Supplementary material

The fractured landscape of RNA-seq alignment: The default in our STARs

Sara Ballouz, Alexander Dobin, Thomas Gingeras, Jesse Gillis*

*Corresponding author

Table of Contents

Table of Supplementary tables

Table of Supplementary figures

Supplementary Tables

Table S1 Studies used in the benchmark meta-assessment

Study	Number of datasets/ samples	Number of methods	Notes
Fraction mapped assessment metric			Fraction of unique reads mapped out of total reads
Bao et al. 2011 (1)	4	11	
Baruzzo et al. 2016 (2)	37	16	
Bonfert et al. 2015 (3)	2	7	
Dillies et al. 2012 (4)	23	3	
Dobin et al. 2012 (5)	1	5	
Engstrom et al. 2013 (6)	4	13	
Gran et al. 2011 (7)	2	11	
Langmead et al. 2012 (8)	7	9	
Li et al. 2009 (9)	3	4	
Correlation assessment metric			Correlations to qPCR
Bray et al. 2016 (10)	4	8	
Chandramohan et al. 2013 (11)	1	4	
Li et al. 2011 (12)	9	4	
Both			Correlations between “known truth” (published or simulated) and estimated abundances
Benjamin et al. 2012 (13)	1	4	
Li et al. 2015 (14)	56	3	

Table S2 Studies used in the database meta-assessment

SRA project ID	GEO series ID	Female samples	Male samples	Unspecified	Total samples
SRP033135	GSE52529			353	353
SRP027383	GSE48865			274	274
SRP051848	GSE64813			188	188
SRP042620	GSE58135			168	168
SRP041538	GSE57148			166	166
SRP011546	GSE36552			124	124
SRP044668	GSE59612			92	92
SRP056733	GSE67427			89	89
SRP042161	GSE57872			84	84
SRP033393	GSE52834			73	73
SRP050272	GSE63646			71	71
SRP028301	GSE49321		7	56	63
SRP051688	GSE64655			56	56
SRP029880	GSE50760			54	54
SRP045352	GSE60216			54	54
SRP041751	GSE57395			53	53
SRP043162	GSE58434			53	53
SRP055390	GSE66117			52	52
SRP041179	GSE56796			42	42
SRP021193	GSE46224			40	40
SRP030041	GSE50893			36	36
SRP042286	GSE57982			31	31
SRP016568	GSE41716			29	29
SRP018525	GSE44183			29	29
SRP049068	GSE62526			29	29
SRP042153	GSE57866			28	28
SRP041620	GSE57253			26	26
SRP022133	GSE46665			25	25
SRP041675	GSE57299			25	25
SRP051765	GSE64741	17		24	41
SRP033466	GSE52934			24	24
SRP033569	GSE53094			24	24
SRP043080	GSE58335			24	24
SRP046226	GSE61141			24	24
SRP047233	GSE61491			22	22
SRP043085	GSE58375			21	21
SRP032789	GSE52194			20	20
SRP041159	GSE56785			20	20
SRP027258	GSE48812		24	12	36
SRP050036	GSE63452	13		12	25
SRP041162	GSE56788		40		40
SRP051083	GSE64098	40			40
SRP040418	GSE56066	30			30
SRP042616	GSE58111	24			24
SRP045421	GSE60296	16		8	24
*SRP029262	GSE50244	35	54		89
*SRP043108	GSE58387	12	9		21
*SRP044917	GSE59810	9	9	9	27
*SRP043368	GSE58608	12	12		24
*SRP028336	GSE49379	15	15		30
*SRP045666	GSE60590	16	17		33
*SRP043221	GSE56787	19	19		38
*SRP035599	GSE54308	19	21		40
*SRP049593	GSE63055	30	27		57
*SRP047476	GSE61742	41	31		72
*SRP033566	GSE53080	9	35		44
*SRP026042	GSE47944	44	40		84

* Experiments used in the X-Y alignment assessment

Supplementary Text

Additional materials and methods

We ran STAR version 2.4.2a on an ENCODE dataset which consists of 17 samples from different cell lines. We used genome version GRCh38.p2 and GENCODE version 22 (15). The parameters changed were the minimum alignment score (minAS, parameter: --outFilterScoreMinOverLread), number of mismatches (numMM, --outFilterMismatchNmax), length of reads (lenR, --clip3pNbases), and read downsampling. The minimum alignment score was varied to range between 0.55 and 0.99. The number of mismatches allowed was varied to range between 0 and 10. Length of reads (trimming reads from the 3' end) was varied to range between 20 and 76. We downsampled reads with a local script by sampling across the mapped reads at random (95% to 50% by increments of 5%). RSEM version 1.2.28 was run to quantify the expression levels of the ENCODE dataset. We considered both FPKM and TPM.

To perform a network enrichment analysis, we first generate co-expression networks using all samples at each alignment parameter. Briefly, we calculate a weight between gene pairs by using the Spearman correlation coefficient which is then rank standardized. To then measure the information content of the network, we use the performance of the *n*-fold cross validation task of a neighbor voting algorithm. If we can hide known information about genes in a gene set and then “learn” this information from the network, then our network has, to a degree, information that is reflective of the known biology of that GO term. This is based on the "guilt-by-association” principle, which states that genes with shared functions should be connected preferentially in the network. The reported performance metric from this task is the averaged AUROC (area under the ROC curve) for each group across the n-folds. We used the Bioconductor package EGAD (16) and GO (17) to perform this analysis on the individual co-expression networks.

ENCODE dataset reproduces the parameter choice paradigm

We repeated all the same analyses on a second, dataset with fewer samples but greater depth, across a larger number of parameters within STAR, summarized in **Figure S10**. We first characterize the effects of the choice of parameter on the read depth and gene coverage. Even though read depth ranged between ~38M and 73M reads, gene coverage only changed between 14.5K and 15.5K (**Figure S10A**), with some parameters changing only a few hundred genes at most. We then calculate the replicability scores for each of the parameters. Under default parameters, most samples have good replicability scores (below 0, **Figure S10B**). As in the GEUVADIS dataset, we find very little effects on the replicability score (**Figure S10D**) across all the parameters. Although not significant, for most of the parameters the more stringent parameter (grey distribution in the violin plots), has better average scores and heavier negative tails.

