File S1

Supplemental Methods

Simulation parameters

RNA-seq reads were simulated from the CAST inbred strain and from a reconstructed DO individual using the Flux Simulator (version 1.2) and the parameters below.

Command line argument: flux-simulator –lsp Parameter_filename.txt

Single-end sequence parameters

```
REF_FILE_NAME      path/to/Gene_annotations.gtf
GEN_DIR            path/to/Genome.fa
LIB_FILE_NAME      filename.lib
SEQ_FILE_NAME      filename.bed
PRO_FILE_NAME      filename.pro
RT_PRIMER                   PDT
READ_NUMBER        10000000 (or 30000000)
READ_LENGTH        100
FILTERING                   true
SIZE_DISTRIBUTION   N(280,50)
FASTA              true
TSS_MEAN                    NaN
POLYA_SCALE        NaN
POLYA_SHAPE        NaN
ERR_FILE           76
```

Paired-end sequence parameters

```
REF_FILE_NAME      path/to/Gene_annotations.gtf
GEN_DIR            path/to/Genome.fa
LIB_FILE_NAME      filename.lib
SEQ_FILE_NAME      filename.bed
PRO_FILE_NAME      filename.pro
RT_PRIMER                   PDT
READ_NUMBER        60000000
READ_LENGTH        100
PAIRED_END                  YES
FILTERING                   true
SIZE_DISTRIBUTION   N(280,50)
FASTA              true
TSS_MEAN                    NaN
POLYA_SCALE        NaN
POLYA_SHAPE        NaN
ERR_FILE           76
```

Table S1 Isoform-level summary of read alignment in the simulated CAST data

		Aligned to CAST					
	Read Class	Incorrect Unique Reads	Incorrect Multireads	Unmapped Reads	Correct Multireads	Correct Unique Reads	**Total**
Aligned to NCBIM37	Incorrect Unique Reads	1,378	1	4	11,721	2,725	15,829
	Incorrect Multireads	3	5,842	2	8,713	492	15,052
	Unmapped Reads	48	52	1,709,356	191,919	222,222	2,123,597
	Correct Multireads	15	62	145	4,378,338	10,739	4,389,299
	Correct Unique Reads	1	2	150	5,075	3,450,918	3,456,146
	Total	1,445	5,959	1,709,657	4,595,766	3,687,096	9,999,923

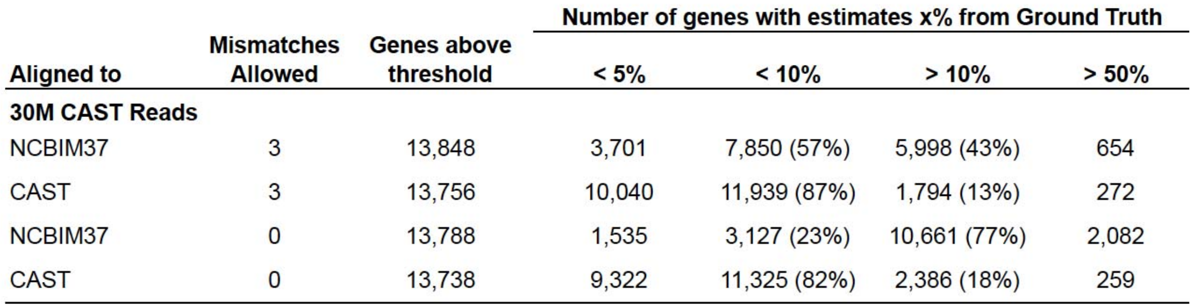
The simulated reads were aligned to the NCBIM37 and CAST transcriptomes. Reads that improve by alignment to CAST are highlighted in green, with those that improve by two or more categories are highlighted in dark green. Reads that improve by alignment to NCBIM37 are highlighted in red, with those that improve by two or more categories highlighted in dark red. Reads on the diagonal align equivalently by both strategies.

Table S2 List of genes from the CAST simulation that were affected by read misalignment or alignment failure from the reference alignment strategy

Three lists of genes are included in the attached table. The first list shows genes for which simulated CAST reads align uniquely but falsely in the NCBIM37 transcriptome. Alignment to the CAST transcriptome rescues these reads to their correct, unique origin (second list). The third list shows genes from which reads fail to align at all in the NCBIM37 transcriptome but align to the correct, unique position in the CAST transcriptome.

Table S2 is available for download as a MS Excel file at
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165886/-/DC1

Table S3 Comparison of gene-level abundance results from alignment of 30 million simulated CAST reads to NCBIM37 and CAST transcriptomes



			Number of genes with estimates x% from Ground Truth			
Aligned to	Mismatches Allowed	Genes above threshold	< 5%	< 10%	> 10%	> 50%
30M CAST Reads						
NCBIM37	3	13,848	3,701	7,850 (57%)	5,998 (43%)	654
CAST	3	13,756	10,040	11,939 (87%)	1,794 (13%)	272
NCBIM37	0	13,788	1,535	3,127 (23%)	10,661 (77%)	2,082
CAST	0	13,738	9,322	11,325 (82%)	2,386 (18%)	259

Alignment of 30 million simulated CAST reads to the individualized CAST transcriptome (≤3 mismatches) results in nearly three times as many gene estimates (N= +6,339) that fall within 5% of ground truth value and fewer than a third as many gene estimates (N= -4,204) that deviate more than 10% from the ground truth. Gene-level abundance results for perfect matching reads (i.e. 0 mismatches) are also shown.

Table S4 List of genes from the DO simulation that were affected by read misalignment or alignment failure from the reference alignment strategy

Three lists of genes are included in the attached table. The first list shows genes for which simulated DO reads align uniquely but falsely in the NCBIM37 transcriptome. Alignment to the DO transcriptome rescues these reads to their correct, unique origin (second list). The third list shows genes from which reads fail to align at all in the NCBIM37 transcriptome but align to the correct, unique position in the DO transcriptome.

Table S4 is available for download as a MS Excel file at
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165886/-/DC1

Table S5 Comparison of gene-level abundance results from alignment of 30 million simulated DO reads to NCBIM37 and individualized DO transcriptomes

			Number of genes with estimates x% from Ground Truth			
Aligned to	**Mismatches Allowed**	**Genes above threshold**	**< 5%**	**< 10%**	**> 10%**	**> 50%**
30M DO Reads						
NCBIM37	3	13,260	7,371	10,995 (83%)	2,265 (17%)	355
DO IRG	3	13,209	9,829	11,696 (89%)	1,501 (11%)	262
NCBIM37	0	13,222	2,301	4,800 (36%)	8,422 (64%)	728
DO IRG	0	13,196	9,136	11,169 (85%)	2,012 (15%)	249

Gene estimates in the simulated DO sample are improved by read alignment to the individualized transcriptome (≤3 mismatches), yielding 33% more gene estimates (N= +2,458) within 5% of the ground truth value and 34% fewer estimates (N= -764) that deviate more than 10% from the ground truth. Gene-level abundance results for perfect matching reads (i.e. 0 mismatches) are also shown.

