Walter+Eliza Hall
Institute of Medical Research

From reads to genes in less than 10 *R* commands

Wei Shi

Walter and Eliza Hall Institute of Medical Research
Melbourne, Australia

July 19, 2013

FIAT LUX

Walter+Eliza Hall
Institute of Medical Research

A bioconductor pipeline for RNA-seq analysis

- **Rsubread** package
  - Read mapping
  - Read summarization
  - Exon-exon junction detection
  - Indel detection
  - *Fusion detection*
  - *SNP calling*

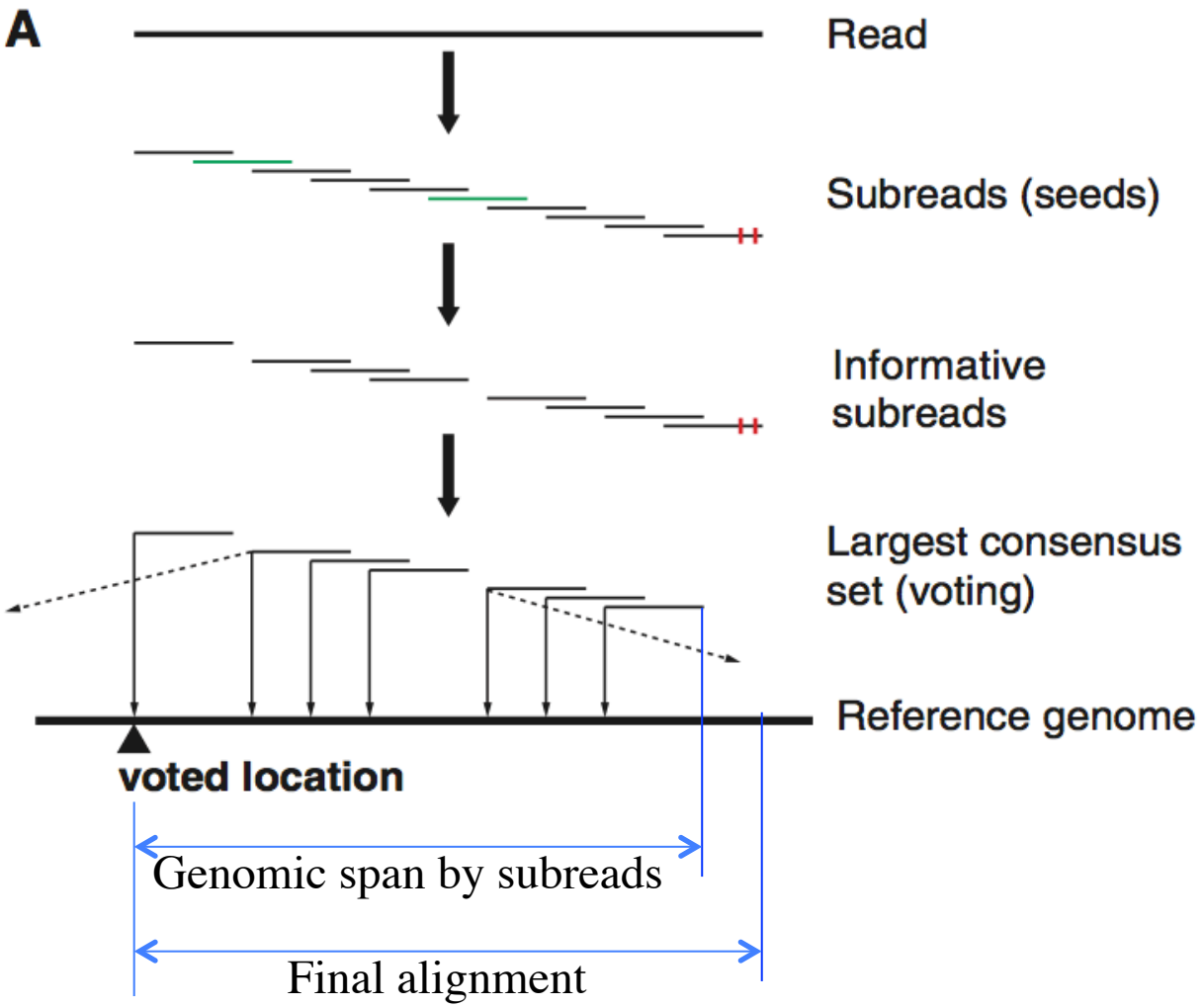
  …. …
- **Limma** package
  - Voom: modelling mean-variance relationship
  - Differential expression analysis
  - Gene set test

  … …

Read mapping (Subread)

Seed-and-vote: a new mapping paradigm

A number of seeds (16bp mers) are extracted from each read and they vote for mapping location of the read.



