SuperTranscript: a reference for analysis and visualization of the transcriptome

Anthony DK Hawkins[1], Alicia Oshlack[1,2*], Nadia M Davidson[1*]

[1]Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia

[22]School of BioSciences, University of Melbourne, Victoria, Australia

[*]Corresponding authors: alicia.oshlack@mcri.edu.au, nadia.davidson@mcri.edu.au

Abstract

Transcriptomes are tremendously diverse and highly dynamic; visualizing and analysing this complexity is a major challenge. Here we present superTranscript, a single linear representation for each gene. SuperTranscripts contain all unique exonic sequence, built from any combination of transcripts, including reference assemblies, de novo assemblies and long-read sequencing. Our approach enables visualization of transcript structure and provides increased power to detect differential isoform usage.

High throughput sequencing has revolutionized transcriptomics because it allows cDNA sequence to be read and expression levels quantified using a single, affordable assay[1,2]. RNA sequencing (RNA-seq) can examine expression at the gene level as well as infer transcript abundances and differential isoform usage. Alternative splicing can alter gene function and contributes to the overall transcriptional diversity in eukaryotes[3,4] but the dynamic and complex splicing structure of genes complicates RNA-Seq analysis. Traditionally, methods to analyse RNA-Seq data required reads to be mapped to a reference genome leading to the development of aligners capable of mapping reads across exon boundaries[5–7]. Generally, splice-aware aligners either determine the position of splice junctions from reference transcripts, or infer novel splice junctions using the sequences of the reads themselves. The expression level of a feature, such as a gene or exon, can then be quantified by counting the number of reads that overlap the feature. Recently, there has been a move towards quantifying RNA-Seq reads using a reference transcriptome rather than genome[8–10]. This approach has been facilitated by the availability of more accurate and complete reference annotations for model organisms (e.g. ENCODE[4]) and has enabled very fast feature quantification, which is necessary for the high volume of data produced in contemporary studies.

Despite the enormous advances in RNA-seq analysis thus far, the use of both the genome and transcriptome reference have limitations in terms of visualization and analysis. In particular, visualizing the reads used to infer transcript abundance on a reference genome is often ineffective, as the distribution of exonic sequences across the genome is sparse. Additionally, annotated exons are not necessarily the most appropriate feature for the quantification of alternative splicing[11]. Moreover, for most non-model organisms, a reference genome is not available. Using a reference transcriptome also presents difficulties for analysis, because genes are represented by multiple transcripts, which makes gene-level visualization of read alignments difficult. The use of a transcriptional reference also relies on accurate gene-models, which are not available for non-model organisms.

To provide a method for visualization of transcriptome data for model and non-model organisms we introduce the concept of a superTranscript. In addition, we show that superTranscripts enable the accurate detection of differential isoform usage, even in non-model organisms, for which this has not been previously possible.

SuperTranscripts act as a reference for transcriptome data, where each gene is represented by a single sequence that contains the union of all the exons in their transcriptional order (Figure 1A). Although superTranscripts do not necessarily represent any true biological molecule, they can be used in the same way as a reference genome. For example, reads can be aligned to the superTranscriptome using a splice

aware aligner, and subsequently visualized using standard tools such as IGV[12]. Quantification can also be performed with existing software by counting the reads that overlap superTranscript features. Analogous to exons in the genome, superTranscripts are annotated with blocks where a block is a contiguous sequence without splice junctions (Figure 1B). Isoforms are represented by different combinations of blocks. Hence, a block may correspond to one exon, multiple exons, or part of an exon, depending on the splicing structure of the alternative transcripts within a gene.

SuperTranscripts can be constructed by concatenating the exonic sequence of a gene. This process is simple if a reference genome and annotation is available (we provide the superTranscriptome for human on https://github.com/Oshlack/superTranscript_paper_code). If no reference genome is available, superTranscripts can be built from any set of transcripts, including de novo assembled transcripts, using an overlap assembly method. We have implemented this algorithm in a Python program called Lace (available from https://github.com/Oshlack/Lace/wiki). For each gene, Lace finds overlaps between transcripts using BLAT[13], builds a splicing graph[14], then topologically sorts nodes in the graph using Kahn's Algorithm[15] (methods, Supplementary Fig. 1). Topological sorting requires a directed acyclic graph, so any loops are broken using a similar approach to that described by Pevzner et al.[16].