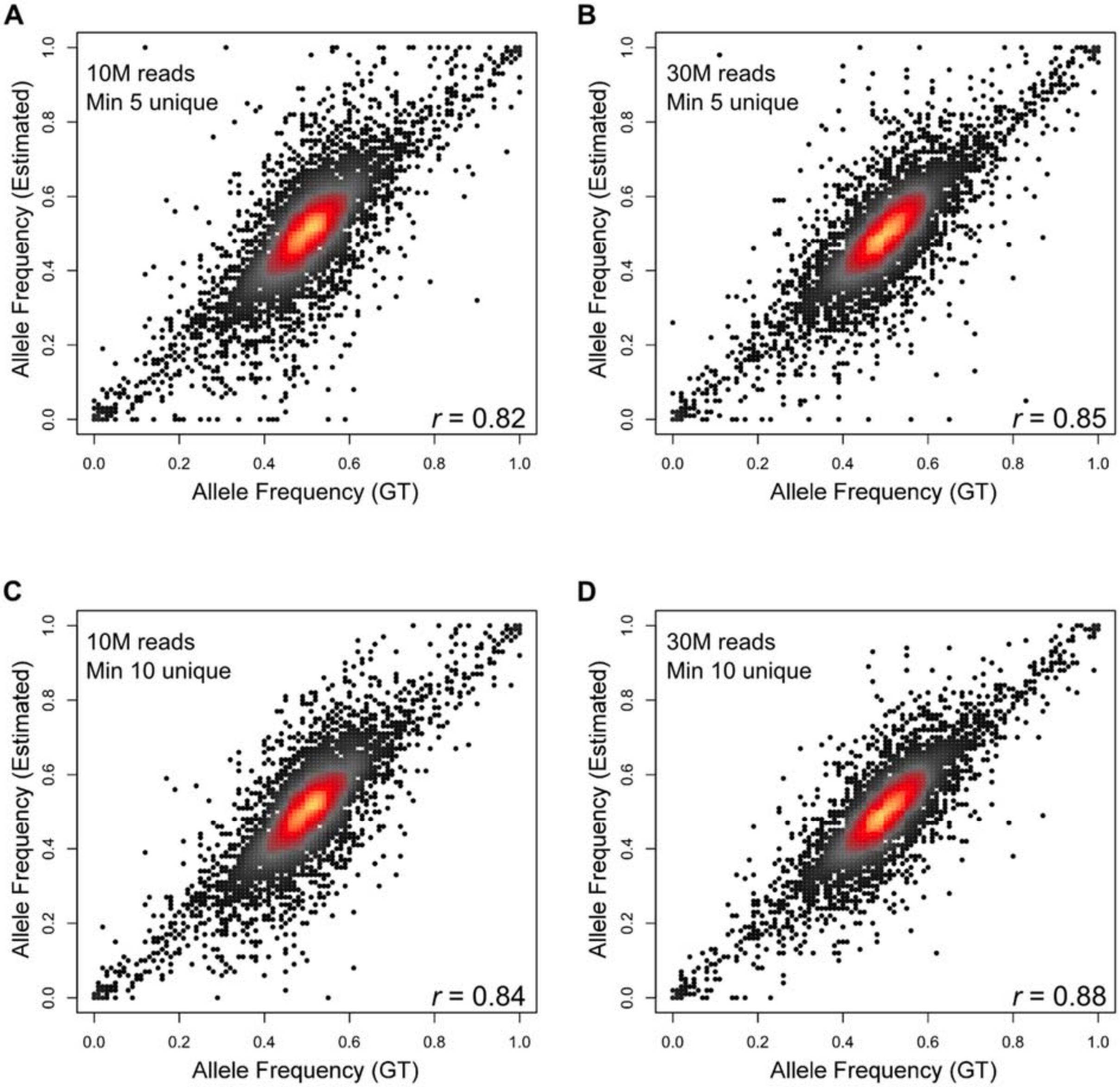
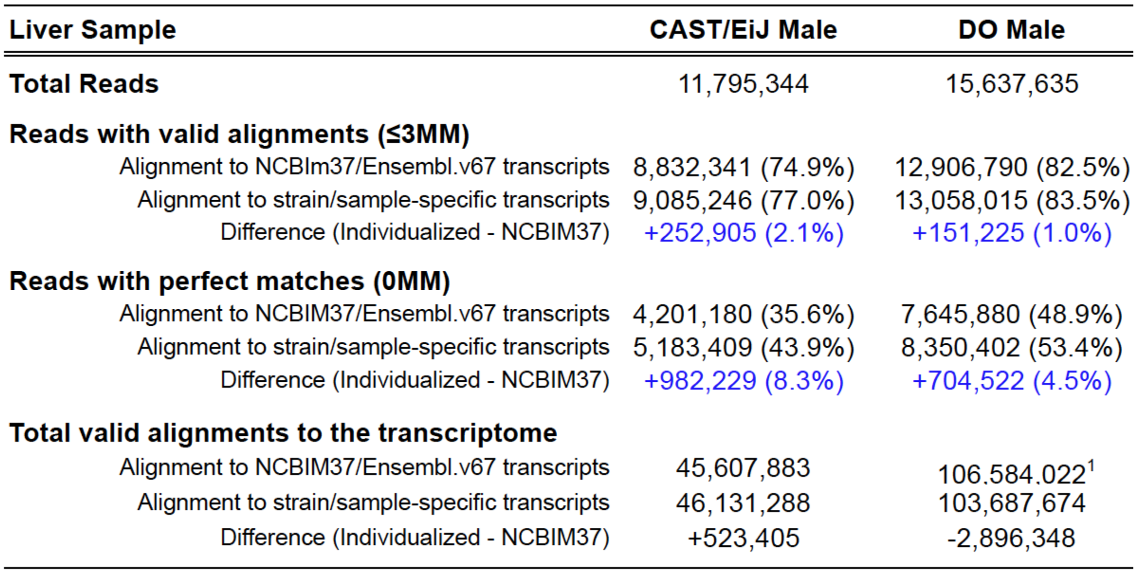


Figure S1 Characterization of sequencing depth and unique read threshold on estimation of allele-specific expression. Estimated allele frequency (y-axis) is plotted in panels A-D against the ground truth allele frequency (x-axis) for robustly expressed genes (sum of allele counts ≥ 100) in the simulated DO dataset. Allele frequency estimates are improved by increasing the read depth from 10 million (panels A and C) to 30 million reads (panels B and D) and by increasing the gene inclusion stringency to require ten (panels C and D) rather than five (panels A and B) reads with unique allele alignments.