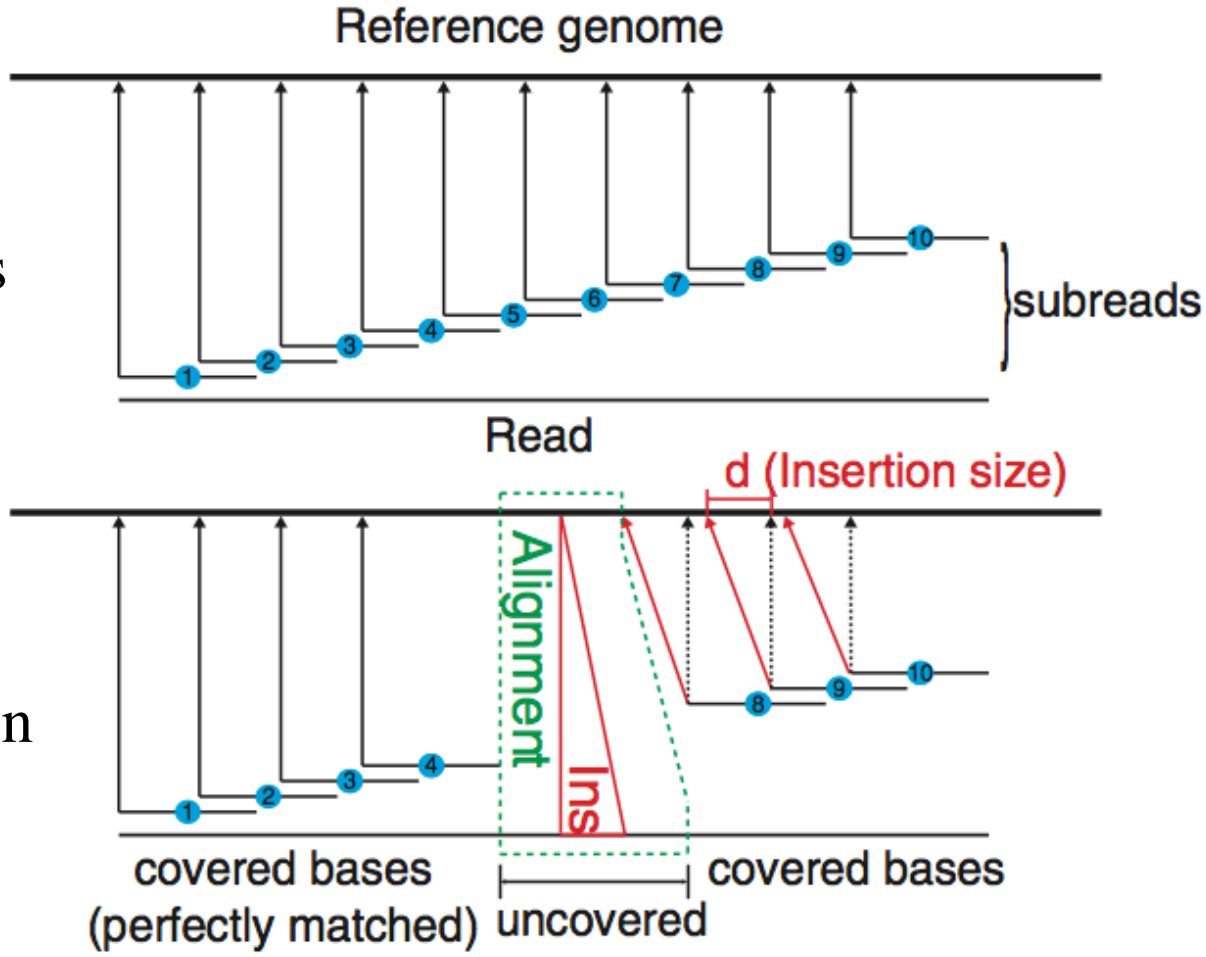
Liao, Smyth and Shi, NAR, 2013

Walter+Eliza Hall
Institute of Medical Research

Indel detection and final alignment

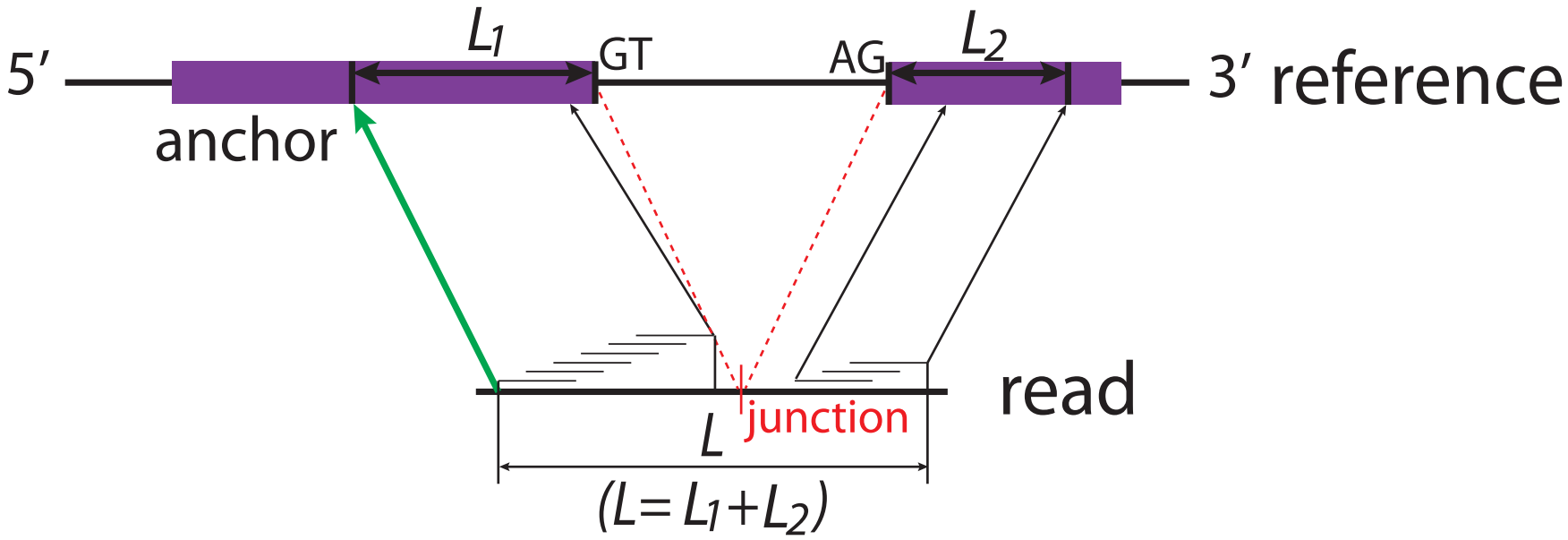
No indels

An insertion



Liao, Smyth and Shi, NAR, 2013

Exon-exon junction detection (Subjunc)



All reads are re-aligned after junctions are identified.

Liao, Smyth and Shi, NAR, 2013

Walter+Eliza Hall
Institute of Medical Research

Multi-mapping

When more than one location receives the highest number of votes, we choose the one which has

- largest genomic span by consensus subreads
- highest mapping quality score (optional)

$$\mathrm{MQS} = 100 + \frac{100}{l}\left\{\sum_{i \in b_m}(1 - p_i) - \sum_{i \in b_{mm}}(1 - p_i)\right\}$$

- smallest edit distance (optional)

