http://gbsc.stanford.edu

RNA-Seq Best Practices Workflows

Ramesh Nair

RNA-Seq Workflows

Variant Calling & Gene Expression

A high-level overview of NGS data processing

Base calling
(vendor tool)

FASTQ file:
raw NGS reads

Data
Pre-processing

Alignment or
assembly

SAM/BAM file:
aligned NGS reads

Gene Expression

FPKM/TPM file:
Transcript quantification

.sam file: Uncompressed text file
.bam file: Compressed and indexed file

Variant Calling

VCF file:
genomic variation

VCF Files can be block-compressed and indexed.

Data Formats

FASTQ: Raw unaligned reads

- Simple extension from traditional FASTA format.
- Each block has 4 elements (in 4 lines):
  - Sequence Name (read name, group, etc.)
  - Sequence
  - + (optional: Sequence name again)
  - Associated quality score.
- Example record:

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

Identifier

Sequence

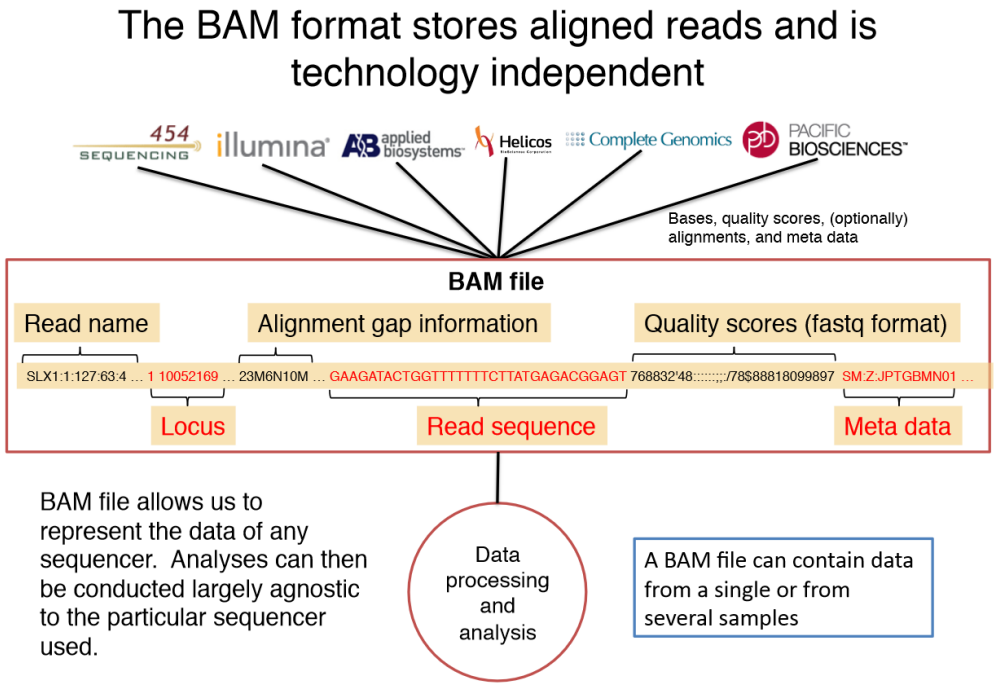
Base Qualities
(ASCII 33 + Phred scaled Q)

Sanger format
(standard FASTQ)

Sanger: Phred+33
Solexa/Illumina 1.0: Solexa+64
Illumina 1.3+: Phred+64
Illumina 1.5+: Phred+64
Illumina 1.8+: Phred+33

FASTQ Format Specification: http://maq.sourceforge.net/fastq.shtml

SAM/BAM: Aligned reads



SAM: Sequence Alignment/Map
SAM Format Specification: http://samtools.github.io/hts-specs/SAMv1.pdf

VCF: Variant calls

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

VCF: Variant Call Format

VCF Format Specification: http://samtools.github.io/hts-specs/VCFv4.2.pdf

RPKM/FPKM/TPM: Transcript quantification

- RPKM (single-end reads): Reads Per Kilobase of transcript per Million reads mapped
- FPKM (paired-end reads): Fragments Per Kilobase of transcript per Million reads mapped
- TPM: Transcripts Per Million

RNA-Seq Tools

RNA-Seq Alignment Tools

- **STAR**: Spliced Transcripts Alignment to a Reference

  Refs:
  https://code.google.com/p/rna-star/
  http://bioinformatics.oxfordjournals.org/content/29/1/15.long

- **TopHat2**: splice junction mapper for RNA-Seq reads

  Refs:
  http://ccb.jhu.edu/software/tophat/index.shtml
  http://genomebiology.com/2013/14/4/R36

STAR vs. TopHat2

Runtime and memory usage of RNA-Seq aligners

Program	Runtime (wall time)	Peak memory (GB)	Parameters
TopHat2 2.0.8 (Transcriptome only mapping)	8h 29m	4.9	-G --transcriptome-only --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (Default: genome and spliced mapping)	17h 1m	5.4	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (With transcriptome mapping)	17h 31m	5.2	-G --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (Realignment with realignment edit distance of 0)	29h 55m	5.6	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60 --read-realign-edit-dist 0
GSNAP 2013-01-23	55h 26m	7.6	--max-mismatches=3 -N 1
RUM 1.12_01	26h 34m	*36.4	
MapSplice 1.15.2	44h 50m	3.7	min_missed_seg = 0
STAR 2.3.0e	32m	27.8	--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3

Mapping statistics for simulated RNA-Seq data

Mapping parameters:

STAR:

1-pass (line 1): all default

2-pass (line 3): `--alignSJDBoverhangMin 1`

TopHat2 (with Bowtie2):

no realignment (line 2): `--mate-inner-dist 145 --mate-std-dev 50`

realignment (line 4): `--mate-inner-dist 145 --mate-std-dev 50 --read-realign-edit-dist 0`

Inner distance between mates and its standard deviation were calculated from the actual distribution for the simulated reads.

	Method		Correctly mapped 200 bases	>=150 bases correctly mapped	Unmapped	True positive junctions		False positive junctions	
						Number	Sensitivity	Number	FDR
Line number	Realignment	Aligner	1	2	4	5	6	7	8
1	no (1-step)	STAR	81.3%	95.0%	4.82%	148,487	92.7%	409	0.3%
2	no	TopHat2	82.6%	83.7%	6.70%	135,006	84.3%	1,228	0.9%
3	yes (2-step)	STAR	98.5%	99.5%	0.40%	148,789	92.9%	453	0.3%
4	yes	TopHat2	96.4%	96.8%	1.22%	136,416	85.2%	3,132	2.2%

Refs:

http://genomebiology.com/2013/14/4/R36

http://biorxiv.org/content/biorxiv/early/2013/11/22/000851.full.pdf

Transcript Quantification Tools

- **Cufflinks**: Transcript assembly, differential expression, and differential regulation for RNA-Seq (FPKM)

  Refs:
  http://cufflinks.cbcb.umd.edu/manual.html
  http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html

- **RSEM**: RNA-Seq by Expectation-Maximization (TPM/FPKM)

  Refs:
  http://www.biomedcentral.com/1471-2105/12/323
  http://deweylab.biostat.wisc.edu/rsem/

- **Sailfish**: alignment-free transcript quantification (TPM/RPKM)

  Refs:
  http://www.nature.com/nbt/journal/v32/n5/full/nbt.2862.html
  http://www.cs.cmu.edu/~ckingsf/software/sailfish/

SAM/BAM Processing Tools

- Picard Tools

a) Reorder SAM/BAM: match the chromosome ordering in a provided reference file
b) Sort SAM/BAM: sort reads in coordinate order for each chromosome
c) Mark Duplicates: remove duplicate reads

Ref: http://picard.sourceforge.net/

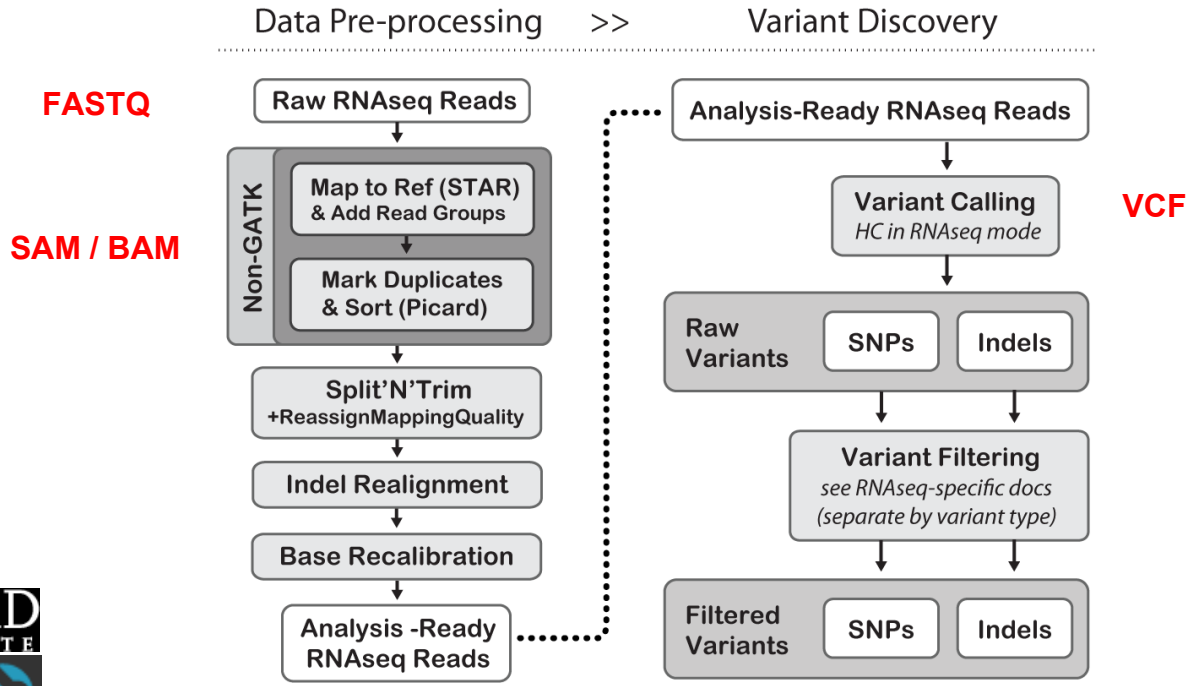
- SAMtools

a) Index BAM
b) Sort BAM
c) SAM to BAM and vice versa

Ref: http://samtools.sourceforge.net/

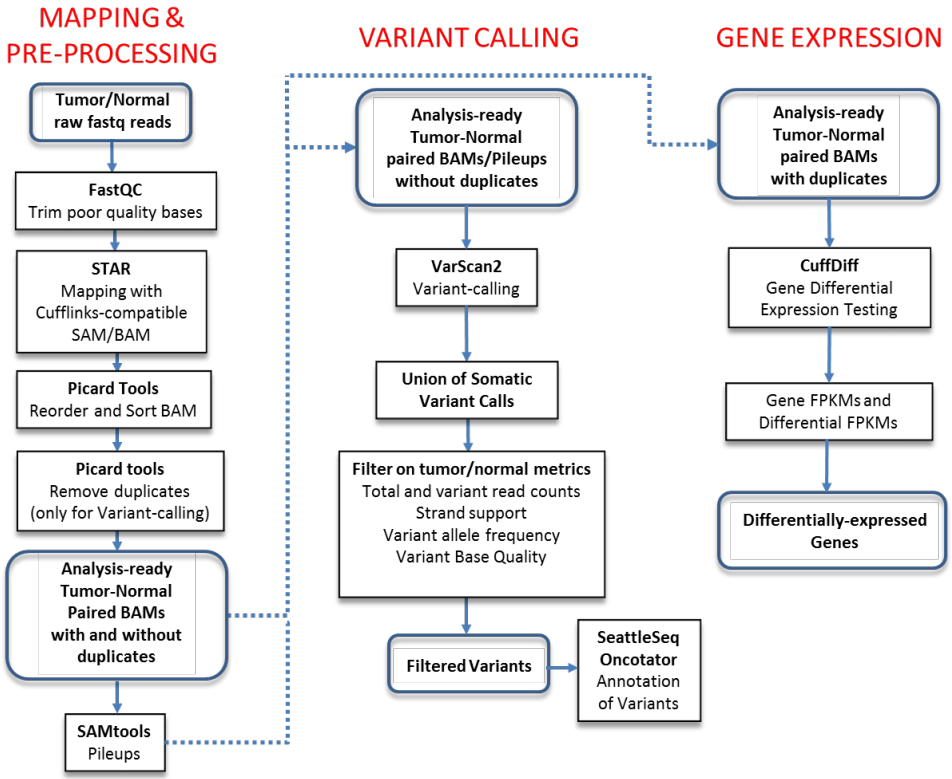
RNA-Seq Pipelines

Broad GATK Best Practices workflow for variant calling on RNA-Seq data



Workflow in practice: http://www.broadinstitute.org/gatk/guide/article?id=3891

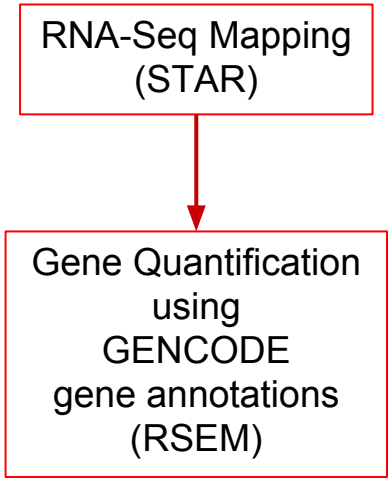
Stanford CCSB RNA-Seq pipeline



Stanford CCSB

ENCODE RNA-Seq pipeline (work-in-progress)

ENCODE is in the process of developing a set of uniform processing pipelines for the data generated by the Consortium



Ref:
https://github.com/ENCODE-DCC/long-rna-seq-pipeline

RNA-Seq Galaxy Workflow Demo

https://gbsc-galaxy.stanford.edu/u/rvnair/p/galaxy-rna-seq-workflow-demo

Acknowledgments

- BioTeam
- SCGPM
- CCSB (Sylvia Plevritis, Andrew Gentles)