Table S6 Alignment statistics for real CAST and DO liver RNA-seq data



Liver Sample	CAST/EiJ Male	DO Male
Total Reads	11,795,344	15,637,635
Reads with valid alignments (≤3MM)		
Alignment to NCBIm37/Ensembl.v67 transcripts	8,832,341 (74.9%)	12,906,790 (82.5%)
Alignment to strain/sample-specific transcripts	9,085,246 (77.0%)	13,058,015 (83.5%)
Difference (Individualized - NCBIM37)	+252,905 (2.1%)	+151,225 (1.0%)
Reads with perfect matches (0MM)		
Alignment to NCBIM37/Ensembl.v67 transcripts	4,201,180 (35.6%)	7,645,880 (48.9%)
Alignment to strain/sample-specific transcripts	5,183,409 (43.9%)	8,350,402 (53.4%)
Difference (Individualized - NCBIM37)	+982,229 (8.3%)	+704,522 (4.5%)
Total valid alignments to the transcriptome		
Alignment to NCBIM37/Ensembl.v67 transcripts	45,607,883	106.584.022[1]
Alignment to strain/sample-specific transcripts	46,131,288	103,687,674
Difference (Individualized - NCBIM37)	+523,405	-2,896,348

Bowtie (version 0.12.8) parameters: -v 3 -a -m --best --strata

[1] For comparison to the diploid transcriptome alignments in DO samples, the total number of alignments to NCBIM37 were scaled by 2x.

Alignment of real data to individualized CAST- or DO-specific transcriptomes yields more reads with valid alignments (≤ 3 mismatches (MM)), and significantly more reads with perfect (0 MM) alignments. Reads align with greater specificity (i.e. fewer alignments per mapped read) to individualized transcriptomes than to NCBIM37.

Table S7 eQTL simulation summary showing the classification of eQTL calls that differ between alignment strategies differentiated by gene biotype

	Correct Calls			Incorrect Calls		
				False	False Positive	False Positive
Gene Biotype	True Local	True Distant	True No eQTL	Negative	Local	Distant
Antisense	2	3	15	-2	-16	-2
IG_C_gene	0	0	1	0	-1	0
lincRNA	6	3	15	-7	-15	-2
misc_RNA	2	0	2	-2	-1	-1
Mt_rRNA	0	0	0	0	0	0
non_coding	0	0	0	0	0	0
polymorphic pseudogene	1	0	0	-1	0	0
processed_transcript	3	0	2	-3	-2	0
protein_coding	336	94	981	-353	-980	-78
pseudogene	23	3	32	-10	-7	-41
retrotransposed	3	-1	1	-2	1	-2
sense_intronic	0	0	1	0	-1	0
sense_overlapping	0	0	0	0	0	0
snoRNA	0	0	0	0	0	0
Total	**376**	**102**	**1050**	**-380**	**-1022**	**-126**

Choice of read alignment strategy affects ten percent of genes (n = 1,528/15,027 total) in our simulation study. Alignment to individualized DO transcriptomes yields the correct eQTL assignment for all but one gene with a discordant call. Many gene biotypes yield incorrect eQTL calls after alignment to GRCm38 but pseudogenes in particular appear to be sensitive to false positive distant associations.

Table S8 Gene-level summary of eQTL simulation results

Columns 1-7 give information for the expressed gene, columns 8-10 show the SNP identifier and location for the marker with the highest LOD score in the simulation, and columns 11-13 provide details of the simulated eQTL including LOD score, p-value, and eQTL class (e.g., significant local or distant eQTL, no eQTL). Columns 14-19 show the eQTL mapping results after alignment of the simulated reads to the GRCm38 reference transcriptome. Column 18 shows the eQTL assignment relative to the simulated ground truth, and Column 19 lists whether the peak SNP associated with gene expression after alignment to GRCm38 matches the simulated peak SNP. Columns 20-25 show the same classes of eQTL data but after alignment of the simulated reads to individualized DO transcriptomes.

TableS8 is available for download as a MS Excel file at
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165886/-/DC1

Table S9 List of eQTL from alignment to individualized DO transcriptomes

Columns 1-6 give information for the expressed gene, columns 7-9 show the SNP identifier and location for the marker with the highest LOD score, and columns 10-13 provide details of the eQTL including LOD score, raw p-value, adjusted q-value, and position relative to the controlled transcript (i.e. local or distal eQTL).

TableS9 is available for download as a tab-delimited text file at
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165886/-/DC1

Table S10 List of eQTL from alignment to NCBIM37

Columns 1-6 give information for the expressed gene, columns 7-9 show the SNP identifier and location for the marker with the highest LOD score, and columns 10-13 provide details of the eQTL including LOD score, raw p-value, adjusted q-value, and position relative to the controlled transcript (i.e. local or distal eQTL).

TableS10 is available for download as a tab-delimited text file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165886/-/DC1

