Genome Biology

METHODOLOGY

Open Access

CrossFilt: a cross-species filtering tool that eliminates alignment bias in comparative genomics studies of primates

Kenneth A. Barr[1] and Yoav Gilad[1*]

*Correspondence: gilad@uchicago.edu

[1] Department of Medicine, University of Chicago, Chicago, IL 60637, USA

Abstract

Comparative functional genomic studies are often affected by biased read mapping across species due to inter-species differences in genome structure, sequence composition, and annotation quality. We developed CrossFilt, a filtering strategy that retains only sequencing reads that map reciprocally between genomes, ensuring that quantification of read counts is based on directly comparable genomic features. Using both real and simulated RNA-sequencing data from primates, we show that CrossFilt outperforms five alternative approaches that are commonly used, resulting in more accurate inference of gene expression differences. Our results underscore how different preprocessing choices can shape downstream conclusions in cross-species functional genomics analyses.

Keywords: Cross-species RNA-seq, Read mapping bias, Differential expression, Comparative genomics

Background

The field of functional genomics has adopted sequencing-based technologies to measure cellular activity, such as gene expression, chromatin accessibility, methylation, splicing, and other molecular phenotypes. The first step in nearly any sequencing-based functional genomics study is to align sequencing reads to a reference genome to identify their location. In most studies, genomic annotations are used to categorize and contextualize the sequencing read counts. For example, reads that map to annotated coding regions may be counted and compared across samples to identify inter-individual differences in the gene expression levels.

For studies involving samples from the same species, the alignment and annotation steps are relatively straightforward. Although polymorphisms can affect the mapping probability of sequencing reads to specific genomic regions [1], effective approaches exist to address this issue [2], and overall, the sequences of different genomes are highly similar between individuals of the same species. Moreover, functional annotations and

context of the mapped reads are, by definition, identical between individuals of the same species, such that even errors in annotation do not bias the results of the comparisons, because they are shared across the entire sample.

Cross-species functional genomics comparisons introduce more nuanced challenges. In addition to the alignment and mapping steps, which must be done for multiple species, one must identify orthologous and syntenic genomic features [3]. For example, to meaningfully compare gene expression levels between humans and other primates, one must identify and annotate orthologous genes in the studied species. Once these orthologous and syntenic regions are identified – a task that is by no means trivial – one must also account for different read alignment properties across species. The problem is that if the probability of sequence alignment to annotated features differs between species (for example, because genes have different length in different species), it will introduce bias when comparing estimated quantities (e.g., gene expression levels) across species.

Alignment bias emerges when there is inter-species variation in the length, copy-number, or structural organization of annotated features. A genomic feature that is longer in one species than another, whether due to differences in exon count, repetitive sequences, or expansion of regulatory elements, provides a larger target for mapping sequencing reads. This can result in an artificially higher measured signal, even if the underlying molecular activity is identical. In non-comparative studies, standard normalizations can easily correct for length bias across features. In comparative studies, it is more difficult. For instance, using RNA-sequencing data, one might observe apparent higher transcript abundance in a species with longer exons, despite no actual difference in gene expression levels across species. Conceptually, correcting for differences in the length of corresponding features across species is straightforward if precise structural information is available. In practice, however, most comparative studies typically define corresponding regions across species without detailed information on potential differences in length and structural organization.

Inter-species differences in genomic duplications also contribute to alignment bias. For example, a chromatin accessibility assay could indicate lower read counts in an accessible region in one species simply because the corresponding enhancer is within a perfectly duplicated region elsewhere in the genome. Some regions exist in a single copy in one species but have duplicates in another. When this occurs, sequencing reads from duplicated regions may be removed from the alignment data due to mapping ambiguity, making it difficult to know whether observed inter-species differences reflect true changes in molecular activity or artifacts of mapping redundancy. This issue is particularly relevant for gene families that have expanded in certain lineages, as well as for conserved regulatory sequences that have been duplicated and repositioned across the genome [4, 5].

Differences in annotation quality between genomes add further complication. Some species, such as human and mouse, have highly curated genome annotations that comprehensively define genes, transcripts, and regulatory elements [6]. Other genomes are often annotated using incomplete or lower-confidence annotations. When species-specific annotations are projected from one genome coordinate system to another, researchers may inadvertently compare a well-annotated genome to one with missing features, mistaking absent annotations for true biological differences. This issue is particularly problematic for regulatory sequences, which are often annotated based on

indirect evidence (such as sequence conservation or chromatin state) and may be underrepresented in species with less extensive functional genomics data.

To address these challenges, we developed CrossFilt (Cross-species Filtering Tool), an alignment-based method that mitigates mapping biases by excluding sequencing reads that do not map reciprocally between species. In this study, we applied CrossFilt to both real and simulated RNA-sequencing data from humans, chimpanzees, and rhesus macaques [7] to identify differentially expressed genes between the species. Our aim was to assess how different alignment and feature definition strategies influence the outcomes of comparative genomics analyses.

Compared to five alternative methods commonly used for cross-species sequencing data, CrossFilt reduced the number of false positives and improved the consistency of inferred gene expression differences between species. By restricting analysis to reciprocally mappable reads, CrossFilt ensures that feature quantification is based only on directly comparable genomic regions. More broadly, we demonstrate that the methodological choices made during data processing can substantially affect the set of molecular differences identified between species, with some existing and commonly used approaches producing markedly inflated false positive rates.

Results

CrossFilt, a novel read filtering strategy for comparative genomic studies

CrossFilt is a cross-species filtering approach that is designed to facilitate unbiased comparisons of sequencing-based data obtained from different species. It uses a reciprocal liftover [8] strategy to retain only the sequencing reads that map accurately and consistently to the genomes of each species (see Additional file 1: Fig. S1). In the following vignette, we describe CrossFilt as applied to a typical case: a pairwise comparison of RNA-sequencing data between two species (e.g., human and chimpanzee). Here, the 'target' and 'query' genomes respectively correspond to the human and chimpanzee reference genomes. CrossFilt also requires genome annotations for each species but is agnostic to their source, allowing flexible integration with any annotation set of choice.

CrossFilt's reciprocal liftover approach relies on externally available chained alignments between species, which are needed to accurately convert coordinates from the genome of one species to the other. A chain is a sequence of pairwise alignments that are arranged in a single order and orientation; by definition, this arrangement must be consistent with the order of genomic sequence in both species (thus, chains do not include large-scale translocations, duplications, or inversions). The CrossFilt process begins by aligning sequencing reads derived from the target samples to the target reference genome. For each aligned read, CrossFilt uses the reference genome coordinate system to identify its start and end positions. It then retrieves all chains that fully encompass the read (see Methods). If more than one chain fully encompasses a single read, CrossFilt selects the chain with the highest UCSC alignment score [9], which reflects the length and quality of the alignment, favoring long, gap-free regions with few mismatches. Next, each sequence of nucleotides in the read that matches the target genome is replaced with the corresponding sequence from the query genome, accounting for insertions and deletions. CrossFilt excludes reads if their start or end coordinates coincide with insertions or deletions in the query genome.

The lifted reads are then converted into FASTQ format and independently realigned to the query genome using the query species' annotations. A custom script (*crossfilt-filter*) verifies that the realigned reads match exactly with the lifted BAM alignment in terms of start position, read name, and concise idiosyncratic gapped alignment report (CIGAR) string. Reads that fail this verification are discarded. The reads passing this step are reciprocally lifted back to the target genome, and the script *crossfilt-filter* is used again to confirm that each read maps precisely to its original genomic location. Only reads successfully passing both reciprocal liftover steps are retained.

The next step is to define orthologous features between the two species. As mentioned, CrossFilt is compatible with any number of methods for orthology inference, whether based on sequence identity, predictive models, or external data. As our study includes data from humans, chimpanzees, and rhesus macaques, we elected to use the Comparative Annotation Toolkit (CAT, [8]), which provides comprehensive annotations for all three species (see Methods). After separately mapping reads from each species to the orthologous features in the corresponding reference genome, CrossFilt keeps only the reads that map consistently to the same genomic feature in both species. Reads mapping inconsistently between species are removed. Through this rigorous reciprocal filtering process, CrossFilt effectively mitigates alignment and annotation biases, ensuring robust comparative analyses.

