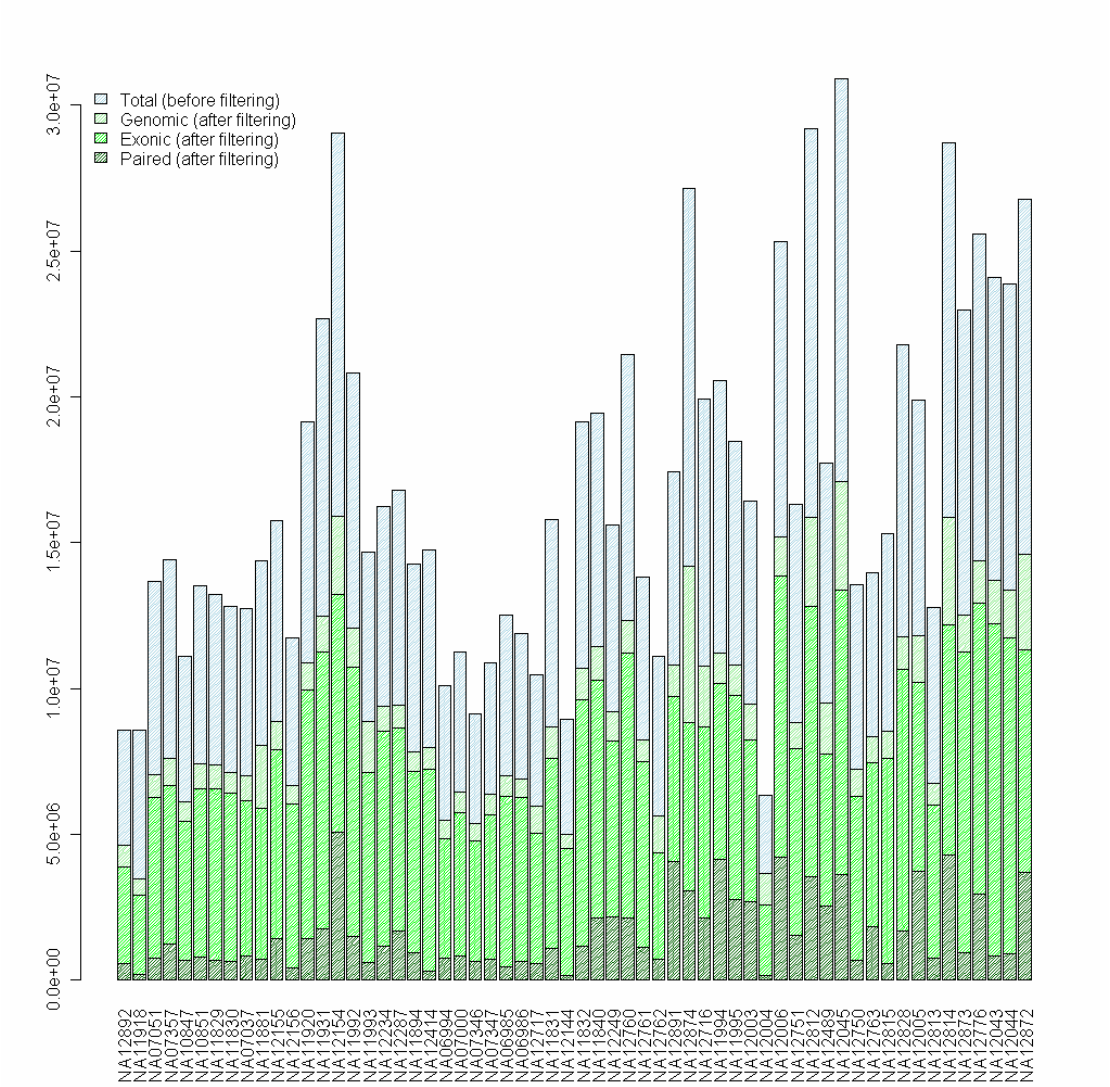
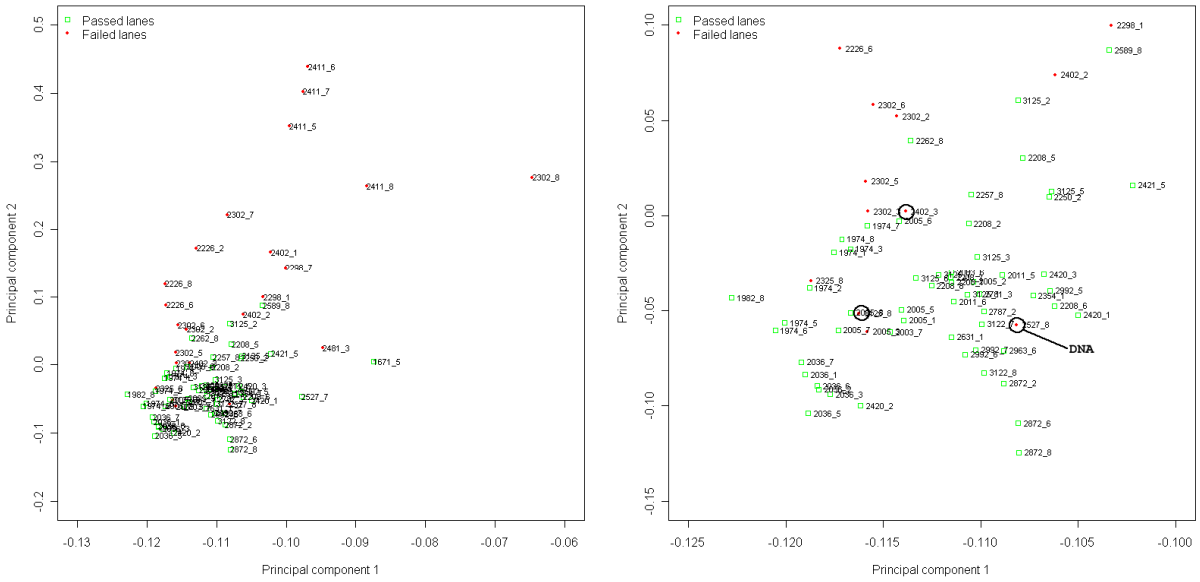
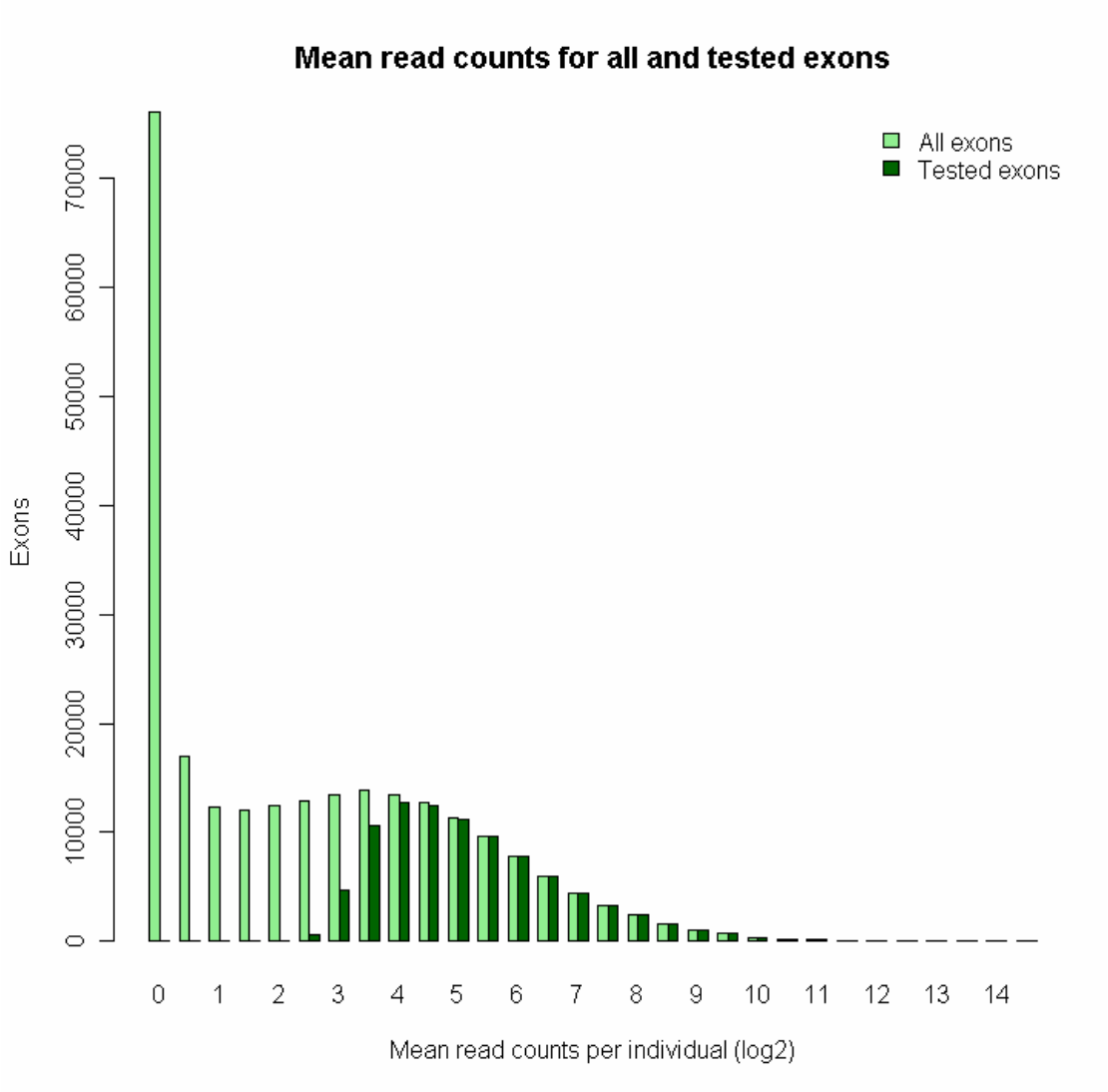


Supplemental Figure 1: Total reads mapped by sample.
The total number of mapped reads across each individual is plotted as "Total". The "Genomic" fraction is the number of reads after mapping quality filter of 10, location and paired orientation filtering. The "Exonic" fraction is the number of reads mapping to known exons and where both paired-end reads map to the same annotated transcript. The "Paired" fraction is the proportion of paired end reads that map to different exons for the same transcript requiring that the physical distance between exons is 200bp away. The proportion of reads that were excluded by location from "Total" to "Genomic" that were unassembled or mitochondrial DNA is ~5-7%.