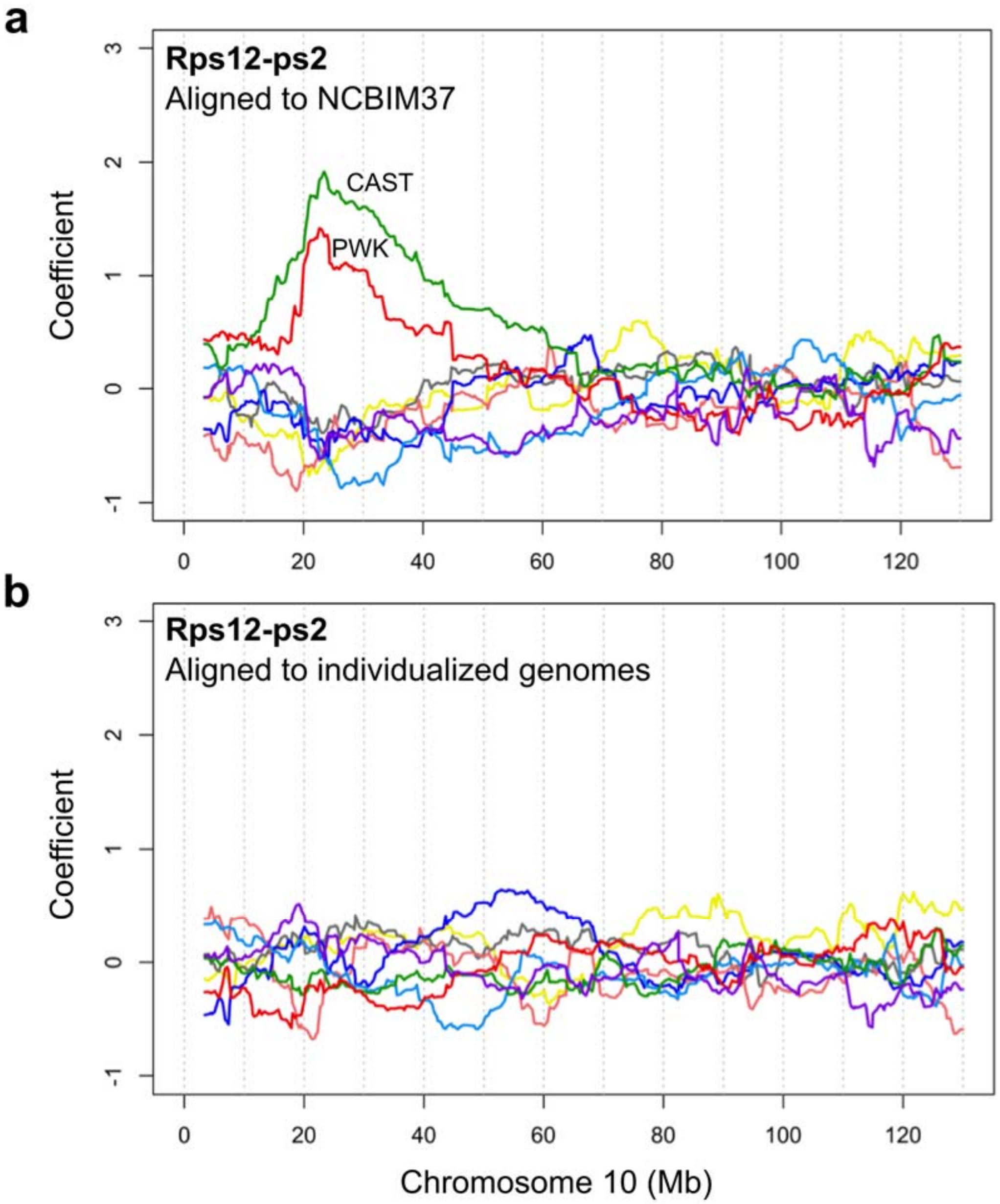


Figure S2 Comparison of Chromosome 10 founder coefficient plots for *Rps12-ps2* expression derived from alignment to NCBIM37 or individualized DO transcriptomes. Read alignment to individualized DO transcriptomes ameliorates spurious alignments to pseudogenes. When DO reads are aligned to the NCBIM37 reference transcriptome (A), it appears that DO animals that derive the Chr 10 region from CAST or PWK have higher expression of the pseudogene *Rps12-ps2*. When individual genetic variation is accounted for in the alignment (B), the CAST- and PWK-derived reads align preferentially to the parent protein coding gene *Rps12*, and the spurious *Rps12-ps2* eQTL is eliminated.

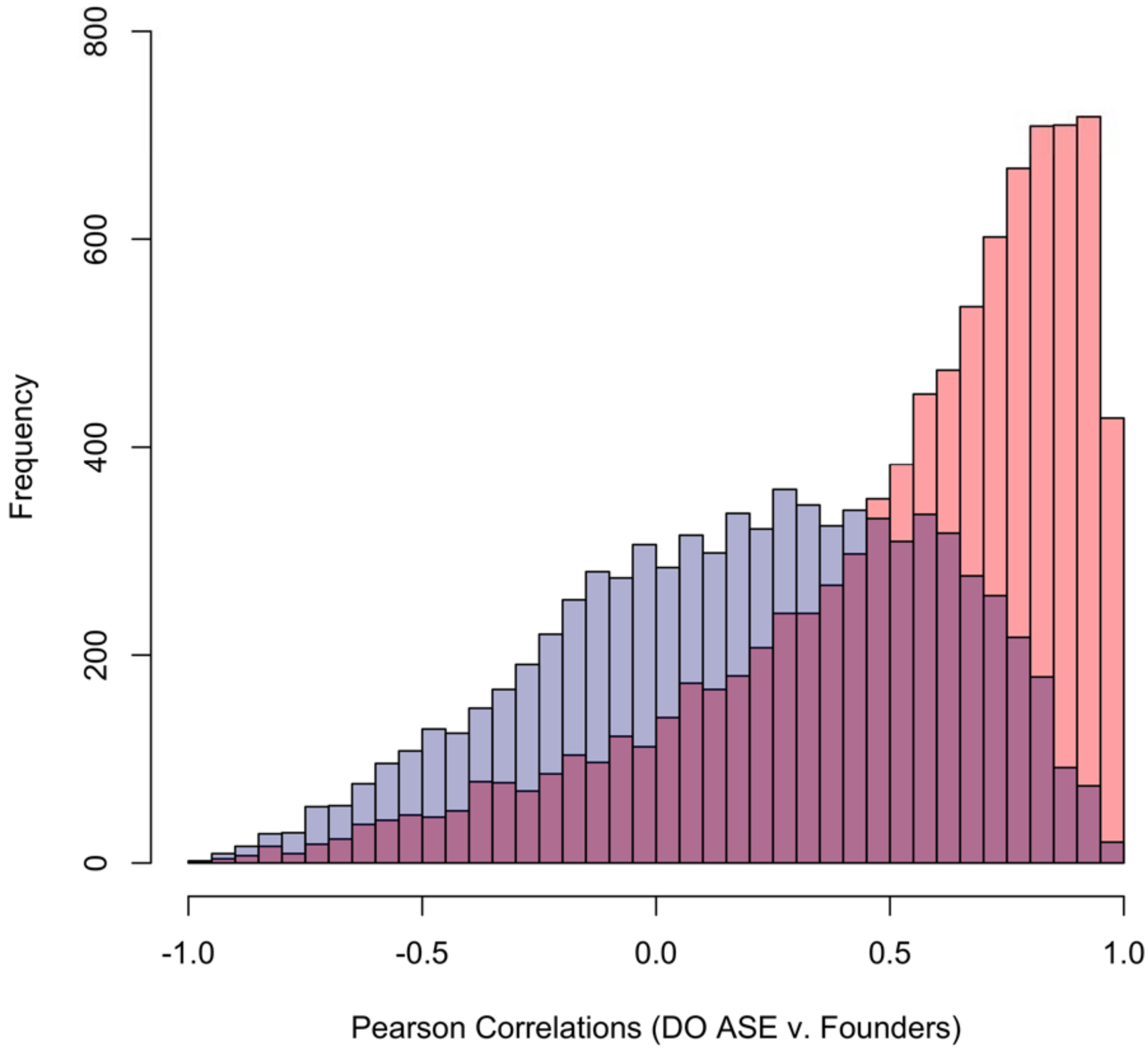
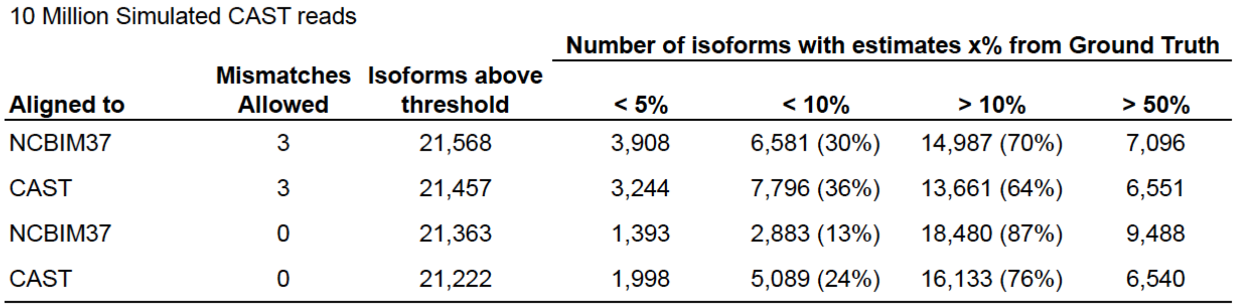