SuperTranscripts in model organisms

The superTranscript representation has several advantages over alternative references. Firstly, read alignment is simplified with respect to a reference genome because there is less sequence and fewer splice junctions (Supplementary Table 1). Secondly, visualization with widely used programs such as IGV is enabled for the first time for de novo assembled transcriptomes. Furthermore, visualization is drastically improved for reference genomes because intronic sequence is excluded, giving a compact view of the mapped reads (Figure 2A). Finally, there is more power to detect differential splicing because there are generally fewer blocks than exons, and the average block contains more counts (Figure 1C, Supplementary Fig. 2). Critically, blocks do not need to be defined using a fixed annotation ("Annotated Blocks"), but can be defined dynamically ("Dynamic Blocks") based on the splicing present in a particular dataset (Figure 1B), further increasing the counts per block (Supplementary Fig. 2).

SuperTranscripts in non-model organisms

For studies involving non-model organisms where a reference genome is not available, visualization of read coverage across a gene, and differential isoform detection, are particularly challenging. A standard approach to analysing such data would involve *de novo* assembly of reads into a set of transcripts that can then be used as a reference transcriptome. While transcript level abundances can be quantified using

inference methods such as Kallisto[8] or Salmon[10], the read coverage can only be visualised by selecting a representative transcript for each gene, such as the longest, but some exons from alternative transcripts might be missed. An elegant alternative is to use superTranscripts built by Lace (Figure 2B). In addition, differential isoform usage can be detected using existing methods such as DEXseq[17] on the defined blocks. In order to test this approach for detecting differential isoform usage we assembled data from Trapnell *et al*[18] using Trinity[19] and compared to a genomic reference approach (see methods). We found that indeed counting reads in superTranscript blocks performed better than transcript counts from inference methods (Figure 2C). In particular, dynamically defined blocks were able to detect splicing events in genes where only a single contig was assembled. This resulted in a substantial increase in performance with the detection of 45% more true positives at $p<0.05$. Another application of superTranscripts in non-model organisms is the ability to call variants from the data after alignment to superTranscripts using, for example, GATK's best practices workflow (Supplementary Fig. 3). In general, most RNA-Seq analyses that use a reference genome can now be performed equally well for non-model organisms using superTranscripts as a reference.

Novel applications for superTranscripts

Lace can produce superTranscripts from any source, not just the reference or *de novo* assembled transcriptomes. By integrating data from multiple sources, Lace can reveal unique insights into the complexities of transcriptomes and identify novel expressed sequence. As an example, we created superTranscripts for published chicken gonads[20] by combining four different transcriptomes: the Ensembl annotation, RefSeq annotation, a Cufflinks[21] genome-guided assembly and a Trinity[19] *de novo* assembly (see methods). The resulting superTranscriptome was compact, containing less than 100Mbp. However, none of the four contributing transcriptomes contained of all the sequence; 88%, 77%, 47% and 17% of bases were covered by Trinity, Cufflinks, Ensembl and Refseq, respectively. Critically, 3% of the bases in the chicken superTranscriptome could not be found in the any of the reference chicken genome. This novel sequence included superTranscripts with protein coding sequence either completely (135 superTranscripts) or partially (1295 superTranscripts) absent from the reference genome e.g. *C22orf39* (Figure 2d, Supplementary Fig 4). The incomplete annotation of the chicken transcriptome is largely the result of gaps in the chicken genome reference (Supplementary Fig. 5). This analysis demonstrates the utility of superTranscripts and Lace to construct comprehensive transcriptome sequences in an automated way. It also highlights the major benefits of exploiting superTranscripts, even for a reasonably complete genome.

