

Closing the Gap in Time: From Raw Data to Real Science (Science as a Service - SaaS)

Anjana Varadarajan¹, Justin H. Johnson¹, Angelo Scorpio² and David Deshazer³

¹- EdgeBio, Gaithersburg, MD, 20877, ²-National Biodefense Analysis and Countermeasures Center, Frederick MD, 21702, ³- U.S. Army Medical Research Institute of Infectious Diseases, Frederick, 21702

Introduction

Science as a Service (SCaaS) leverages commercial grade software with the ability to integrate open source tools and proprietary internal algorithms to build pipelines for a vast variety of applications such as

- Targeted Resequencing (Amplicon, Exome, etc)
- Transcriptome and miRNA
- Methylation & Protein Binding Site Analysis
- Whole Genome Resequencing

EdgeBio uses Life Technology's SOLiD platform for Transcriptome sequencing which allows for:

- Conservation of strandedness of cDNA, allowing you to discern between overlapping RNAs transcribed from the sense or antisense strand
- Multiplexing and sequencing of multiple RNA libraries simultaneously, reducing the cost of analysis per sample

Sequencing and QC

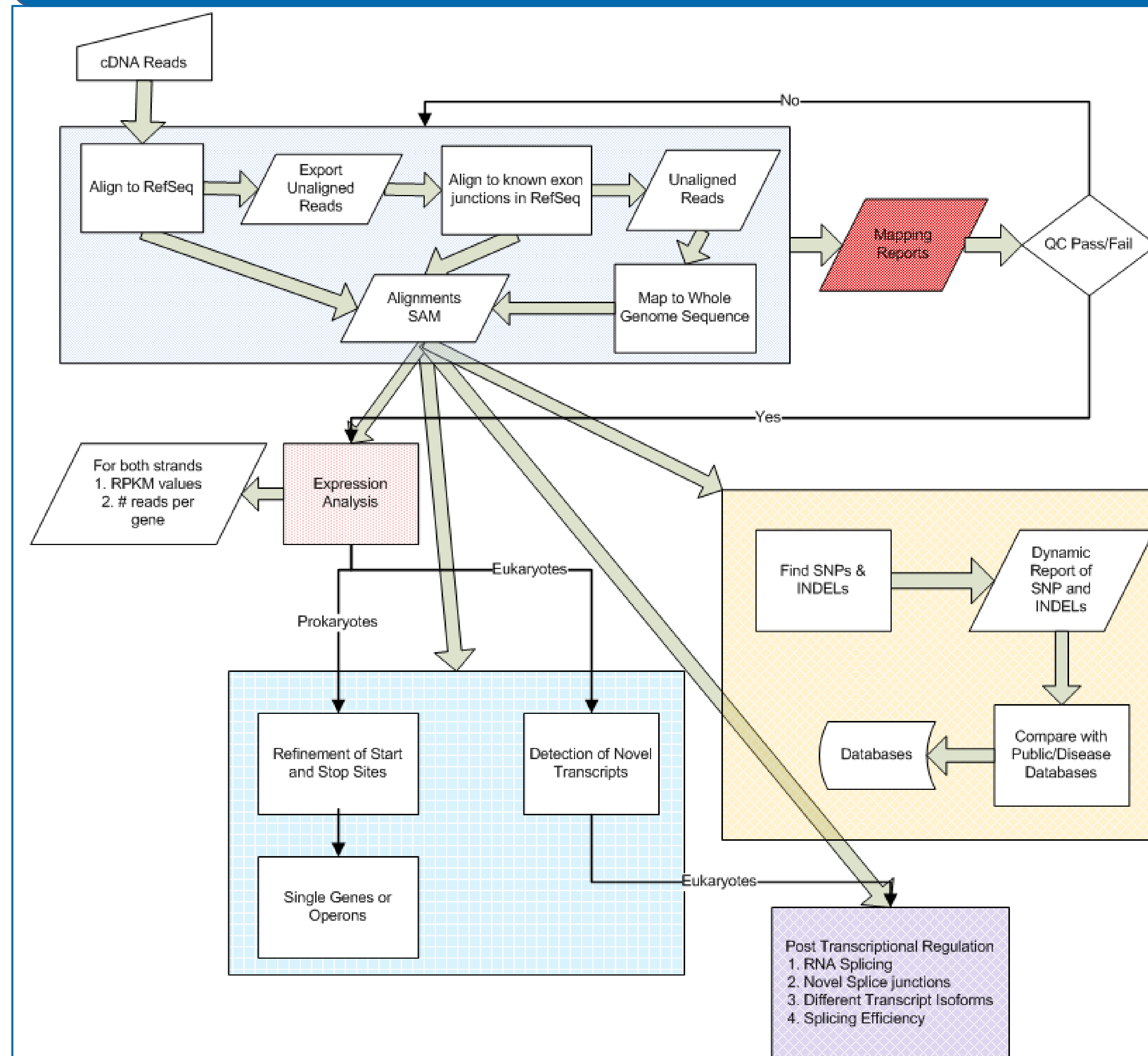
- RNA-Seq provides a largely unbiased method to define comprehensively and systematically the transcriptome of an organism in a manner that is significantly more sensitive than microarray hybridization approaches
- Seven samples each for *Burkholderia pseudomallei* bacteria strain AI were sequenced.
 - AIPBHR22-24 contains pBHR2 plasmid
 - AIPHR2VIRAG1-4 contains the pBHR2-virAG plasmid (virAG - regulatory system)
- A single ended fragment library with conserved strandedness of the cDNA was constructed.
- ~ 1 billion reads were sequenced producing 45Gb of usable sequence (Table 1)