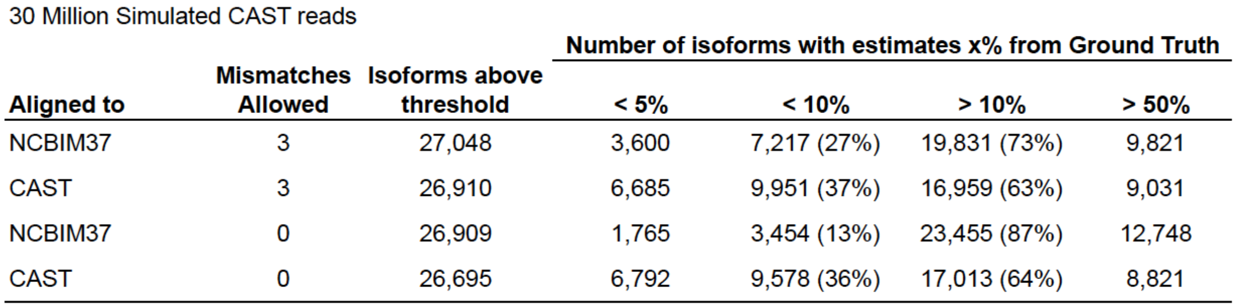
Figure S3 Comparison of gene-level expression in Founder strain samples and founder allele-level estimates in the DO samples for genes with and without significant local eQTL after alignment to individualized genomes. Pearson correlations between founder strain expression and founder allele estimates in the DO population are plotted as a histogram above. Founder allele estimates for genes with significant local eQTL (n=8,981 genes, shown in pink) exhibit higher concordance to gene-level liver expression in Founder strain samples compared to genes that do not have significant local eQTL (n=7,893 genes, shown in blue).

Table S11 Isoform abundance results in CAST simulation study



10 Million Simulated CAST reads

			Number of isoforms with estimates x% from Ground Truth			
Aligned to	**Mismatches Allowed**	**Isoforms above threshold**	**< 5%**	**< 10%**	**> 10%**	**> 50%**
NCBIM37	3	21,568	3,908	6,581 (30%)	14,987 (70%)	7,096
CAST	3	21,457	3,244	7,796 (36%)	13,661 (64%)	6,551
NCBIM37	0	21,363	1,393	2,883 (13%)	18,480 (87%)	9,488
CAST	0	21,222	1,998	5,089 (24%)	16,133 (76%)	6,540

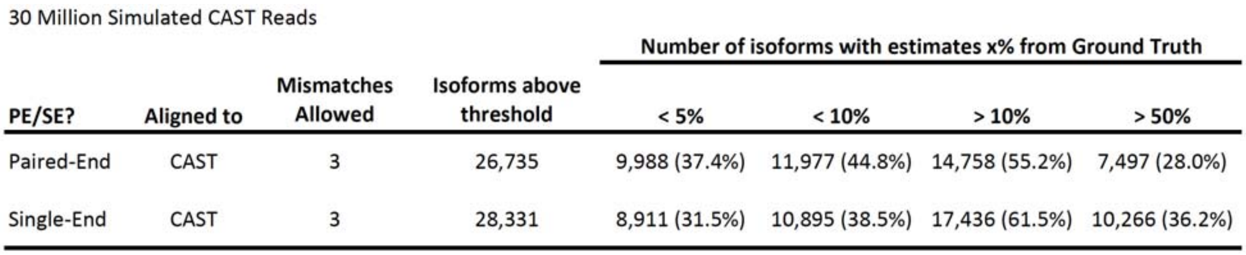


30 Million Simulated CAST reads

			Number of isoforms with estimates x% from Ground Truth			
Aligned to	**Mismatches Allowed**	**Isoforms above threshold**	**< 5%**	**< 10%**	**> 10%**	**> 50%**
NCBIM37	3	27,048	3,600	7,217 (27%)	19,831 (73%)	9,821
CAST	3	26,910	6,685	9,951 (37%)	16,959 (63%)	9,031
NCBIM37	0	26,909	1,765	3,454 (13%)	23,455 (87%)	12,748
CAST	0	26,695	6,792	9,578 (36%)	17,013 (64%)	8,821

Alignment of simulated CAST reads to the individualized CAST transcriptome (≤3 mismatches) improves estimates of isoform abundance compared to alignment to NCBIM37. Increasing the sequencing depth from 10 to 30 million single-end reads significantly does not improve isoform resolution – more isoform estimates fall within five percent of the simulated ground truth but the total number of isoforms expressed above threshold increases too, causing no relative improvement in the accuracy of isoform abundance estimates. Isoform-level abundance results for perfect matching reads (i.e. 0 mismatches) are also shown.

Table S12 Comparison of isoform abundance results in CAST simulation study from using paired-end or single-end sequencing