Network analysis as an assessment metric

As the co-expression between gene pairs can be used to generate weighted gene-gene networks, we can perform a network analysis task that measures the amount of information in a network using a “guilt-by-

association" predictor of Gene Ontology (GO) annotations. As expected from the previous results, the average performance across all GO groups is very similar across the different parameter choices (average AUROC~0.61), and correlations across the individual GO groups near 0.88 (**Figure S11A-B**).The node degrees of the networks generated are also highly correlated (**Figure S11C-D**). Additionally, one could look at the change in co-expression of a gene to all other genes (e.g., *XIST* **Figure S11E**). For the approximately 30K transcripts, co-expression pairs remain highly correlated compared to the default minAS (average rs=0.90, **Figure S11F**). Protein-coding genes were also more correlated (average rs=0.94). The lowest correlations were again to the most conservative minAS, with scores per gene (minAS 0.66 vs 0.99, average rs=0.78).

Supplementary Figures

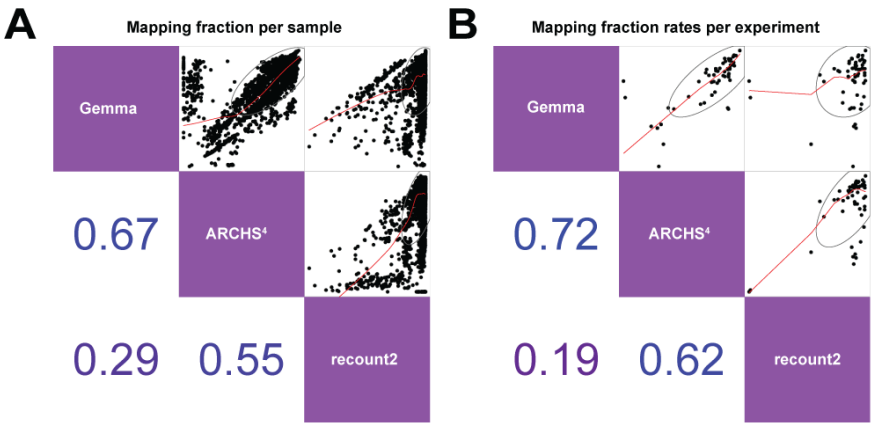

Figure S1. Meta-assessment across the three gene expression databases: Gemma, ARCHS[4] and recount2.

(A) Comparing fraction mapping rates per samples (B) and then averaged per experiment. There are some experiments that are outliers. Input reads differed mainly due to PE/SE counting and QC filtering that was not described which may have affected mapping rate calculations for some samples.

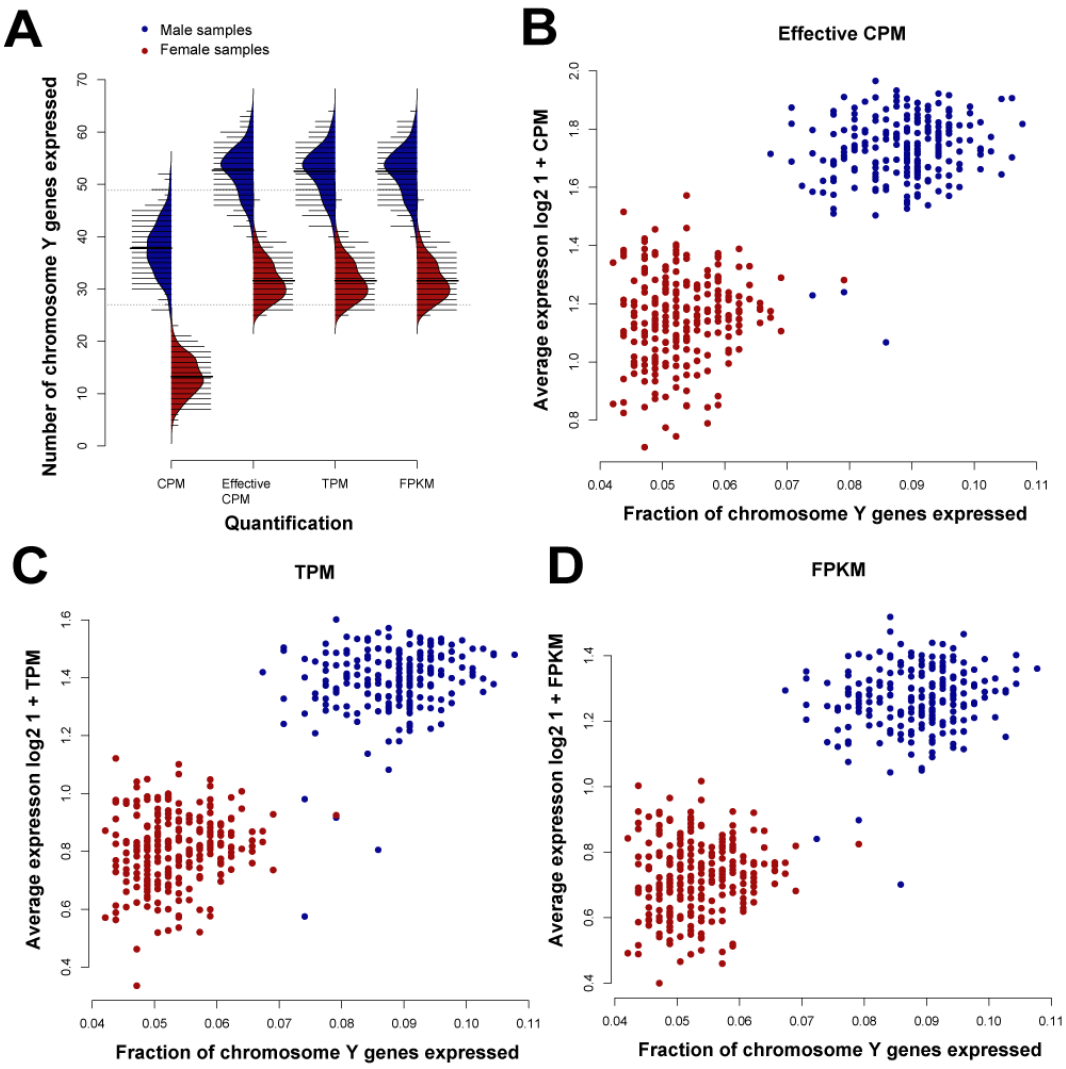

Figure S2. Quantification and X-Y gene misalignment.

(A) Quantification exacerbates the problem with Y gene expression in female samples. Here we used RSEM under default parameters. (B) Comparing fraction of Y genes expressed compared to the effective counts as reported by RSEM (C) for TPM (D) and FPKM.