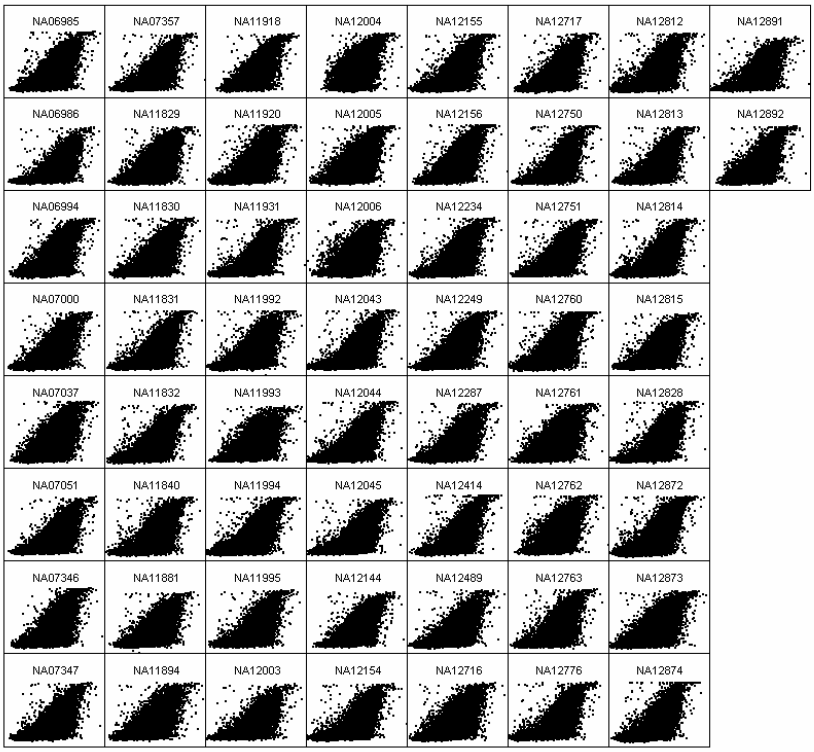


Supplemental Figure 2: Principal component analysis of lane quality and composition metrics

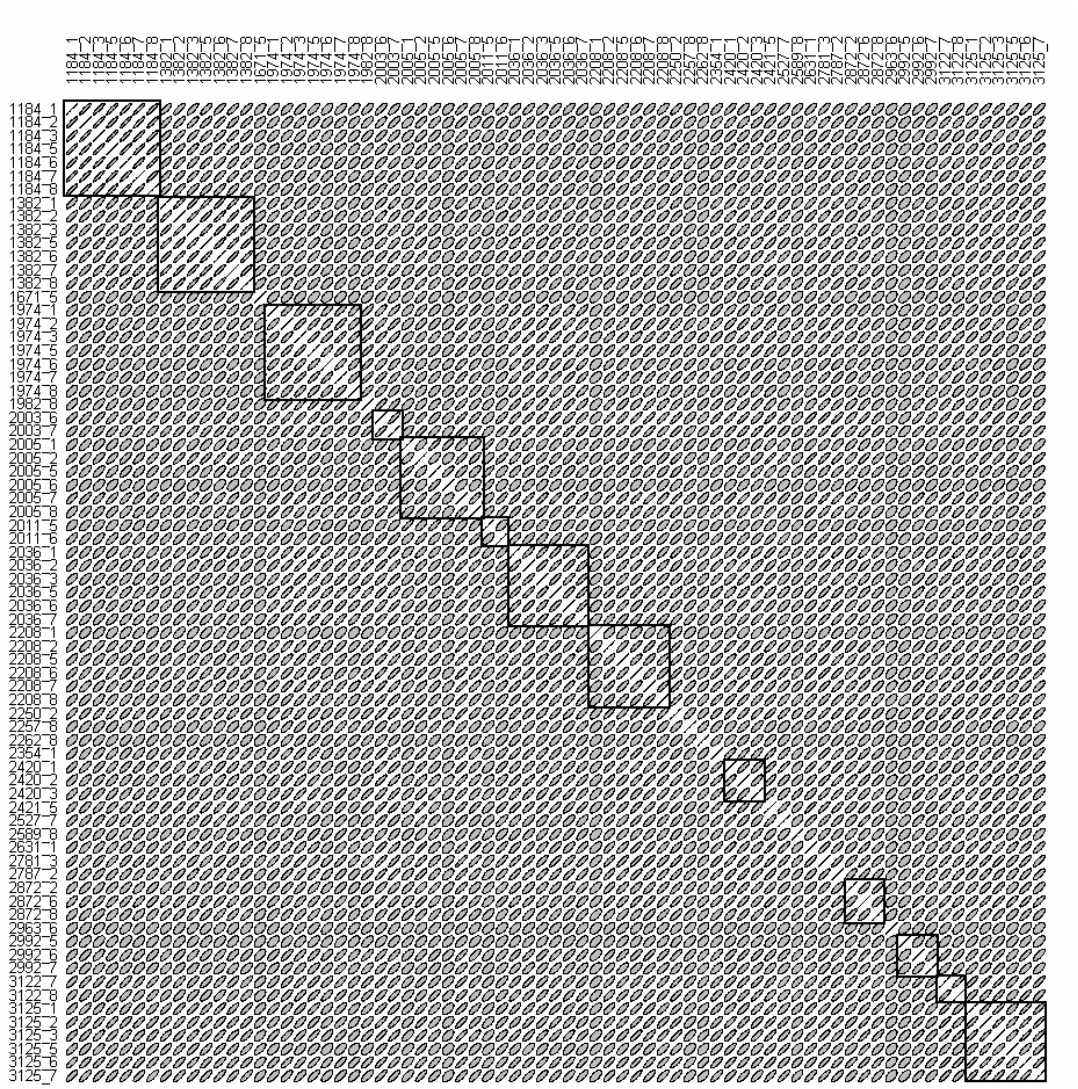
During sequencing several samples were rerun when lanes had poor quality, low yield or poor fraction of mappable reads to either the genome or the known exome. We tracked and tested various properties normalizing for sequencing depth including mean quality per base (for the whole fastq file and at both the top 10% and bottom 10% of the file), the median quality value per base, the standard deviation of the quality value per base, the distribution of quality values, the mean quality of dinucleotides, the mean quality value by read position, the dinucleotide percentage, the number of uncalled bases, the mean number of uncalled bases by position in read, the mean GC content (at the top 10% and bottom 10% of the fastq file) and the standard deviation and distribution of GC content. Incorporating these metrics, we generated PCA plots to help us diagnose lanes. As can be seen in the plot the passing lanes (green) cluster tightly, while the failed lanes (red) are spread out. PC1 and PC2 are predominantly weighted by base quality metrics (PC1 describing 99% of the variance). Right panel is zoomed-in version of the left panel. Lanes marked as failed within the cluster are due to either DNA being erroneously introduced into the sequencing (3 cases highlighted), or, for 2325_8 were repeated with a better yielding lane, 2005_3 had low exonic mapping, or 2302_3, 2302_5 and 2302_6 were redone because of low yield and poor mapping percentage. Despite a higher proportion of uncalled bases, lane 1671_5 (NA11918) was included in the analysis because it had good yield (8.5 million reads) and 88% of them mapped with the expected proportion mapping to the known exome.



Supplemental Figure 3: Mean read counts by exon (all 60 individuals).
The mean read counts for all the exons and the association tested fraction from the 60 individuals is plotted. In total there were 254,955 exons analyzed of which 90,064 exons for 10,777 genes were tested (having no more than 10% missing data). The *X*-axis is the log2 number of counts where 1 has been added to prevent taking the log of 0. We begin to see quantification of the tested fraction around 5 read counts.