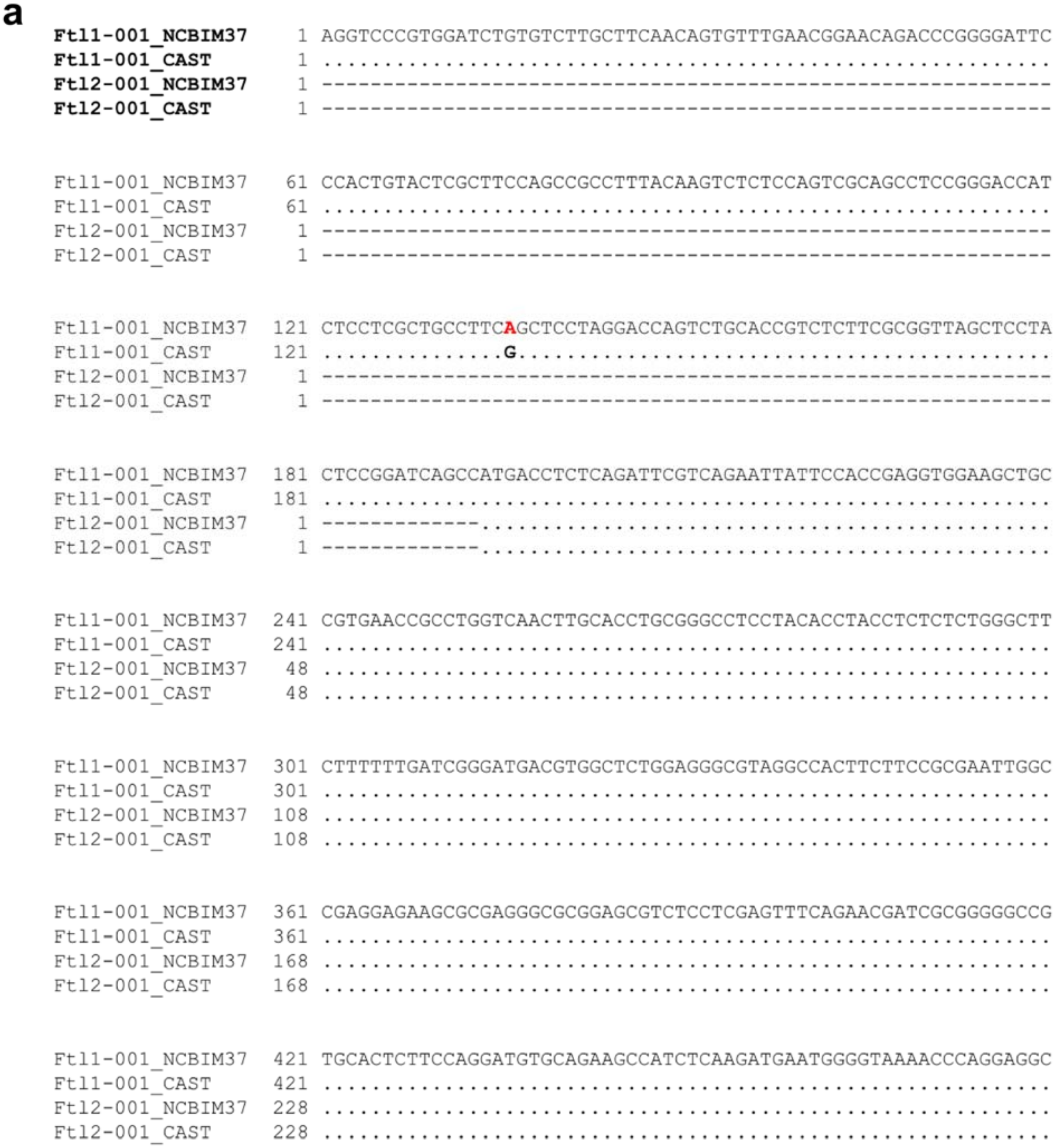


30 Million Simulated CAST Reads

				Number of isoforms with estimates x% from Ground Truth			
PE/SE?	Aligned to	Mismatches Allowed	Isoforms above threshold	< 5%	< 10%	> 10%	> 50%
Paired-End	CAST	3	26,735	9,988 (37.4%)	11,977 (44.8%)	14,758 (55.2%)	7,497 (28.0%)
Single-End	CAST	3	28,331	8,911 (31.5%)	10,895 (38.5%)	17,436 (61.5%)	10,266 (36.2%)

Paired-end sequencing yields modest improvements in isoform abundance estimation relative to single-end reads. For example, 45% of isoform estimates fall within ten percent of the simulated ground truth value in the analysis of paired-end reads, compared to 39% for single-end reads.

Figure S4



```
Ftl1-001_NCBIM37    1 AGGTCCCGTGGATCTGTGTCTTGCTTCAACAGTGTTTGAACGGAACAGACCCGGGGATTC
Ftl1-001_CAST       1 ............................................................
Ftl2-001_NCBIM37    1 ------------------------------------------------------------
Ftl2-001_CAST       1 ------------------------------------------------------------

Ftl1-001_NCBIM37   61 CCACTGTACTCGCTTCCAGCCGCCTTTACAAGTCTCTCCAGTCGCAGCCTCCGGGACCAT
Ftl1-001_CAST      61 ............................................................
Ftl2-001_NCBIM37    1 ------------------------------------------------------------
Ftl2-001_CAST       1 ------------------------------------------------------------

Ftl1-001_NCBIM37  121 CTCCTCGCTGCCTTCAGCTCCTAGGACCAGTCTGCACCGTCTCTTCGCGGTTAGCTCCTA
Ftl1-001_CAST     121 ...............G............................................
Ftl2-001_NCBIM37    1 ------------------------------------------------------------
Ftl2-001_CAST       1 ------------------------------------------------------------

Ftl1-001_NCBIM37  181 CTCCGGATCAGCCATGACCTCTCAGATTCGTCAGAATTATTCCACCGAGGTGGAAGCTGC
Ftl1-001_CAST     181 ............................................................
Ftl2-001_NCBIM37    1 -------------...............................................
Ftl2-001_CAST       1 -------------...............................................

Ftl1-001_NCBIM37  241 CGTGAACCGCCTGGTCAACTTGCACCTGCGGGCCTCCTACACCTACCTCTCTCTGGGCTT
Ftl1-001_CAST     241 ............................................................
Ftl2-001_NCBIM37   48 ............................................................
Ftl2-001_CAST      48 ............................................................

Ftl1-001_NCBIM37  301 CTTTTTTGATCGGGATGACGTGGCTCTGGAGGGCGTAGGCCACTTCTTCCGCGAATTGGC
Ftl1-001_CAST     301 ............................................................
Ftl2-001_NCBIM37  108 ............................................................
Ftl2-001_CAST     108 ............................................................

Ftl1-001_NCBIM37  361 CGAGGAGAAGCGCGAGGGCGCGGAGCGTCTCCTCGAGTTTCAGAACGATCGCGGGGGCCG
Ftl1-001_CAST     361 ............................................................
Ftl2-001_NCBIM37  168 ............................................................
Ftl2-001_CAST     168 ............................................................

Ftl1-001_NCBIM37  421 TGCACTCTTCCAGGATGTGCAGAAGCCATCTCAAGATGAATGGGGTAAAACCCAGGAGGC
Ftl1-001_CAST     421 ............................................................
Ftl2-001_NCBIM37  228 ............................................................
Ftl2-001_CAST     228 ............................................................
```

Figure S4 (continued)

```
                                          *
Ftl1-001_NCBIM37   481 CATGGAAGCTGCCTTGGCCATGGAGAAGAACCTGAATCAGGCCCTCTTGGATCTGCATGC
Ftl1-001_CAST      481 ...................T........................................
Ftl2-001_NCBIM37   288 ...................C........................................
Ftl2-001_CAST      288 ...................C........................................

                                         *
Ftl1-001_NCBIM37   541 CCTGGGTTCTGCCCGCGCGGACCCTCATCTCTGTGACTTCCTGGAAAGCCACTATCTGGA
Ftl1-001_CAST      541 ..................C.........................................
Ftl2-001_NCBIM37   348 ......C...........C..................................TC.....
Ftl2-001_CAST      348 ......C...........C..................................TC.....

Ftl1-001_NCBIM37   601 TAAGGAGGTGAAACTCATCAAGAAGATGGGCAACCATCTGACCAACCTCCGCAGGGTGGC
Ftl1-001_CAST      601 ............................................................
Ftl2-001_NCBIM37   408 ............................................................
Ftl2-001_CAST      408 ............................................................

                       *                                   *
Ftl1-001_NCBIM37   661 GGGGCCACAACCAGCGCAGACTGGCGCGCCCCAGGGGTCTCTGGGCGAGTATCTCTTTGA
Ftl1-001_CAST      661 A...................................A.......................
Ftl2-001_NCBIM37   468 A...................................A.......................
Ftl2-001_CAST      468 A...................................A.......................

Ftl1-001_NCBIM37   721 GCGCCTCACTCTCAAGCACGACTAGGAGGCCTCTGTACCTTCCAAGGGGCTCCCCCCTCT
Ftl1-001_CAST      721 ............................................................
Ftl2-001_NCBIM37   528 .........................-----------------------------------
Ftl2-001_CAST      528 .........................-----------------------------------

Ftl1-001_NCBIM37   781 GCTCTGCACCAGCCCGCCCTGGGACCTCCACCTGAATGAACCTCTCAAGCCACTAGGCAG
Ftl1-001_CAST      781 ............................................................
Ftl2-001_NCBIM37       ------------------------------------------------------------
Ftl2-001_CAST          ------------------------------------------------------------

Ftl1-001_NCBIM37   841 CTTTGTAACCGCCCTGGAGCCTCTGTCAAGTCTTGGACCAAGTAAAAATAAAGCTTTTTG
Ftl1-001_CAST      841 ............................................................
Ftl2-001_NCBIM37       ------------------------------------------------------------
Ftl2-001_CAST          ------------------------------------------------------------

Ftl1-001_NCBIM37   901 AGACAGC
Ftl1-001_CAST      901 .......
Ftl2-001_NCBIM37       -------
Ftl2-001_CAST          -------
```