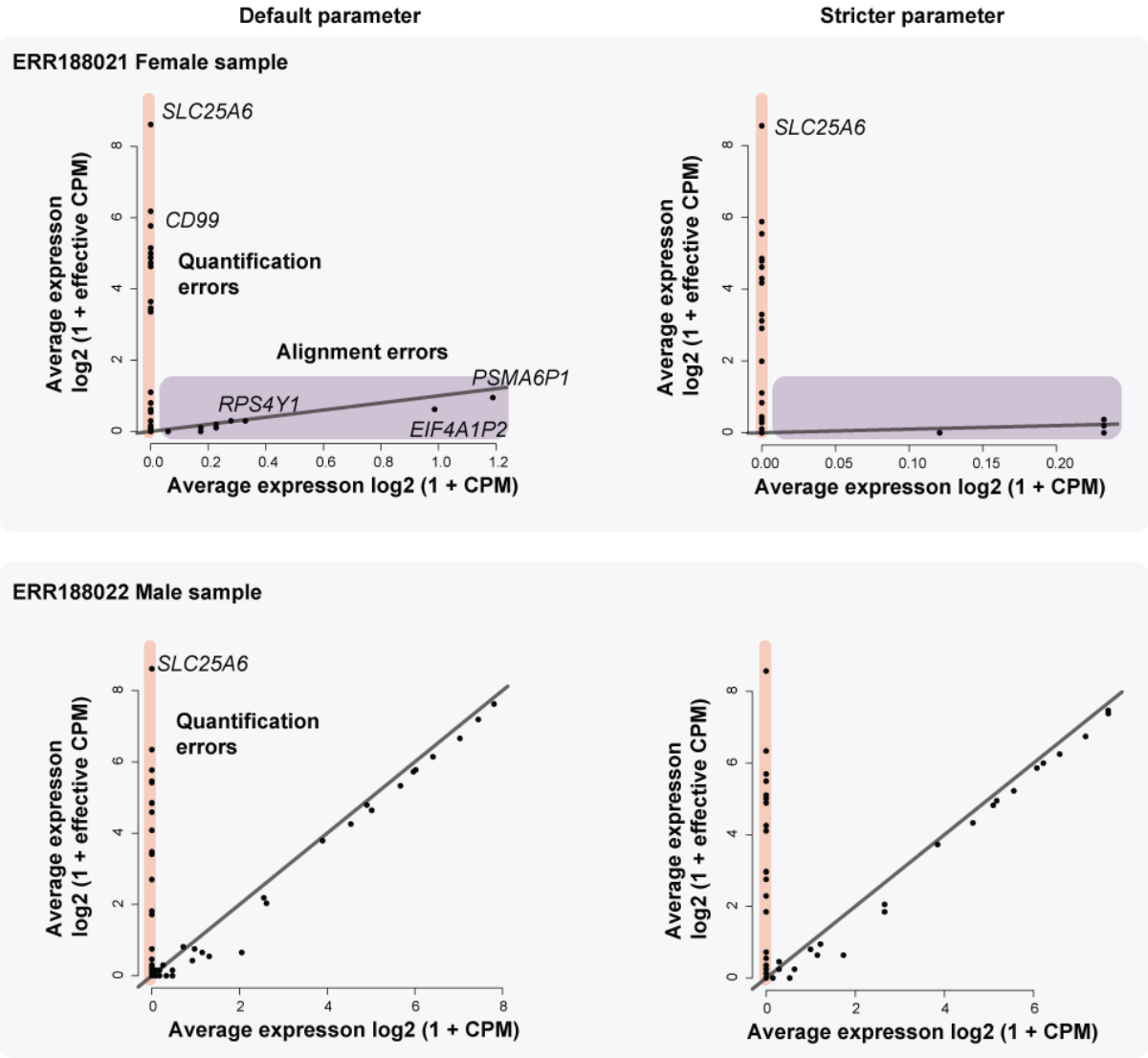

Figure S3. Distinguishing between alignment errors and quantification errors.

Comparing alignment CPM to Effective CPM distinguishes alignment errors from quantification based errors. Here we've shown a representative female and male sample at default parameters and a stricter parameter (minAS=0.99). There are genes that are not expressed (counts based) but appear as expressed once quantified (effective counts), labelled as quantification errors. Errors of alignment, on the other hand, appear as both counts and effective counts, and can be distinguished in the female samples here.