Commonly used strategies for comparing sequencing-based data across species

Comparative functional genomics studies have used different strategies to define orthologous features and quantify molecular phenotypes across species using sequencing-based data. These strategies differ in three key respects: (***i***) whether sequencing reads from all species are aligned to a single reference genome or to species-specific genomes, (***ii***) how functional annotations are assigned to define genomic features such as enhancers, genes, or exons in each genome, and (***iii***) whether steps are taken to correct for read mapping biases introduced by sequence divergence, structural variation, or differences in annotation quality across genomes of different species.

Approaches that align all reads to a single genome bypass the need to define orthologous features explicitly but are prone to alignment bias when mapping efficiency is affected by sequence divergence between the species (e.g., when chimpanzee-derived reads are aligned to the human genome). In contrast, approaches that align reads to the corresponding genome of each species (e.g., reads from human samples are aligned to the human genome, and reads from the chimpanzee samples are aligned to the chimpanzee genome) require definition of orthologous sequences or functional features, either by name matching or via formal annotation transfer methods. Additional bias correction strategies such as read- or feature-level filtering, further distinguish methods in terms of their robustness and reliability (as we show in our analysis below).

We compared CrossFilt to five commonly used approaches that span these design choices. We refer to these approaches as single-reference, consensus, dual-reference, annotation transfer, and orthologous exons. Below we briefly describe each approach and highlight their primary differences. A comparison of relevant features of these approaches is available in Table 1.

Table 1 Summary of compared approaches

Approach/method	Using a single genome as reference, or the genomes for all species considered in the study	Human annotation source	Alternative source for annotation (or annotation transfer) in other species	Method applied to correct for annotation or mapping bias
Single-reference	Single genome	Ensembl	None	None
Consensus	Single genome	Ensembl	None	Masking differences between genomes with *N*
Dual-reference	The genomes for all species in the study	Ensembl	RefSeq	TPM
Annotation transfer	The genomes for all species in the study	Ensembl	Ensembl transferred with CAT	None
Orthologous exons	The genomes for all species in the study	Ensembl	Ensembl transferred with CAT	Only include exons with high sequence identity across species
CrossFilt	The genomes for all species in the study	Ensembl	Ensembl transferred with CAT	CrossFilt read filtering based on reciprocal chained alignments

Single-reference

In this approach, used by [10] to build a single-cell atlas of primate brain development, sequencing reads from multiple species are aligned to the genome of a single species, and the corresponding functional annotation (e.g., Ensembl [11]) is used. No attempt is made to correct for mapping bias or annotation mismatches. This approach is straightforward and does not require ortholog definitions, but the alignment is sensitive to sequence divergence between species.

Consensus

Used by [10, 12, 13], a synthetic reference genome is created by masking all nucleotide differences between the studied species with 'N's, to create a reference genomic sequence. Reads from all species are then aligned to this masked consensus genome using a single genomic annotation set. This removes alignment bias at divergent sites but does not address broader structural differences or annotation mismatches. The use of a single annotation also avoids the need to explicitly define orthologous features.

Dual-reference

In this approach, used by [14–23] and others, reads from each species are aligned to their respective reference genomes. In a comparative study involving human and chimpanzee samples, for example, it would be reasonable to use the best available annotation set for each species, even if they differ in terms of gene structure and completeness. For instance, human reads could be aligned to the Ensembl-annotated human genome, while chimpanzee reads could be aligned to the panTro6 assembly using RefSeq annotations. Transcripts Per Million (TPM) normalization is typically

applied to normalize gene expression levels, as this accounts for differences in gene length between samples. Orthology can then be defined implicitly, by matching gene names across species. This approach is simple but may be biased when annotation quality differs substantially between species.

Annotation transfer

Used by [24–27] and others, annotations from one genome are transferred to the other species using genome alignment tools such as CAT [28] or TOGA (Tool to Infer Orthologs from Genomic Alignments) [29]. This produces consistent gene models across species and avoids the limitations of low-quality species-specific annotations. However, no further steps are taken to address alignment bias, so mapping artifacts may still contribute to false positives.

Orthologous exons

Several studies, including from our group, have used a more conservative variant of annotation transfer [30–35] in order to reduce alignment and annotation bias. In the 'orthologous exons' approach, exons are filtered after the annotation transfer step to ensure that only highly comparable features are included in the analysis. For example, one might retain only the exons with high sequence identity (e.g., > 85% for a comparison of human and chimpanzee) and minimal length differences (e.g., < 10%) between species.

CrossFilt

Our method, CrossFilt, uses a reciprocal alignment strategy to filter reads prior to quantification. Reads are retained only if they map precisely and consistently to both species' genomes after accounting for insertions, deletions, and mismatches. This is combined with ortholog definitions derived from annotation transfer (e.g., CAT annotations). By filtering at the read level and not just at the feature level, CrossFilt minimizes both alignment and annotation bias, yielding reliable cross-species comparisons.

How did we compare CrossFilt with other approaches?

To systematically compare these approaches against CrossFilt, we evaluated all six methods in the context of a comparative gene expression study using RNA sequencing data from human, chimpanzee, and rhesus macaque. Our goal was to assess the performance of each approach not only in terms of the number of genes tested, but also the quality and reliability of the resulting inferences, focusing on identifying differentially expressed genes between the species (Additional file 1: Table S1). While maximizing gene inclusion is desirable, doing so without accounting for differences in annotation quality or mapping efficiency may introduce bias. The most effective methods are therefore those that strike an optimal balance between inclusiveness and accuracy.

We structured our comparison around three distinct analyses. We began by assessing the extent and symmetry of genome and transcriptome coverage across species for each method. These metrics reflect both the number of genes that can be reliably tested and the completeness of their annotations. For example, an annotation set with more transcripts in one species than another may lead to asymmetric read mapping, potentially biasing estimates of gene expression (because when using RNA sequencing data,

gene expression is a relative measurement [36, 37]). We evaluated coverage in two ways: First, by comparing the number of genes that passed method-specific filtering criteria and would be included in each pipeline. Each method specifies different criteria for inclusion. For example, the 'single-reference' method is the most inclusive in this context, as it uses all annotated human genes for the comparison. Other methods include filtering criteria such as 'minimum sequence identity' or 'length similarity' (required by the'orthologous exons' method), or 'reciprocal mapping' (required by CrossFilt). Second, we measured the degree of annotation completeness across species, using the proportion of each gene's sequence that can be covered by aligned RNA-sequencing reads. For example, in the 'consensus' approach, the number of 'N' sequence replacements in certain genomic regions is large enough to exclude unique read mapping to these regions.

Next, to evaluate the accuracy and robustness of each method, we simulated RNA-sequencing data with known inter-species differences in gene expression. This allowed us to directly measure statistical power and false discovery rates across approaches, as well as to detect alignment and mapping induced species-specific biases in apparent gene expression levels or effect size estimates. We performed differential expression analysis and compared the proportion of true and false positives identified using each method. We assessed how well each method recovered the simulated inter-species differences in gene expression, and quantified bias in the direction and magnitude of false positive signals.

Finally, we applied each method to real data, using previously collected bulk RNA-sequencing data from human, chimpanzee, and macaque tissues [7]. Using consistent quality control and normalization procedures across all pipelines, we identified differentially expressed genes in pairwise species comparisons and evaluated consistency across methods (because, using real data, we do not have a gold standard set of differentially expressed genes). Specifically, we assessed the overlap in differentially expressed gene sets, the concordance of effect size estimates, and the inferred sharing of inter-species differentially expressed genes across tissues. We also performed gene set enrichment analysis to determine whether methodological differences influenced downstream biological interpretation. By examining patterns of gene sharing, effect size correlation, and enrichment among functional categories, we evaluated how methodological choices impact the interpretation of interspecies differences.

This three-part evaluation framework allowed us to compare the strengths and limitations of each approach and identify sources of bias that may distort downstream analyses. In the sections that follow, we present the results of these comparisons and highlight the distinct advantages of CrossFilt relative to existing approaches.

