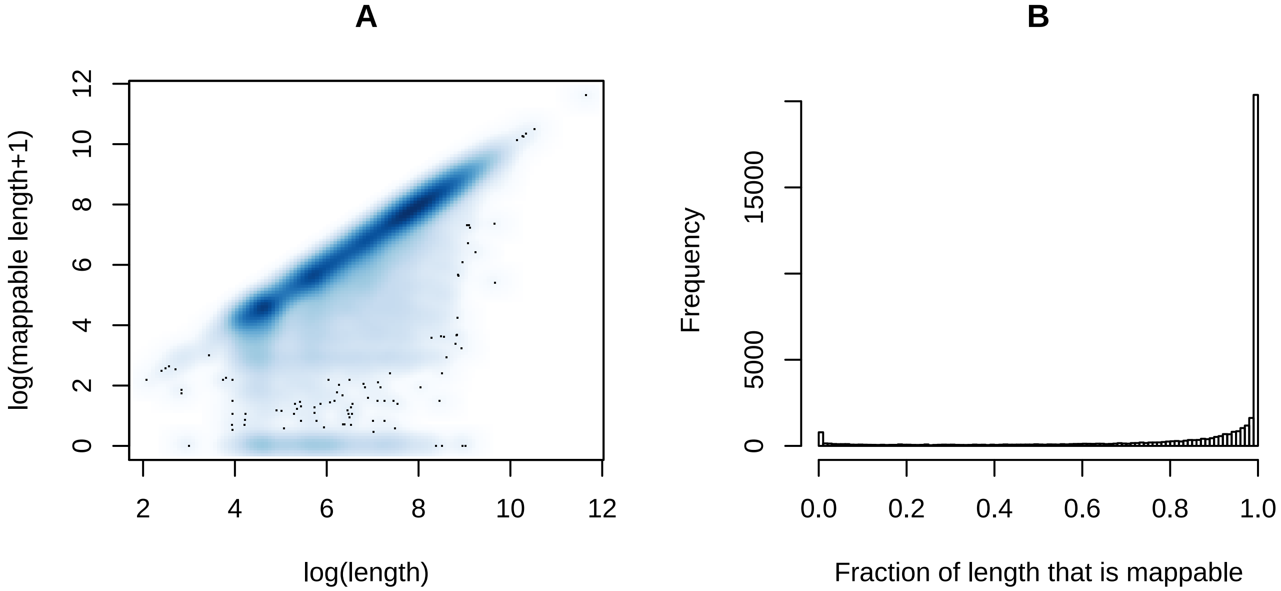

Figure 11: **"Mappability" of genes. A. Mappable versus true length**. For each gene, we calculated the "mappable" length of the gene by simulating 35bp sequence reads tiling the gene and determining what fraction are mapped back to the correct location. Plotted is the log of the true length of the gene (the sum of the lengths of all exons) versus the log of the "mappable" length. One base was added to the mappable length to avoid infinite values in the log. **B. Histogram of the fraction of a gene that is mappable**. For each gene, we took the mappable length of the gene and divided it by the true length. Plotted is the histogram of these fractions.

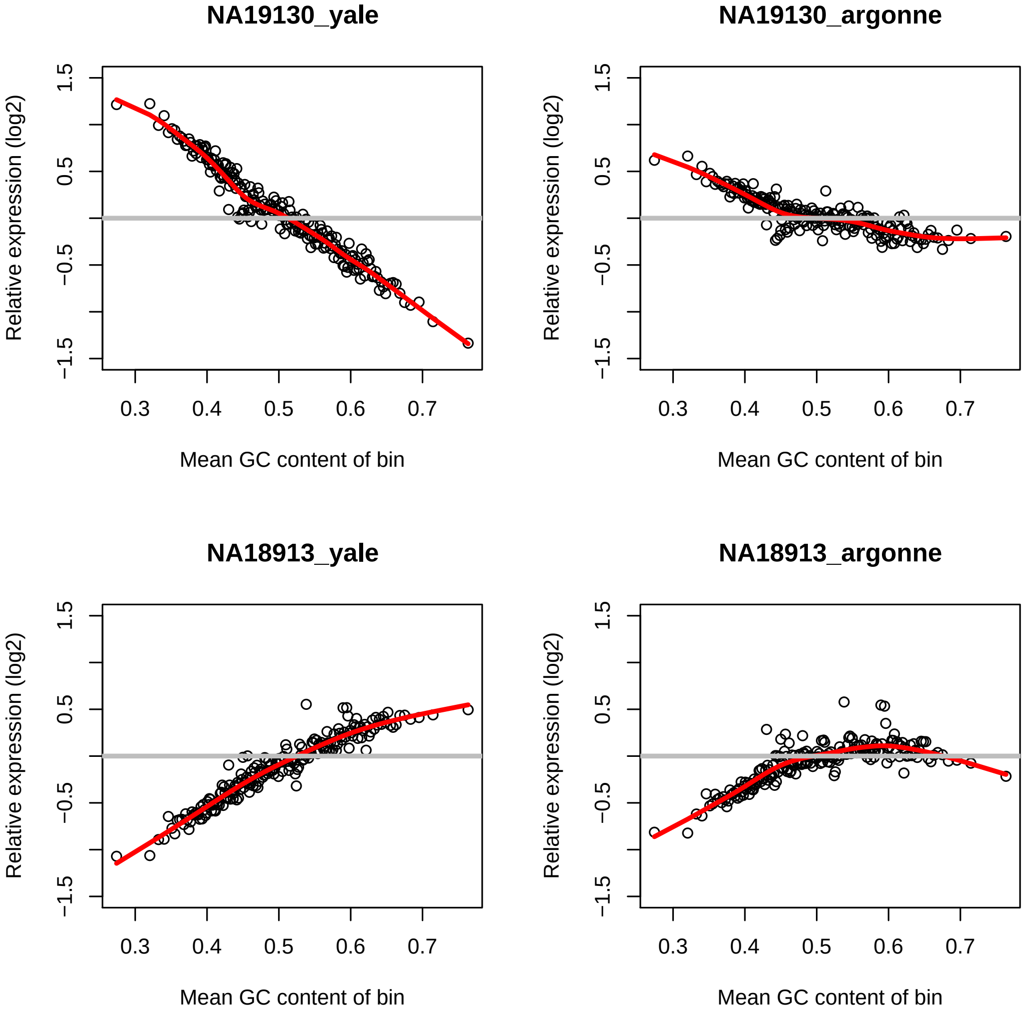

Figure 12**: GC content influences inference of expression levels from RNA-Seq.** We split all exons into 200 approximately equally-sized bins based on GC content. Then, for each lane, we counted the number of reads falling into exons in that bin, then divided this by the number of reads falling into that bin summing across all lanes. If there were no effect of GC content on sequencing depth, this fraction should be approximately equal across GC bins. Plotted for four lanes is the log2 of this ratio divided by the fraction expected if GC content played no role in sequencing. In each plot, the grey line is the expectation under the null. There is a strong effect of GC content in these lanes. The red line shows a spline fitted to the points; as described in the Supplementary Methods, the counts are adjusted based on this spline fit.

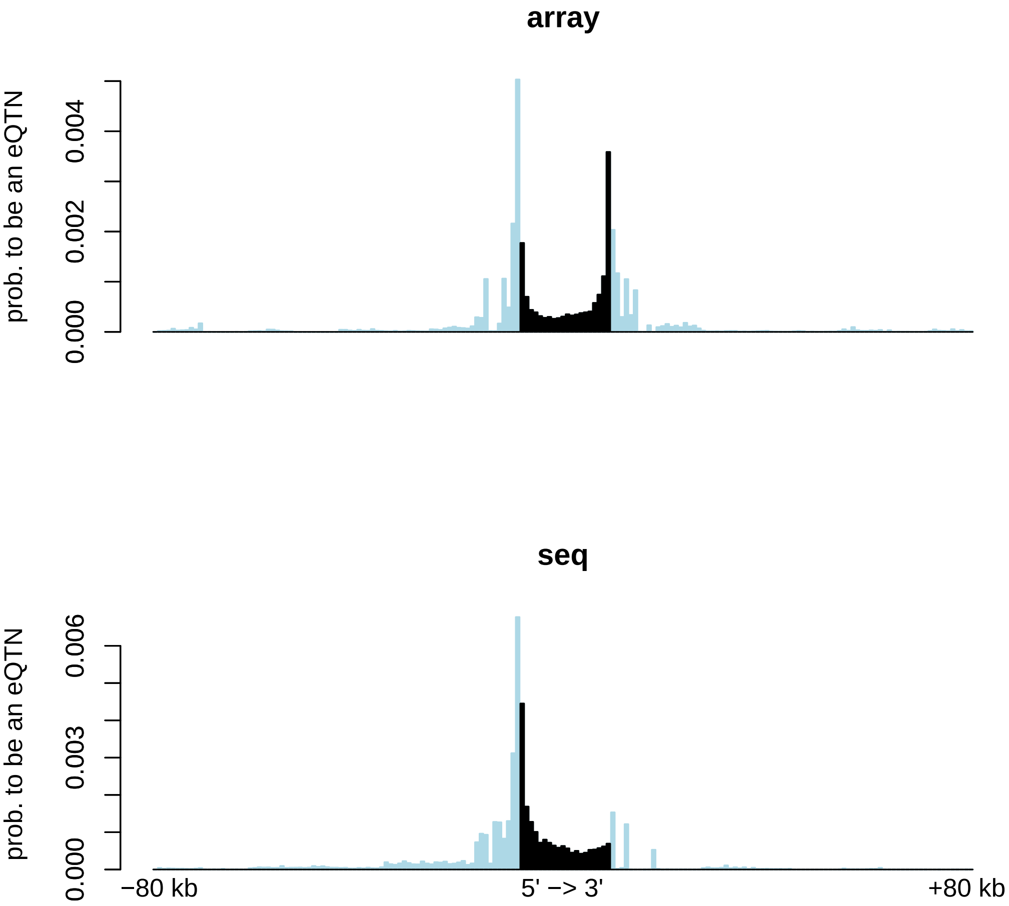

Figure 13**: Spatial distribution of eQTLs around genes.** We performed "local" eQTL studies using expression levels estimated from either RNA-Seq data or Illumina expression array data. We then used the hierarchical model of Veyrieras et al. (2008) to refine the spatial distribution of eQTLs around genes. We divided regions outside genes into bins of 1 kb, and genic regions into 20 bins, and estimated in each bin the probability that a SNP in that bin is a QTL. In black are genic bins, and in blue are non-genic bins. In the top panel is this distribution estimated from an eQTL study using expression levels estimated from Illumina microarray data, and the bottom panel is the distribution estimated from an eQTL study using expression levels estimated using RNA-Seq data. We believe that the higher 3' peak in the array data is due to sQTLs affecting 3' probes in that data set.

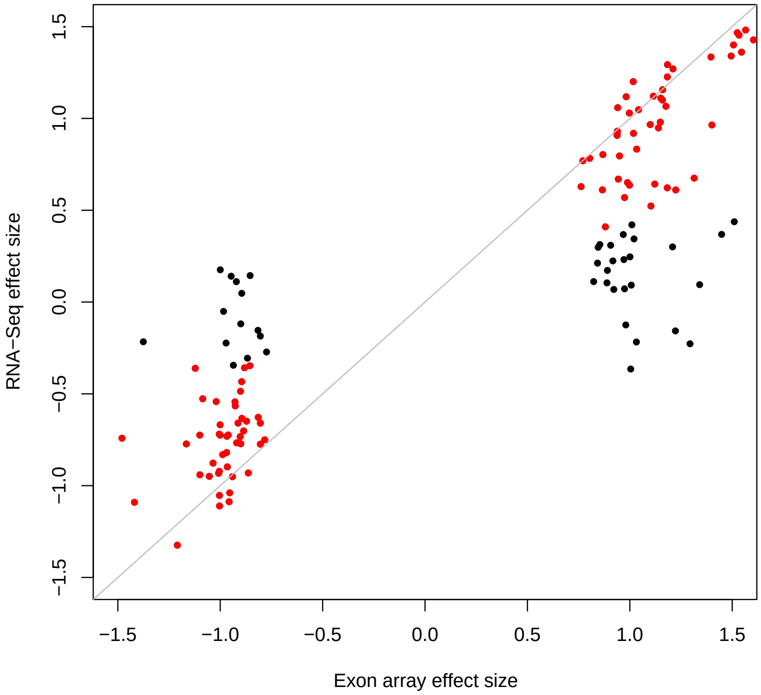

Figure 14: **Replication of eQTLs identified using exon arrays.** We performed eQTL studies using expression levels estimated using exon microarray and sequencing data from the same 53 individuals to identify eQTLs. For all eQTLs identified at an FDR of 10% in the eQTL study using arrays, we compared the effect sizes in the two technologies. In black are the effects for all QTLs identified in the study exon using arrays, and in red are those which also have $p < 0.05$ in the study using sequencing.

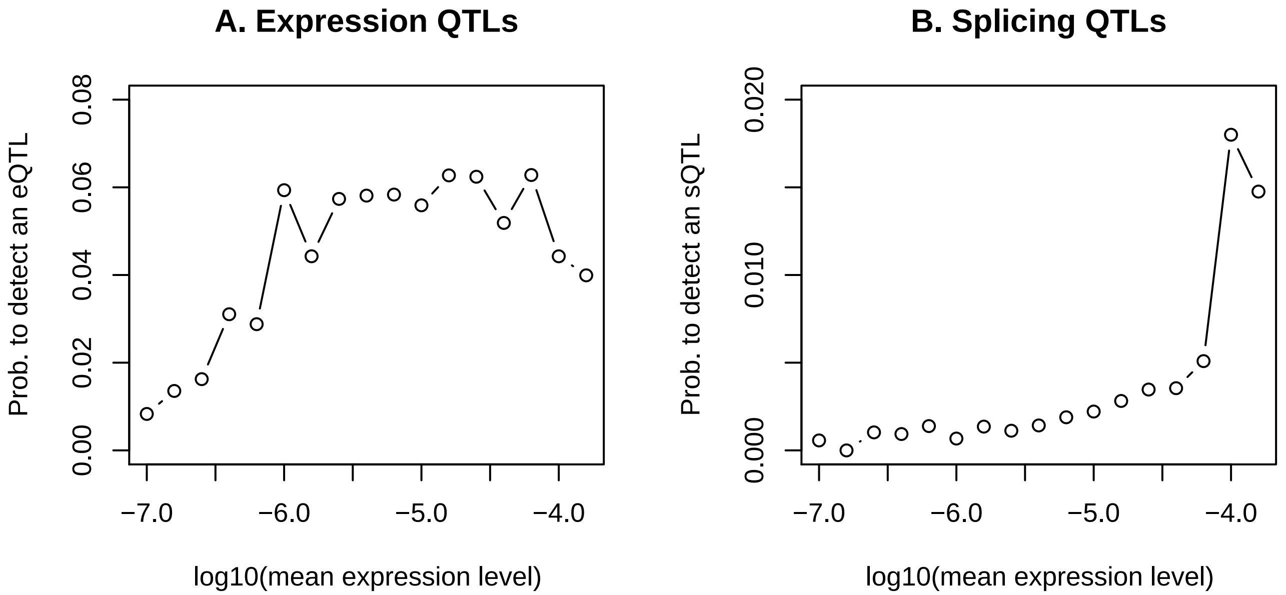

Figure 15: **The power to detect an eQTL varies as a function of expression level. A. Overall gene expression QTLs**. We split genes into bins based on mean expression level, and calculated the fraction of genes in each bin with an eQTL (at an FDR threshold of 10%). The probability to detect an eQTL approaches a plateau at an average rate of about 1 read/million. This corresponds to about 10 reads/individual, given the sequencing depth in our study (a minimum of two lanes per individual, and about 5 million exonic reads per lane). 65% of genes that show evidence of expression (ie. that have reads in more than half the individuals) are expressed above this level. **B. Splicing QTLs**. We split exons into bins based on mean expression level, and calculated the fraction of exons in each bin with an sQTL (at an FDR threshold of 10%). The probability to detect an sQTL peaks at an average rate of about 1 read/10,000. 0.5% of exons that show evidence of expression are expressed at this level.

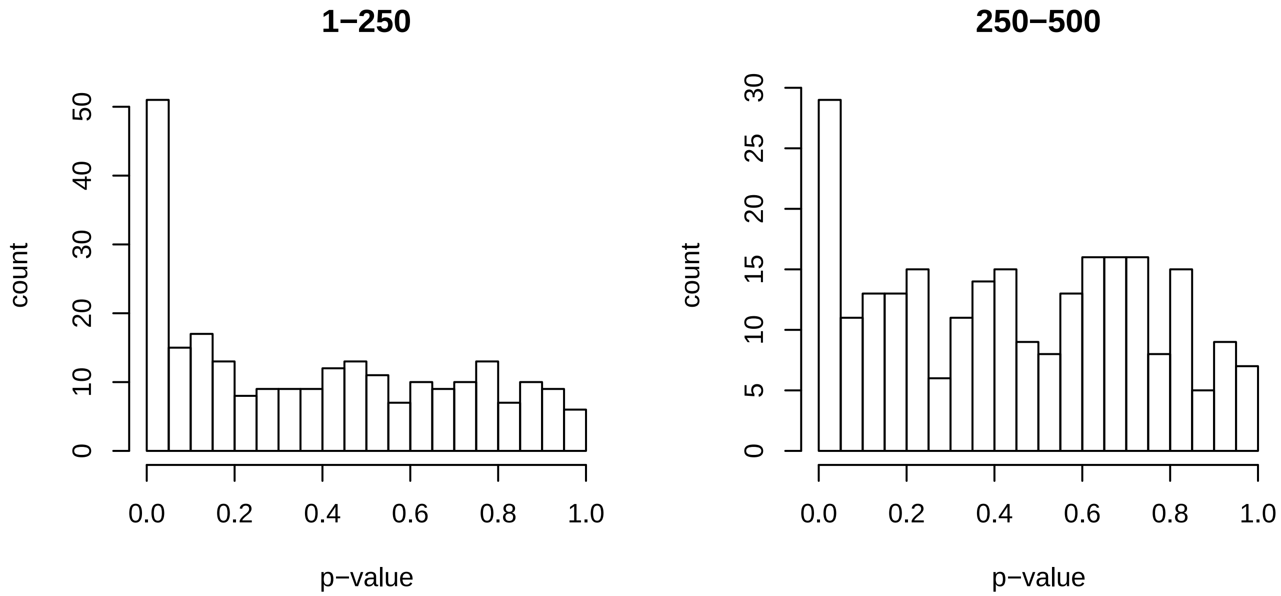

Figure 16**: Replication in the Nigerian population of eQTLs identified in a European population.** We obtained P-values from the top 500 SNP-gene pairs from an RNA-Seq eQTL experiment done using LCLs derived from a European population (Montgomery et al., 2009), and calculated P-values for the same associations in our data. Plotted are histograms of the P-values in our data for the top 250 and 250-500th ranked associations in the European population (after excluding the 40 SNPs which are not segregating in our population). There is a clear enrichment for low P-values, especially for the strongest associations identified in the European population.

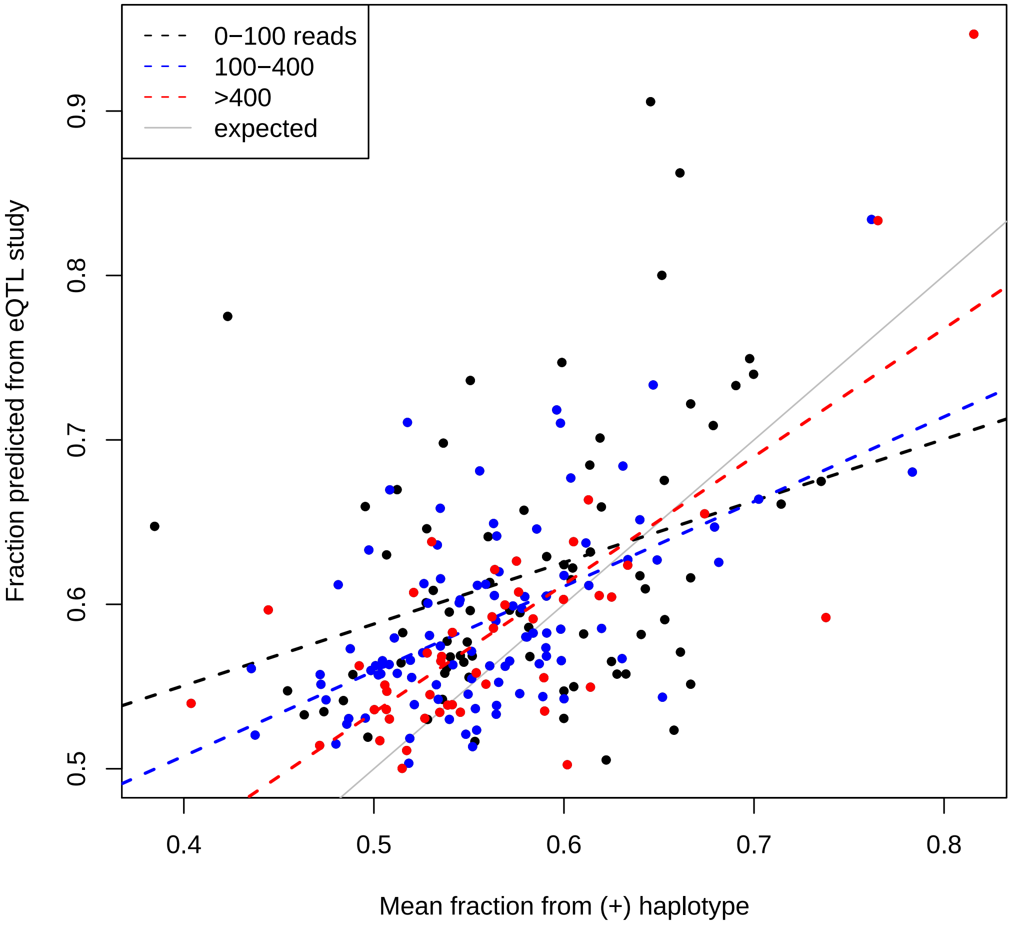

Figure 17: **The correlation in effect sizes between allele-specific expression and eQTL studies increases as a function of read depth.** This is a replotting of Figure 2C in the main text. Here, all points are labeled according to how many total reads could be assigned to either the high- or low-expression haplotype across all individuals. Genes with more reads (and thus more confidence in the effect sizes) show a stronger correlation between the two estimates.

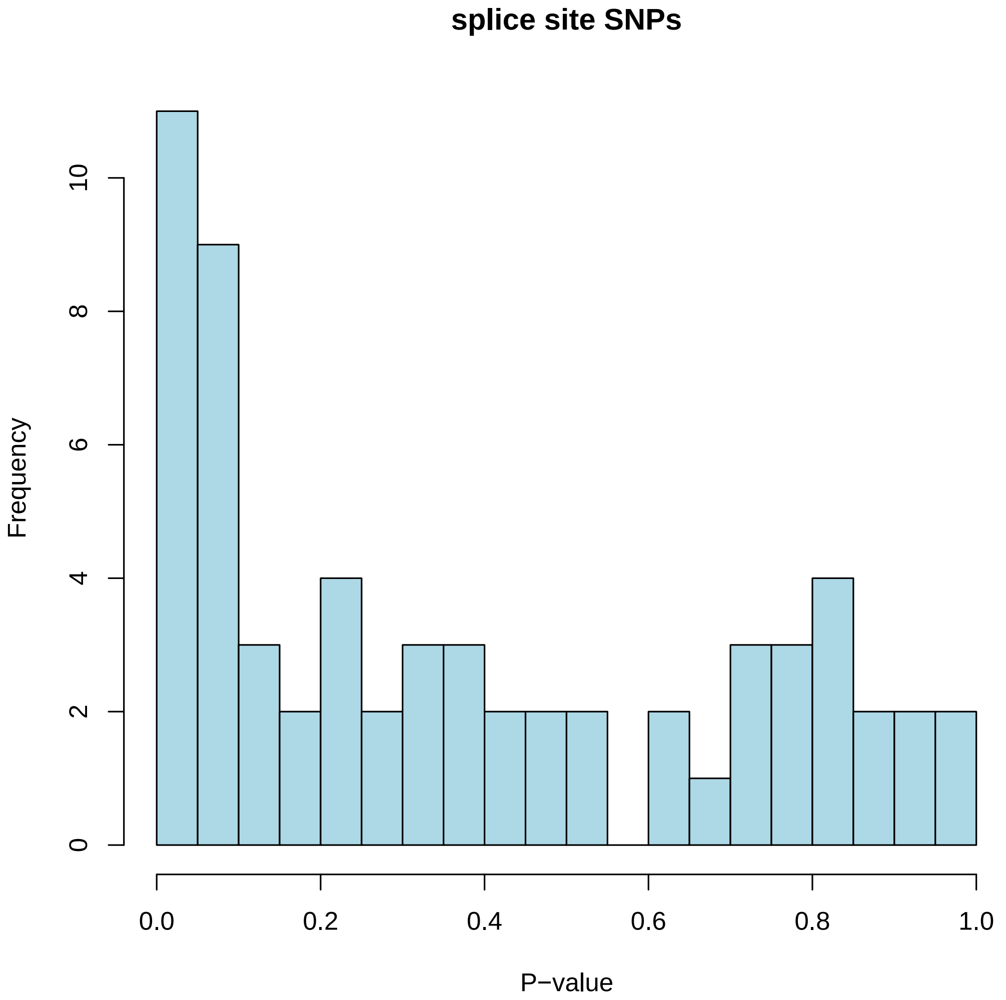

Figure 18**: SNPs in splice sites show a skew towards significant P-values**. We examined the P-values for SNP-exon pairs in the sQTL study where the SNP falls in the two bases of the consensus splice sites on either side of the exon. Plotted is the histogram of these P-values. There is a modest skew towards low values.

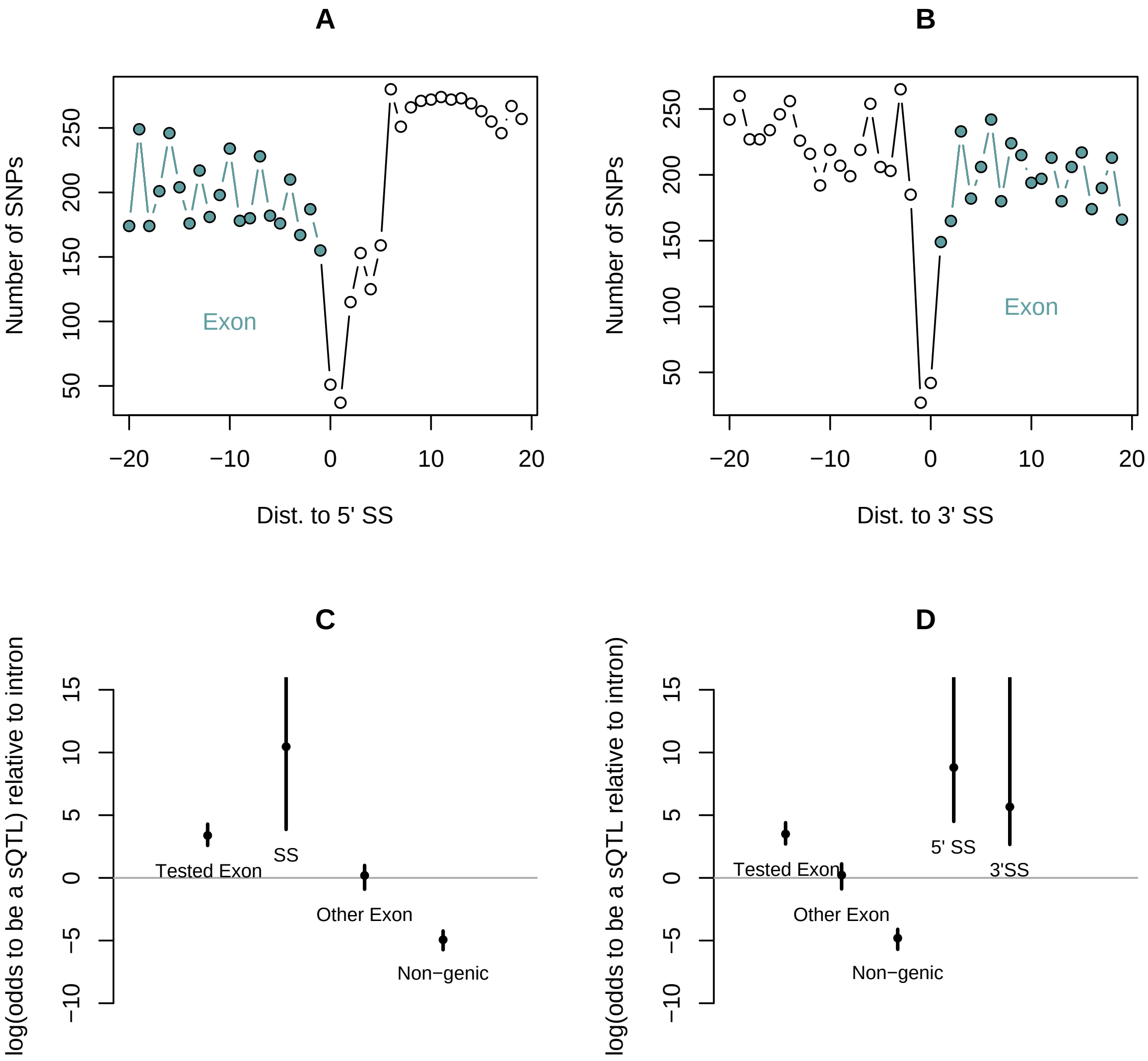

Figure 19**: SNPs in splice sites are rare but enriched for SNPs influencing exon inclusion. A, B. SNPs in splice sites are rare.** We counted the number of SNPs at different positions for the 5' and 3' splice sites. At both ends, there is a strong depletion of SNPs in the canonical two bases, and somewhat less depletion around them. **C. SNPs in the canonical splice sites are enriched among sQTLs**. We used the hierarchical model to estimate the log (base e) odds ratio for SNPs in different annotations. In this figure only the canonical splice sites of two base pairs are labeled as splice sites. **D. Equal enrichment of sQTLs at the 3' and 5' splice sites**. As in C, but here the splice sites are split into two annotations–one for the 3' splice site and one for the 5' splice site. In this figure, the 5' splice site is defined as all SNPs falling two bases exonic to six bases intronic of the 3' end of an exon, and the 3' splice site is defined as all SNP falling one base exonic to 20 bases intronic of the 5' end of an exon. In both C. and D., the 95% confidence intervals for the splice site annotations extend above 15, but have been truncated for visualization.