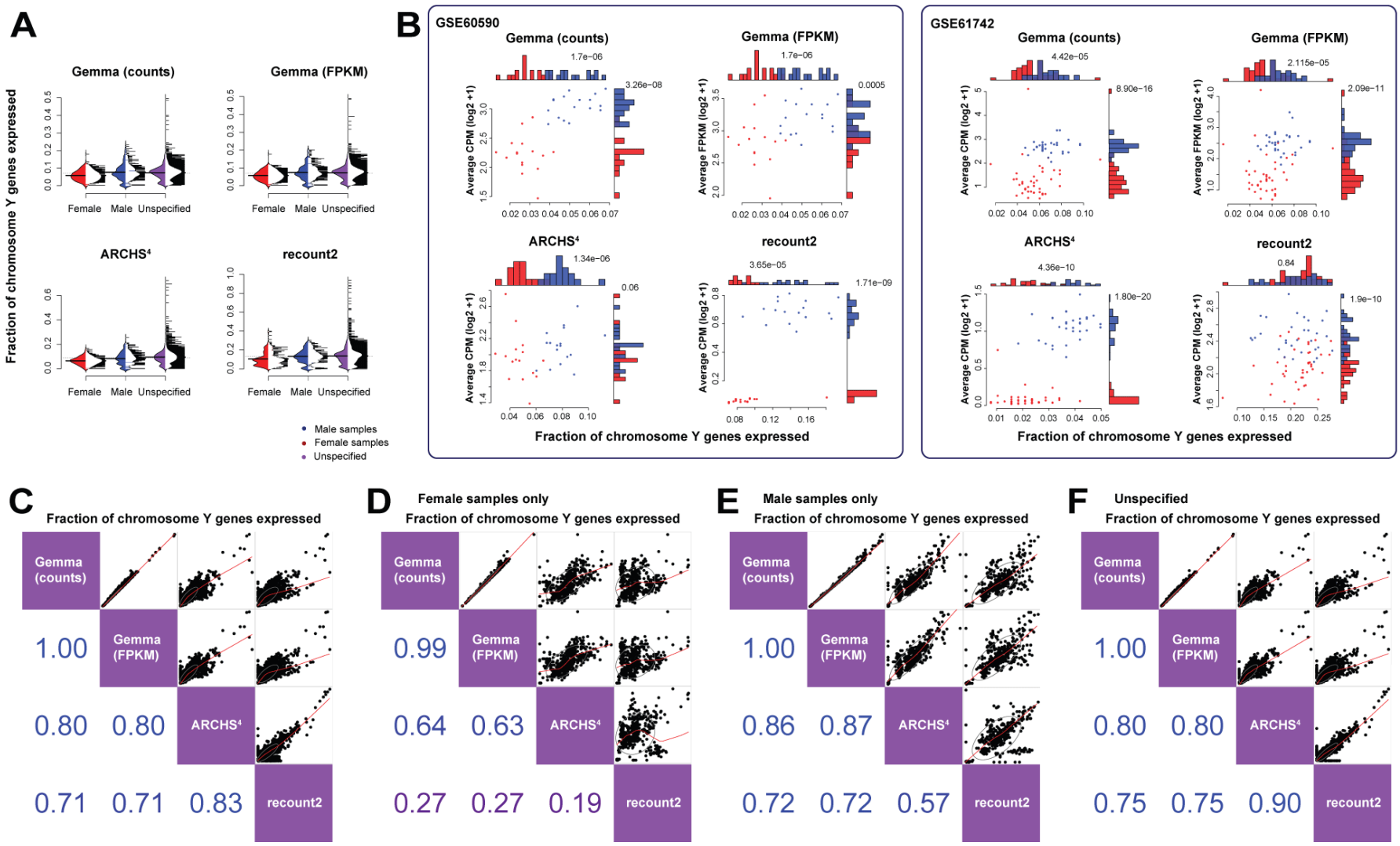

Figure S4. X-Y alignment assessment across three gene expression databases: Gemma, ARCHS^4 and recount2.

(A) Violinplots of the fraction of Y genes mapped by each of the four pipelines (in the three databases). These comparisons are once again between 3,405 samples in 57 experiments. Recount2 had a few samples missing from the analysis. (B) Two example experiments GSE60590 and GSE61742. (C) Comparison of the fraction mapped to Y genes for all samples (D) those labelled as female, (E) those labeled as male and (F) those unspecified.

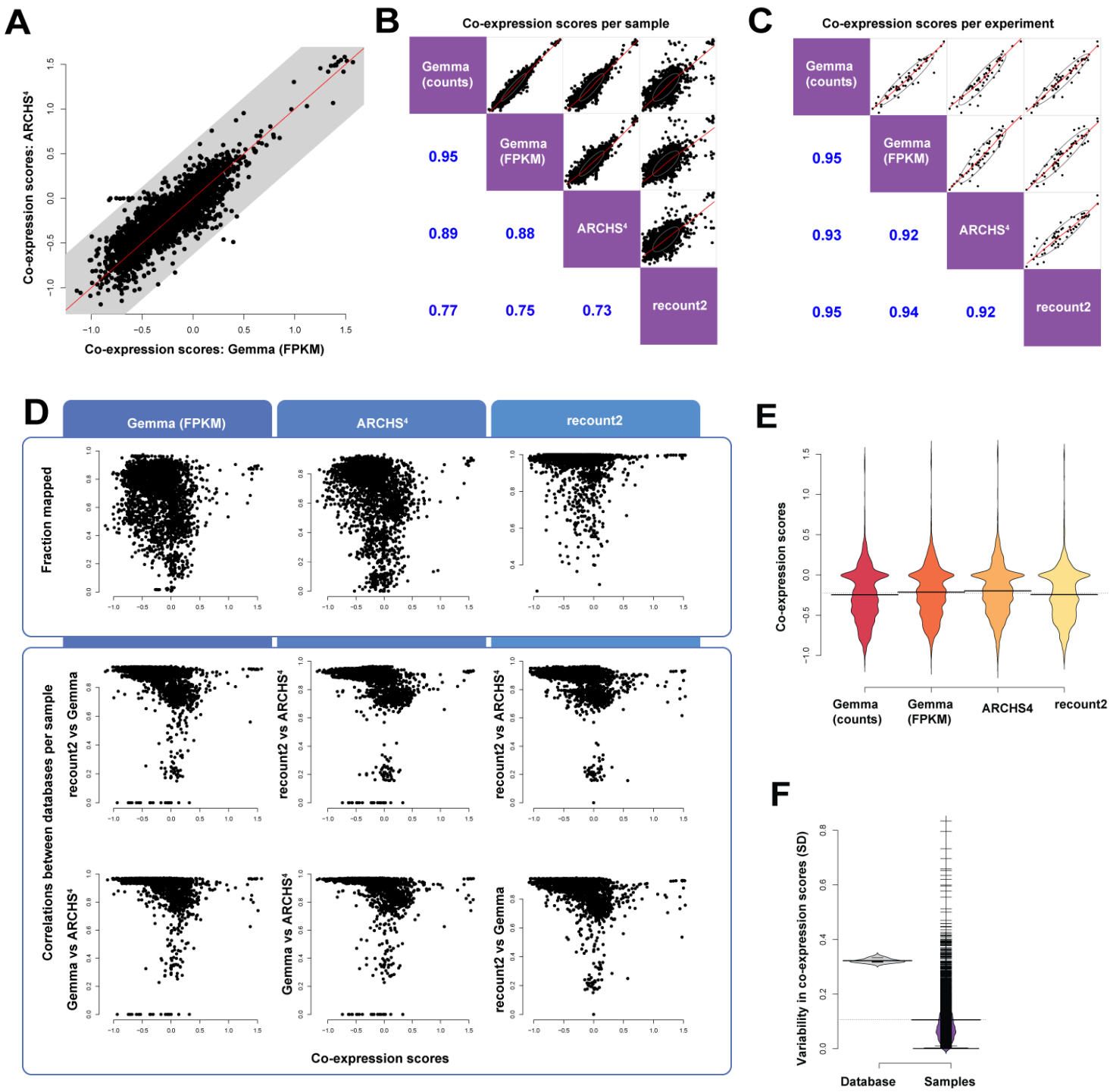

Figure S5. Co-expression scores across three gene expression databases: Gemma, ARCHS[4] and recount2.