Sample	Reads (in M)	Avg. QV	Bases sequenced (in M)
AIPBHR22	126.5	22.4	6325
AIPBHR23	173.2	21.8	8660
AIPBHR24	151.6	22.1	7580
AIPBHR2VIRAG1	100.6	22.4	5030
AIPBHR2VIRAG2	94.1	22.5	4705
AIPBHR2VIRAG3	104.5	22.2	5225
AIPBHR2VIRAG4	167.7	21.8	8385

Table 1: Per sample QC details

Pipeline



The transcriptome pipeline consists of the following steps:

- **Mapping**: The reads are mapped to the RefSeq and the whole genome using CLC Genomics workbench, output alignment files are exported in both .sam and .bam format.
- **Expression Analysis**: The expression values of genes are captured using reads per kilobase of exon model per million mapped reads (RPKM). RPKM values for both sense and anti-sense strands are obtained.
- **Transcript annotation and refinement**: For prokaryotes, co-operonic genes and transcript start sites are identified using the expression of the genome in intergenic regions. Novel transcripts and exons (in eukaryotes) are identified using the transcribed regions of the genome.
- **Variant analysis**: SNPs and INDELS are identified and annotated using databases such as dbSNP.
- **Post transcriptional regulation**: A largely eukaryotic step consisting of identification of alternate splice junctions, fusion transcripts, transcript isoforms and splicing efficiency.

Contact

EdgeBio - www.EdgeBio.com
205 Perry Parkway, Suite 5 Gaithersburg, MD 20877
E-mail : bioserv@edgebio.com Phone # : 301-990-2685