Figure S4 (continued)

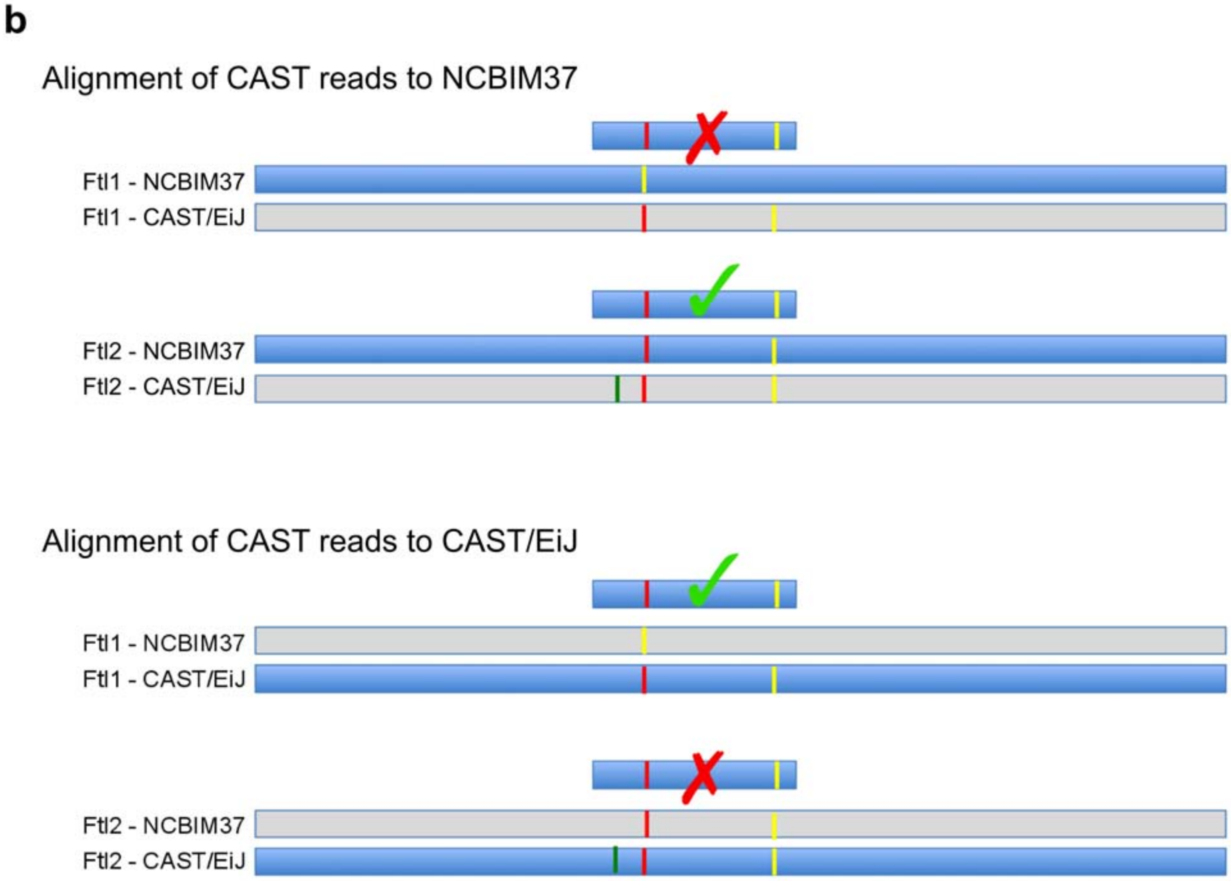


Figure S4 Strain polymorphisms between NCBIM37 and CAST in *Ftl1* and *Ftl2* transcript sequences can bias alignment of CAST-derived *Ftl1* reads. (A) Multiple alignment of *Ftl1-001* and *Ftl2-001* transcript sequences from NCBIM37 and the individualized CAST genomes. Variation in *Ftl1*/*Ftl2* abundance estimates in CAST liver RNA-seq stems mainly from 3-4 SNPs (starred). (B) Schematic showing how CAST polymorphisms in RNA-seq reads can cause misalignments in NCBIM37. CAST *Ftl1* reads that overlap any of these SNPs will align preferentially to *Ftl2* if aligned to NCBIM37 (upper panel). Accounting for CAST strain variation in *Ftl1* reduces spurious alignments to the *Ftl2* pseudogene (lower panel).