(A) Comparison between Gemma and ARCHS[4] data (light grey is 1 SD=0.3 from the identity line, rs=0.88). (B) Correlations of co-expression scores for the four pipelines (in the three databases). These comparisons are between 3,405 samples. Recount2 had a few samples missing from the analysis. (C) Correlations of the scores once summarized (averaged) per experiment (57 experiments listed in Table S2). (D) Comparisons to fraction mapped metrics and correlations between databases (per sample). (E) Distribution of scores by database (F) and the variability of samples by database and samples across databases.

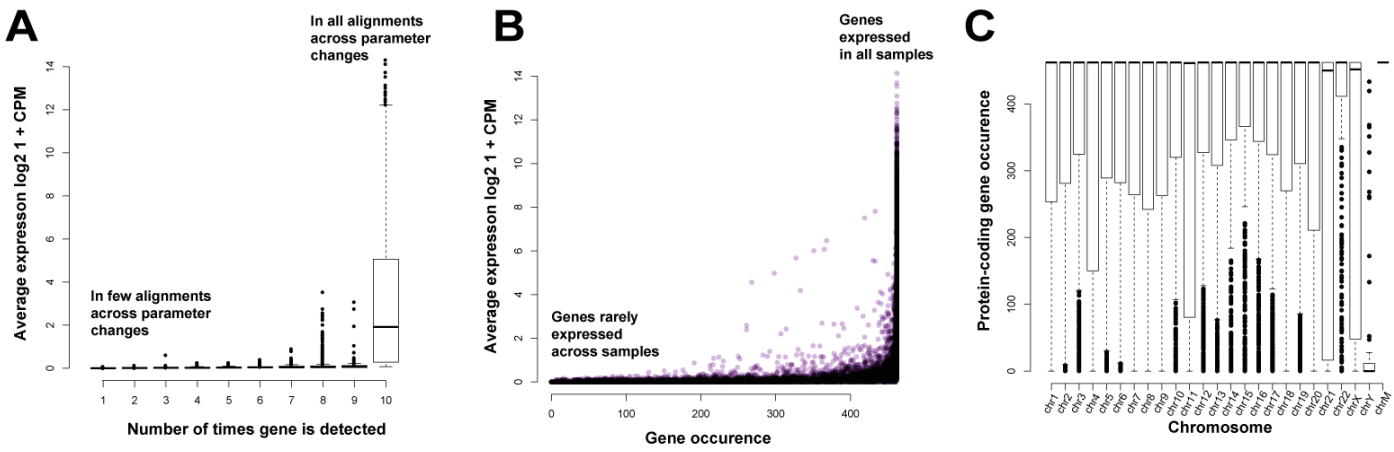

Figure S6. Gene detection differences and expression levels.

(A) Dropouts between parameters versus average expression, in an example sample (ERR188479). Most differences across parameters are from low expressing genes, with a few exceptions. (B) Occurrence vs expression for all samples at the default parameter (minAS=0.66). Most samples express most genes. (C) Occurrence across samples per chromosome for default parameter (minAS=0.66). Majority of dropouts between samples (bottom of boxplots) are the Y chromosome genes (medians are the lowest).

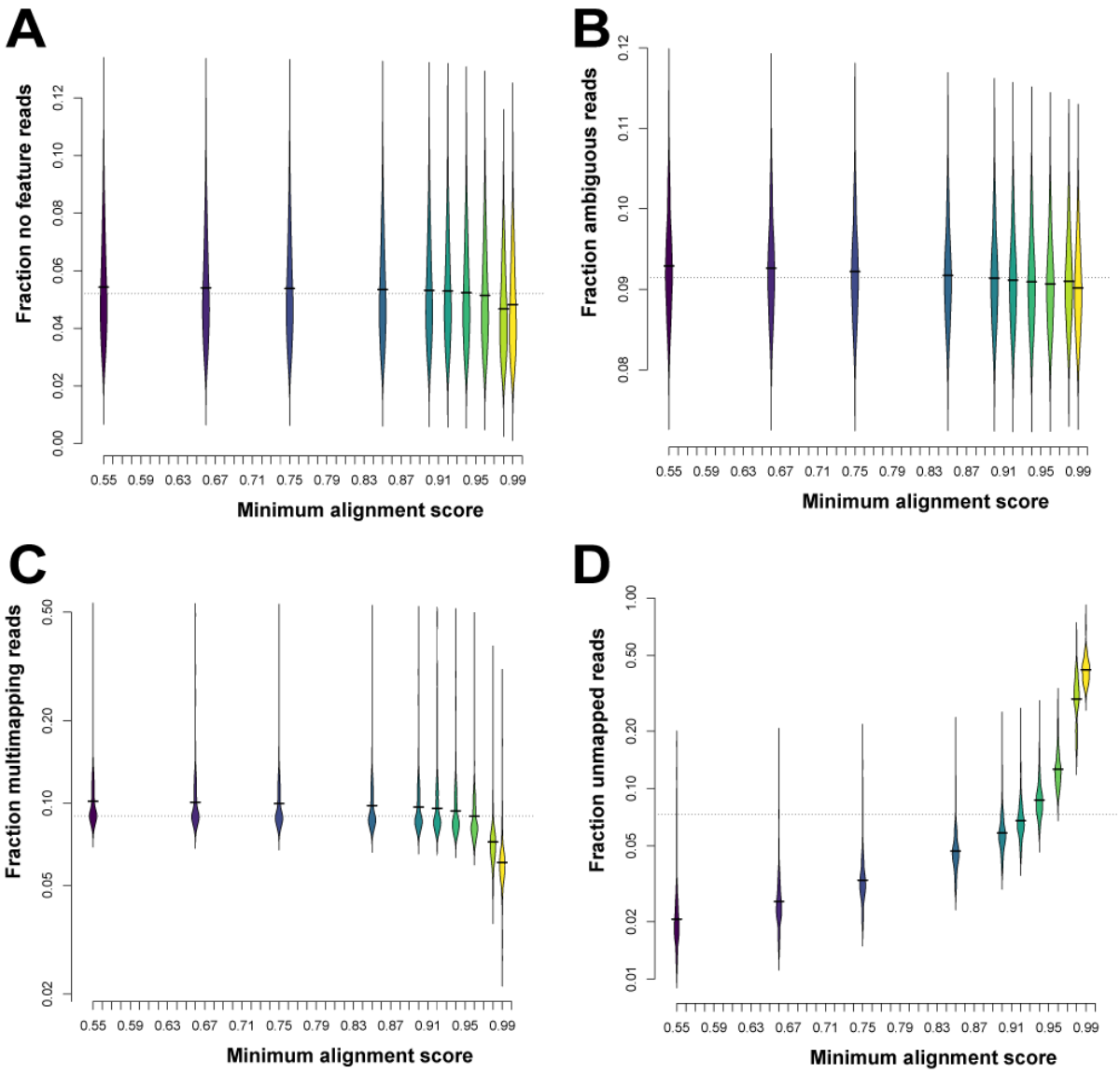

Figure S7. Other mapping statistics and metrics

(A) Fraction of uniquely mapped reads with no features. (B) Fraction of ambiguous mapped reads (cross feature boundaries). (C) Fraction of multimappers in total. (D) Fraction of unmapped reads.