Results

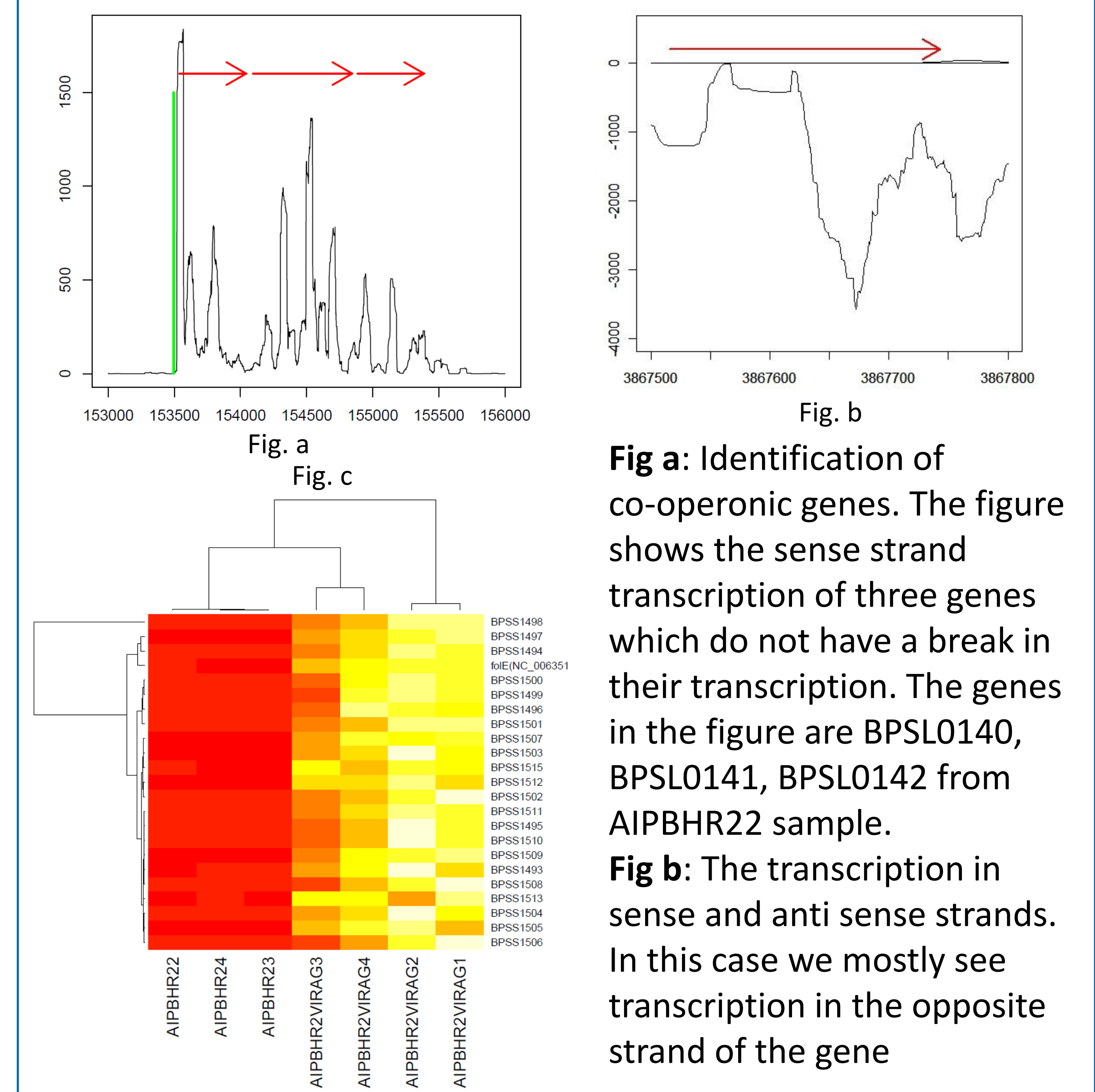


Fig a: Identification of co-operonic genes. The figure shows the sense strand transcription of three genes which do not have a break in their transcription. The genes in the figure are BPSL0140, BPSL0141, BPSL0142 from AIPBHR22 sample.

Fig b: The transcription in sense and anti sense strands. In this case we mostly see transcription in the opposite strand of the gene

Fig c: Heatmap of gene expressions (RPKM) of genes in cluster 1 of T6SS cluster under different conditions for all 7 samples. Samples containing plasmid pBHR2-virAG show high gene expression

Table 2: gene expression with varied transcription

Gene ID	Sense RPKM	Antisense RPKM	Ratio Antisense/Sense
BPSL3254A	0.163	51,925.88	318563.7
BPSS0074	0.41	44,738.39	109118
BPSSt06	9,057.09	30.477	0.003365
BPSLs05	30,217.21	62.602	0.002072

References

- CLC Genomics workbench: www.clcbio.com
- Passalacqua et al. Structure and complexity of a bacterial transcriptome. J.Bacteriol 2009 191 (10) 3203-3211
- Moratzavi et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. methods 2008 5 (7) 621-628

Funding - This work received support from Agreement No. HSHQDC-07-C-00020 awarded by the U.S. Department of Homeland Security for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC)