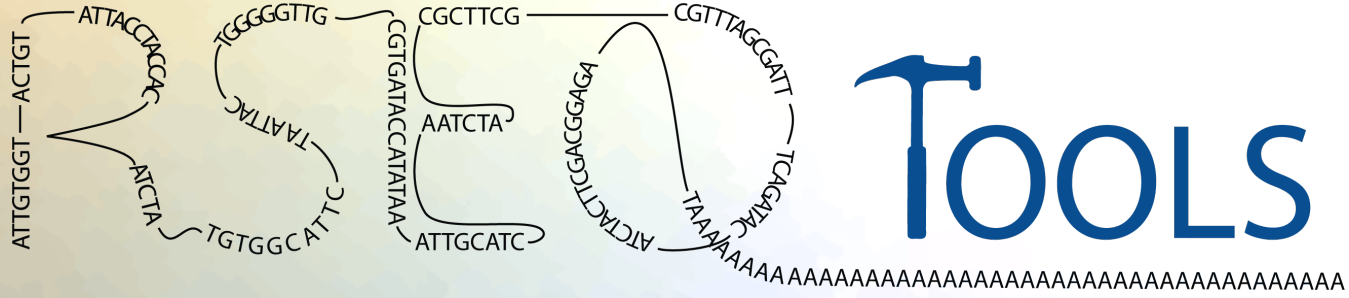


A modular framework to analyze RNA-Seq data using compact and anonymized data summaries

Lukas Habegger*, Andrea Sboner*, Tara Gianoulis,
Joel Rozowsky, Ashish Agarwal, Michael Snyder,
Mark Gerstein

Yale University

Accepted for publication in *Bioinformatics*


nature
THE BITER BIT
TROPICAL CYCLONES
BLACK HOLE PHYSICS
BIG DATA
SCIENCE IN THE PETABYTE ERA

storage

processing

storage

processing

sharing

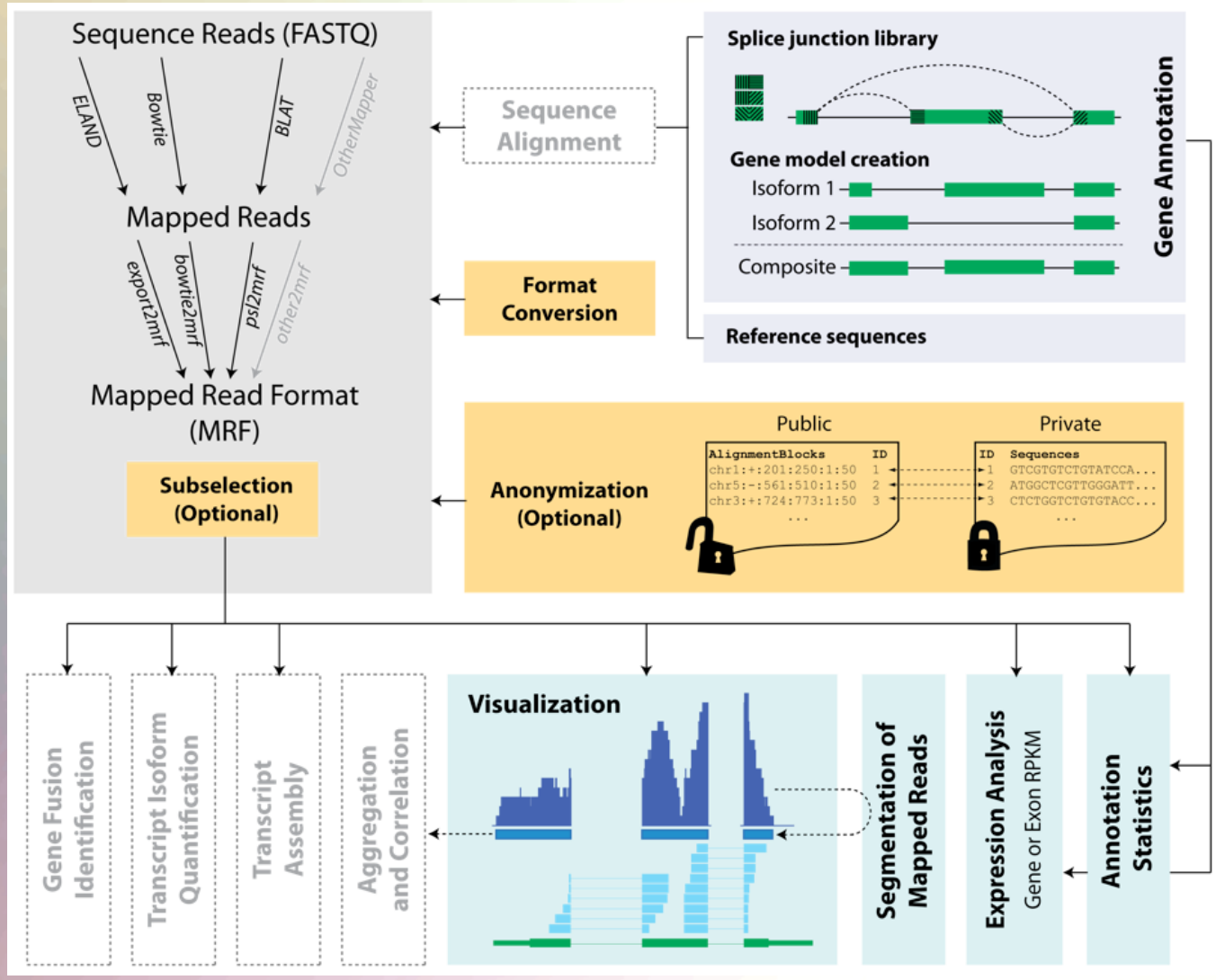
storage

sharing

processing

privacy

RSEQtools - schematic



http://rseqtools.gersteinlab.org/

RSEQtools modules

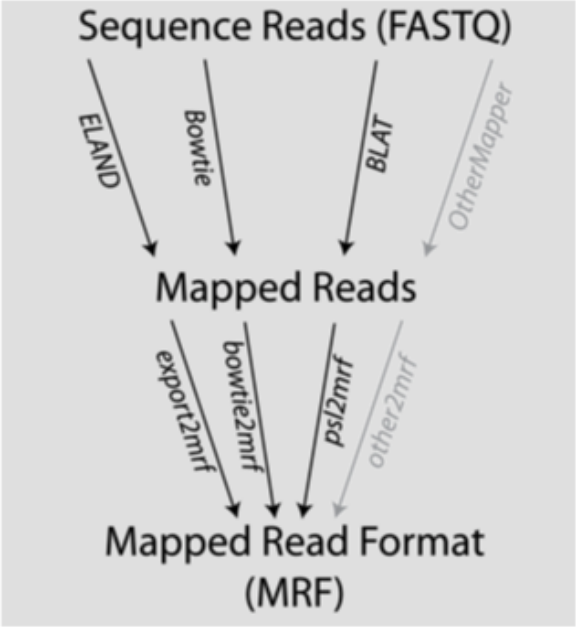
Core modules

- Format conversion utilities
- Genome annotation tools
- Expression analysis
- Visualization tools
- Segmentation of mapped reads
- Annotation statistics tools
- MRF selection utilities
- Auxiliary utilities

RSEQtools modules

Core modules

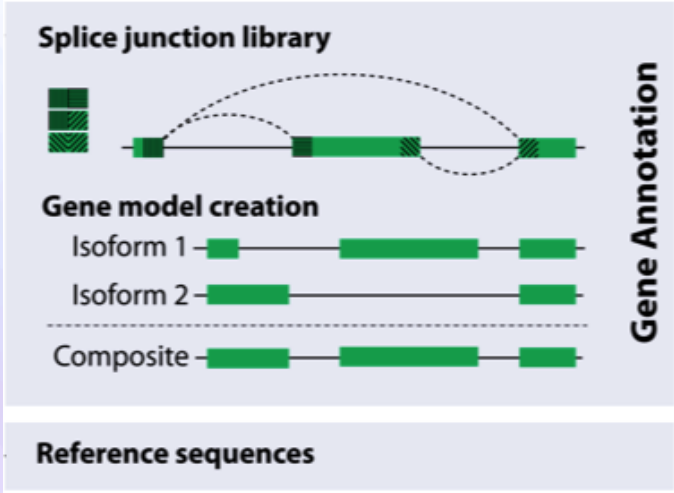
- *Format conversion utilities*
- Genome annotation tools
- Expression analysis
- Visualization tools
- Segmentation of mapped reads
- Annotation statistics tools
- MRF selection utilities
- Auxiliary utilities



RSEQtools modules

Core modules

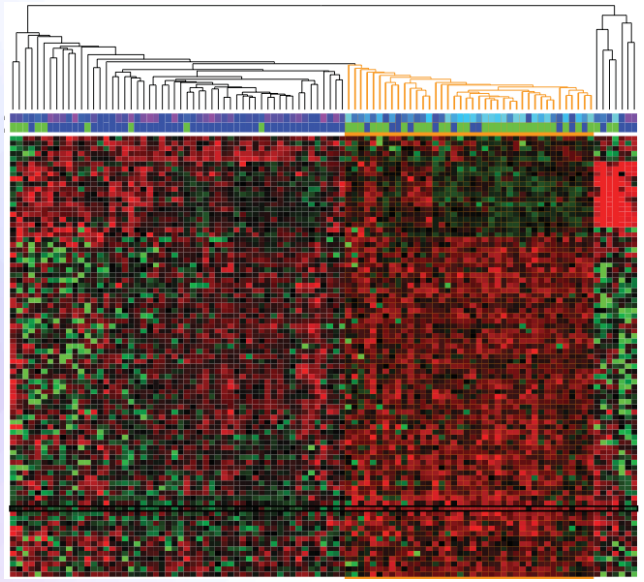
- Format conversion utilities
- *Genome annotation tools*
- Expression analysis
- Visualization tools
- Segmentation of mapped reads
- Annotation statistics tools
- MRF selection utilities
- Auxiliary utilities



RSEQtools modules

Core modules

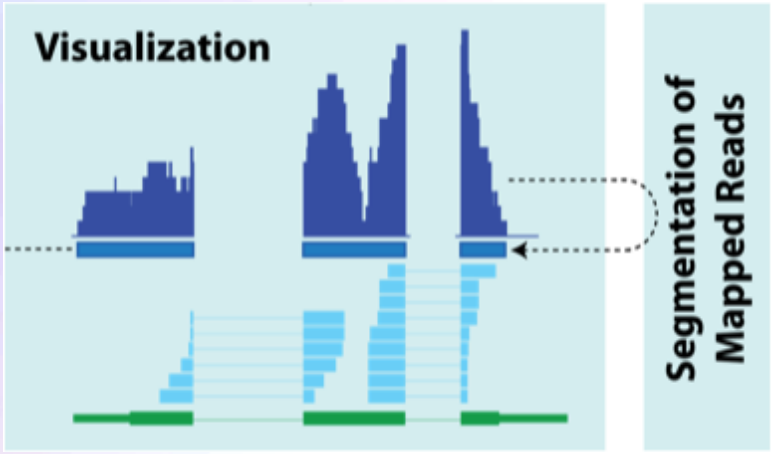
- Format conversion utilities
- Genome annotation tools
- ***Expression analysis***
- Visualization tools
- Segmentation of mapped reads
- Annotation statistics tools
- MRF selection utilities
- Auxiliary utilities



RSEQtools modules

Core modules

- Format conversion utilities
- Genome annotation tools
- Expression analysis
- *Visualization tools*
- *Segmentation of mapped reads*
- Annotation statistics tools
- MRF selection utilities
- Auxiliary utilities



RSEQtools modules

Core modules

- Format conversion utilities
- Genome annotation tools
- Expression analysis
- Visualization tools
- Segmentation of mapped reads
- Annotation statistics tools
- MRF selection utilities
- Auxiliary utilities

Associated tools

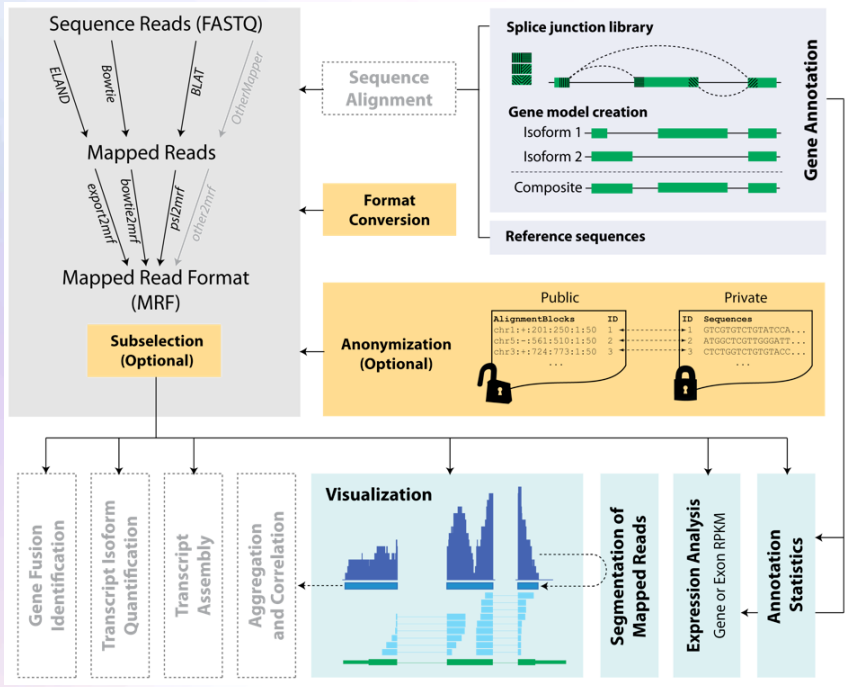
- **IQSeq**: transcript isoform quantification
  [Du et al, PloS Comp Biol, submitted]
  (http://rnaseq.gersteinlab.org/IQSeq)
- **ACT**: Aggregation and Correlation Toolbox
  [Jee et al. Bioinformatics, in press]
  (http://act.gersteinlab.org)
- **FusionSeq**: Identification of chimeric transcripts [Sboner et al. Genome Biol, 2010;11:R104]
  (http://rnaseq.gersteinlab.org/fusionseq)

Associated projects

- Transcript assembly (RGASP, ongoing project)
- Transcribed pseudogenes (PseudoSeq, ongoing project)
- incRNA: Identification of non-coding RNAs
- Calibration framework for RNA-Seq and tiling arrays transcriptome data
  [Agarwal et al., BMC Genomics, 2010]

Need for common formats

- Modular framework
- Standardized data format that is:
  - compact
  - functional
  - privacy “aware”



Mapped Read Format (MRF)

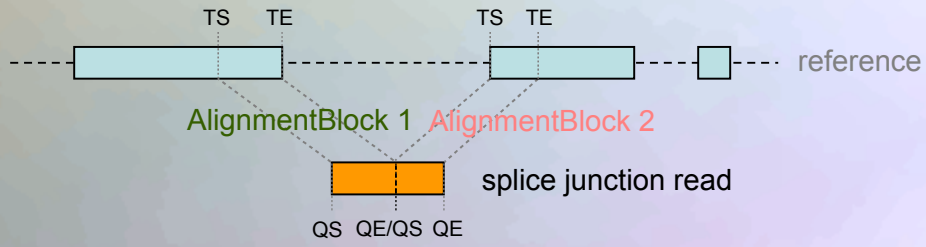
- Compact data summary format for short, long, single and paired-end read alignments
- Enables the anonymization of confidential information
- Still possible to carry out most functional genomics analyses

- MRF has three components:
  - comment lines
  - header line
  - mapped reads
- Header line:
  - AlignmentBlocks
  - Sequences
  - QualityScores
  - queryIDs

MRF: some examples

AlignmentBlock = TargetName:Strand:TargetStart:TargetEnd:QueryStart:QueryEnd

Comment lines	`# Example 1`		
Header line	`AlignmentBlocks`	`Sequence`	`QualityScores`
Mapped reads	`chr2:+:601:630:1:30,chr2:+:921:940:31:50`	`AGGCTCAAGCTTC…`	`effcghffegeggec…`

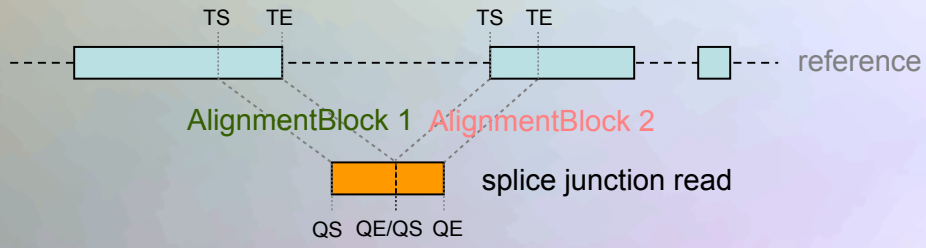


Legend
TS = TargetStart, TE = TargetEnd,
QS = QueryStart, QE = QueryEnd

MRF: some examples

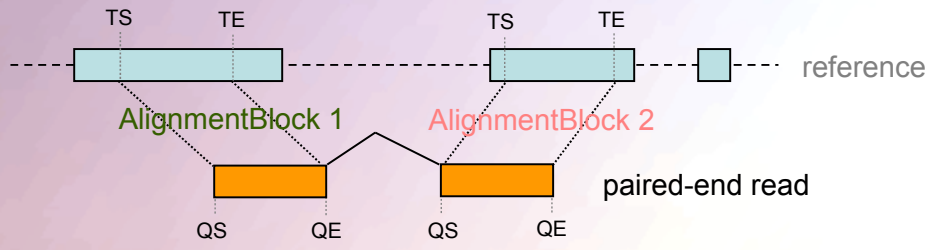
AlignmentBlock = TargetName:Strand:TargetStart:TargetEnd:QueryStart:QueryEnd

Comment lines	`# Example 1`		
Header line	`AlignmentBlocks`	`Sequence`	`QualityScores`
Mapped reads	`chr2:+:601:630:1:30,chr2:+:921:940:31:50`	`AGGCTCAAGCTTC...`	`effcghffegeggec...`



Legend
TS = TargetStart, TE = TargetEnd,
QS = QueryStart, QE = QueryEnd

Comment lines	`# Example 2`		
Header line	`AlignmentBlocks`	`Sequence`	`QualityScores`
Mapped reads	`chr9:+:431:480:1:50|chr9:+:945:994:1:50`	`AGGCTC...|TGCTTC...`	`effcgh...|fgggec...`

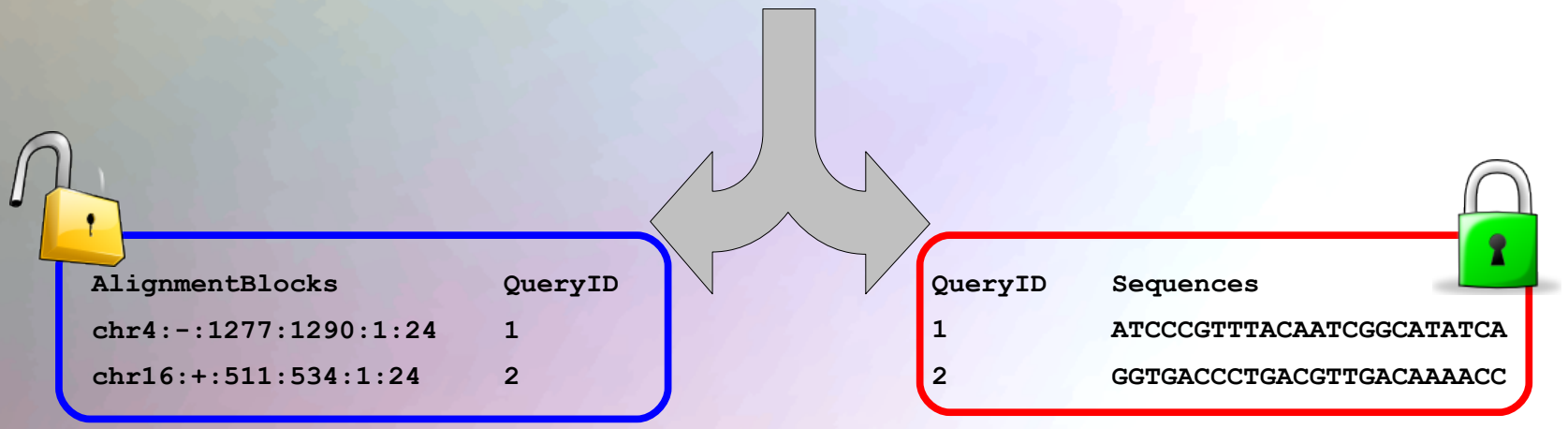


Anonymization of confidential sequence information

AlignmentBlocks	Sequences	QueryID
chr4:-:1277:1290:1:24	ATCCCGTTTACAATCGGCATATCA	1
chr16:+:511:534:1:24	GGTGACCCTGACGTTGACAAAACC	2

Anonymization of confidential sequence information

AlignmentBlocks	Sequences	QueryID
chr4:-:1277:1290:1:24	ATCCCGTTTACAATCGGCATATCA	1
chr16:+:511:534:1:24	GGTGACCCTGACGTTGACAAAACC	2



AlignmentBlocks	QueryID
chr4:-:1277:1290:1:24	1
chr16:+:511:534:1:24	2

Public

QueryID	Sequences
1	ATCCCGTTTACAATCGGCATATCA
2	GGTGACCCTGACGTTGACAAAACC

Private

Run Statistics for 1 Lane (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 minutes	MRF: 400Mb	Number of mapped reads: ~12 million
Quantification	*mrfQuantifier*	45 seconds	Gene expression values: 3.5Mb	~22,000 gene models
Visualization	*mrf2wig*	~2 minutes	One WIG file per chromosome: 1Mb - 150Mb	Signal track of mapped reads normalized per million mapped reads
	mrf2gff	45 seconds	One GFF file per chromosome: 100Kb - 16Mb	To visualize splice junction reads

Conclusions

- Developed a format for mapped reads (MRF)
  - Compact data summaries
  - Enables the anonymization of sequence information
  - Decouples the alignment step from the downstream analyses
- Implemented a suite of tools that uses this format for the analysis of RNA-Seq data sets
- Utilized this framework for several RNA-Seq projects

http://rseqtools.gersteinlab.org/