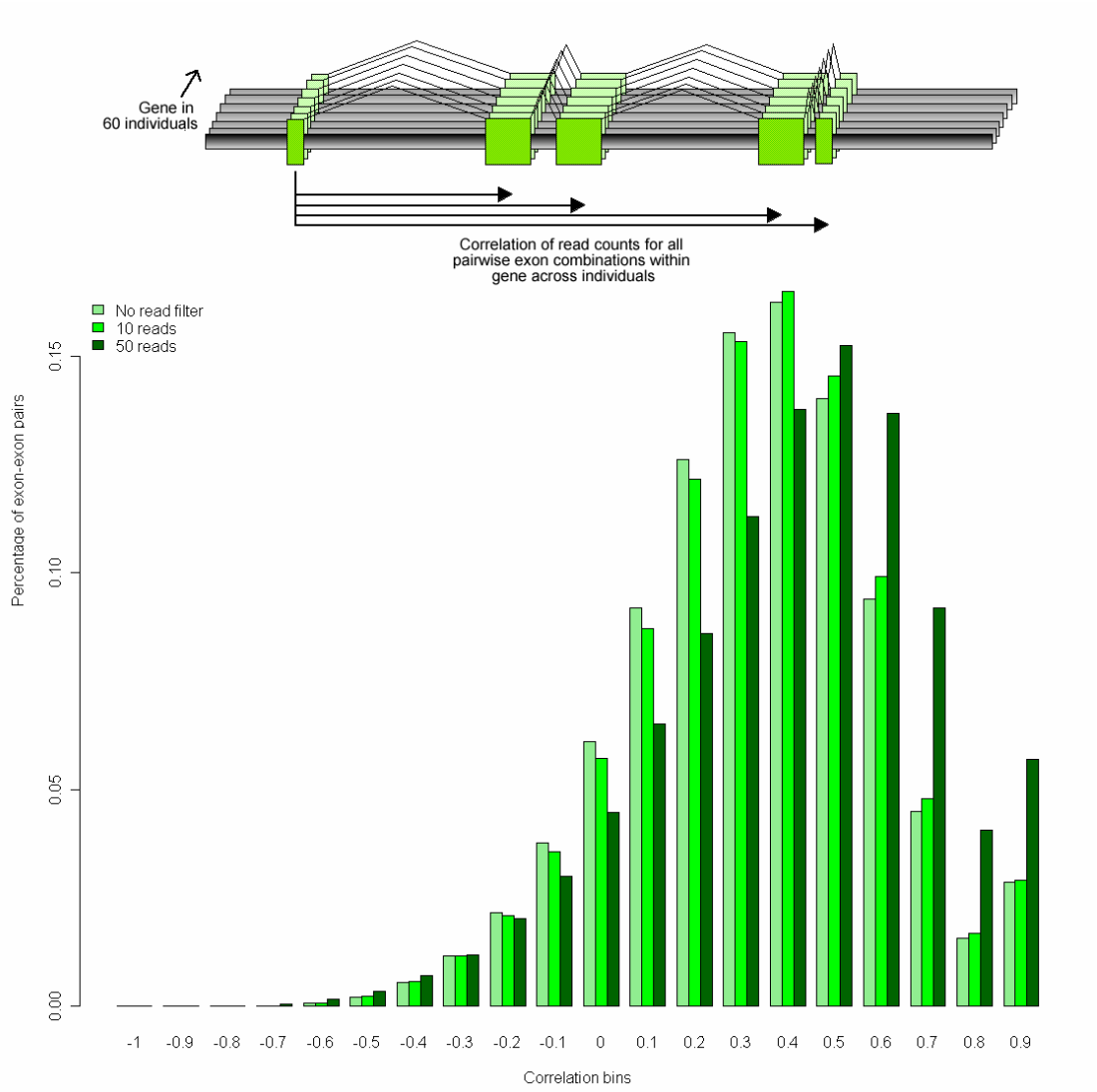


Supplemental Figure 4: Comparison of gene RPKM (Reads Per kb per Million Reads) quantification (X-axis) versus array intensities (Y-axis).
For each individual, each gene's read count assessed via sequencing is transformed to RPKM values and compared to array quantification. In total, 16,892 genes were compared for each individual to the mean values obtained for array-based probes for that respective gene. The mean spearman correlation coefficient between the samples was 0.80 +/- 0.018 (mean +/- SD) (min: 0.72, max: 0.83)



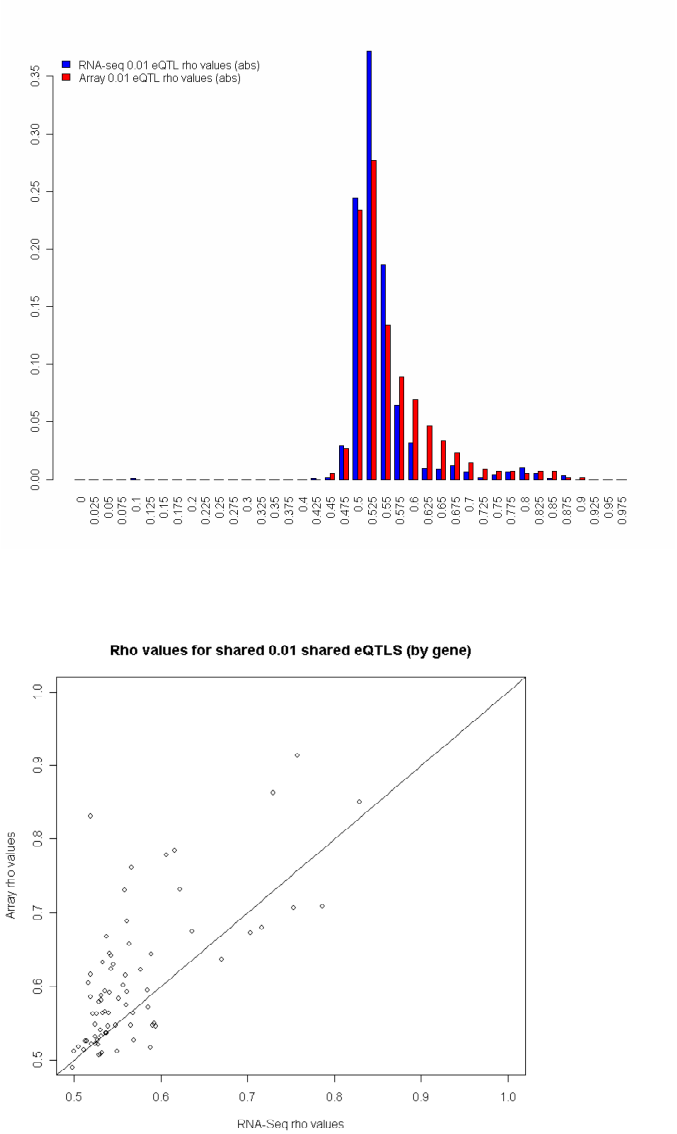
Supplemental Figure 5: Correlation matrix of all sequencing lanes for tested exon data.

Pairwise correlation of exon read count data for the tested 90,064 exons is plotted using the R ellipse package *plotcorr* correlation plot. Here diagonal lines represent perfect correlation and full circles represent no correlation. All shared runs are highlighted by a bounding box. The first such box in the top left corner is a Yoruban individual (NA19238) that was sequenced in 7 lanes. The next such box is for a CEPH individual (NA12892) that was also sequenced in 7 lanes. When comparing identical samples which were sequenced on the same run we obtained an inter-lane spearman correlation of 0.919 ± 0.006 (this was 0.913 ± 0.003 for the 7 lanes of NA12892 and 0.925 ± 0.003 for the 7 lanes of NA19238). When different samples were sequenced in the same run the spearman correlation was 0.829 ± 0.086, or in a different run: 0.754 ± 0.090. We also looked at mean correlation between CEU individuals and a CEU individual sequenced in 7 lanes and between CEU individuals and a YRI individual sequenced in 7 lanes and found correlations of 0.750 ± 0.070 and 0.730 ± 0.063 (tested). This reduced correlation between CEU+YRI individuals compared to CEU+CEU individuals was significant (p=8.44e-06; t-test) and reinforces what we would expect due to population differentiation of gene expression. Lanes are identified by a 4 digit run number followed by a lane number.

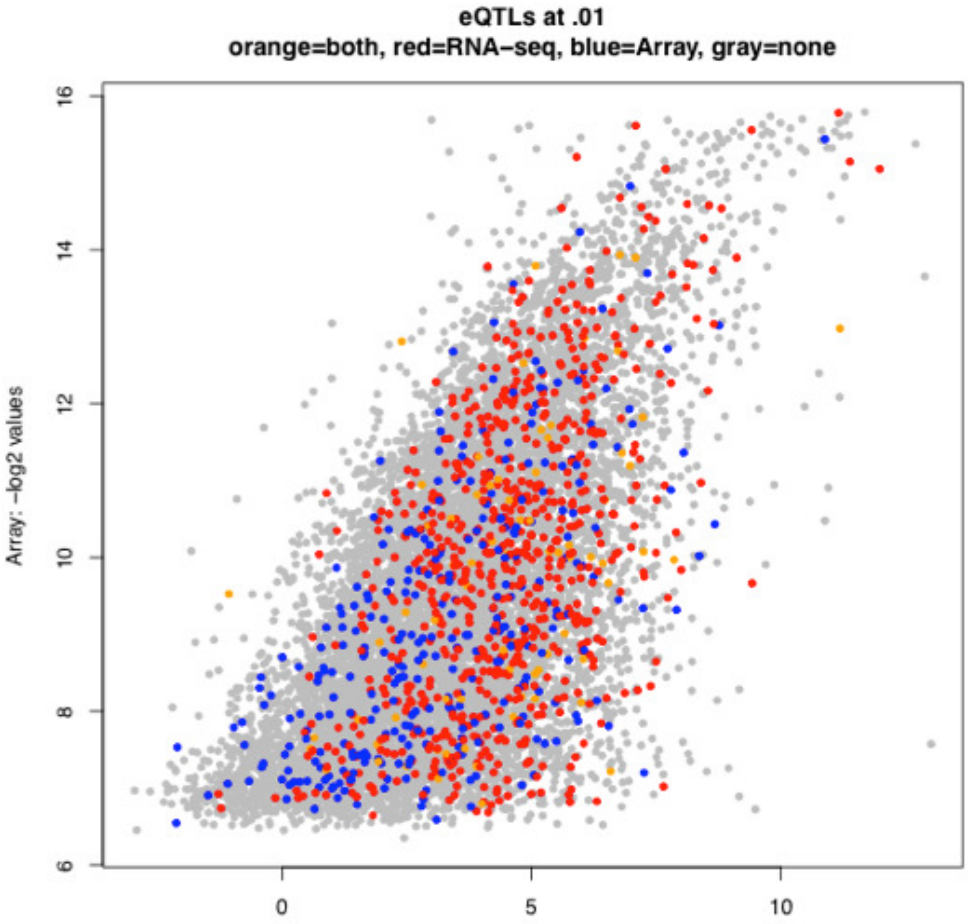


Supplemental Figure 6: Average pairwise exon-exon correlation within gene across individuals.

