A modular framework for RNA-Seq data analysis using compact, anonymized data summaries

L. Habegger[1,2], A. Sboner[1,2], R. Alexander[1,2], T.A. Gianoulis[3,4], J.S. Rozowsky[2], A. Agarwal[2,5], D.Z. Chen[1], W.M. Hussain[6,7], D. Pflueger[6], S. Terry[6], F. Demichelis[6,7], M.A. Rubin[6], M. Snyder8, and M.B. Gerstein[1,2,5]

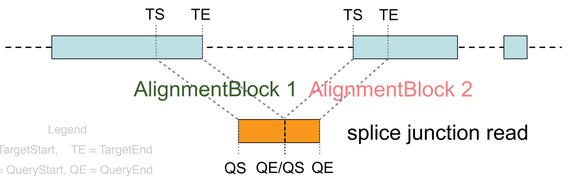
1 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
2 Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
3 Wyss Institute for Biologically-Inspired Engeneering, Harvard University, Boston, MA, USA
4 Department of Genetics, Harvard Medical School, Boston, MA, USA
5 Department of Computer Science, Yale University, New Haven, CT, USA
6 Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA
7 Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA
8 Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

Introduction

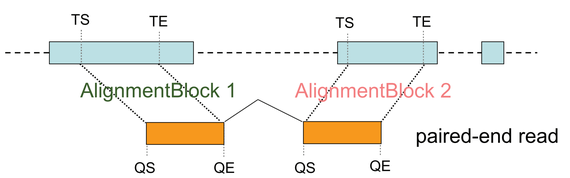
- Next-generation sequencing for functional genomics has given rise to large quantities of sequence information.
- Sequence reads from a specific individual can potentially identify and genetically characterize that person, raising privacy concerns.
- *RSEQtools* address these issues with *Mapped Read Format (MRF)*, a compact data summary format for short, long and paired-end read alignments.
- *RSEQtools* also provide a set of programs for RNA-seq data analysis.

Mapped Read Format (MRF)

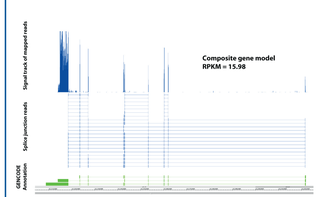
```
Comment Lines:  # Example 1
Header Line:    AlignmentBlock                                  Sequence
Mapped reads:   chr2:+:601:630:1:30,chr2:+:921:940:31:50        AGGCTCAAGCTTC...
```




```
Comment Lines:  # Example 1
Header Line:    AlignmentBlock                                  Sequence
Mapped reads:   chr9:+:431:480:1:50|chr9:+:945:994:1:50         AGGCTCCG...|TGCTTCA...
```