Figure S5

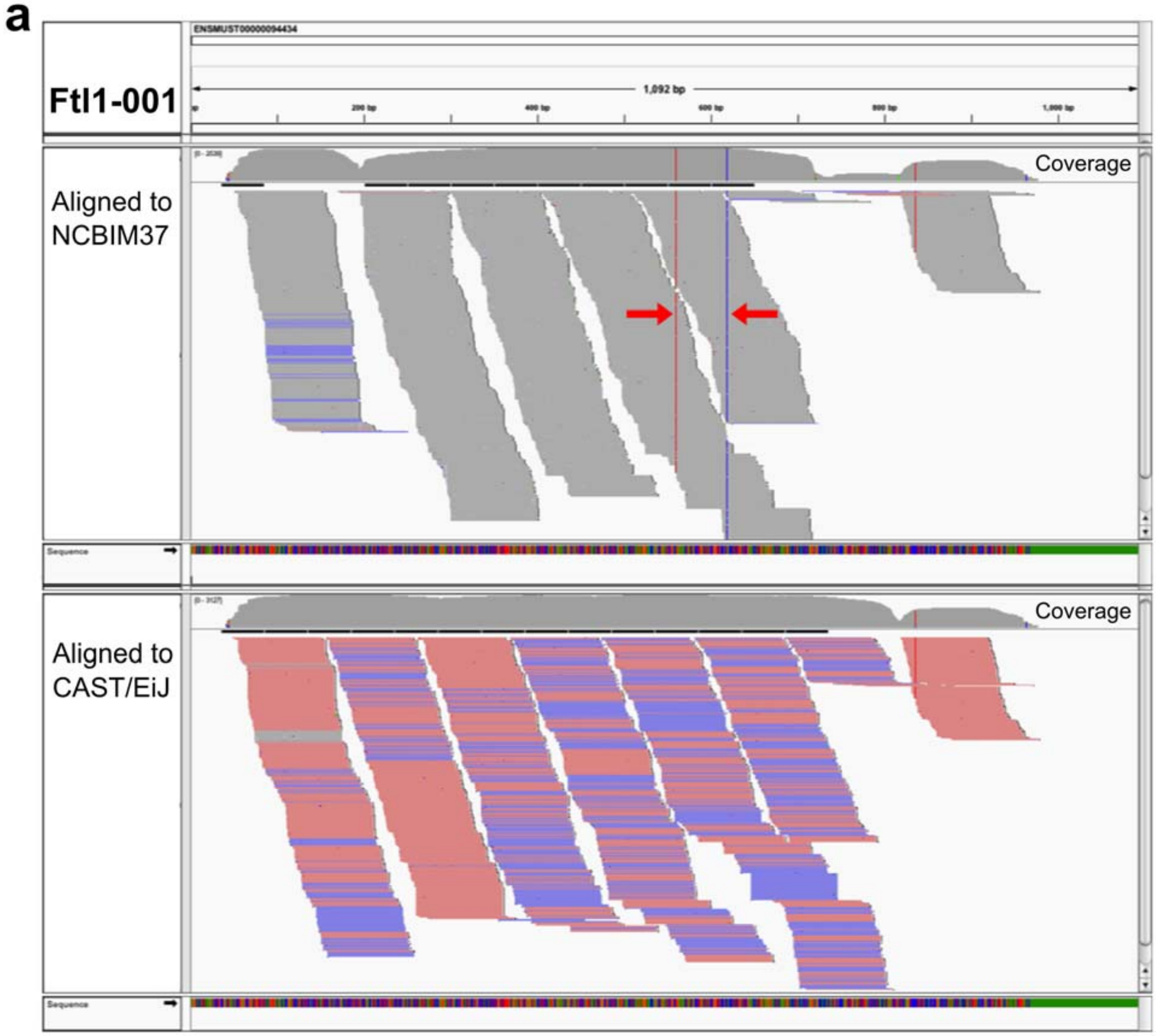

a
Ftl1-001
ENSMUST00000094434
1,092 bp
200 bp
400 bp
600 bp
800 bp
1,000 bp
Aligned to NCBIM37
Coverage
Sequence
Aligned to CAST/EiJ
Coverage
Sequence

Figure S5 (continued)

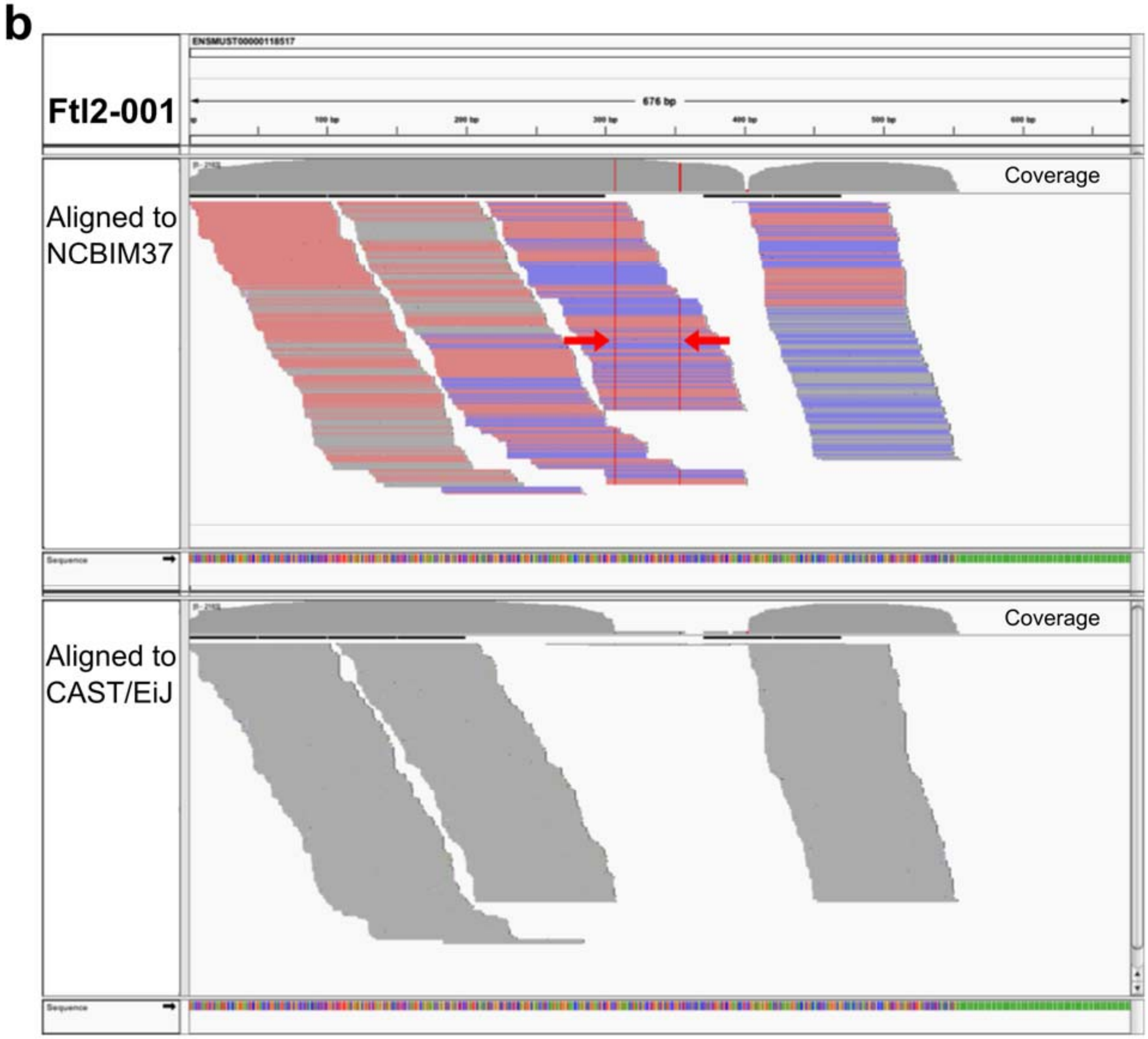


Figure S5 Coverage of CAST reads to *Ftl1* and *Ftl2* transcript sequences derived from the NCBIM37 reference genome and individualized CAST genome. Coverage plots show the distribution of CAST RNA-seq read alignments to *Ftl1-001* (A) and *Ftl2-001* (B) from alignment to each of the NCBIM37 reference and individualized CAST transcriptomes. Read coverage density (log transformed) is displayed at the top of each panel. For individual aligned reads, read color corresponds to orientation (red = forward strand, blue = reverse strand) and posterior probability. Gray reads have low probability of being transcribed from the aligned transcript location (as estimated by RSEM), while blue/red indicates reads that have been assigned high posterior probabilities. The red arrows point to SNPs in the CAST reads that differ from NCBIM37. Accounting for these CAST SNPs in the alignment diverts many reads from the *Ftl2* pseudogene to the parent protein-coding gene *Ftl1*.