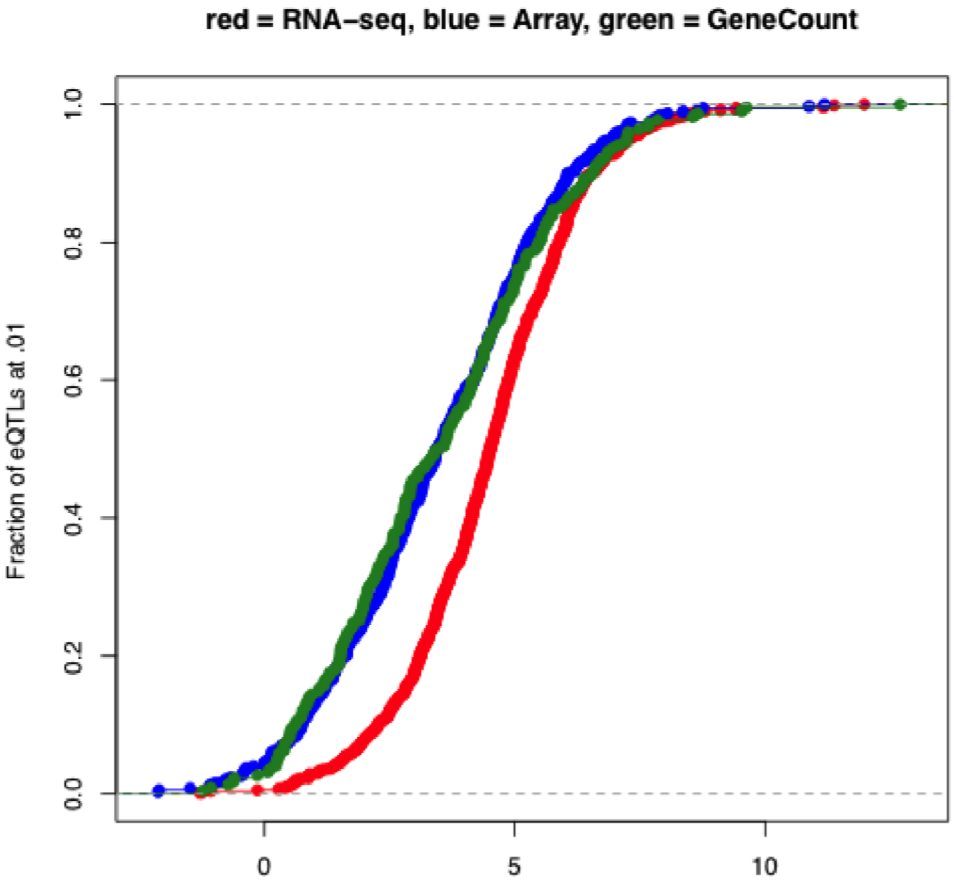
The distribution of R^2 correlation statistic for reads for exon-exon comparisons filtered for those exons with minimum average reads across individuals of 0, 10 and 50. This demonstrates that as quantification increases the ability for one exon to inform the values in another also increases. This highlights that single exon quantification does not go far enough to recover the correlation structure of transcripts within a gene but suggests that some intermediate quantification will inform quantification in the majority of transcripts.



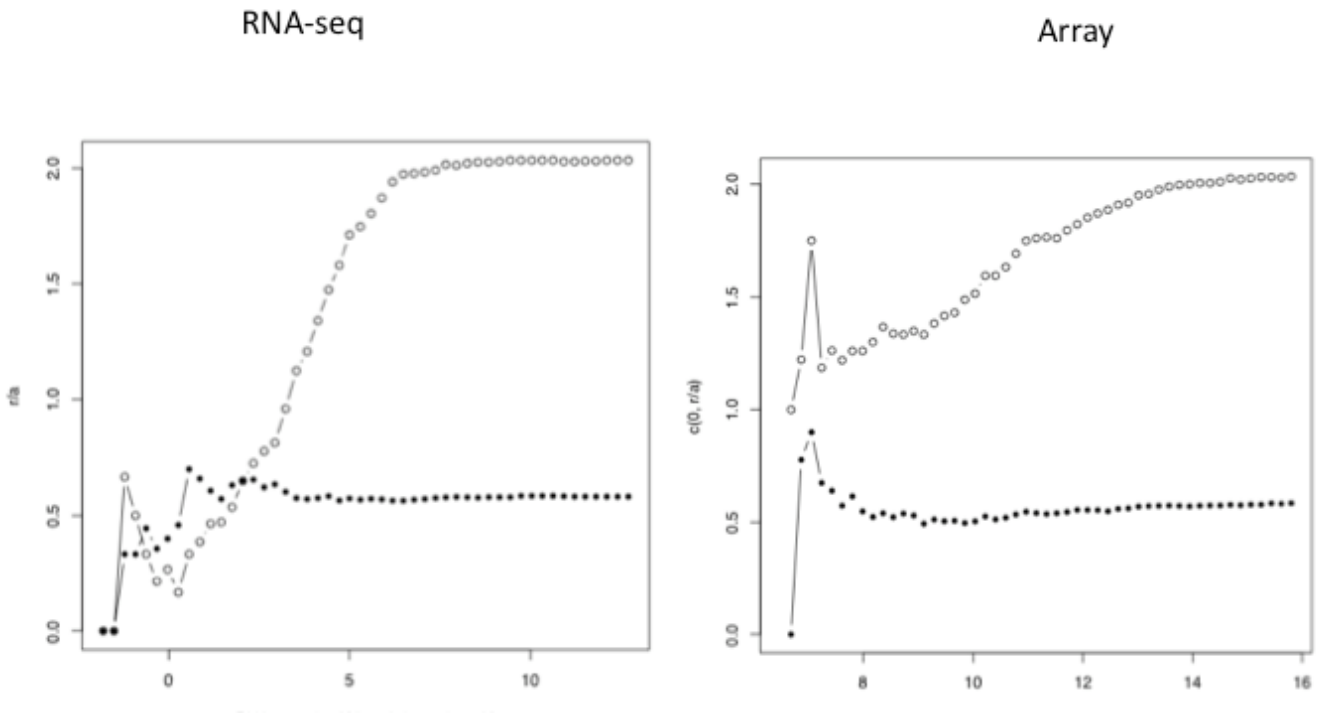
Supplemental Figure 7: eQTL effect sizes between array and sequencing
For 0.01 eQTLs the best association per exon or probe was determined. The corresponding rho values were then transformed by taking the absolute value and average across all the associations detected for the same gene (i.e. if there were multiple probes or exons possessing 0.01 eQTLs per gene). These values were compared for genes sharing an eQTL between the array and the RNA-seq. The distribution of array rho values was higher than the sequencing rho values (top panel). There was significant correlation between the rho values for the shared eQTLs (bottom panel - Pearson correlation between the rho values was 0.68).



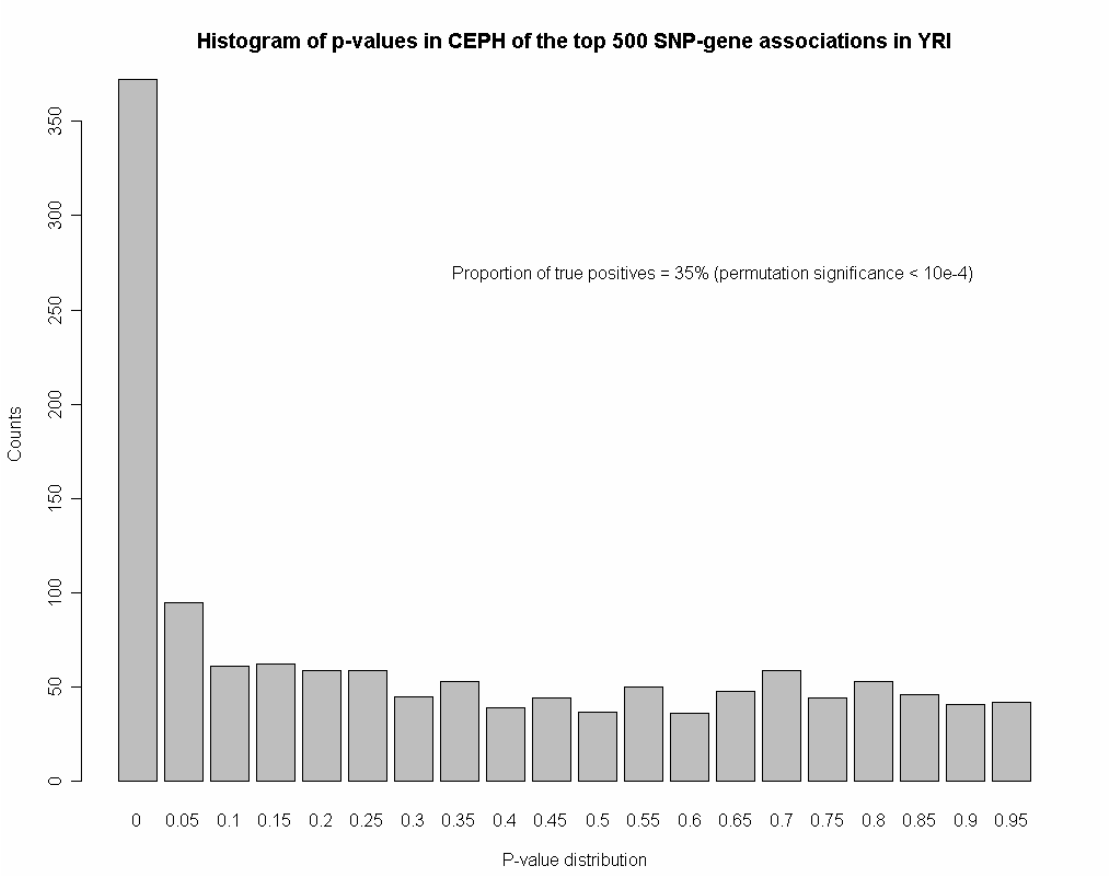
Supplemental Figure 8: Dynamic range of tested genes and genes with eQTLs. This plot shows the dynamic range of genes in the array data (Y-axis) and RNAseq data (X-axis – RPKM). Coloured dots indicate genes for which eQTLs (0.01 perm threshold) were found in RNA-Seq exon analysis only (red), array only (blue) and both (orange) at the 0.01 permutation threshold (see also Suppl Table 1 below).