Comparison of genomic and genic coverage across approaches

The number of genes retained for differential expression analysis varied substantially across approaches (Fig. 1A, B; Additional file 1: Table S2). The 'single-reference' and 'consensus' approaches use complete, unfiltered human genome annotations. The 'single-reference' approach thus includes nearly all genes annotated in the human genome (38,584 genes). The 'consensus' approach includes slightly fewer genes (Fig. 1C), with the exact number depending on the threshold used to define a gene's 'mappability'. For example, if we use 50% sequence masking as the cutoff beyond which of a gene is

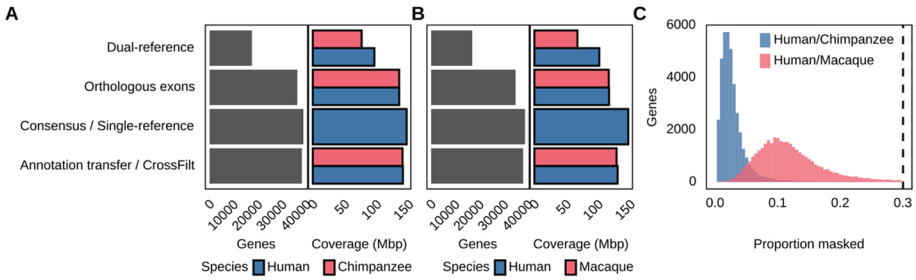

Fig. 1 Genomic and gene coverage. **A** The number of genes and megabases of genomic coverage for each annotation approach when comparing human and chimpanzee RNA expression. **B** The number of genes and megabases of genomic coverage when comparing human and macaque RNA expression. **C** Histogram of the proportion of each gene masked by 'N' when using a consensus genome for the indicated species comparisons, excluding genes with greater than 30% masking, indicated with a dashed line (488 genes from the human-chimpanzee comparison and 2,575 from the human-macaque comparison)

considered unmappable, 316 and 1,163 genes would be excluded from the human-chimpanzee and human-macaque comparisons, respectively. These masked regions reflect sites of sequence mismatch (divergence), insertion, or deletion between species, and are therefore expected to contribute to alignment bias in the 'single-reference' approaches. As expected, given the evolutionary distance between species, we found that genes in the human-macaque consensus genome contain many more masked positions (14.7%) than the human–chimpanzee genome (3.6%). This observation explains why divergence-driven alignment bias is expected to be more severe in more distantly related species.

The 'dual reference' approach resulted in the smallest set of comparable genes (17,443 in the human-chimpanzee comparison and 16,885 in the human-macaque comparison) due to the more limited annotations available for the chimpanzee and macaque genomes. We also observed substantial inter-species differences in read coverage under this approach (Fig. 1A, B). Notably, the 'dual reference' method differs from the others in that orthologous genes are inferred by matching gene names between species. As a result, it retains primarily protein-coding genes, since other transcripts often lack standardized gene names across species and are excluded. The other approaches ('annotation transfer', 'orthologous exons', and CrossFilt) use sequence-based definitions of orthology that do not depend on gene names. The 'annotation transfer' and 'orthologous exons' methods performed better in terms of gene inclusion, though the 'orthologous exon' approach excluded a larger number of genes due to its strict sequence similarity filters. In our implementation of CrossFilt, we used CAT for comparative annotations, which resulted in identical gene counts as 'annotation transfer'.

Assessing differential expression accuracy using simulated data

To evaluate how alignment and annotation choices influence differential expression (DE) analysis, we simulated bulk RNA-sequencing data from human, chimpanzee, and rhesus macaque tissues (heart, liver, lung, and kidney). Using gene-specific expression means and dispersion estimates derived from empirical data we collected in a previous study [7], we simulated the expression of 34,693 genes in humans and chimpanzees and 31,807 genes in humans and macaques in four biological replicates per species and tissue combination (see Methods for details). In each pairwise species comparison, we randomly

selected 10% of genes to be differentially expressed, with half showing higher expression in human and the other half in the comparison species (simulated inter-species log fold-changes were set to 2). We then applied each of the six approaches to align and process the simulated reads, followed by DE testing using a standardized limma-voom framework (see Methods for implementation and comparison to DEseq2). When we used the 'dual-reference' method, we applied TPM normalization instead of counts-per-million (CPM) to account for gene length differences between species.

We identified differentially expressed genes using each method at a 5% false discovery rate (FDR) and assessed the quality of the binary classification (Additional file 1: Figs. S2-S4). With respect to the true proportion of false positives (true false discovery rate; tFDR), all six methods performed as well or better in the human–chimpanzee comparison than in the more divergent human–macaque comparison (Fig. 2A, B; Additional file 1: Fig. S5). CrossFilt exhibited the strongest overall performance, achieving the lowest empirical false discovery rate (tFDR of 4.0% in both species comparisons). The 'consensus' approach ranked second in accuracy by these metrics (tFDR of 5.4% and 8.9% in the comparisons of human-chimpanzee and human-macaque, respectively). Among the remaining approaches, performance varied depending on the comparison. For example, in the human–chimpanzee comparison, the 'single-reference' approach produced fewer

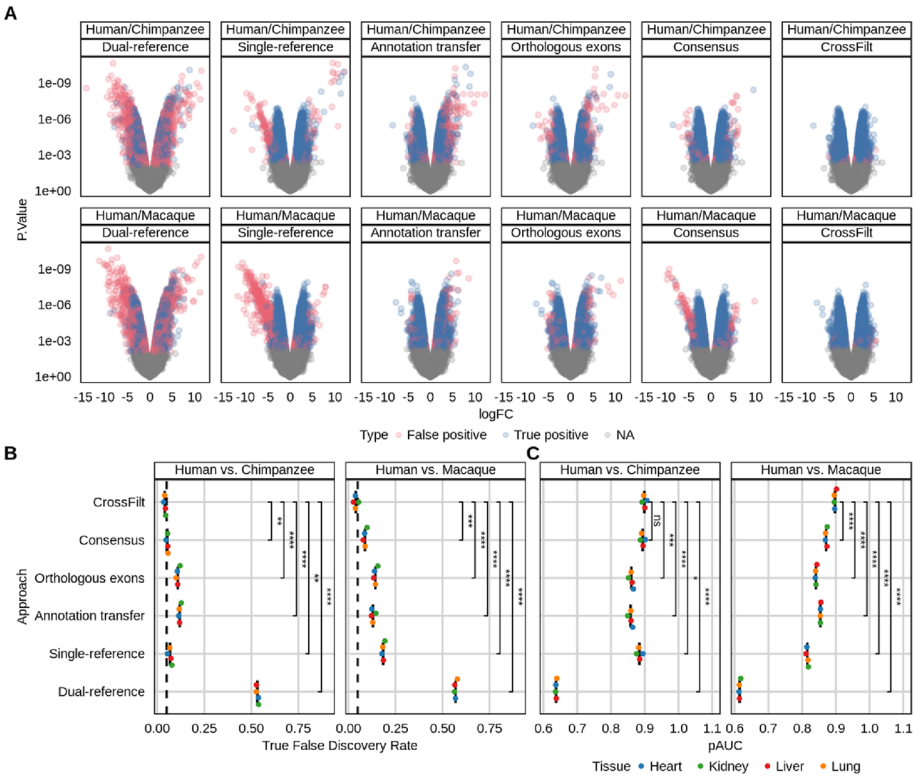

Fig. 2 Simulation performance. **A** Volcano plots for each comparison and approach. Genes with FDR ≤ 5% are highlighted in red (false positives) and blue (true positives). Positive values of logFC indicate genes higher expression in humans relative to the comparison species. **B** The true number of false positives among genes with FDR ≤ 5% in each approach, tissue, and species comparison. *P*-values are from Student's *t*-test. **C** Area under the partial receiver operator characteristic (pAUC) of specificities between 1 and 0.95 in each approach, tissue, and species comparison. (ns = not significant; * = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$; **** = $P < 0.0001$)

false positives (tFDR of 6.8%) than the 'annotation transfer' or 'orthologous exons' methods (tFDR of 12.0% and 10.9%, respectively). However, in the human-macaque comparison, the 'single-reference' approach resulted a marked increase in false positives (tFDR of 18.5%). The 'dual reference' approach, which is commonly used in comparative genomics studies in primates [14–23], resulted in the worst performance in both comparisons, with tFDR values of 53% for human-chimpanzee and 57% for human-macaque.

The relatively stringent read-level filters in CrossFilt could lead to fewer reads aligning to the transcriptome, potentially resulting in lower sensitivity. When we inspected the percent of simulated reads aligned after CrossFilt filters, the loss relative to other approaches was modest (Additional file 1: Fig. S6). Compared to 'annotation transfer' – which had the highest fraction of aligned reads – CrossFilt had 2% and 5% fewer aligned reads in the chimpanzee and macaque comparison, respectively. In the more distant human-macaque comparison, CrossFilt retained more reads than the 'consensus', 'single reference', and 'dual reference' approaches.