Table 1**: Numbers of QTLs at different FDR thresholds**. For each type of QTL (splicing QTL or expression QTL), we give the number of genes with evidence for a QTL at different gene-level FDR thresholds. Thresholds were determined by permutation (see Section 6.3)

	1% FDR	10% FDR	20% FDR
eQTLs	411	929	1353
sQTLs	111	187	237

Table 2**: Summary of data analysed**. For each lane (indexed in the first column), we give the sequencing center where the lane was sequenced, the cell line ID, the total number of reads, the number that map (uniquely or not) to the genome (excluding splice junctions), the number that failed to map to the genome (this number includes those that later mapped to splice junctions), the number that map uniquely (MAQ quality score > 10), the number of unique matches to exons, and the number that map to exon-exon junctions. Cell line IDs with a “_2” appended indicate that this was a second library prepared for this individual. All data will be available at http://eqtl.uchicago.edu.

Index	Center	Ind	Total	Mapped	Unmapped	Unique	Exons	Junctions
1	yale	GM19210	9956027	9121764	834263	7000757	6027329	523086
2	yale	GM19209	10301610	9427778	873832	7188718	6293180	568979
3	yale	GM19098	10280860	9414645	866215	7275600	6377479	611530
4	yale	GM19201	9408711	8639716	768995	6462860	5542133	493631
5	yale	GM19153	8474542	7740449	734093	5816765	5095746	471178
6	yale	GM19144	8327861	7599754	728107	5955437	5322540	501437
7	yale	GM18909	9130603	8221992	908611	6168587	5497691	485255
8	yale	GM19152	10247210	9365185	882025	7095188	6051552	570723
9	yale	GM18511	10547619	9554796	992823	7244277	6371359	583744
10	yale	GM19108	10627854	9644337	983517	7324437	6418129	580037
11	yale	GM18498_2	9739474	8796843	942631	6788485	6204043	584575
12	yale	GM18499	5782532	5265454	517078	3985735	3471763	326223
13	yale	GM18520	7695310	6990931	704379	5459155	4915610	477229
14	argonne	GM19238	7218325	6165771	1052554	4938010	4149331	589841
15	argonne	GM19239	7738999	6694790	1044209	5423972	4561020	667990
16	argonne	GM19098	8128155	7048150	1080005	5752368	5017657	738741
17	argonne	GM19144	8476906	7317427	1159479	6059447	5384173	791630
18	argonne	GM19201	8540954	7489161	1051793	5924925	5050654	695532
19	argonne	GM19210	8432368	7401488	1030880	5991812	5100610	673856
20	argonne	GM19153	7125012	6172680	952332	4910951	4275796	622273
21	argonne	GM18909	9798209	8451674	1346535	6668302	5929889	773970
22	argonne	GM19147	10054713	8595867	1458846	6741998	6073542	776037
23	argonne	GM19152	10618007	9275215	1342792	7400400	6309094	859193
24	argonne	GM18499	10614370	9225356	1389014	7367668	6439366	854033
25	argonne	GM19209	10706267	9398941	1307326	7532355	6572236	859191
26	argonne	GM19143	10115208	8862148	1253060	7119353	5876990	772722
27	yale	GM19257	7469692	6670565	799127	4863191	4305079	381559
28	yale	GM18861	4642527	4268333	374194	3228806	2707867	276474
29	yale	GM19131	8414765	7705984	708781	5905823	5007738	475434
30	yale	GM19192	8552871	7667518	885353	5807836	5079639	455853
31	yale	GM18916	7559177	6877582	681595	5203985	4421816	406726
32	yale	GM19222	4567036	4207450	359586	3165854	2577827	244420
33	yale	GM19225	8249884	7222871	1027013	5166613	4501171	361846
34	yale	GM18913	9006101	7985185	1020916	6094254	5394769	519078
35	yale	GM18853	9717230	8992384	724846	6706468	4517073	303699
36	yale	GM18862	8619671	7825936	793735	5961325	4769737	392085

Index	Center	Ind	Total	Mapped	Unmapped	Unique	Exons	Junctions
37	yale	GM19147	7578019	6789755	788264	5012080	4528700	388164
38	yale	GM19143	8381930	7652587	729343	5819895	4852784	446158
39	yale	GM19190	6729781	6068707	661074	4438392	3959957	356918
40	yale	GM18501	6422000	5858443	563557	4583877	3964512	406574
41	yale	GM18856	5267997	4713897	554100	3592544	2980711	283438
42	yale	GM18912	4211405	3922569	288836	2787784	2228699	173436
43	yale	GM19102	4025513	3761590	263923	2626390	2097345	151618
44	yale	GM19119	3816178	3523702	292476	2554531	1709033	137370
45	yale	GM19171	6154300	5550470	603830	4255000	3689054	367607
46	yale	GM19200	6213671	5664154	549517	4236623	3669605	358187
47	yale	GM18517	9367811	8537621	830190	6691939	5953135	567691
48	yale	GM19128	9498521	8879533	618988	6385615	5287926	370872
49	yale	GM19130	8914041	8240501	673540	6156110	5201677	473464
50	yale	GM19238	8406706	7711358	695348	5738323	4815662	439082
51	yale	GM19239	9085407	8330518	754889	6315716	5291029	499192
52	yale	GM18504	7113128	6395759	717369	4900926	4276159	393638
53	yale	GM18516	7818182	7187650	630532	5458447	4590102	398880
54	yale	GM18522	7691810	7001604	690206	5399961	4777886	456879
55	yale	GM19093	6997219	6365558	631661	4961534	4346717	412851
56	yale	GM19172	7086910	6413209	673701	4837168	4177112	377074
57	yale	GM19140	8970177	8090500	879677	5932405	5140808	453660
58	yale	GM18508	8701858	7892549	809309	5918297	5112250	483474
59	yale	GM18519	8838591	8005087	833504	5939852	5133797	473576
60	yale	GM19127	9353324	8553933	799391	6247433	4888659	362927
61	yale	GM18505	8849477	8056290	793187	6038849	4970405	442502
62	yale	GM19138	9041280	8045028	996252	6231520	5456418	558903
63	yale	GM18502	8909231	7974623	934608	5904514	5106499	482832
64	yale	GM19114	8969212	7962131	1007081	5567206	5028496	380492
65	yale	GM18507	10329429	9249468	1079961	6949451	5946587	562540
66	yale	GM18504_2	8729513	7809568	919945	5911828	5163744	467580
67	yale	GM19193	9370200	8492872	877328	5935598	4931809	340311
68	yale	GM18516_2	7602729	6893859	708870	5231672	4491184	417772
69	argonne	GM18511	8968585	7789041	1179544	6215977	5447017	732044
70	argonne	GM18520	9176125	8051008	1125117	6534641	5853051	802342
71	argonne	GM18498_2	8699228	7626730	1072498	6045089	5505361	712491
72	argonne	GM19131	9200019	8059266	1140753	6583416	5568512	774192
73	argonne	GM19108	9033582	7897772	1135810	6297058	5496814	720205
74	argonne	GM19190	8965588	7816579	1149009	6040516	5335611	712061
75	argonne	GM18861	8705183	7515118	1190065	6175818	5313901	804192
76	argonne	GM19257	8287438	6845762	1441676	5406030	4824945	632270
77	argonne	GM19192	7650417	6414542	1235875	5196193	4582571	594097
78	argonne	GM19222	8249437	6961569	1287868	5618603	4770907	668587
79	argonne	GM18916	8167882	6871544	1296338	5571326	4802559	650169

Index	Center	Ind	Total	Mapped	Unmapped	Unique	Exons	Junctions
80	argonne	GM18862	9436367	8216981	1219386	6671875	5307009	645247
81	argonne	GM18853	9648196	8430593	1217603	6801848	4533072	440460
82	argonne	GM18913	8291212	7019537	1271675	5591618	4892969	704865
83	yale	GM19203	9630700	8730129	900571	6300704	4988789	380447
84	yale	GM19101	9320429	8556624	763805	6453666	5420576	495348
85	yale	GM19116	9693662	8917973	775689	6783335	5618018	504349
86	yale	GM19099	9955015	8996444	958571	6965741	5939733	581438
87	yale	GM18517_2	10007597	9053289	954308	7003362	5955312	566479
88	yale	GM18498	9971094	8978277	992817	6789530	5980146	550784
89	yale	GM19160	9662020	8736354	925666	6722735	5525217	489294
90	argonne	GM19140	8067824	6865408	1202416	5370156	4644358	611174
91	argonne	GM19127	8570023	7466185	1103838	5834141	4563887	503161
92	argonne	GM18505	7132811	6191462	941349	4874997	3972916	527485
93	argonne	GM18502	8344912	6936263	1408649	5412538	4657657	650309
94	argonne	GM18508	8454404	7119598	1334806	5666294	4889027	679966
95	argonne	GM19138	8690202	6793511	1896691	5541884	4797358	731950
96	argonne	GM18519	8634607	7347164	1287443	5760231	4944469	675359
97	argonne	GM18858	7917503	6771204	1146299	5353299	4624623	630949
98	argonne	GM19225	8448726	6816830	1631896	5217382	4594078	527568
99	argonne	GM18504_2	8197125	6985762	1211363	5728894	4953045	627623
100	argonne	GM19160	9208450	7846485	1361965	6259641	5198198	634017
101	argonne	GM18507	8330364	6983286	1347078	5507691	4681165	620232
102	argonne	GM19114	8677053	7343705	1333348	5555657	5005472	537090
103	argonne	GM18516_2	9088080	7857707	1230373	6395267	5363250	728211
104	yale	GM19204	6353169	5663615	689554	4410901	3962139	434402
105	yale	GM18510	6248122	5600379	647743	4244746	3703474	367575
106	yale	GM18871	6261611	5529754	731857	4206834	3673798	385869
107	yale	GM18486	5930518	5238224	692294	4165337	3680065	361735
108	yale	GM18870	6833330	6103176	730154	4727155	4053282	392882
109	yale	GM19159	6382899	5717368	665531	4510193	4113029	401651
110	yale	GM19160_2	6515037	5811434	703603	4447812	3686468	328529
111	argonne	GM19101	8870597	7496430	1374167	6113831	5163101	726940
112	argonne	GM19128	9988733	8978485	1010248	6993734	5806133	613057
113	argonne	GM19130	10349515	9037537	1311978	7227185	6177664	830191
114	argonne	GM19203	9338220	8067073	1271147	6219082	4931719	555461
115	argonne	GM18517	10116924	8802465	1314459	7315397	6473398	888056
116	argonne	GM18517_2	9127309	7831347	1295962	6441933	5449522	766332
117	argonne	GM19099	9474797	8078694	1396103	6663680	5665072	829477
118	argonne	GM18504	5363063	4303002	1060061	3049747	2641834	319251
119	argonne	GM18516	5528980	3877649	1651331	2363163	1965419	215001
120	argonne	GM18522	9185184	7730490	1454694	6217851	5475846	760097

Index	Center	Ind	Total	Mapped	Unmapped	Unique	Exons	Junctions
121	argonne	GM19093	11714315	10091916	1622399	8258579	7115109	1022337
122	argonne	GM19172	11662307	10048372	1613935	8001354	6846782	923465
123	argonne	GM18498	11630601	9955243	1675358	8032567	7043748	972880
124	argonne	GM19116	10528137	9220686	1307451	7551936	6234101	851491
125	argonne	GM18871	7201594	6107752	1093842	4904199	4256022	624062
126	argonne	GM19204	8512123	7315680	1196443	6012341	5339940	831963
127	argonne	GM19159	6511480	5588092	923388	4628648	4181760	572495
128	argonne	GM18486	7749699	6589147	1160552	5475058	4783469	665098
129	argonne	GM19171_2	7677962	6468687	1209275	5228082	4559507	667316
130	argonne	GM18912_2	6848584	5895181	953403	4500683	3594833	405842
131	argonne	GM18523	5919358	4829663	1089695	3900792	3331278	460872
132	argonne	GM19193	6340598	5277427	1063171	4061047	3376781	335365
133	argonne	GM19160_2	5152078	4235522	916556	3427215	2847121	355728
134	argonne	GM18852	5649346	4781203	868143	3869576	3366738	463310
135	argonne	GM18855	6928515	5463263	1465252	4443657	3935800	538637
136	argonne	GM18870	6042345	5097747	944598	4172144	3565614	489162
137	argonne	GM18510	6373677	5420620	953057	4362898	3799046	528971
138	yale	GM19200_2	9279557	8258122	1021435	6345161	5610031	553069
139	yale	GM18912_2	9081942	8214308	867634	5921002	4841708	389631
140	yale	GM18501_2	8097072	6987268	1109804	5558866	4817720	517161
141	yale	GM18856_2	8946747	8005738	941009	6283089	5469719	539203
142	yale	GM18505_2	7677809	6995732	682077	5436042	4857629	494048
143	yale	GM18502_2	8277356	7375622	901734	5719473	5110116	531274
144	yale	GM18855	7930075	6972243	957832	5383556	4783877	456648
145	yale	GM18852	9606138	8599424	1006714	6654315	5814495	571014
146	yale	GM18858	8800059	7903672	896387	6056067	5315096	509672
147	yale	GM19171_2	6077346	5336917	740429	4130026	3661495	361784
148	yale	GM19137	6574947	5862646	712301	4555920	4036064	402163
149	yale	GM18523	4528349	3867749	660600	2909873	2523694	242956
150	argonne	GM19137	8012428	6718005	1294423	5510072	4849310	658141
151	argonne	GM19200_2	7227700	5937544	1290156	4737945	4194635	531457
152	argonne	GM18502_2	6988357	5959134	1029223	4862630	4317666	648810
153	argonne	GM18856_2	6087872	5208619	879253	4317440	3743774	528054
154	argonne	GM18505_2	7950616	6945039	1005577	5669009	5023360	738492
155	argonne	GM18501	3545671	3105596	440075	2439412	2035888	244615
156	argonne	GM18856	1256211	1076337	179874	799362	635045	65973
157	argonne	GM18912	3996364	3659112	337252	2631561	2071471	187452
158	argonne	GM19102	3989379	3647559	341820	2587256	2014135	175871
159	argonne	GM19119	2326743	2092136	234607	1483912	953518	80277
160	argonne	GM19171	2021638	1752899	268739	1299517	1096290	116862
161	argonne	GM19200	2091947	1877598	214349	1346952	1129660	117858

References

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M., and Burdick, J. T., 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**(7063):1365–9.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K., 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**(24):3207–12.

Fairbrother, W. G., Yeh, R.-F., Sharp, P. A., and Burge, C. B., 2002. Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**(5583):1007–13.

Fraser, H. B. and Xie, X., 2009. Common polymorphic transcript variation in human disease. *Genome Res*, **19**(4):567–75.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., *et al.*, 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164):851–861.

Gough, B., 2003. *GNU Scientific Library Reference Manual - 2nd Edition*. Network Theory Ltd.

Guan, Y. and Stephens, M., 2008. Practical issues in imputation-based association mapping. *PLoS Genet*, **4**(12):e1000279.

Heap, G. A., Yang, J. H. M., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., Franke, L., Dubois, P. C., Mein, C. A., Dobson, R. J., *et al.*, 2009. Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. *Hum Mol Genet*, .

Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W. N., Welsh-Bohmer, K. A., Hulette, C. M., Denny, T. N., and Goldstein, D. B., *et al.*, 2008. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol*, **6**(12):e1.

Huang, R. S., Duan, S., Bleibel, W. K., Kistner, E. O., Zhang, W., Clark, T. A., Chen, T. X., Schweitzer, A. C., Blume, J. E., Cox, N. J., *et al.*, 2007. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci U S A*, **104**(23):9758–63.

Hull, J., Campino, S., Rowlands, K., Chan, M.-S., Copley, R. R., Taylor, M. S., Rockett, K., Elvidge, G., Keating, B., Knight, J., *et al.*, 2007. Identification of common genetic variation that modulates alternative splicing. *PLoS Genet*, **3**(6):e99.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2):249–64.

Kang, H. M., Ye, C., and Eskin, E., 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**(4):1909–25.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., *et al.*, 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A*, **106**(28):11667–72.

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T. J., Sladek, R., and Majewski, J., 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet*, **40**(2):225–31.

Lee, J.-H., Park, I.-H., Gao, Y., Li, J. B., Li, Z., Daley, G. Q., Zhang, K., and Church, G. M., 2009. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet*, **5**(11):e1000718.

Leek, J. T. and Storey, J. D., 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**(9):1724–35.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14):1754–60.

Li, H., Ruan, J., and Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, **18**(11):1851–8.

Lin, M. F., Deoras, A. N., Rasmussen, M. D., and Kellis, M., 2008. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol*, **4**(4):e1000067.

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., *et al.*, 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, **17**(12):1797–808.

Montgomery, S., Sammeth, M., Lach, R., C.Ingle, Nisbett, J., Guigo, R., and Dermitzakis., E., 2009. Second generation transcriptome sequencing in a population of European descent. *submitted*, .

Scheet, P. and Stephens, M., 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, **78**(4):629–44.

Servin, B. and Stephens, M., 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, **3**(7):e114.

Siepel, A., Diekhans, M., Brejová, B., Langton, L., Stevens, M., Comstock, C. L. G., Davis, C., Ewing, B., Oommen, S., Lau, C., *et al.*, 2007. Targeted discovery of novel human exons by comparative genomics. *Genome Res*, **17**(12):1763–73.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., *et al.*, 2007. Population genomics of human gene expression. *Nat Genet*, **39**(10):1217–1224.

Tian, B., Hu, J., Zhang, H., and Lutz, C. S., 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*, **33**(1):201–12.

Trapnell, C., Pachter, L., and Salzberg, S. L., 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9):1105–11.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K., 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, **4**(10):e1000214.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221):470–6.

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M., and Losick, R., 2008. *Molecular Biology of the Gene, Sixth Edition*. Benjamin Cummings, 6 edition.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**(8):1586–91.

Yassour, M., Kaplan, T., Fraser, H. B., Levin, J. Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., *et al.*, 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A*, **106**(9):3264–9.

Zhang, K., Li, J. B., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J.-H., Aach, J., Leproust, E. M., *et al.*, 2009a. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods*, **6**(8):613–8.

Zhang, W., Duan, S., Bleibel, W. K., Wisel, S. A., Huang, R. S., Wu, X., He, L., Clark, T. A., Chen, T. X., Schweitzer, A. C., *et al.*, 2009b. Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum Genet*, **125**(1):81–93.