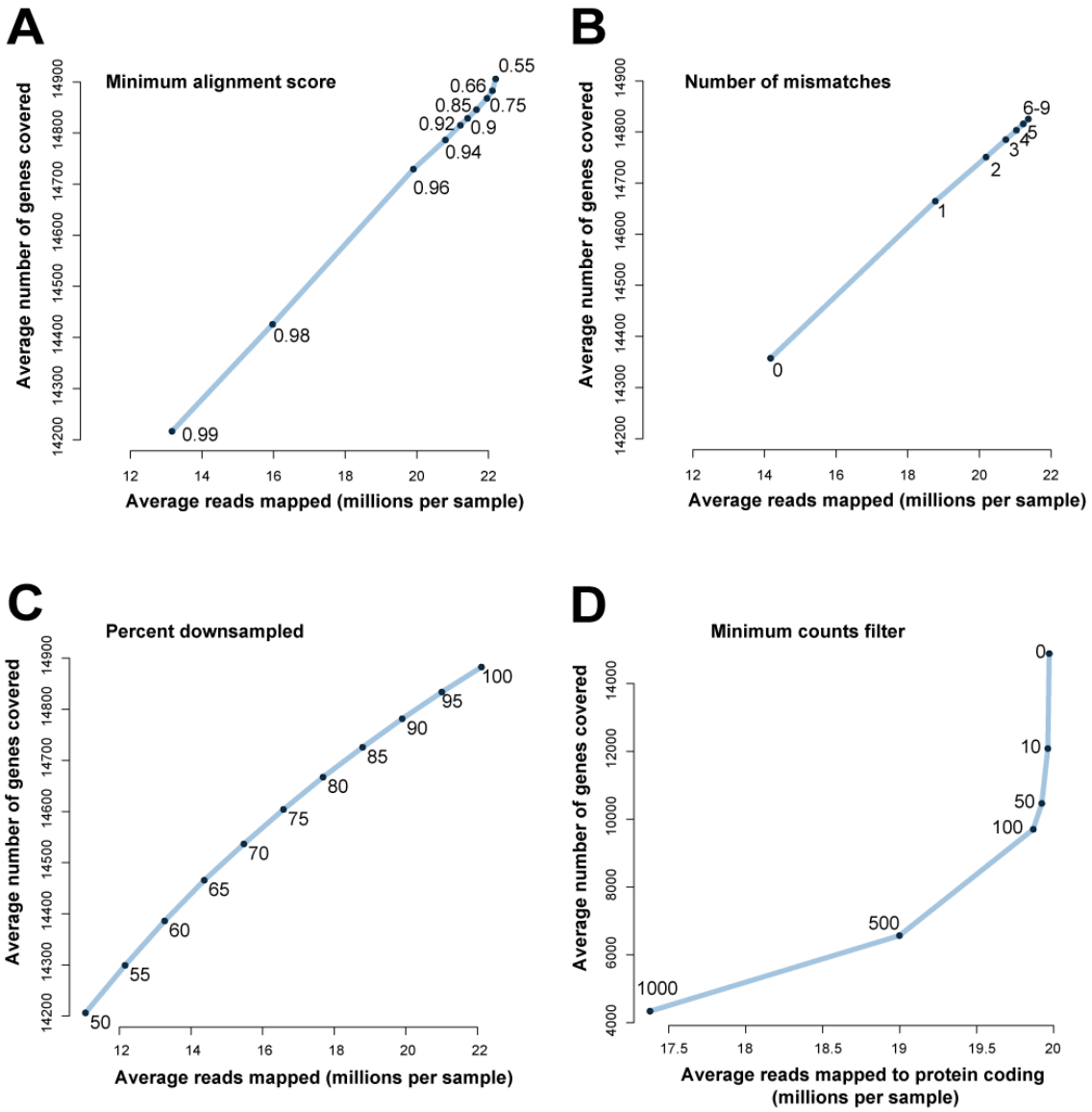

Figure S8. The effects of varying parameters on gene detection

Effects of gene detection on the GEUVADIS dataset when (A) varying minimum alignment scores, (B) number of mismatches, (C) downsampling and (D) filtering low expressing genes