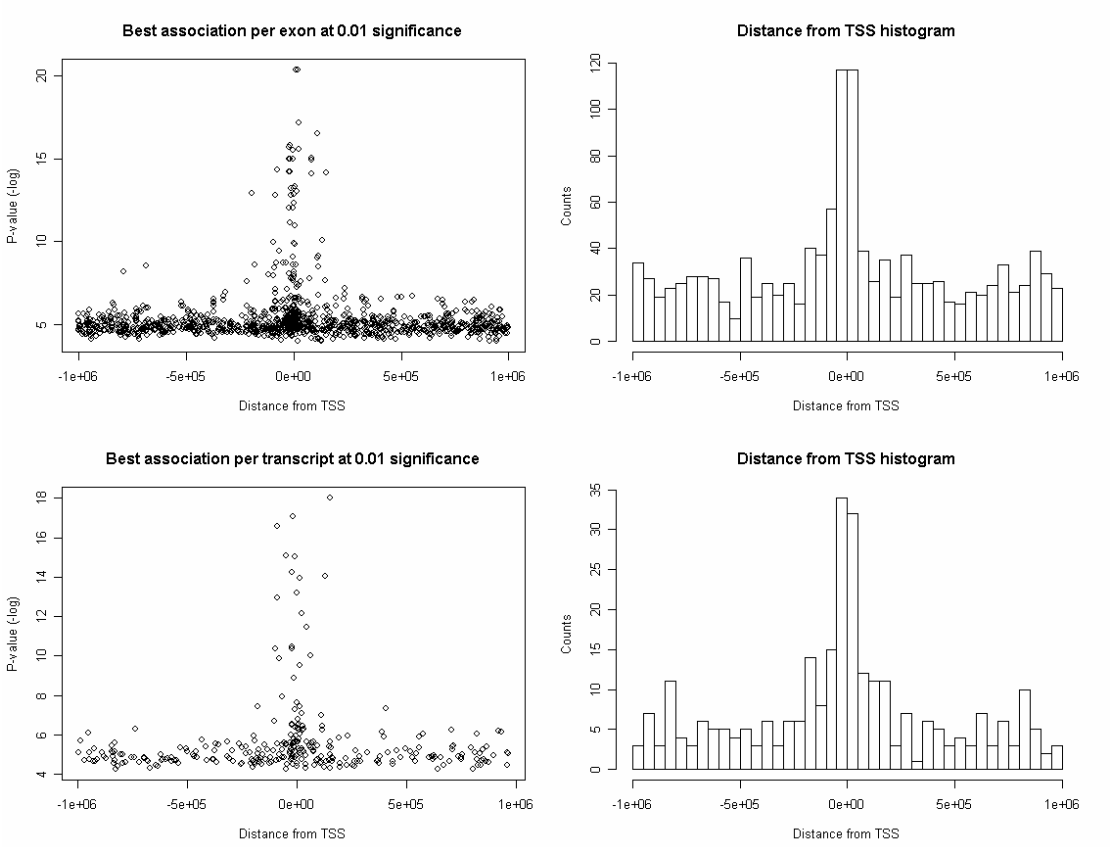
Supplemental Figure 9: Cumulative plots of eQTL discovery
Cumulative plots of eQTL discovery (0.01 perm threshold) for RNA-Seq exons (red), array data (blue) and whole gene RNA-Seq (green) along the dynamic range of RNAseq data (defined as RPKM) on the X-axis.



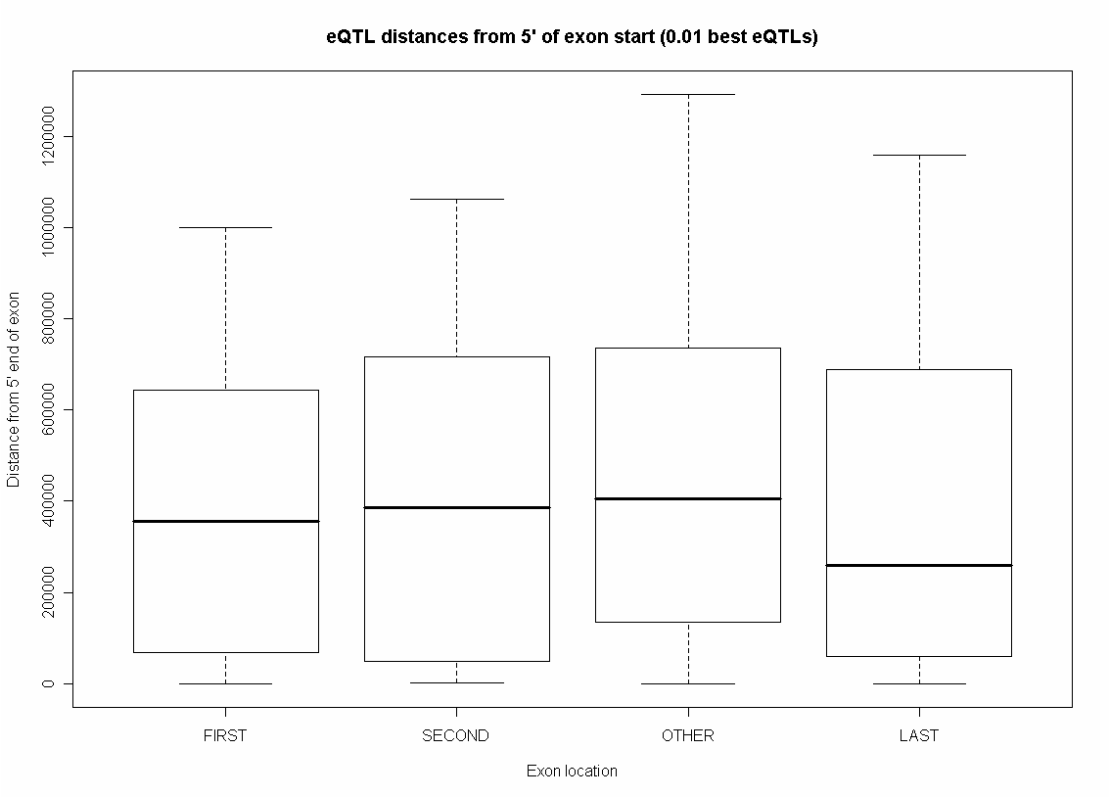
Supplemental Figure 10: Relative eQTL discovery
Relative eQTL discovery (0.01 perm threshold) for RNA-Seq exons/arrays (grey) and RNA-Seq whole genes/arrays (black) along the dynamic range in RNA-Seq RPKM (left panel) and array log2 (right panel).



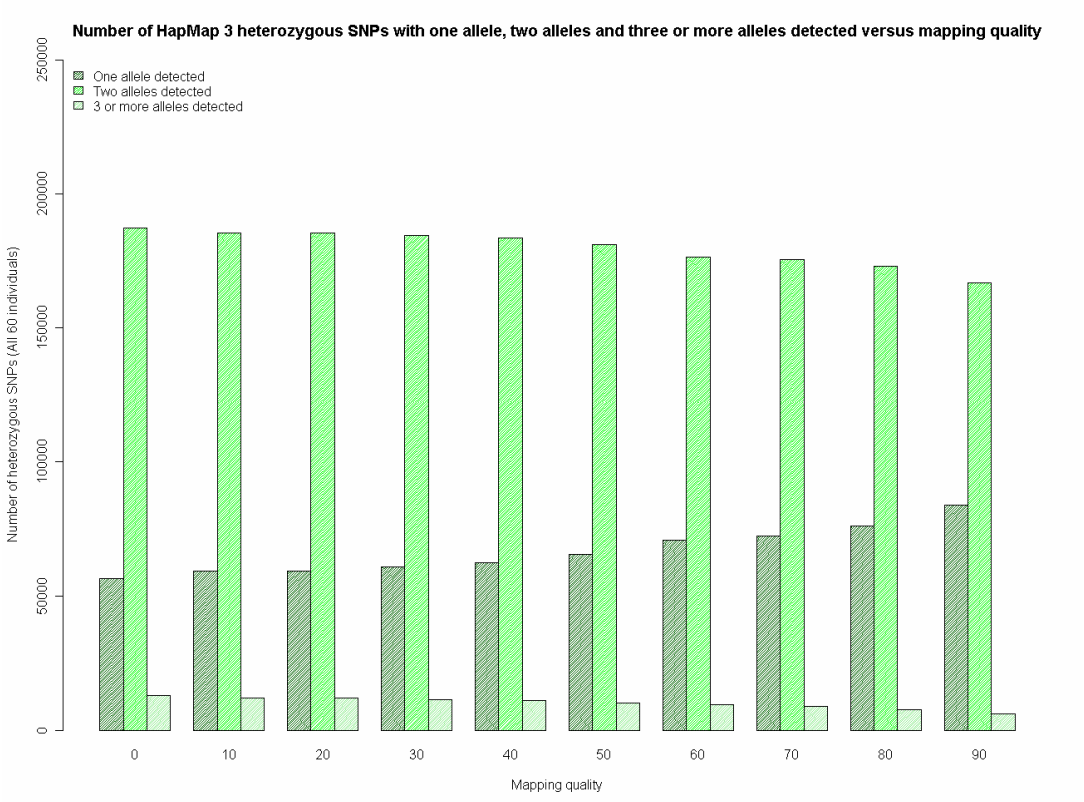
Supplemental Figure 11: Replication of YRI eQTLs in CEU.
The top 500 SNP-gene associations from population sequencing and eQTL discovery in Yoruban individuals were compared in our CEU sequencing. We plotted the p-value distribution for all matching SNP-exon pairs in CEU. This corresponded to 222 similarly tested SNPs and 358 genes which 1345 SNP-exon associations. Here, we see a strong enrichment in the significant p-values in the tail of the p-value distribution in CEU. Using the 1-pi0 q-value statistic, this corresponds to 35% of the shared signals are true positives. This was evaluated by permuting the same number of shared SNPs and genes and was found to be significant to at least 10,000 iterations (by example the median percentage of true positives across the permutations was 0.7% and the mean was 3.5%). Indicating that such replication between the studies by chance is very unlikely.



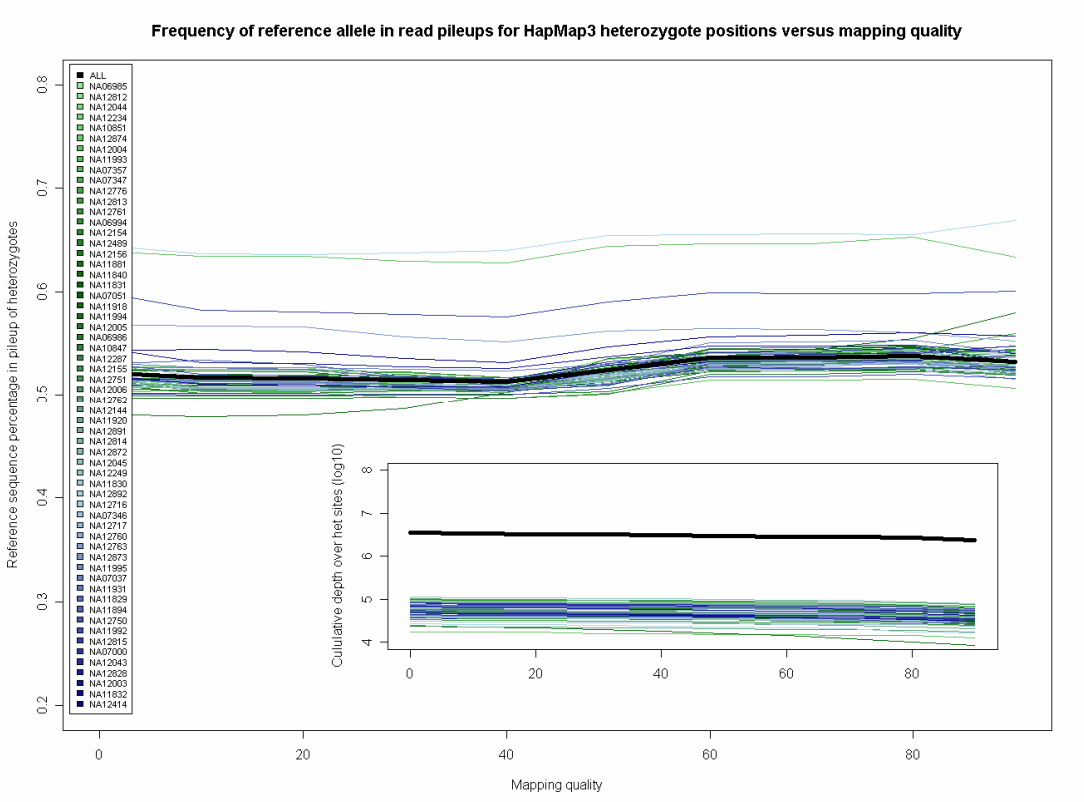
Supplemental Figure 12: eQTL distribution around transcription start site
The best association per exon (top panels) and transcript (bottom panels) at a minimum permutation threshold of 0.01 is plotted. Strong enrichment in significance and number of effects is centred on the transcription start site indicating that the majority of significant eQTLs are proximal to the gene.



Supplemental Figure 13: eQTL distances from 5' exon start.
The proximity of the best association per exon at a minimum permutation threshold of 0.01 is plotted with respect to the 5' end of the exon and its location in a multi-exonic gene. Here we see that eQTLs discovered for last exons are proportionally the closest to their target exon followed by first exons, second exons and other exons (p=0.0058; t-test).

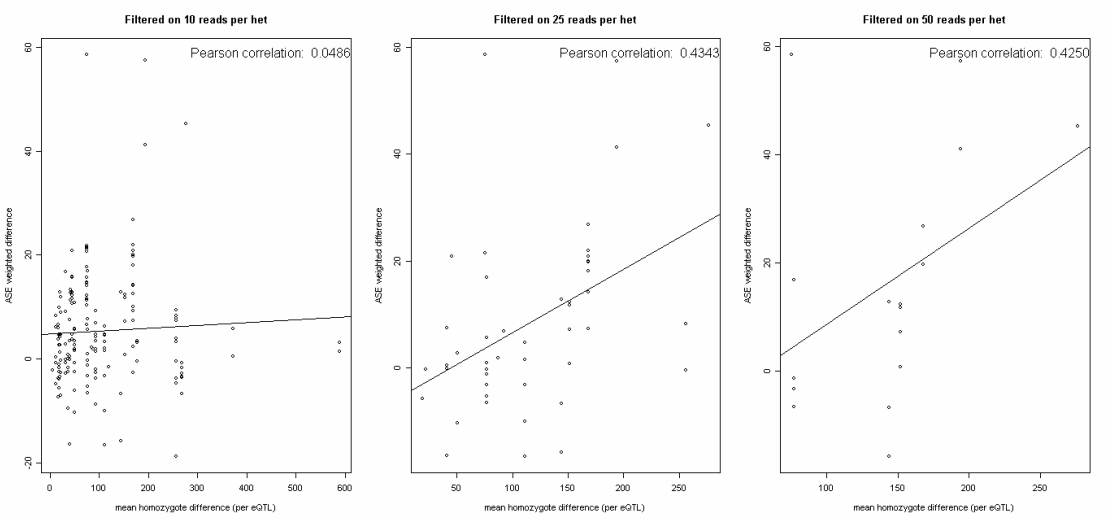


Supplemental Figure 14: Number of heterozygous alleles detected as a function of mapping quality.
For heterozygote positions from all 60 individuals, we plotted the number of alleles that were discovered as a function of MAQ mapping quality. Here the majority of heterozygotes have both alleles discovered. A small fraction of the heterozygotes have three alleles discovered. As mapping quality increases the fraction where only one allele is detected increases due to decreasing tolerance for mismatches.



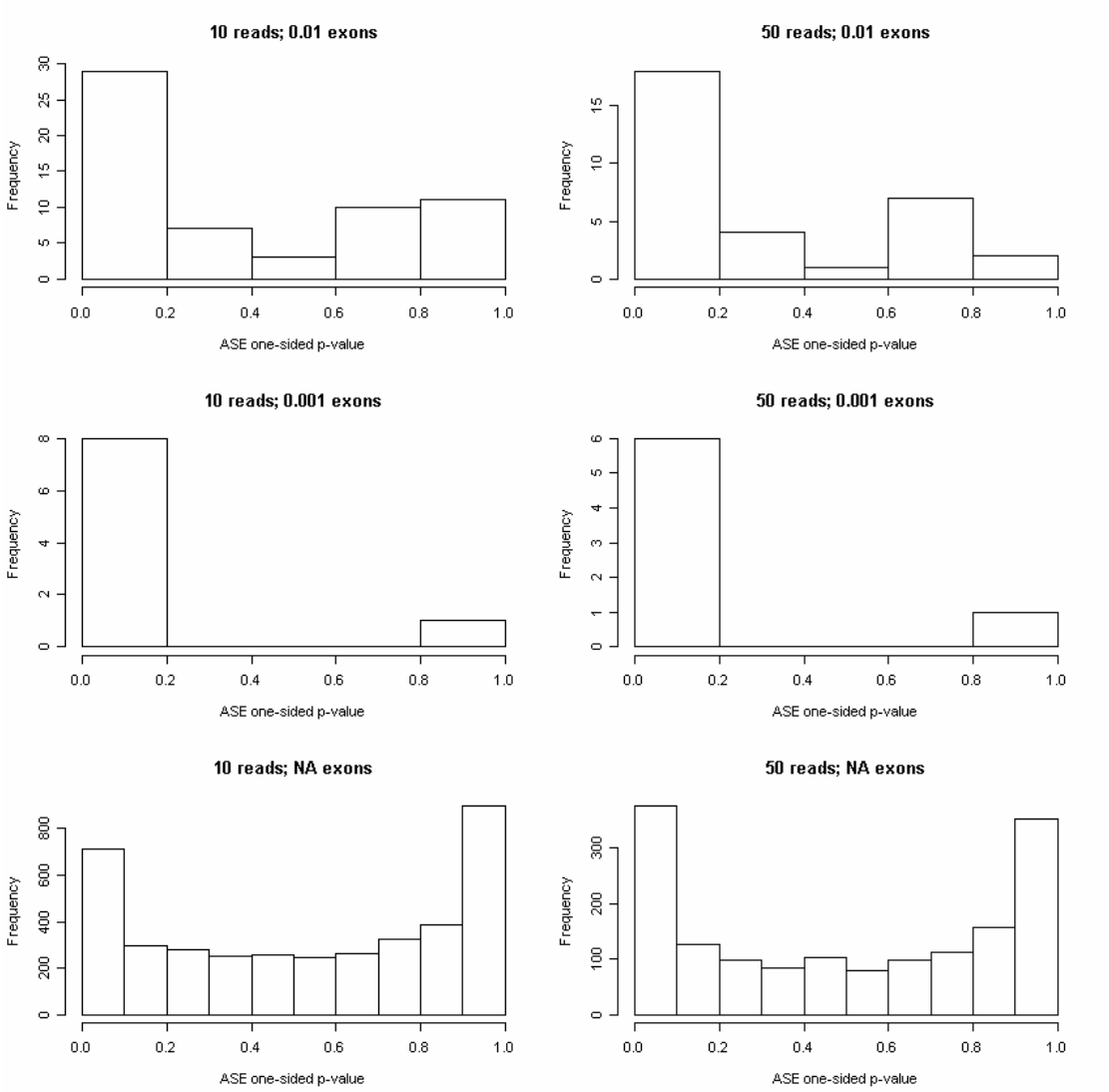
Supplemental Figure 15: Reference allele frequency in read pileups by mapping quality.

We investigate the bias towards the reference allele in heterozygote positions in each individual. We observe a tendency for the reference allele to be overrepresented in pileups over a heterozygote. We used this frequency as the success rate when assessing the binomial probability of allele-specific expression. The three individuals which had noticeably higher proportion of reference to non-reference mapping at MAQ10 were NA12815 (0.58), NA12004 (0.63) and NA12892 (0.64). INSET: The depth of sequencing over all heterozygote sites per individual is plotted.

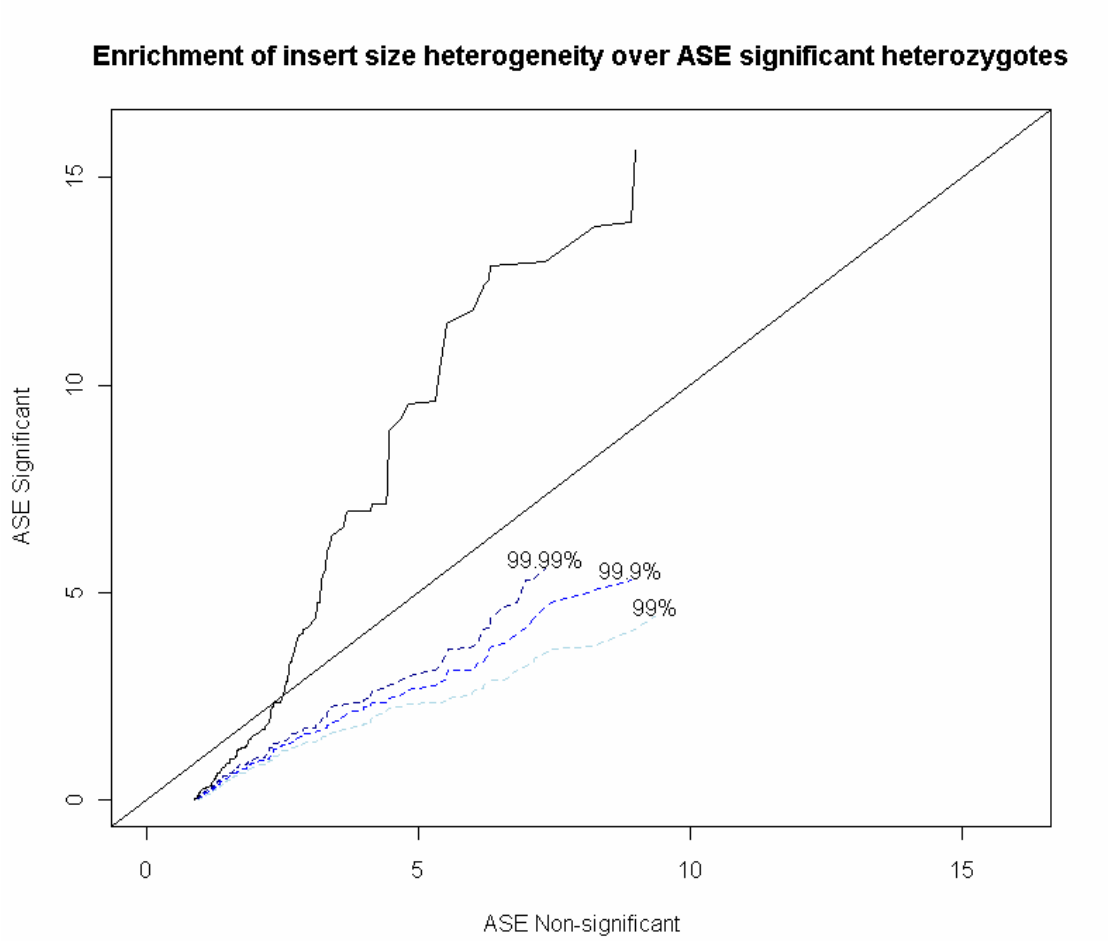


Supplemental Figure 16: Correlation between eQTL significance and ASE effect ratio by coverage.

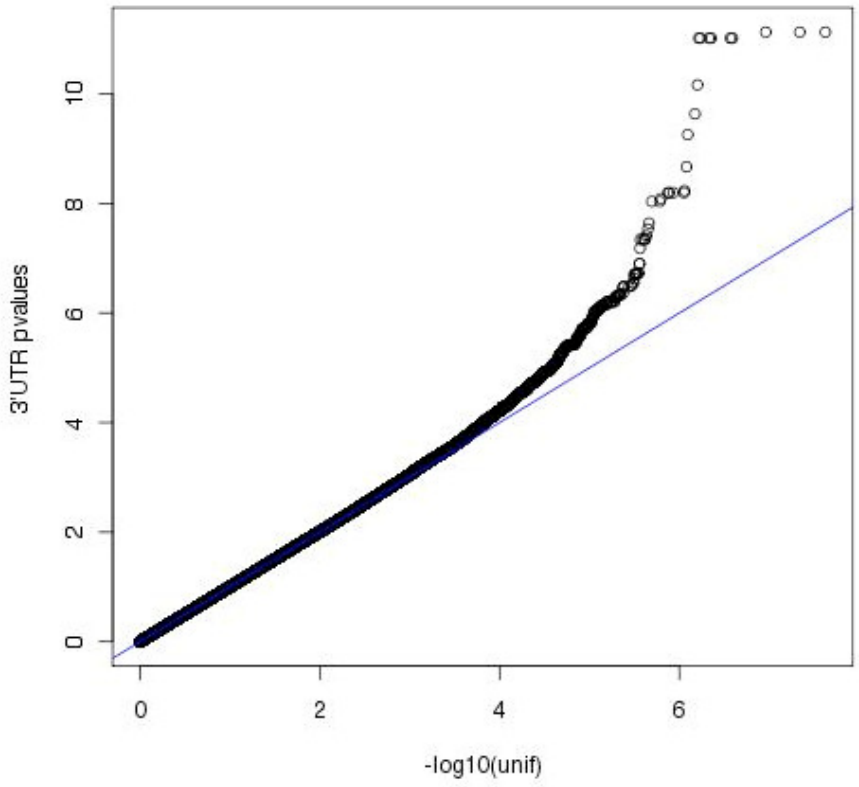
The correlation of mean homozygote difference per 0.01 exon eQTL and the corresponding phased ASE heterozygote's weighted ASE ratio is plotted. It is apparent that as read depth over the heterozygote increases the correlation improves. Intermediate correlation indicates that the ASE effect and eQTL effect are in the same direction and support one another. However, a challenge with this type of quantification is that coverage over heterozygotes is a function of transcript length meaning that some exons may be powered for eQTL discovery but have insufficient coverage for ASE quantification.



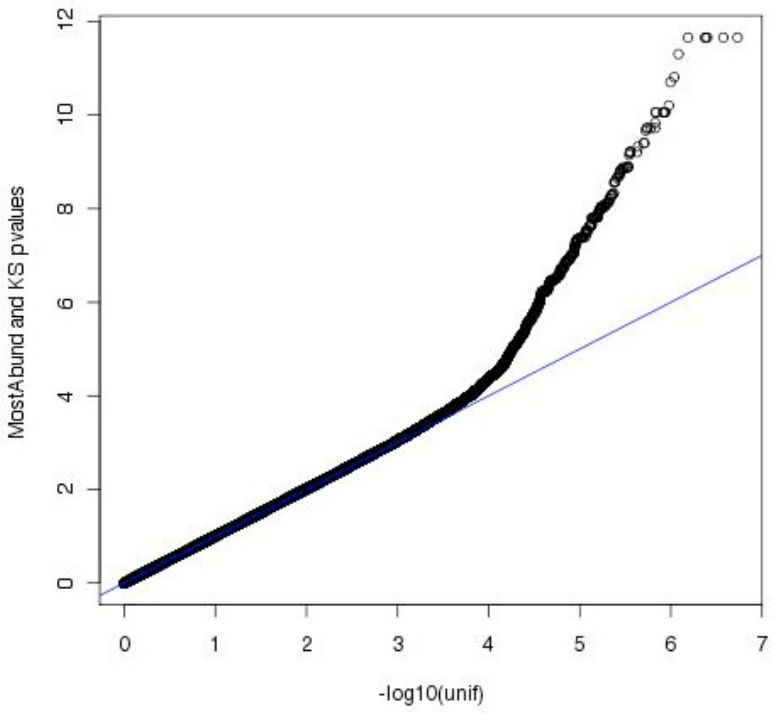
Supplemental Figure 17: One-sided ASE p-value distribution for exon eQTLs
A one-sided ASE binomial p-value weighting by reference allele discovery at mapping quality 10 for reads at mapping quality 10 was assessed relative to the phasing for significant and non-significant eQTLs. For significant eQTLs, we find enrichment in the tail of the p-value distribution indicating that the ASE is in the same direction as the eQTL. For non-significant eQTLs, we see enrichment at both tails indicating as one would expect that the chosen eQTL SNP is not-informative for the direction of ASE signal. This enrichment at both tails also highlights that there are regulatory haplotypes that are not captured by association.



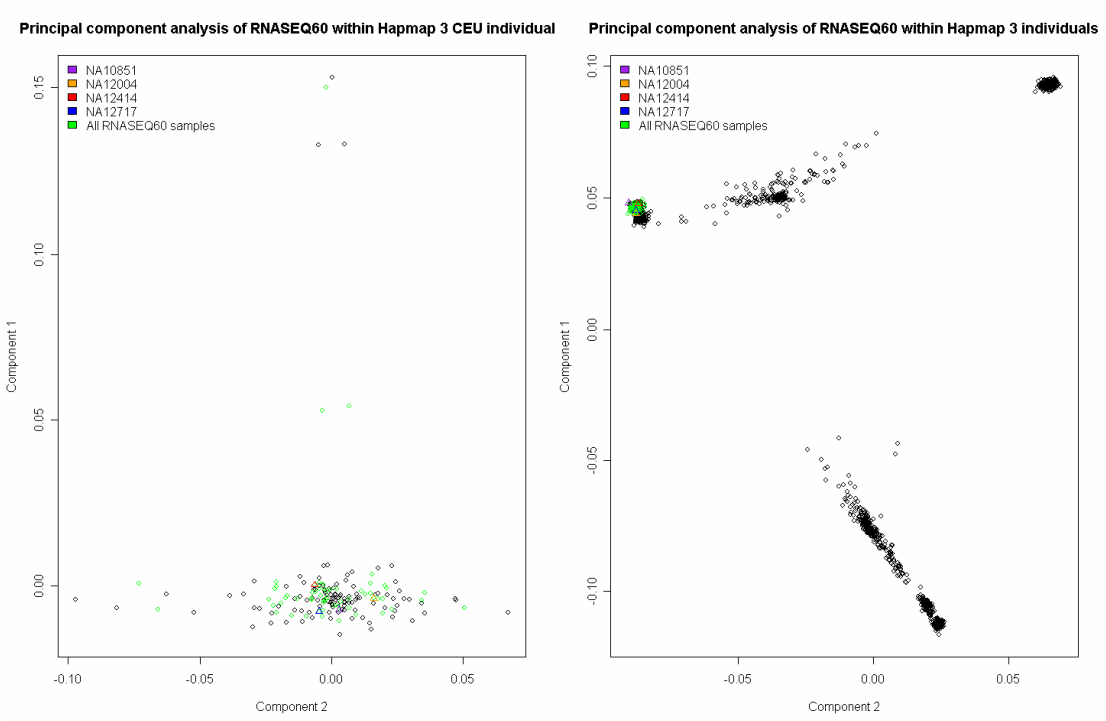
Supplemental Figure 18: Insert size heterogeneity over significant ASE heterozygotes. We sought to measure the degree to which genetics influences transcript-specific expression. To do this we looked at the insert size distribution of paired end reads over each heterozygote. Our expectation was that the heterogeneity of inserts sizes over significant ASE heterozygotes between each of their alleles would be increased relative to that between alleles of non-significant ASE heterozygotes. This is because if one haplotype is increasing the expression of a particular transcript relative to another, the insert size distribution over that allele would be changed relative to the other allele. This increase could indicate an allele-specific expression of one isoform relative to another. We tested this by conditioning for heterozygotes with a minimum of 50 reads for both alleles. For each heterozygote for an individual, we ran a bootstrapped Kolmogorov-Smirnov test (1000 permutation) for the respective insert size distributions. We then separated the p-values given the heterozygote was significant for ASE or not. Of the 901 heterozygotes, 235 were significant for ASE and of those 105 had significant transcript distribution heterogeneity (KS p-value < 0.05); this corresponded to 72 of 105 genes which contained an ASE significant heterozygote. The plot above shows the resulting QQ plot of –log10 transformed KS-test p-values partitioned by the significant versus non-significant ASE heterozygotes. A feature of this comparison is that it is not biased by technical insert size heterogeneity as each individual is processed separately. To further assess the significance of this deviation, we permuted the assignment of significant ASE 10,000 times within the heterozygotes and plotted the 99.99%, 99.9% and 99% QQ plots. As expected, because of the increased relative pool of non-significant to significant ASE heterozygotes, these permuted plots showed skew towards non-significant p-values. None of these permuted plots include the real signal indicating that the significance of this enrichment is at least 1 in 10,000 and that the ASE significant heterozygotes have more insert size heterogeneity than the ASE non-significant heterozygotes. This enrichment highlights that transcript-specific genetic control is a feature of regulatory complexity.