Mapping genomic DNA sequencing data

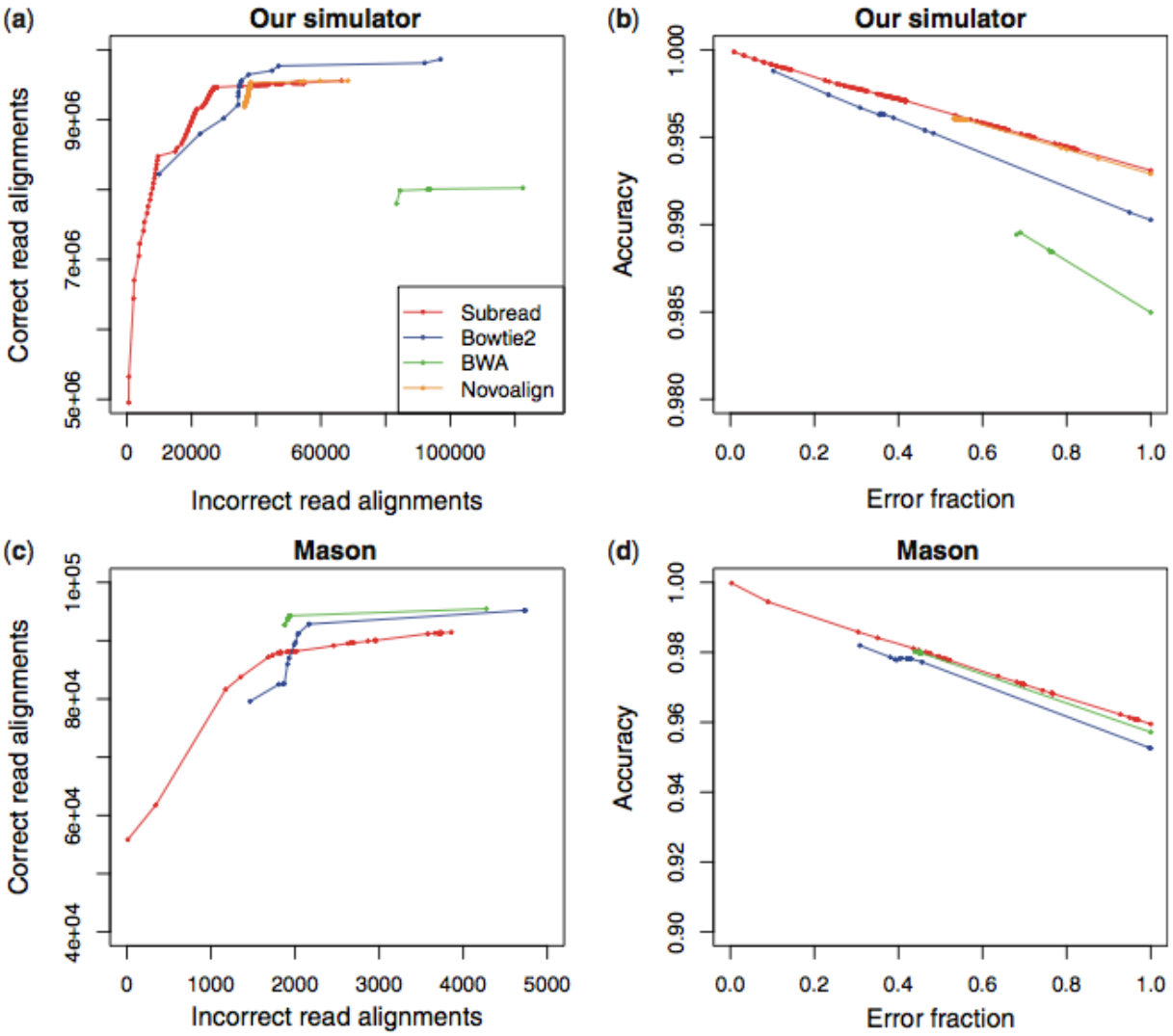
Table 1. Performance of aligners in mapping genomic DNA reads from the 1000 Genomes project

Aligner	Mapped (%)	Rabema intervals (%)	Time (h)	Memory (Gb)
Subread (default)	97.7	86.7	1.6	7.6
Subread (low memory)	97.7	86.7	2.9	4.3
Bowtie2	99.1	87.2	6.0	3.3
BWA	95.6	82.6	15.2	3.3
Maq	98.1	86.3	48.3	19.1
Novoalign	93.9	68.9	18.7	8.2
MrsFast	70.3	73.8	48.2	25.8

Liao, Smyth and Shi, NAR, 2013

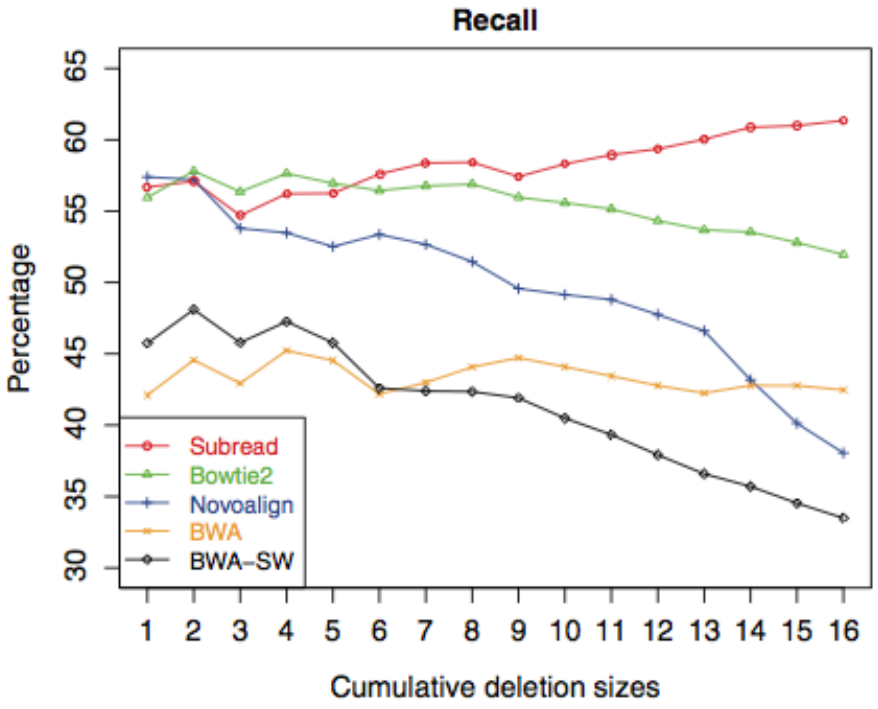
Walter+Eliza Hall
Institute of Medical Research