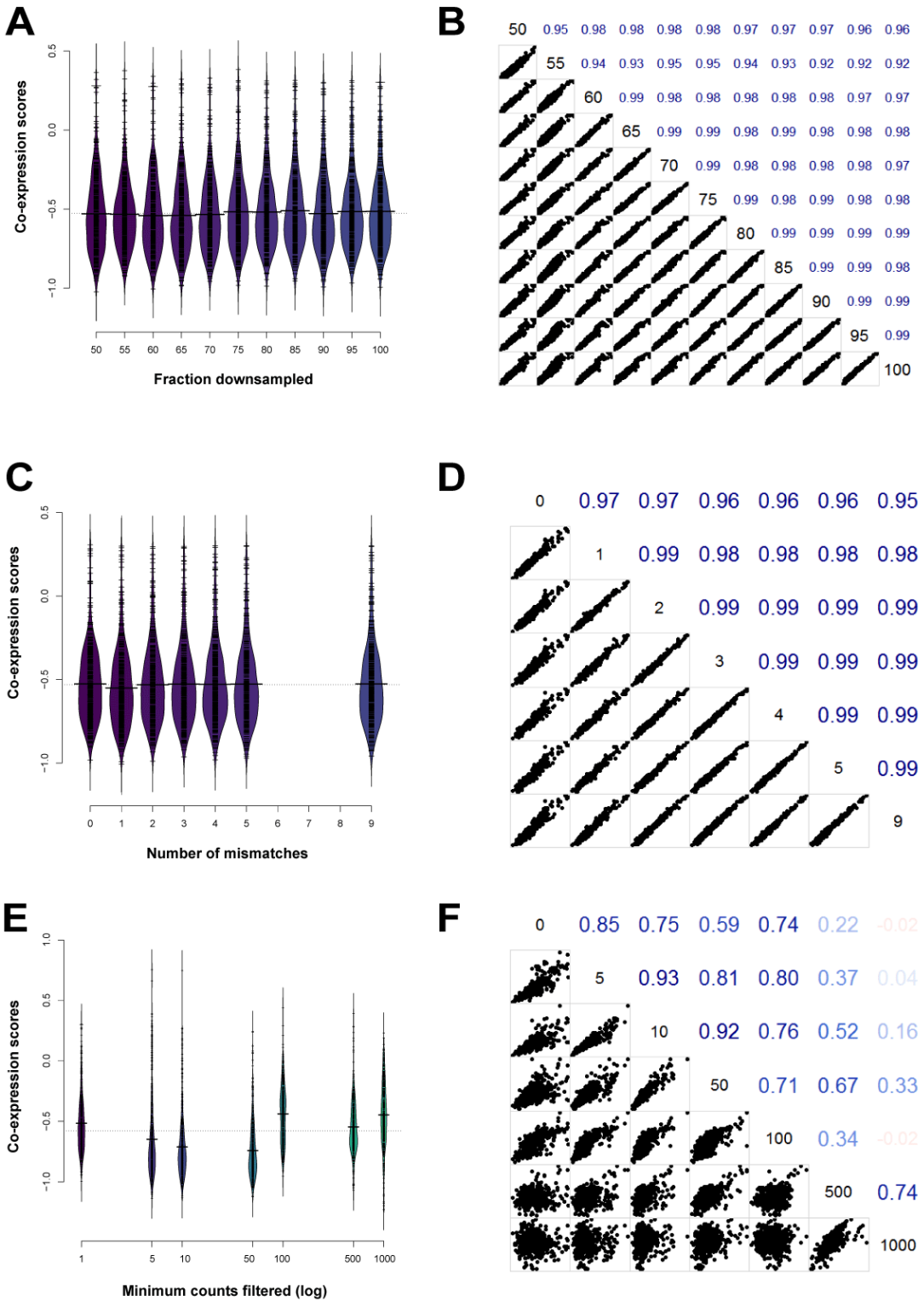

Figure S9. The effects of varying parameters on co-expression scores

(A) and (B) downsampling. (C) and (D) number of mismatches allowed. (E) and (F) counts filter.

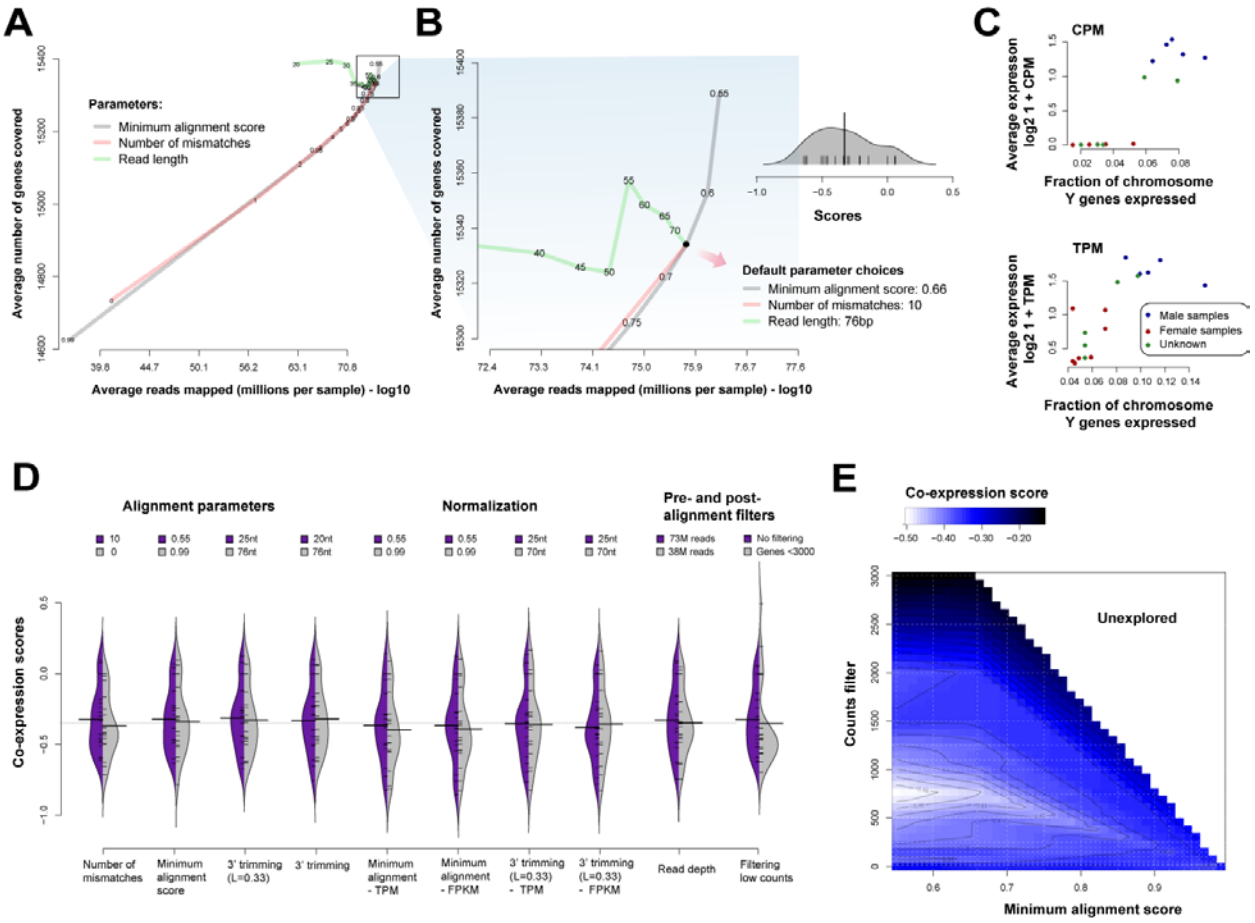

Figure S10. Varying the parameter space of STAR in an ENCODE dataset.

(A) Mapping statistics and coverage across the parameter space of STAR for the ENCODE dataset (B) Default parameter mapping statistics and inset showing the replicability scores. (C) XY misalignment is still a problem (CPM, top), that is increased with quantification (TPM, bottom). (D) Distributions of scores for the extreme parameters tested on an ENCODE dataset of 17 samples. The purple distribution shows the scores per sample of the most permissive parameter and the grey distribution the most stringent. As in the GEUVADIS dataset, the distributions are consistent showing little change, with the filtering of low counts showing the most change in distribution (last violin plot). (E) Interpolated co-expression scores showing alignment and post-alignment filter hotspots (light blue to white). Dark areas are least replicable. These results are averaged over all 17 samples for 1000 runs, and interpolated between the dashed lines. Contours define interpolated score boundaries. The unexplored regions could not be interpolated from tested data.

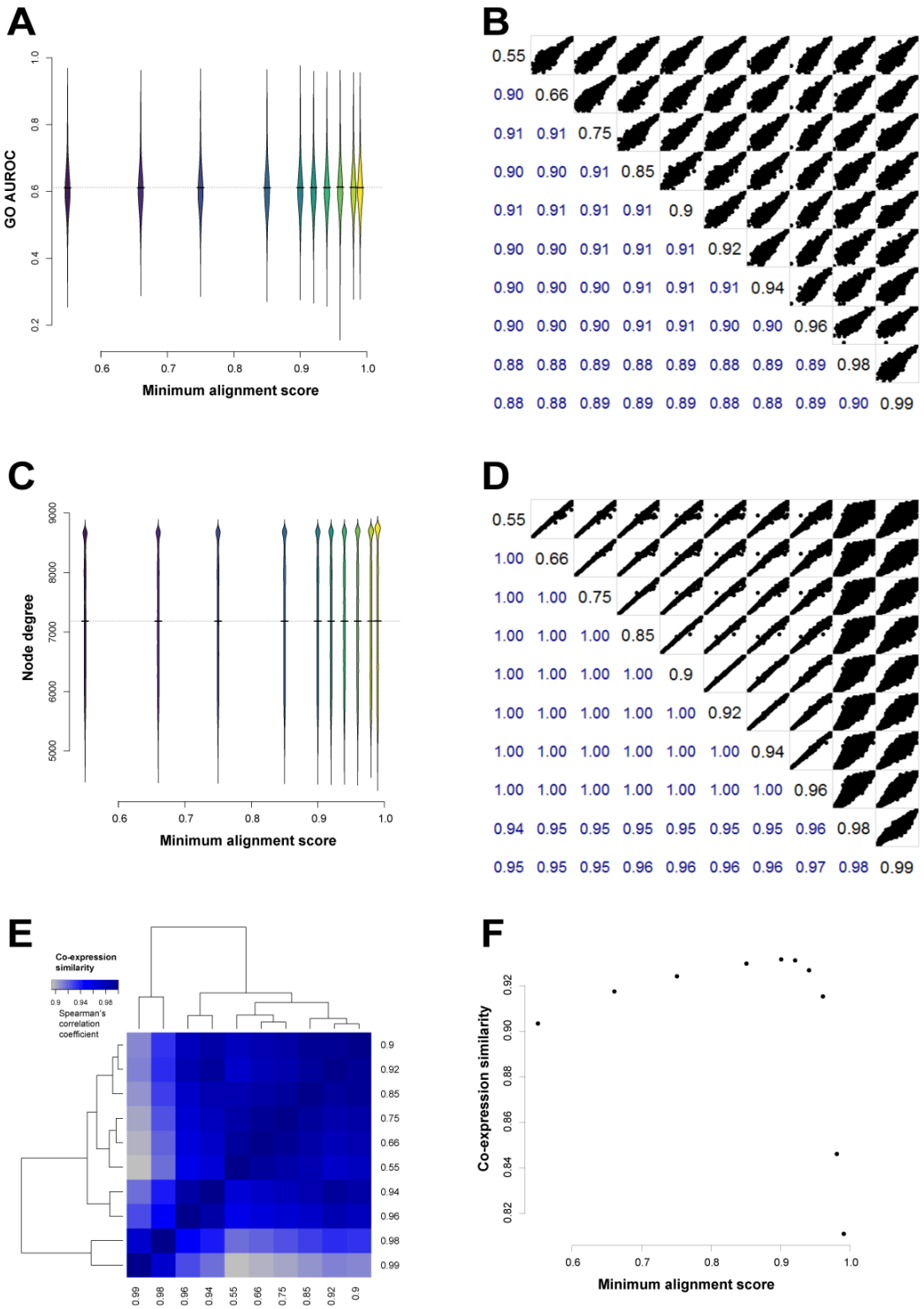

Figure S11. Parameter impact on downstream biological interpretation.

(A) Distribution of AUROC GO term scores (B) and their correlations. (C) Distribution of node degrees of network (D) and their correlations. (E) Similarity of co-expression values of *XIST* to all other genes. (F) Average similarity between co-expression values of all genes to all others.

Additional References

1. Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X. and Song, Y.-Q. (2011) Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet*, **56**, 406-414.
2. Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A. and Grant, G.R. (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Meth*, **14**, 135-139.
3. Bonfert, T., Kirner, E., Csaba, G., Zimmer, R. and Friedel, C.C. (2015) ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics*, **16**, 122.
4. Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, **14**, 671-683.
5. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*.
6. Engstrom, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., The, R.C., Ratsch, G., Goldman, N., Hubbard, T.J., Harrow, J. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Meth*, **10**, 1185-1191.
7. Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B. and Pierce, E.A. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518-2528.
8. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth*, **9**, 357-359.
9. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
10. Bray, N., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, **34**, 525.
11. Chandramohan, R., Wu, P.-Y., Phan, J.H. and Wang, M.D. (2013) Benchmarking RNA-Seq quantification tools. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, **2013**, 647-650.
12. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**, 323.
13. Benjamin, A.M., Nichols, M., Burke, T.W., Ginsburg, G.S. and Lucas, J.E. (2014) Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genomics*, **15**, 570.
14. Li, P., Piao, Y., Shon, H.S. and Ryu, K.H. (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, **16**, 347.
15. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760-1774.
16. Ballouz, S., Weber, M., Pavlidis, P. and Gillis, J. (2017) EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*, **33**, 612-614.
17. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25-29.