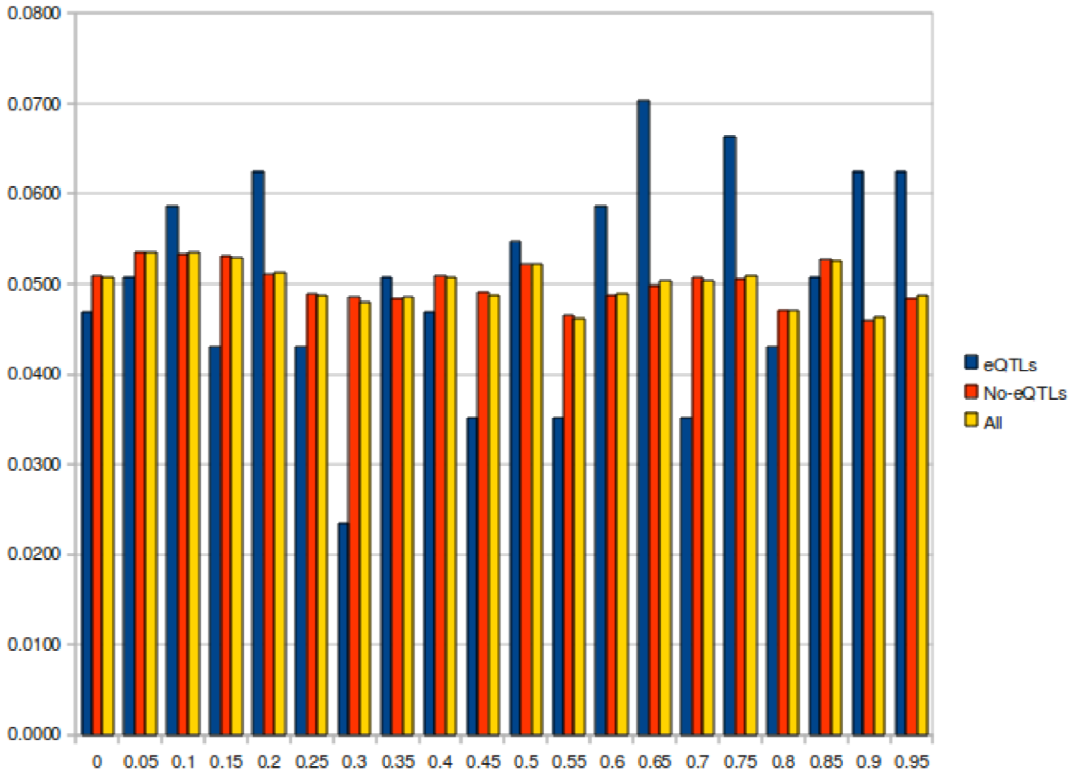
Supplemental Figure 19: QQ plot of associations of SNPs tested against the lenght of the 3' UTR of genes (alternative endings).
The QQ plot shows large enrichment for genetic effects influencing the length of the 3' UTR of genes (Spearman Rank correlation between inferred length of 3' UTR and SNP genotypes within 1 Mb)



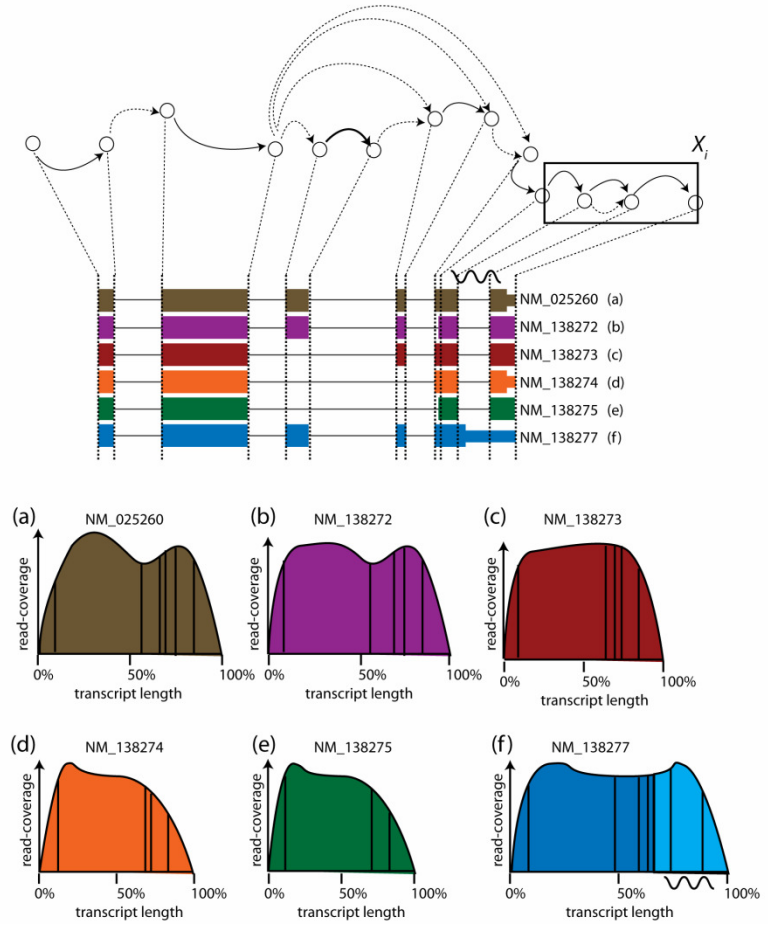
Supplemental Figure 20: QQ plot of associations of SNPs tested against insert size distributions (alternative transcript forms).
The QQ plot shows large enrichment for genetic effects influencing the internal structure of genes as this is represented by the distribution of insert sizes (Spearman Rank correlation between mean insert size and SNP genotypes within 1 Mb)



Supplemental Figure 21: PCA of RNASEQ60 samples within HapMap3
The first two principal components of a PCA within CEPH and within HapMap3 populations for the sequenced samples are shown. Here the imputed SNPs are highlighted. It can be seen that there is no special clustering of the imputed SNPs which would indicate high-error during imputation.



Supplemental Figure 22: Correlation of gene expression with read depth
The histogram (Y-axis: frequency; X-axis: p-value) shows the p-value distribution of Spearman Rank Correlation tests between read depth of a given sample with normalized gene expression abundance estimates for all genes (yellow), genes with eQTL (blue) and gene without and eQTL(red). As is clear from the plot, the distributions are uniform, as expected, so read depth does not have an overall effect on gene expression quantification, and furthermore there is no difference between those genes with an eQTL and genes without an eQTL (Mann-Whitney, P = 0.98)



Supplemental Figure 23: FluxCapacitor outline
Information about overlapping segments in alternative transcripts (center) is non-redundantly described by a form of splice graphs, with edges for each segment of an exon included in a different number of transcripts (top). Each node represents a splice site (respectively, transcription start and poly-adenylation). Intronic edges are depicted as dashed arrows, whereas edges that represent exonic segments are solid arrows. Subsequently, reads (zig-zag line) get mapped to sets of these edges called *k*-super edges X_i. For each such *k*-super edge X_i holds that the sum of expression contribution (i.e., "flux") from all transcripts including the edge, meets the observed read flux at the edge (i.e., the reads mapped to the edge) plus some error Δ_i. In order to account for biases in the read distribution along the transcript (bottom), capacity correction factors b_i^j are estimated for each edge and transcript. Indeed read distribution profiles are calculated in non-overlapping transcripts, binned by several transcript lengths and expression levels such that given a transcript, the most appropriate profile given the transcript length and expression is used to calculate the area under the profile between the edge limits. This area is the correction capacity for the edge and transcript. The collection of linear constraints along all exonic edges of the labelled splice graph forms a system that is solved by linear programming, minimizing the sum of all edge noise levels

$\sum_i |\Delta_i|$.

Supplemental Table 1: eQTL results for the overlapping set of 9319 genes

ASSOCIATIONS	PERMUTATION THRESHOLDS*		
	0.05	0.01	0.001
Exon quantification	3003	768	85
Transcript quantification	914	217	36
Whole gene quantification	711	206	44
Array-based quantification	961	348	127

- gene level thresholds

For the overlapping set of 9319 genes we compare eQTL discoveries for the 3 RNA-Seq based quantifications and arrays.