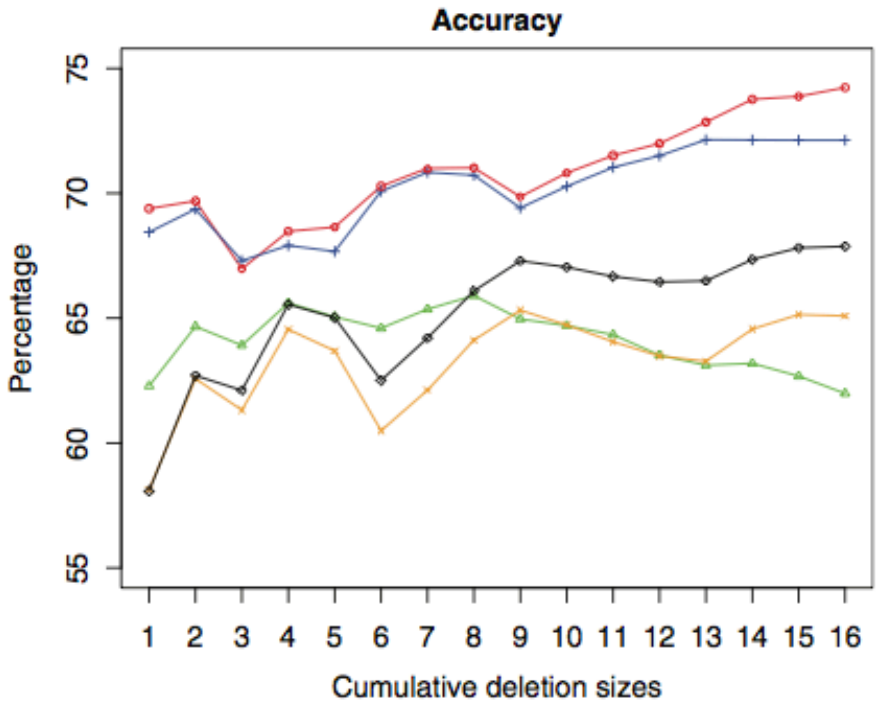
Mapping accuracy



Liao, Smyth and Shi, NAR, 2013

Indel detection





Liao, Smyth and Shi, NAR, 2013

Mapping RNA sequencing data

Both Subread and Subjunc can align RNA-seq reads.

The main difference between them is that Subjunc performs full alignments for exon-spanning reads, but Subread reports only the largest mappable region in such reads.

However, Subread is sufficient if the purpose of the RNA-seq experiment is to carry out an expression analysis.

Liao, Smyth and Shi, NAR, 2013

Mapping SEQC/MAQC III RNA-seq reads

Table 2. Performance of aligners in mapping RNA-seq reads from the SEQC project

Aligner	Mapped (%)	Time (min)	Memory (Gb)
Subread (default)	96.9	23	7.6
Subread (low memory)	96.9	40	4.3
Bowtie2	85.7	90	3.3
BWA	78.6	284	3.3
Maq	66.4	685	5.2
Novoalign	78.4	361	8.1
MrsFast	46.2	398	7.4

Liao, Smyth and Shi, NAR, 2013

Walter+Eliza Hall
Institute of Medical Research

Detecting exon-exon junctions

Aligners	# junctions ('000)	% known junctions	% reads supporting known junctions	Speed (Hrs)	Memory (GB)
Subjunc	152	84.4	95.8	1.4 (1.9)	8.4 (4.7)
MapSplice	171	78.3	94.4	5.6	4.3
TopHat	156	82.5	93.8	9.2	2.9
TopHat 2	152	83.8	94.1	9.9	3.5