Here we have presented the idea of superTranscripts as an alternative reference for RNA-Seq. We also introduce Lace, a software program to construct superTranscripts. Lace and superTranscripts can potentially be applied in a broad range of scenarios, some of which have been presented herein, including: aiding visualization, detection of differential splicing and allowing transcriptomes from a variety of sources to be merged into a comprehensive reference. Another application may include calling variants from RNA-Seq mapped to superTranscripts. Lace is capable of assembling transcripts from any source, but existing de novo assemblers could also be modified to produce superTranscripts as additional output. This would simply require the assembly graph to be topologically sorted. Read lengths are increasing and future technologies promise to accurately sequence full-length transcripts. Assembling such data into a superTranscript is convenient for exploring the structure of genes, without bias from a reference (Supplementary Figure 6 shows an example for PacBio data). However, one of the most powerful applications of superTranscripts is to unlock analytical approaches developed for reference genomes, in species where no reference is available.

Methods

SuperTranscript construction - Lace

The Lace software for creating superTranscripts and their corresponding annotation can be found on GitHub: https://github.com/Oshlack/Lace/wiki. The results presented here used Lace version 0.75. The Lace algorithm takes two input files: (1) transcript sequences (2) a text file with the clustering information that groups each transcript into a gene or cluster. Lace then creates a superTranscript for each gene. The Lace assembly is conceptually described in Supplementary Figure 1 and includes the following steps:

- For each gene, all pairwise alignments between transcripts are performed using BLAT[13].
- The BLAT output is parsed to determine the sequences common to each transcript pair.
- A single directed graph is constructed per gene, where each node is a base in one of the transcripts and the directed edge retains the ordering of the bases in each transcript. Bases from overlapping sequence are merged based on the BLAT output.
- The graph is simplified by concatenating sequences of nodes along non-diverging paths. Then all cycles are removed in order to create a Directed Acyclic Graph.
- The nodes are topologically sorted (each node becomes a string of bases from the original graph) using Khan's algorithm, which gives a non-unique sorting of the bases.
- Each superTranscript is annotated with blocks and transcripts. Block positions can be defined by forks or divergences in the graph (with Lace) or can be defined dynamically using splice junctions from the reads mapped back to the superTranscript. Transcripts are annotated against the superTranscript using BLAT. Figure 1B demonstrates the annotation of a superTranscript with blocks and transcripts.

SuperTranscript construction - using a reference genome

When a reference genome was available we constructed superTranscripts by concatenating exonic sequence rather than using Lace. Doing so is more accurate as it does not rely on BLAT alignment or resolving loops in a splicing graph. The genome and annotation we used for human were taken from https://github.com/markrobinsonuzh/diff_splice_paper. As in Soneson et al.[11], the genome annotation was flattened, such that transcripts were merged into disjoint blocks. The sequence of each block was then extracted and concatenated for each flattened gene using the gffread -w command from the cufflinks suite. To annotate, we projected the genomic coordinates of transcripts onto the superTranscripts, then flattened the transcripts into blocks. The resulting human superTranscriptome, its annotation and the scripts used to create them are provided at https://github.com/Oshlack/superTranscript_paper_code.

Datasets

To demonstrate how superTranscripts can be applied for visualization and differential transcript usage we used the public RNA-Seq dataset of human primary lung fibroblasts with an siRNA knock-down of *HOXA1* from Trapnell et al.[18] (GEO accession GSE37704). We also validated superTranscripts on a simulation of differential transcript usage for human created by Soneson et al.[11] (results shown in Supplementary Table 1, Supplementary Fig. 2). Finally, the combined superTrancriptome for chicken was constructed using reads of chicken embryonic gonads from Ayers *et al*.[20] (SRA accession SRA055442).

De novo transcriptome assembly

Trinity[19] version r2013-02-25 was used to assemble the Trapnell dataset into contigs using default options and a minimum contig length of two hundred. Contigs were then clustered by Corset[22] with the test for differential expression turned off. The corset clustering and *de novo* assembled contigs were used as inputs to Lace to construct a superTranscript for each cluster. The superTranscript for each cluster was then assigned to a gene by aligning to the human reference superTranscriptome using BLAT with option -minIdentity=98. Clusters assigned to multiple genes were removed from the differential transcript usage analysis.