We assessed the relative sensitivity of CrossFilt by calculating the partial area under the receiver operator characteristic (pAUC) at specificities between 100 and 95%. CrossFilt performance was on par with the 'consensus' approach in the human-chimpanzee comparison, and significantly better than any other condition, indicating that CrossFilt had the highest sensitivity within a range of acceptable sensitivities (Fig. 2C; Additional file 1: Fig. S7). Compared to other approaches, CrossFilt missed significantly fewer true positives than the number of false positives it correctly rejected (Additional file 1: Fig. S8).

We also evaluated each approach for quantitative errors or biases that are not captured by binary classification metrics. We found that effect size estimates from CrossFilt deviated less from their simulated values than any other method (Fig. 3A). Additionally, we observed species-specific asymmetry in the number, effect sizes, and expression levels of DE genes (Fig. 3B; Additional file 1: Figs. S9-S11). The 'orthologous exon', 'annotation transfer' and 'dual-reference' methods classified significantly more genes as more highly expressed in humans in at least one species comparison. The average effect sizes of false positives in CrossFilt were lower than every approach except 'dual reference' (Fig. 3C). This means that false positives are less likely to show up at near the top of gene-sets when ranked by either effect size (Fig. 3D) or significance (Additional file 1: Fig. S12).

Taken together, our simulation results highlight the susceptibility of some methods, particularly those that do not correct for alignment bias, to elevated false discovery rates and systematic distortions in estimated effect sizes. CrossFilt consistently showed the best overall performance, suggesting that reciprocal read filtering substantially improves the reliability of DE inference in cross-species sequencing-based comparisons (Additional file 1: Table S3).

Assessing differential expression accuracy using real RNA sequencing data

To evaluate how methodological differences affect downstream biological conclusions, we also applied all six pipelines to real bulk RNA-sequencing data from human, chimpanzee, and rhesus macaque tissues [7]. Following our approach for analyzing the simulated data, each pipeline incorporated consistent quality control and normalization

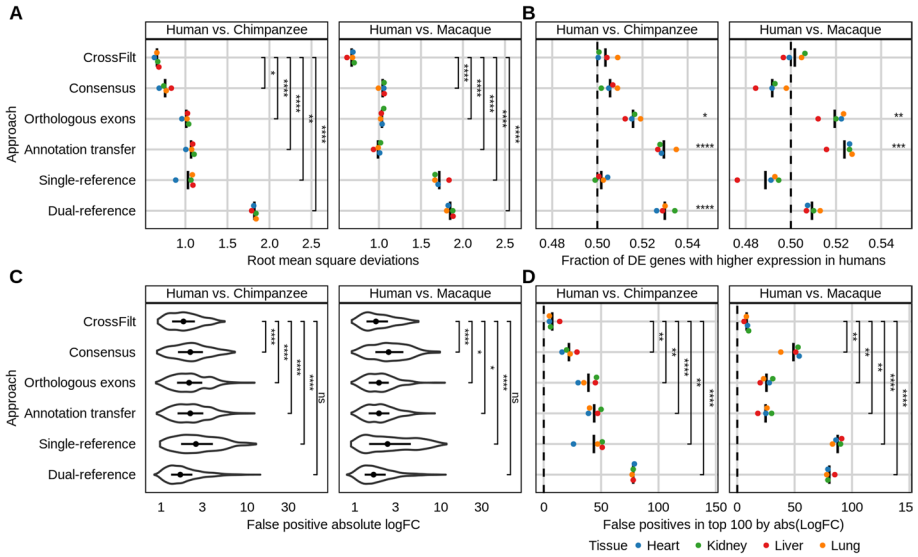

Fig. 3 Quantitative Bias. **A** The standard deviation in the error of effect size estimates for genes with FDR ≤ 5%. *P*-values from Student's *t*-test. **B** The fraction of DE genes (FDR ≤ 5%) with logFC higher in humans relative to the comparison species. *P*-values from binomial test. **C** The distribution in absolute logFC for false positives in each approach. *P*-values are from one-sided Wilcoxon rank sum test. **D** The number of false positives in the top 100 genes in each approach, ranked by absolute logFC. *P*-values from Student's *t*-test. (ns = not significant; * = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$; **** = $P < 0.0001$)

steps (see Methods and Supplementary Information); however, in this case, we do not have a gold standard comparison. Therefore, to assess the concordance of DE results, we first examined the effect sizes of genes identified as DE using at least one approach (at FDR ≤ 5%). We compared these DE genes to those identified using all other approaches, considering each inter-species and inter-tissue comparison independently. In general, effect size estimates were consistent across approaches, with DE genes clustering primarily by tissue (Fig. 4A; Additional file 1: Fig. S13). Consistent with the results from the simulated data, the 'dual-reference' approach resulted in the least agreement with the other pipelines (mean R = 0.48 vs. mean R = 0.79 for all other approaches). We ruled out TPM normalization as the cause of this discrepancy (Additional file 1: Fig. S14) and concluded that alignment and annotation asymmetries specific to the 'dual-reference' method contribute to this pattern. This effect was most pronounced in the human–chimpanzee comparison, where all other approaches perform more similarly.

Across tissues, 68.2% of inter-species DE genes identified by the 'dual-reference' pipeline were not identified as DE by any other method (Fig. 4B). Given the results of the analysis of the simulated data, we considered the DE genes identified only when the 'dual-reference' pipeline is used to be false positives. Although the proportion of false positives was lower in the human–macaque comparisons relative to the human-chimpanzee comparisons, the 'dual-reference' false positive rate was always higher than for any of the other methods. When we considered a comparison of DE genes identified using all other approaches, we noted that, overall, fewer than half of the DE genes were identified by every approach (Fig. 4C), highlighting the profound impact of the choice of alignment and annotation strategy on DE inference.

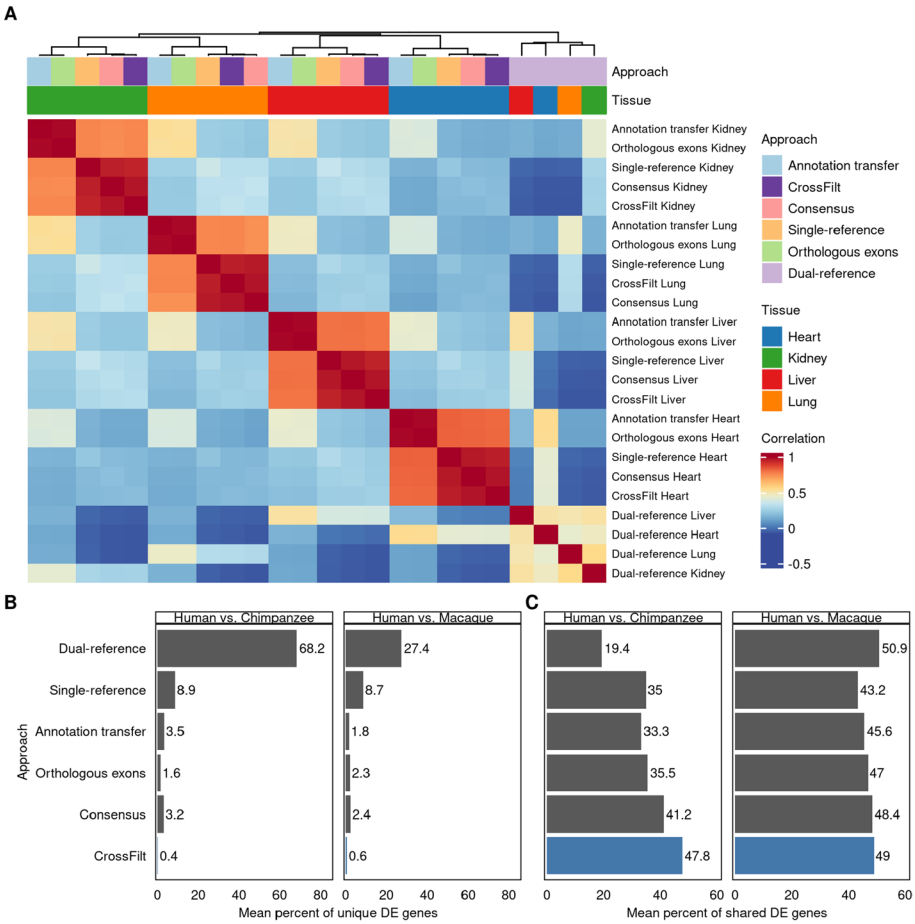

Fig. 4 Concordance of DE gene sets in real data. **A** Heatmap of the Pearson correlation in human-chimpanzee effect size estimates for genes detected as DE in at least one approach and tissue. Order of rows and columns is set by hierarchical clustering. **B** The percent of DE genes identified through each approach that are not classified as DE by any other approach. **C** The percent of DE genes identified through each approach that are also classified as DE by all other approaches

We next used mash [38] to estimate the extent of DE gene sharing across tissues. Because alignment bias should be independent of tissue type, approaches affected by such bias are expected to overestimate the number of interspecies DE genes shared across tissues. Consistent with this expectation, the 'dual-reference' approach resulted in the identification of the highest proportion (75.3% in human-chimpanzee and 27.9% in human-macaque) of inter-species DE genes shared across tissues (Fig. 5A). Using CrossFilt resulted in the lowest estimate (52.4% in human-chimpanzee and 18.7% in human-macaque) of shared interspecies DE genes across tissues. In general, genes identified as interspecies DE in multiple tissues, but not by CrossFilt, were highly enriched for genes identified as false-positives in our simulated data (Fig. 5B).

Finally, we assessed the extent to which methodological differences affect downstream biological interpretations by performing gene set enrichment analysis. Using ranked *t*-statistics for each method, we tested for enrichment of DE genes among Gene Ontology biological process categories. While most gene sets produced broadly concordant enrichment scores across pipelines, we identified cases where alignment artifacts may

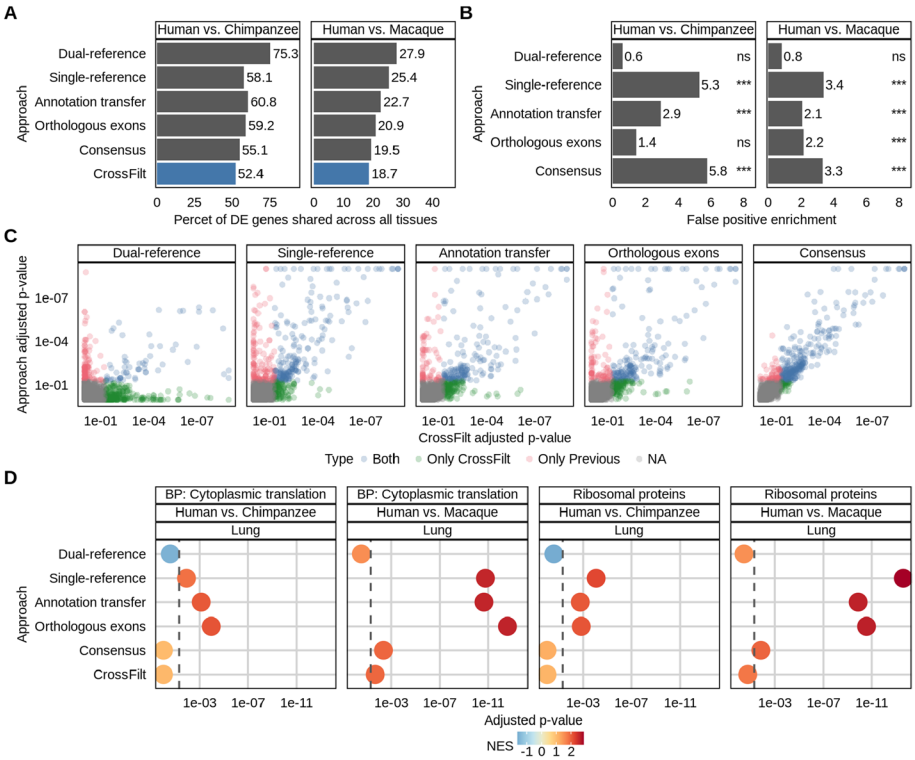

Fig. 5 Cross-tissue sharing of DE and gene-set enrichment in real data. **A** The percent of DE genes with shared effects across all four tissues, estimated by MASH. **B** Enrichment of false positives in simulated data for genes that are identified as shared by the indicated method, but not CrossFilt, relative to genes identified as shared by both approaches. *P*-values are from one-sided Fisher's exact test (ns = not significant; *** = $P < 0.001$). **C** Gene-set enrichment (GSEA) significance of Gene Ontology: Biological Process terms from the indicated approach (y-axis) and CrossFilt (x-axis). Terms are colored based on whether the term was significant in only the indicated approach (red), only CrossFilt (green) or both (blue). **D** GSEA significance of the term 'Cytoplasmic translation' and ribosomal proteins (RPS and RPL genes) in lung. Points are colored based on normalized enrichment score (NES)

have inflated statistical significance (Fig. 5C). For example, the category 'cytoplasmic translation' was significantly enriched among genes identified as inter-species DE using the 'single-reference', 'annotation-transfer', and 'orthologous-exons' pipelines. Further inspection revealed that this enrichment was driven by a group of ribosomal genes (Fig. 5D), where alignment errors may be more likely due to the large number of processed pseudogenes of ribosomal proteins in primates [39]. When we removed these genes from the category the term was no longer enriched (Additional file 1: Fig. S15).

Taken together, these analyses underscore that methodological choices influence not only which genes are identified as differentially expressed, but also how consistent those results are across contexts (in our case, tissues) and how they are interpreted in a functional context. CrossFilt's design minimizes alignment and annotation bias and avoids many of the artifacts observed in alternative methods.

Discussion

Comparative functional genomics studies have consistently faced the challenge of alignment bias. Whether subtle or pronounced, alignment bias can subsequently affect nearly every step of sequence-based quantification across species. Despite widespread

recognition of this problem and many attempts to mitigate its impact (including our prior work using the 'orthologous exons' approach [40]), residual bias remains a persistent source of false positive identification of interspecies differences, and distorted biological interpretations. As our results demonstrate, CrossFilt substantially improves the accuracy of cross-species quantification by removing sequencing reads that cannot be mapped unambiguously between species.

While some comparative studies have adopted methods to reduce alignment artifacts, many continue to rely on pipelines that are fundamentally flawed. 'Single-reference' and 'dual-reference' approaches, for example, remain in use despite their susceptibility to a high degree of alignment and annotation asymmetries. These approaches tend to inflate expression differences where none exist, particularly when genome annotations differ in quality or completeness between species. There are also many comparative studies that addressed this problem through targeted feature filtering (limiting analyses to orthologous exons with high sequence similarity is one such example), but as we show, CrossFilt offers a more comprehensive and general solution by addressing the issue directly at the level of read mapping.

The importance of resolving alignment bias extends beyond generating a more accurate list of differentially expressed genes between species. The primary value of comparative functional genomics lies in its ability to identify general biological patterns and molecular mechanisms that distinguish species, cell types, or environmental responses. Biases in read alignment and feature annotation can systematically distort this inference, leading to incorrect conclusions about how molecular phenotypes differ across species. For example, in studies of tissue-specific expression, apparent inter-species similarities in gene regulation across tissues may be overestimated because alignment artifacts will be common to the analysis of data from all tissues. Similarly, when comparing gene expression responses to treatment, diet, or environmental exposures, biased mapping can obscure or exaggerate shared species-specific effects and prevent us from finding exposure- or treatment-specific patterns. In all of these contexts, correcting for alignment bias is necessary not only for precision, but for biological validity. Furthermore, functional insights based on gene ontology or pathway enrichment analysis may also be susceptible to alignment bias, because gene function is often correlated with evolutionary constraint. As a result, divergence, and therefore the degree of mapping bias, can vary systematically across functional categories, potentially leading to spurious enrichment signals or the masking of true biological patterns.

The problem is not specific to gene expression. Although we chose RNA sequencing data as the motivating example, and genes as the primary features of interest, the same limitations apply to any sequencing-based molecular phenotype used in a comparative framework. Whether the goal is to compare transcription factor binding, chromatin accessibility, splicing, or other read-based measurements across species, differences in sequence composition, structure, or annotation can lead to asymmetric mapping and biased quantification. CrossFilt addresses this issue in a general way. It does not depend on a particular annotation strategy and can be applied to any analysis involving orthologous genomic features as long as genome sequences and alignment chains are available (see Additional file 1: Fig. S16 and Methods for application to ATAC-seq).

Because CrossFilt operates at the read level and is annotation-agnostic, it can be used flexibly with any method of ortholog definition, including curated annotations, projection tools, or inference-based pipelines. It is compatible with both protein-coding and noncoding features, and readily integrates into existing differential analysis workflows.

Finally, CrossFilt provides a conceptual framework for improving data quality in comparative genomics. As functional genomics expands to incorporate more species, cell types, and molecular phenotypes, the demand for robust, bias-aware preprocessing methods will continue to grow. CrossFilt can help ensure that downstream comparisons are meaningful, that differences reflect biology rather than technical artifacts, and that conclusions drawn from cross-species data are genuinely comparative in nature.

Conclusions

Our findings underscore the importance of methodological choices in comparative genomics. Without careful control for genome assembly and annotation differences, comparisons based on sequencing data can be misleading, attributing biological significance to technical artifacts. By providing a robust and scalable approach for cross-species analysis, CrossFilt enhances the accuracy of comparative genomics and helps clarify the true molecular differences between species.

Methods

RNA-sequencing data

We previously collected 50 bp single-end RNA-sequencing data from human (*Homo sapiens*, all of reported Caucasian ancestry), chimpanzee (*Pan troglodytes*), and Indian rhesus macaque (*Macaca mulatta*) tissues (heart left ventricle, kidney cortex, liver, and lung), which we used to comparatively analyze gene regulation between the species [7]. The data include 48 samples in total, with four individuals per species and four tissues per individual; samples were pooled and sequenced on 26 lanes of four different flow-cells on the Illumina HiSeq 2500 platform [7]. To prepare the data for the current study, we downloaded FASTQ files from the Sequence Read Archive (accession numbers SRR6900765 through SRR6900812) and used TrimGalore (v0.6.10; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), a wrapper based on Cutadapt (v2.6; [41]) to remove sequencing adapters from the reads. We trimmed using a specificity of 3.

Simulated RNA-sequencing data

We simulated RNA-sequencing data to match the structure of the empirical data [7]. We used the mean expression level for each gene in each tissue in the human dataset as a baseline for simulated gene expression levels in both species. We then fit dispersion estimates for every gene using the estimateDisp function in edgeR v4.4.2 [42, 43]. We simulated differential expression for 10% of genes. For 5% of the genes, we increased the mean expression level in humans by factor of 4 (corresponding to a log fold change of 2). For the other 5%, we increased expression in the alternate species by a factor of 4. We then used mean and dispersion estimates to generate data for each species and tissue type. Using the rnbinom function in R, we generated four biological replicates for each species under a negative binomial model. We then simulated reads

from the longest orthologous transcript each of these genes (defined by CAT) using polyester v1.29.1 in R [44].

We also tested simulations where we increased the percent of differentially expressed genes to 20% and 30% of genes. We also tested selecting 1000 random DE genes in each tissue from genes previously reported as DE in [7]. Results for these simulations are presented in Additional file 1: Figs. S17-S18.

Reference genomes, annotation sets, and chain files

We used the hg38, panTro6, and rheMac10 genome builds from the UCSC genome browser [45] as reference genomes for each species. For each pipeline, we performed genome alignments using STAR 2.7.11b [46]. We built each reference with the sjdbOverhang parameter set to 49 to match the 50 bp read-length in our data. For the human annotation set, we used the 10X Genomics Cellranger 2024 GEX annotation set (built from Ensembl v110) [11]. Annotation sets for chimpanzee and rhesus macaque vary according to the approach used, as described below (see also Additional file 1: Table S4). We aligned reads to each of these STAR references using the flags –outSAMmultNmax 1 and –outFilterMultimapNmax 1. We also used –quantMode GeneCounts to generate read-counts for each gene. We obtained the chain files hg38ToPanTro6.over.chain.gz, panTro6ToHg38.over.chain.gz, hg38ToRheMac10.over.chain.gz and rheMac10ToHg38.over.chain.gz from the UCSC genome browser.

CrossFilt

The results in this manuscript were generated with CrossFilt v0.1.2 (available at https://github.com/kennethabarr/CrossFilt, https://pypi.org/project/crossfilt/, and https://anaconda.org/bioconda/crossfilt) [47]. CrossFilt aims to answer the question: If this target read originated from the query species, would it still align to the same feature? Answering this question requires changing the sequence of the read from the target species to the query species. While tools exist that will implement liftOver on bam files [48–50], none of these tools will change the sequence of the read. This functionality is provided by the script crossfilt-lift, which will identify every sequence in a read that matches that of the target genome, and will convert the sequence to that of the query genome. It will update read position, FLAG, and CIGAR string and read sequence.

In more detail, for each read, we identify the start and end coordinates of the read and return all chains that fully encompass that read. Then, for every sequence of nucleotides in the read that matches the target genome, we replace that sequence with the query sequence in that chain, including any insertions, deletions, or more complex variants. We reject any reads whose start or end coordinate is an insertion or deletion in the query genome, as well as any reads with insertions larger than 300 bp. If there are insertions, we use the quality of the last sequenced nucleotide to fill missing data in the quality string. We update strand information and CIGAR strings to reflect the new coordinates in the query genome. If a read successfully maps in multiple chains, we only return the read from the highest-scoring chain. These scores are provided by the chain file, and reflect the quality of alignment, where long, gap-free alignments with few mismatches have higher scores.

In order to filter reads from the target genome alignment, we first lift the reads to the query genome using the target to query chain file. We then convert the lifted bam to FASTQ format using SAMtools v1.22 [51] and we realign the lifted reads in the query genome. We then use crossfilt-filter to check that these aligned reads have the same reference start, name, and CIGAR string as they did in the lifted BAM alignment. We exclude all reads that either failed to lift or did not align to the same location. We then use crossfilt-lift and the query to target chain file to lift these reads back to the target genome and we verify they returned the original aligned location using crossfilt-filter. The previous steps yield BAM files with the reads aligned in the target genome as well as the equivalent reads in the query genome. Using GTF files produced using the Comparative Annotation Toolkit (CAT) [28], we find orthologous features and count reads that map to features in both genomes. Reads that do not map to the same feature in both genomes are removed.

CrossFilt runtime and memory use

We checked the runtime of CrossFilt by simulating datasets with 10 M, 20 M, 50 M, 100 M, and 200 M reads. We simulated using the same parameters as we used for human heart tissue (see section Simulated RNA-sequencing data). We then ran CrossFilt using a single core of an Intel ThinkSystem SR630 V4 base with 2 × Intel 6760P processors, 64 cores per processor, 2.2 GHz, with 2.0 TB RAM and 960 GB NVMe storage. The runtimes are reported in Additional file 1: Table S5 and memory usage in Additional file 1: Table S6. Every other method here requires only a single genome alignment, while CrossFilt requires two alignments, and two calls to each crossfilt-lift and crossfilt-filter. The amount of time required to run CrossFilt is on average about 3.4 × the time it takes to run a single alignment, while the memory required to run CrossFilt is an order of magnitude less than what is required to align reads using STAR.

Consensus genome

We used custom scripts to construct human-chimpanzee and human-macaque consensus genomes from the RefSeq [52] genome builds (available at https://github.com/kennethabarr/ConsensusGenomeTools) [53]. Starting from hg38, we scanned through the UCSC chain file for the other species and masked each mismatched base with an 'N'. We also masked six bases on each side of all insertions and deletions. These were the same parameters used in [10]. We annotated genes in each species using the human annotations.

Annotation transfer

We used CAT v2.2.1 implementations of TransMap [54] and AUGUSTUS [55] to transfer human annotations to the chimpanzee and rhesus macaque genomes. As input, we used genome alignments obtained from the Zoonomia project's 447-way mammalian alignment [56]. The chimpanzee alignments did not include mitochondria; therefore, we ran CAT a second time using –chain-mode with the hg38ToPanTro6.over.chain.gz file obtained from the UCSC genome browser to obtain mitochondrial genome annotations for chimpanzee. We then combined the autosomal and mitochondrial alignments to produce a single chimpanzee annotation set. We then filtered CAT-based annotations to

exclude features that were not shared by both species in each pairwise comparison. We further filtered the data to account for differences in transcript length (see Additional file 1: Table S4).

Orthologous exons

We previously used BLAT [57] to generate lists of orthologous exons with high sequence identity that reciprocally map to the reference genome of each species [7, 32, 34, 35, 58]. Here, instead of using BLAT, we built orthologous exon annotations by filtering exons from the CAT genome annotations, removing exons with less than 85% sequence identity or that differed by more than 10% in sequence length (Additional file 1: Table S4). We specifically chose these parameters to be the same used in [35]. This approach allowed for a straightforward comparison between 'annotation transfer', 'orthologous exons', and CrossFilt.

Differential expression analysis

We performed DE analysis separately in each tissue and species comparison using a standard pipeline. We restricted our analysis to genes that had at least one count in at least one sample from both species (either human/chimpanzee or human/macaque). We further filtered out gene sets using default parameters from the filterByExpr function in edgeR [42, 43]. This function keeps genes that meet a counts-per-million (CPM) threshold in a number of samples equal to the smallest group size. We calculated library sizes and standardized the data using trimmed-mean of M-values [59] and generated log-transformed CPM and precision weights using the voom function in edgeR v4.4.2 [60]. We used these CPM values for differential expression analysis, except in the 'dual reference' approach. Here, we generated log-transformed transcripts-per-million (TPM) in place of CPM to account for gene length differences. Using the limma v3.6.2 function lmFit [61], we fit a linear model using a single fixed effect for species and moderated the t-statistics using the eBayes functions in limma. To rule out normalization as a potential factor explaining differences between the 'dual reference' approach and the other pipelines, we repeated these steps using TPM normalization in all of the approaches (Additional file 1: Fig. S14).

Differential expression analysis comparison with DESeq2

We also checked whether our results replicated using DESeq2 v1.46.0 [62] as an additional DE testing framework. As we did for limma, we only considered genes with at least one count in both species. We also filtered for genes that had at least 3 samples with at least 10 total counts. We then ran DESeq2 using default parameters. We regenerated Figs. 2 and 3 from the DESeq2 results (Additional file 1: Figs. S19-S20). We excluded the 'dual reference' condition because DESeq2 is built on a negative binomial model and requires raw counts. This means that it is incompatible with TPM normalization. When using DESeq2 instead of limma we maintain the lowest tFDR, highest pAUC, highest accuracy in logFC effect size estimation, smallest effect size in false positives, and the fewest false positives in the top 100 genes.

Sharing of differentially expressed genes between tissues

We used *mash* [38] – implemented in the R package mashr v0.2.79 – to estimate sharing of DE effects between tissues in each species separately. Mash takes the DE effect sizes and standard errors as input. We used effect size estimates from limma and computed standard errors from the effect sizes and *p*-values. We ran *mash* using both canonical and data-driven covariance matrices, using four principal components for the latter. We first checked for sharing of effects between each pair of tissues. We considered a gene to have shared differential expression if it had a false sign rate below 5% in at least one tissue, and if posterior estimates of the effect sizes were within a factor of 1.5 of each other in all pairwise tissue comparisons.

Gene set enrichment analysis

We performed gene set enrichment analysis using the R package fgsea v1.32.4 [63]. We obtained Gene Ontology biological process gene sets from the Molecular Signature Database v7.5.1. We used the ranked t-statistics from each tissue-by-species comparison to perform GSEA for each biological process.

Simulating ATAC-seq reads

While both comparative RNA-seq and comparative ATAC-seq involve the same fundamental task of counting reads that map to orthologous features, there are some differences in the data and methods. To demonstrate the applicability of CrossFilt to ATAC-seq data, we simulated and processed reads with and without CrossFilt filters. First, we used internal ATAC-seq data to identify peaks in the human genome (hg38). We then used liftOver to lift the peaks to the chimpanzee genome (panTro6) and then back to the human genome. We filtered the peaks for the set that returned the original human coordinates. Next, we used edgeR to calculate mean and dispersion estimates across human samples for each of the remaining reciprocally-lifting peaks. As we did with RNA-seq, we simulated 4 biological replicates per-species and selected 10% of peaks to be differentially accessible with a log fold-change of 2. We then simulated counts for each peak in each species using a negative binomial distribution. For each count, we simulated a fragment by selecting a random coordinate uniformly from within the peak, selecting a random orientation, and sampling a fragment length from internal ATAC-seq data. We subtracted 4 bp from the 5' and added 5 bp to the 3' ends to adjust from the center of Tn5 cut sites to the ends.

Aligning and counting ATAC-seq reads without CrossFilt

We aligned the simulated reads to their respective genomes using bwa-mem2 v2.2.1 [64]. We filtered the alignment for mapped paired-end reads with scores greater than or equal to 30. From these high-quality mapped read-pairs we generated a bed file where each line represented a single fragment. We added 4 bp to the start coordinate of the 5' read and subtracted 5 bp from the end coordinate of the 3' read to find the center of Tn5 cut sites. Finally, we generated a count-matrix from the resulting bed file using the FeatureMatrix function in Signac v1.16.0 [65].

Aligning and counting ATAC-seq reads with CrossFilt

To filter ATAC-seq reads with CrossFilt we first mapped reads to their target genomes with bwa-mem2 v2.2.1. We lifted these reads to their query genome using crossfilt-lift. We realigned the lifted reads in their query genomes and filtered for reads that mapped with high quality (score ≥ 30) to the same location that they lifted. We then lifted reads back to the target genome and filtered for reads that returned their original location. Using filtered reads mapped to the human genome, we generated a fragment bed file and generated a count-matrix using FeatureMatrix in Signac.

Differential accessibility

We used the same workflow to assess differential accessibility as we did to assess differential expression. We filtered for peaks with sufficient expression using the function filterByExpr in edgeR. We used TMM-normalized log-CPM values as input to limma. We called peaks true-positives if they were simulated as differentially accessible (DA) and they had a FDR-adjusted p-value less than 5%. We called peaks false-positive if they met the 5% FDR threshold but were not simulated as DA. CrossFilt read filtering removed a set of false positives that showed highly significant increased accessibility in humans relative to chimpanzees (Additional file 1: Fig. S16).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04082-2.

Additional file 1.

Acknowledgements
We thank reviewer 2 for their feedback on our previous paper [35]. Their questions regarding the influence of alignment bias inspired this work. We thank Genevieve Housman for helpful discussion. We thank the creators of CrossMap [57], whose implementation of liftOver helped us get started with this project. We also acknowledge Natalia Gonzales for her assistance writing and revising the manuscript.

Peer review information
Chuan-Yun Li and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
KB conceptualized the study, designed the method, generated results, and performed the analysis. YG provided funding for the project and wrote the manuscript, with input and edits from KB. All authors read and approved the final manuscript.

Funding
This work was supported by NIH grants R35GM131726 and R01HG010772, to YG.

Data availability
All data used in this work are publicly available. Human, chimpanzee, and rhesus tissue expression data were obtained from GSE112356. The methods we have implemented are available in github repositories https://github.com/kennethabarr/CrossFilt and https://github.com/kennethabarr/ConsensusGenomeTools.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 May 2025 Accepted: 14 April 2026
Published online: 23 April 2026

References

1. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009;25(24):3207–12. https://doi.org/10.1093/bioinformatics/btp579.
2. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12(11):1061–3. https://doi.org/10.1038/nmeth.3582.
3. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39(1):309–38. https://doi.org/10.1146/annurev.genet.39.073003.114725.
4. Yousaf A, Liu J, Ye S, Chen H. Current progress in evolutionary comparative genomics of Great Apes. Front Genet. 2021;11:12. https://doi.org/10.3389/fgene.2021.657468.
5. Demuth JP, Hahn MW. The life and death of gene families. Bioessays. 2009;31(1):29–39. https://doi.org/10.1002/bies.080085.
6. Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Res. 2025;53(D1):D966-75. https://doi.org/10.1093/nar/gkae1078.
7. Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, et al. A comparison of gene expression and DNA methylation patterns across tissues and species. Genome Res. 2020;30(2):250–62. https://doi.org/10.1101/gr.254904.119.
8. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H. The UCSC genome browser database: update 2006. Nucleic Acids Res. 2006;34:D590-8. https://doi.org/10.1093/nar/gkj144.
9. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003;100(20):11484–9. https://doi.org/10.1073/pnas.1932072100.
10. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature. 2019;574(7778):418–22. https://doi.org/10.1038/s41586-019-1654-9.
11. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acids Res. 2025;6(D1):D948–57. https://doi.org/10.1093/nar/gkae1071.
12. He Z, Bammann H, Han D, Xie G, Khaitovich P. Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. RNA. 2014;20(7):1103–11. https://doi.org/10.1261/rna.043075.113.
13. Mora-Bermúdez F, Badsha F, Kanton S, Camp JG, Vernot B, Köhler K, et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife. 2016;26:e18683. https://doi.org/10.7554/elife.18683.
14. Khrameeva E, Kurochkin I, Han D, Guijarro P, Kanton S, Santel M, et al. Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. Genome Res. 2020;1(5):776–89. https://doi.org/10.1101/gr.256958.119.
15. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature. 2021;598(7879):111–9. https://doi.org/10.1038/s41586-021-03465-8.
16. Hahn J, Monavarfeshani A, Qiao M, Kao AH, Kölsch Y, Kumar A, et al. Evolution of neuronal cell classes and types in the vertebrate retina. Nature. 2023;624(7991):415–24. https://doi.org/10.1038/s41586-023-06638-9.
17. Jorstad NL, Song JHT, Exposito-Alonso D, Suresh H, Castro-Pacheco N, Krienen FM, et al. Comparative transcriptomics reveals human-specific cortical features. Science. 2023;382(6667):eade9516. https://doi.org/10.1126/science.ade9516.
18. Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, et al. An early cell shape transition drives evolutionary expansion of the human forebrain. Cell. 2021;15(8):2084-2102.e19. https://doi.org/10.1016/j.cell.2021.02.050.
19. Jiang W, Chen L. Tissue specificity of gene expression evolves across mammal species. J Comput Biol. 2022;1(8):880–91. https://doi.org/10.1089/cmb.2021.0592.
20. Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. Nature. 2019;571(7766):505–9. https://doi.org/10.1038/s41586-019-1338-5.
21. Zaremba B, Fallahshahroudi A, Schneider C, Schmidt J, Sarropoulos I, Leushkin E, et al. Developmental origins and evolution of pallial cell types and structures in birds. Science. 2025. https://doi.org/10.1126/science.adp5182.
22. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;21(6114):1587–93. https://doi.org/10.1126/science.1230612.
23. Geirsdottir L, David E, Keren-Shaul H, Weiner A, Bohlen SC, Neuber J, et al. Cross-species single-cell analysis reveals divergence of the primate microglia program. Cell. 2019;179(7):1609-1622.e16. https://doi.org/10.1016/j.cell.2019.11.010.
24. Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, et al. Spatiotemporal transcriptomic divergence across human and macaque brain development. Science. 2018;362(6420):eaat8077. https://doi.org/10.1126/science.aat8077.
25. Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, et al. Establishing cerebral organoids as models of human-specific brain evolution. Cell. 2019;176(4):743-756.e17. https://doi.org/10.1016/j.cell.2019.01.017.
26. Murat F, Mbengue N, Winge SB, Trefzer T, Leushkin E, Sepp M, et al. The molecular evolution of spermatogenesis across mammals. Nature. 2023;613(7943):308–16. https://doi.org/10.1038/s41586-022-05547-7.

27. Schmitz MT, Sandoval K, Chen CP, Mostajo-Radji MA, Seeley WW, Nowakowski TJ, et al. The development and evolution of inhibitory neurons in primate cerebrum. Nature. 2022;603(7903):871–7. https://doi.org/10.1038/s41586-022-04510-w.
28. Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, et al. Comparative annotation toolkit (CAT)—simultaneous clade and personal genome annotation. Genome Res. 2018;28(7):1029–38. https://doi.org/10.1101/gr.233460.117.
29. Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, et al. Integrating gene annotation with orthology inference at scale. Science. 2023. https://doi.org/10.1126/science.abn3107.
30. Babbitt CC, Haygood R, Nielsen WJ, Wray GA. Gene expression and adaptive noncoding changes during human evolution. BMC Genomics. 2017;18(1):435. https://doi.org/10.1186/s12864-017-3831-2.
31. Ma S, Skarica M, Li Q, Xu C, Risgaard RD, Tebbenkamp ATN, et al. Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. Science. 2022;377(6614):eabo7257. https://doi.org/10.1126/science.abo7257.
32. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 2010;20(2):180–9. https://doi.org/10.1101/gr.099226.109.
33. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, et al. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife. 2015;4:e07103. https://doi.org/10.7554/eLife.07103.
34. Pavlovic BJ, Blake LE, Roux J, Chavarria C, Gilad Y. A comparative assessment of Human and Chimpanzee iPSC-derived cardiomyocytes with Primary Heart Tissues. Sci Rep. 2018;8(1):15312. https://doi.org/10.1038/s41598-018-33478-9.
35. Barr KA, Rhodes KL, Gilad Y. The relationship between regulatory changes in *cis* and *trans* and the evolution of gene expression in humans and chimpanzees. Genome Biol. 2023;24(1):207. https://doi.org/10.1186/s13059-023-03019-3.
36. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018;19(5):776–92. https://doi.org/10.1093/bib/bbx008.
37. Gilad Y. An Intuitive Primer on Effective Functional Genomics Study Design. Yoav Gilad. 187 p.
38. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet. 2019;51(1):187–95. https://doi.org/10.1038/s41588-018-0268-8.
39. Balasubramanian S, Zheng D, Liu YJ, Fang G, Frankish A, Carriero N, et al. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biol. 2009;10(1):R2. https://doi.org/10.1186/gb-2009-10-1-r2.
40. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 2010;20(2):180–9. https://doi.org/10.1101/gr.099226.109.
41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):1. https://doi.org/10.14806/ej.17.1.200.
42. Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Nucleic Acids Res. 2025;53(2):gkaf018. https://doi.org/10.1093/nar/gkaf018.
43. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97. https://doi.org/10.1093/nar/gks042.
44. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015;31(17):2778–84. https://doi.org/10.1093/bioinformatics/btv272.
45. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC genome browser database: 2023 update. Nucleic Acids Res. 2023;51(D1):D1188-95. https://doi.org/10.1093/nar/gkac1072.
46. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
47. Barr KA. kennethabarr/CrossFilt. Zenodo. https://zenodo.org/records/19337619 https://doi.org/10.5281/zenodo.19337619 (2026)
48. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30(7):1006–7. https://doi.org/10.1093/bioinformatics/btt730.
49. Mun T, Chen NC, Langmead B. LevioSAM: fast lift-over of variant-aware reference alignments. Bioinformatics. 2021;37(22):4243–5. https://doi.org/10.1093/bioinformatics/btab396.
50. Chen NC, Paulin LF, Sedlazeck FJ, Koren S, Phillippy AM, Langmead B. Improved sequence mapping using a complete reference genome and lift-over. Nat Methods. 2024;21(1):41–9. https://doi.org/10.1038/s41592-023-02069-6.
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
52. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-745. https://doi.org/10.1093/nar/gkv1189.
53. Barr KA. kennethabarr/ConsensusGenomeTools. Zenodo. https://zenodo.org/records/19339043 https://doi.org/10.5281/zenodo.19339043(2026)
54. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44. https://doi.org/10.1093/bioinformatics/btn013.
55. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(2):W435-9. https://doi.org/10.1093/nar/gkl200.
56. Kuderna LFK, Gao H, Janiak MC, Kuhlwilm M, Orkin JD, Bataillon T, et al. A global catalog of whole-genome diversity from 233 primate species. Science. 2023;380(6648):906–13. https://doi.org/10.1126/science.abn7829.
57. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64. https://doi.org/10.1101/gr.229202.
58. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, et al. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife. 2015;4:e07103. https://doi.org/10.7554/eLife.07103.

59. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.
60. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. https://doi.org/10.1186/gb-2014-15-2-r29.
61. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York, NY: Springer; 2005. p. 397–420. Available from: https://doi.org/10.1007/0-387-29362-0_23. Cited 2025 Mar 25.
62. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
63. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv; 2021. p. 060012. Available from: https://doi.org/10.1101/060012v3, 10.1101/060012 . Cited 2025 Mar 25.
64. Vasimuddin Md, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019. p. 314–24. Available from: https://ieeexplore.ieee.org/document/8820962 https://doi.org/10.1109/IPDPS.2019.00041. Cited 2025 Nov 8.
65. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18(11):1333–41. https://doi.org/10.1038/s41592-021-01282-5.

Publisher's Note