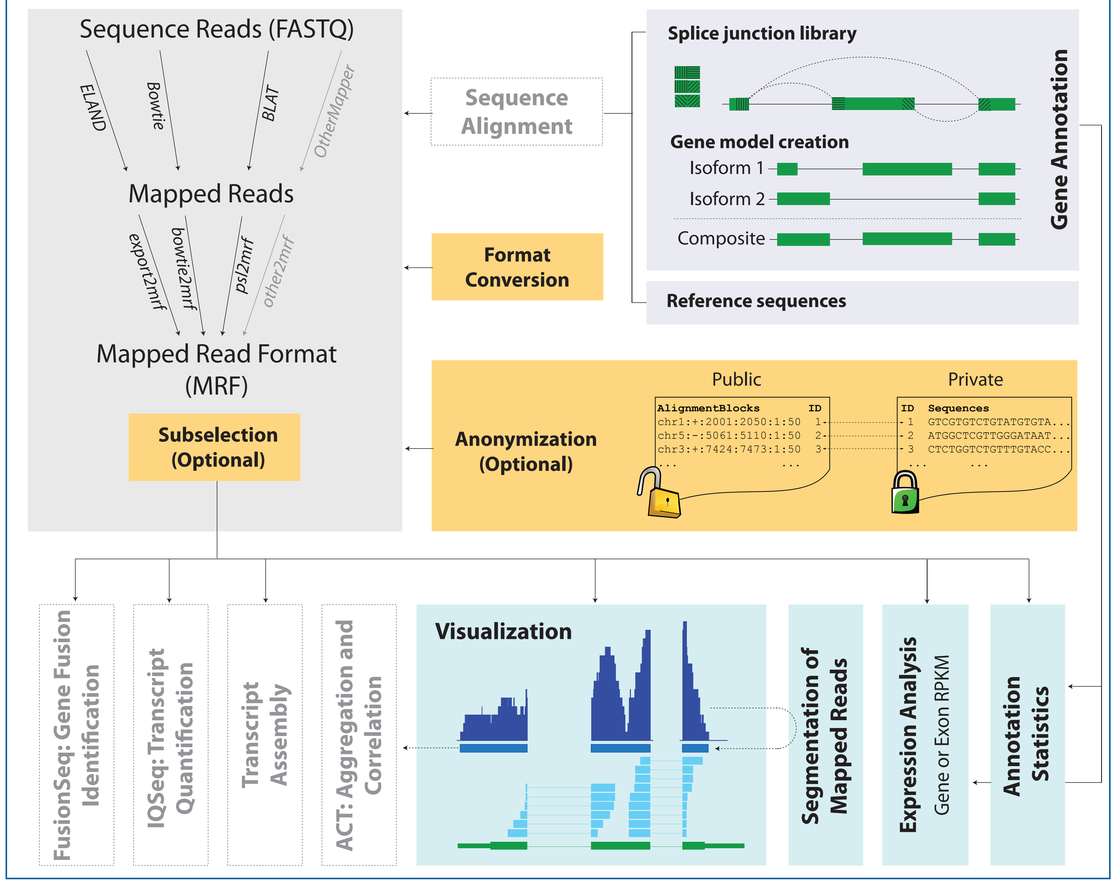
Visualization & Quantification



Programs

Description	Programs	
Format conversion utilities	bowtie2mrf	export2mrf
	psl2mrf	sam2mrf
Genome annotation tools	createSpliceJunction-Library	mergeTranscripts
	interval2sequences	
Expression analysis	mrfQuantifier	bgrQuantifier
Visualization tools	mrf2wig	mrf2gff
	mrf2bedGraph	
Segmentation of mapped reads	wigSegmenter	bgrSegmenter
Annotation statistics tools	mrfAnnotationCoverage	mrfMappingBias
MRF selection utilities	mrfSelectAnnotated	mrfSampler
	mrfSelectRegion	mrfSelectSpliced
Auxiliary utilities	bed2interval	interval2gff
	interval2bed	export2fastq
	gff2interval	mrf2sam

Overall Schematic[1]

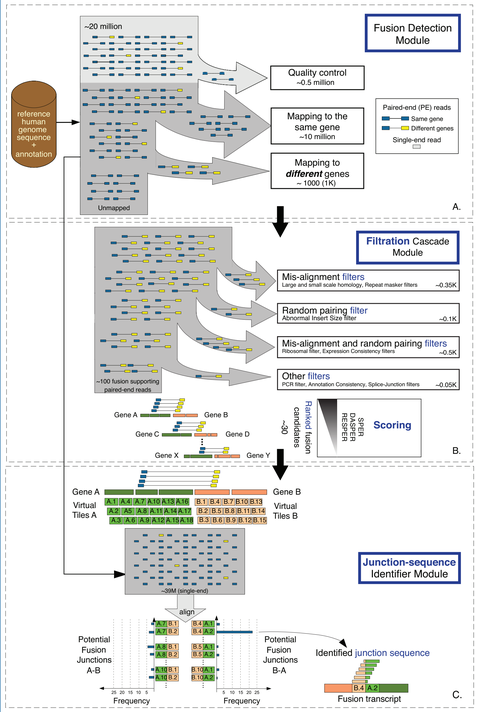


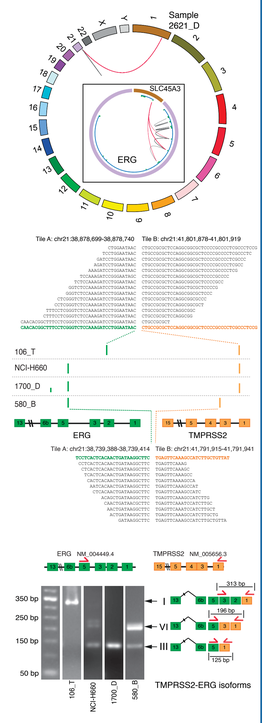
Run Statistics (Illumina GAII data)

Purpose	Program	Time to process	File sizes (uncompressed)	Notes
Alignment + Conversion	ELAND2	~1 day	Export: 4.2Gb	Processing of one flow cell (8 lanes)
	export2mrf	~ 2 min	MRF: 400Mb	Number of mapped reads: ~12M
Quantification	*mrfQuantifier*	~ 45 sec	Gene expression values: 3.5Mb	GENCODE composite gene models (~22,000)
Visualization	*mrf2wig*	~ 2 min	One WIG file per chromosome: 1Mb - 150Mb	Signal track of mapped reads normalized per million mapped reads
	mrf2gff	~ 45 sec	One GFF file per chromosome: 100Kb - 16Mb	To visualize splice junction reads

1. L. Habegger et al. "RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries." Bioinformatics 27, no. 2 (2011): 281-283.
2. A. Sboner et al. "FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data." Genome Biology 11, no. 10 (2010): R104.
3. J. Jee et al. "ACT: aggregation and correlation toolbox for analyses of genome tracks." Bioinformatics 27, no. 8 (2011): 1152 -1154.

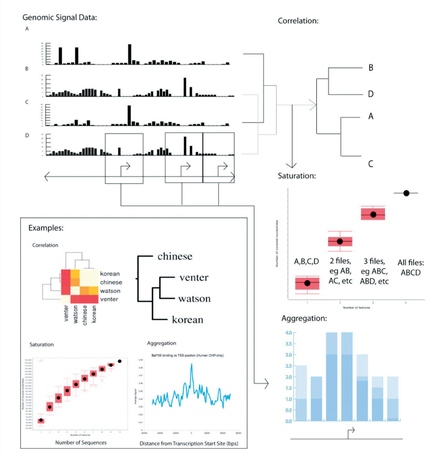
FusionSeq[2]





- Calibrated on eight cancers with and without known rearrangements.
- Discovered known and novel fusion transcripts.

Aggregation and Correlation Tool (ACT)[3]



- *Aggregation*: to determine the distribution of a signal track relative to certain genomic anchors, e.g. transcription start sites (TSS).
- *Correlation*: multiple-related signal tracks are correlated with each other.
- *Saturation*: determining the number of experimental conditions required to achieve high genomic coverage of the biological phenomenon under study.

- **RSEQtools: http://rseqtools.gersteinlab.org**
- **FusionSeq: http://rnaseq.gersteinlab.org/fusionseq**
- **ACT: http://act.gersteinlab.org**
- **IQSeq: http://code.google.com/p/iqseq/**

National Human Genome Research Institute