Read alignment

Reads were aligned to the genome or superTranscriptome using the two-pass mode of STAR[6]. For the superTranscriptome we used the STAR option --outSJfilterOverhangMin 12 12 12 to filter the output splice junctions. Junctions supported by 5 or more reads were used to define dynamic block positions. The annotation was created using Mobius, a python script in the Lace suite.

Counting reads per bin

The featureCounts[23] function from Rsubread R package (v 1.5.0) was used to summarise the number of reads falling in a given genomic region. Reads were counted in paired end mode (-p) requiring that both ends map (-B). We allowed for reads overlapping multiple features (-O) assigning a fractional number of counts, 1/n, depending on how many features, n, the read overlapped (--fraction). The same summarisation procedure was used in all cases.

Visualization

Read coverage and transcript annotation were visualized using either IGV[12] or Gviz[24] . In IGV we loaded the superTranscriptome fasta file using the option “load genome from file” and loaded the mapped reads

and annotation file using "load from file". Our R scripts for Gviz are provided at https://github.com/Oshlack/superTranscript_paper_code.

Differential exon usage comparison

DEXseq[17] R package (v.1.5.0) was used to test for differential isoform usage. DEXSeq takes a table of block level counts from featureCounts and, by default, removes blocks with fewer than 10 counts summed across all samples. DEXseq then produces a per gene q-value as the probability that for a given gene there is at least one differentially used exon, controlling for multiple testing. SuperTranscripts were ranked on their q-values. For the de novo assembly analysis the use of superTranscripts was contrasted with Kallisto[8] and Salmon[10]. Kallisto and Salmon were run on the de novo assembled contigs using the default settings. Estimated counts per contig were then grouped into clusters using the same Corset clustering as was used by the superTranscripts. The count table was processed by DEXseq in the same way as the block counts table for superTranscripts. The "truth" was defined by mapping the reads to the human genome and quantifying differential isoform usage with DEXSeq based on the reference annotation. All genes that had a q-value < 0.05 were considered true positives whilst all genes with a q-value >0.9 were considered true negatives. Where multiple clusters were mapped to the same gene, the cluster with the lowest p-value was chosen and the others discarded. Clusters which mapped to multiple genes were removed from the analysis, and those genes found in a multi-gene cluster were removed from the list of true and false positives.

Constructing a comprehensive chicken superTranscriptome

Ensembl and RefSeq references were downloaded for the chicken genome version galGal4 from UCSC on 24th August 2016. Cufflinks transcripts were assembled using the gonad reads from Ayers et al.[20], mapped to galGal4 using TopHat[5] version 2.0.6. The reference and cufflinks assembled transcripts were then merged into loci based on genomic positions using the cuffmerge command. The resulting annotation was flattened and exonic sequence concatenated to create a genome-based superTranscriptome, similar to that for the human (described above). To supplement these superTranscripts with de novo assembly, we first assembled all reads using Trinity. Trinty contigs were aligned against the genome-based chicken superTranscriptome using blat with options -minScore=200 -minIdentity=98. Contigs that aligned to a single genome-based superTranscript were clustered with it. Contigs matching two or more genome-based superTranscripts were discarded (to remove false chimeric transcripts[25]). Remaining contigs were clustered into genes based on their homology to human superTranscripts (using BLAT with options -t=dnax -q=dnax -minScore=200). Contigs that did not align

to a gene, or those that aligned to multiple genes were removed. Lace was then run on the sequence in each cluster, containing genome-based superTranscripts and Trinity contigs.

In analysing the chicken superTranscriptome, we assessed the coverage from each constituent transcriptome, Ensmebl, RefSeq, Cufflink and Trinity, by aligning their sequence against the superTranscripts using BLAT with options -minScore=200 -minIdentity=98. We determined regions which were not present in the genome by aligning the superTranscripts against the chicken reference genome using BLAT with options -minScore=100 -minIdentity=98. Finally, we looked for regions with homology to human coding sequence by aligning the superTranscriptome against the Ensembl GRCh38 human protein sequence using BLAST[26] with options -evalue 0.00001 -num_alignments 20. For a superTranscript region to be identified as novel protein coding sequence, we required it to be absent from the chicken genome, match a human protein sequence with BLAST e-value<10^{-5}, only be annotated by a Trinity transcript and be 30bp or longer. Scripts used in the chicken superTranscript analysis are provided at https://github.com/Oshlack/superTranscript_paper_code.

Acknowledgements

AO is funded by an NHMRC Career Development Fellowship APP1051481. We would like to thank Ian Majewski, Jovana Maksimovic and Harriet Dashnow for feedback on the manuscript and Michael McLellan for his preliminary contribution.

Author Contributions

N.M.D and A.O. conceived the idea of superTranscripts and Lace. A.D.K.H. developed Lace and performed all data analysis with the exception of the chicken superTranscriptome which was performed by N.M.D. All authors contributed to the writing of the manuscript.

Competing Financial Interests

The authors declare no competing financial interests

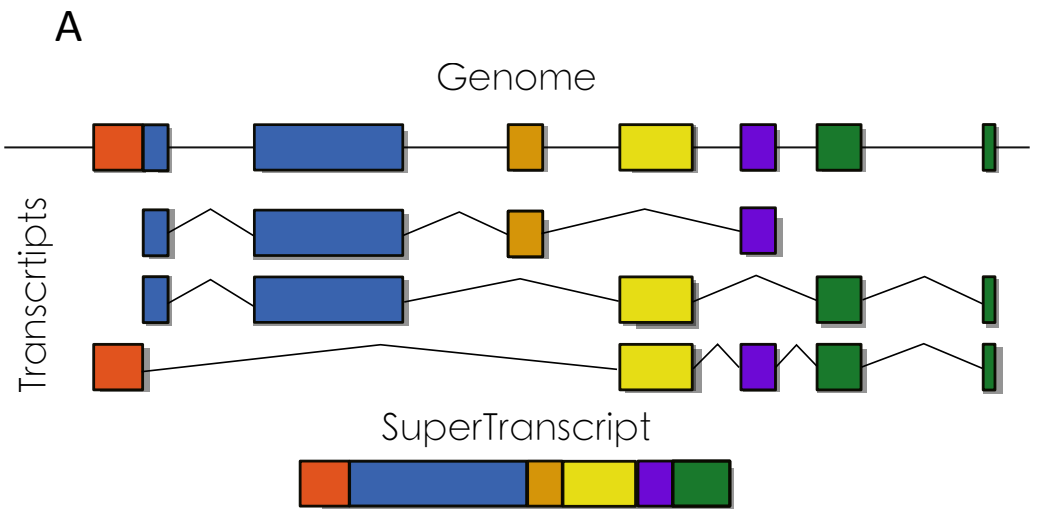
1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
2. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87–98 (2011).
3. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6,** 386–398 (2005).
4. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).
5. Trapnell, C., Pachter, L. & Salzberg, S. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
6. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
7. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–60 (2015).
8. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-Seq quantification. (2015).
9. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
10. Patro, R., Duggal, G. & Kingsford, C. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv* (2015).
11. Soneson, C., Matthes, K. L., Nowicka, M., Law, C. W. & Robinson, M. D. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.* **17,** 12 (2016).
12. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–6 (2011).
13. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–64 (2002).
14. Heber, S., Alekseyev, M., Sze, S.-H., Tang, H. & Pevzner, P. A. Splicing graphs and EST assembly problem. *Bioinformatics* **18 Suppl 1,** S181-8 (2002).
15. Kahn, A. B. & B., A. Topological sorting of large networks. *Commun. ACM* **5,** 558–562 (1962).
16. Pevzner, P. A., Pevzner, P. A., Tang, H. & Tesler, G. De novo repeat classification and fragment assembly. *Genome Res.* **14,** 1786–96 (2004).
17. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–17 (2012).
18. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).
19. Grabherr, M. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).

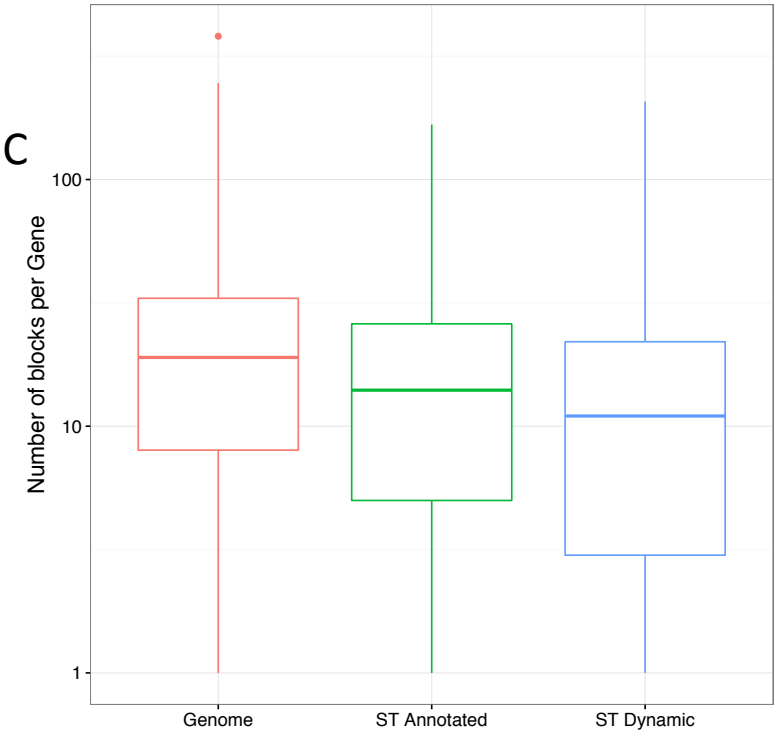
20. Ayers, K. L. *et al.* RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken embryos and allows comprehensive annotation of W-chromosome genes. *Genome Biol.* **14,** R26 (2013).
21. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
22. Davidson, N. M. & Oshlack, A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* **15,** 410 (2014).
23. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–30 (2014).
24. Hahne, F. & Ivanek, R. in 335–351 (2016). doi:10.1007/978-1-4939-3578-9_16
25. Yang, Y. & Smith, S. A. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* **14,** 328 (2013).
26. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

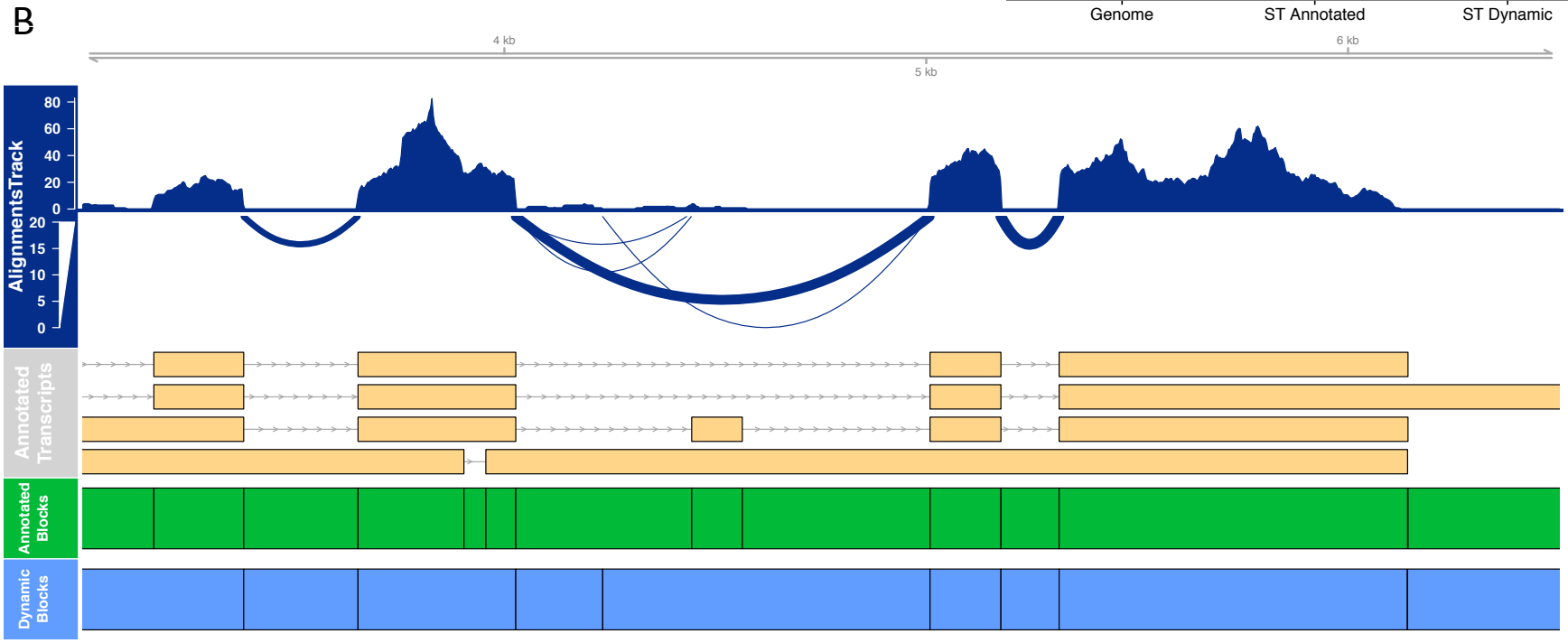
Figure Captions

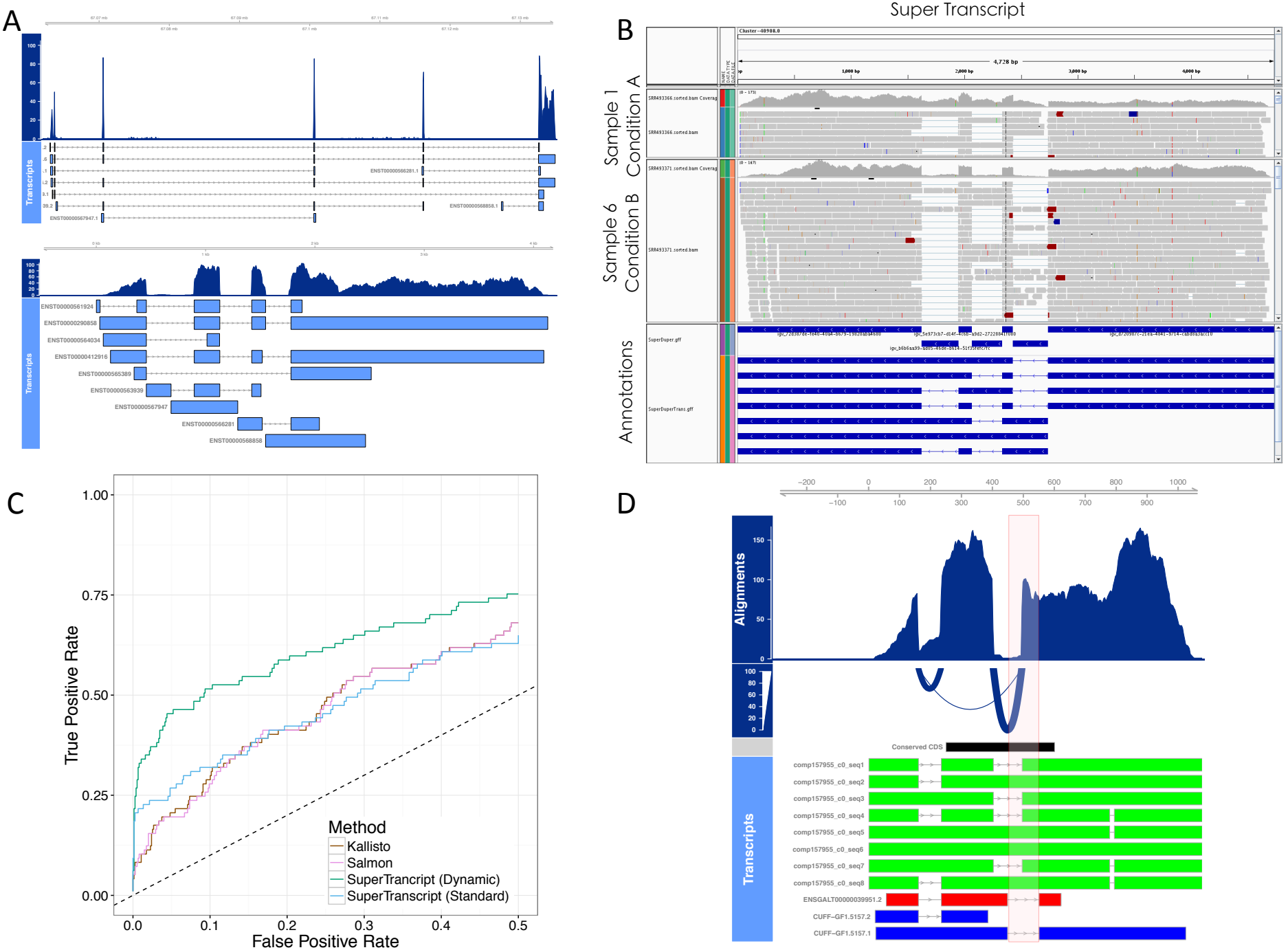
Figure.1 - A) A schematic diagram depicting a gene and its corresponding transcripts in the reference genome (top) compared to the superTranscript for the same gene (bottom). Colours indicate superTranscript blocks. B) An example of the read coverage and junctions when mapping to a superTranscript. The data is human primary lung fibroblasts, taken from Trapnell *et al.*[18]. The two alternative block annotations for the superTranscript – Annotated (green), Dynamic (blue) – are also illustrated underneath. C) The number of blocks per gene using the reference genome, superTranscript with annotated blocks and superTranscript with dynamic blocks. On the same dataset, there are fewer blocks defined for superTranscripts compared to blocks in a reference genome.

Figure.2 - A) Example of the read coverage over human *CBFB* (ENSG00000067955), in the reference genome (top) compared to the superTranscript (bottom). Transcripts are annotated below in light blue. B) An annotated screenshot illustrating how expression of *de novo* assembled transcripts can be visualized in IGV using a superTranscript. Blocks 2, 4 and 6 show differential usage between the two samples of different conditions. C) ROC curve for detecting differential transcript usage using *de novo* assembled transcripts from the Trapnell et al. dataset. True and false positives are defined using a reference genome analysis (See methods). D) Reads aligned back to the superTranscript of chicken *C22orf39 (ENSGALG00000023833*). The region shaded in red is a gap in the reference genome. Transcripts from Ensembl (red) and Cufflinks (blue) miss the gap sequence, whereas the Trinity assembly (green) recovers it.


A
Genome
Transcrtipts
SuperTranscript


C
Number of blocks per Gene
100
10
1
Genome
ST Annotated
ST Dynamic


B
4 kb
5 kb
6 kb
AlignmentsTrack
Annotated Transcripts
Annotated Blocks
Dynamic Blocks


A
Transcripts
ENST00000566281.1
ENST00000568858.1
ENST00000567947.1
ENST00000561924
ENST00000290858
ENST00000564034
ENST00000412916
ENST00000565389
ENST00000563939
ENST00000567947
ENST00000566281
ENST00000568858
B
Super Transcript
Sample 1
Condition A
Sample 6
Condition B
Annotations
4,728 bp
SRR493366.sorted.bam
SRR493371.sorted.bam
SuperDuper.gff
SuperDuperTrans.gff
C
True Positive Rate
False Positive Rate
Method
Kallisto
Salmon
SuperTrancript (Dynamic)
SuperTranscript (Standard)
D
Alignments
Conserved CDS
Transcripts
comp157955_c0_seq1
comp157955_c0_seq2
comp157955_c0_seq3
comp157955_c0_seq4
comp157955_c0_seq5
comp157955_c0_seq6
comp157955_c0_seq7
comp157955_c0_seq8
ENSGALT00000039951.2
CUFF-GF1.5157.2
CUFF-GF1.5157.1