Liao, Smyth and Shi, NAR, 2013

Advantage of seed-and-vote over seed-and-extend

- Candidate mapping locations are determined by virtually the entire read sequence
  - Thus achieving very high mapping accuracy
- Much more computationally efficient
  - In most cases, final read alignments are only required for one genomic location
  - Final alignments are only required for uncovered read bases
- No need to specify the number of mismatches allowed
  - Allow more mismatches while retaining high mapping accuracy at the same time
- Naturally suitable for discovering exon-exon junctions and fusions
  - Using many short subreads enables sensitive detection of multiple mapping locations

Read summarization

To quantify expression levels of genes, mapped reads need to be assigned to genes.

We assign reads to genes by comparing mapped regions of reads with the exonic regions of genes.

We have developed a read summarization program called *featureCounts*, which can be accessed from both Bioconductor Rsubread package and SourceForge Subread package.

Walter+Eliza Hall
Institute of Medical Research

featureCounts

Ultrafast read assignments by feature indexing

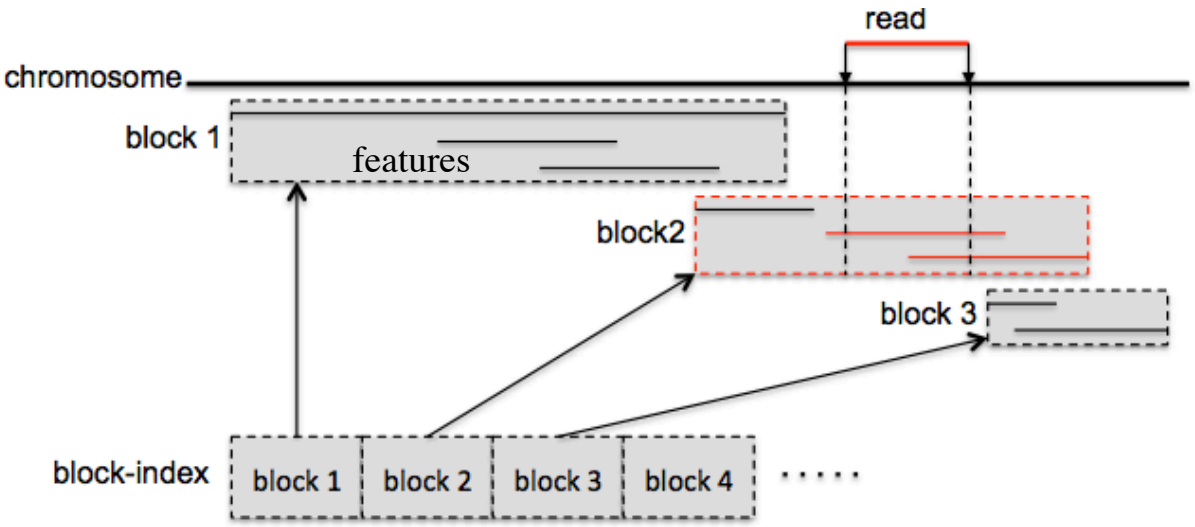


Fig. 1: Indexing blocks of features and identifying features overlapping with the query read.

Liao, Smyth and Shi, submitted

Features and meta-features

featureCounts performs read summarization at feature level or meta-feature level

- A feature is a continuous region in the genome, such as an exon.
- A meta-feature is an aggregation of one or more features. For example, a gene is meta-feature and each of its exons is a feature.

Overlap between reads and features

A read is said to overlap with a feature if there is at least 1 base overlap found between them. A read is said to overlap with a meta-feature if it overlaps with at least one of its features.

Multi-overlapping

A read is a multi-overlapping read if it overlaps with more than one feature when summarization is performed at feature level, or if it overlaps with more than one meta-feature when summarization is performed at meta-feature level.

Summarizing SEQC RNA-seq reads to NCBI RefSeq genes.

Methods	# reads	# fragments	Time (Mins)	Memory (MB)
featureCounts	4,385,354	4,796,948	1.1	33
summarizeOverlaps	4,385,354	3,942,439	12.1 (41.7)	3400 (661)
htseq-count	4,385,207	4,769,913	22.7	101

Liao, Smyth and Shi, submitted

Summarizing reads to features located on a large number of chromosomes/contigs

Table 3. Summarizing RNA-seq reads to genes located on 500,000 contig sequences. *featureCounts* was run using one thread.

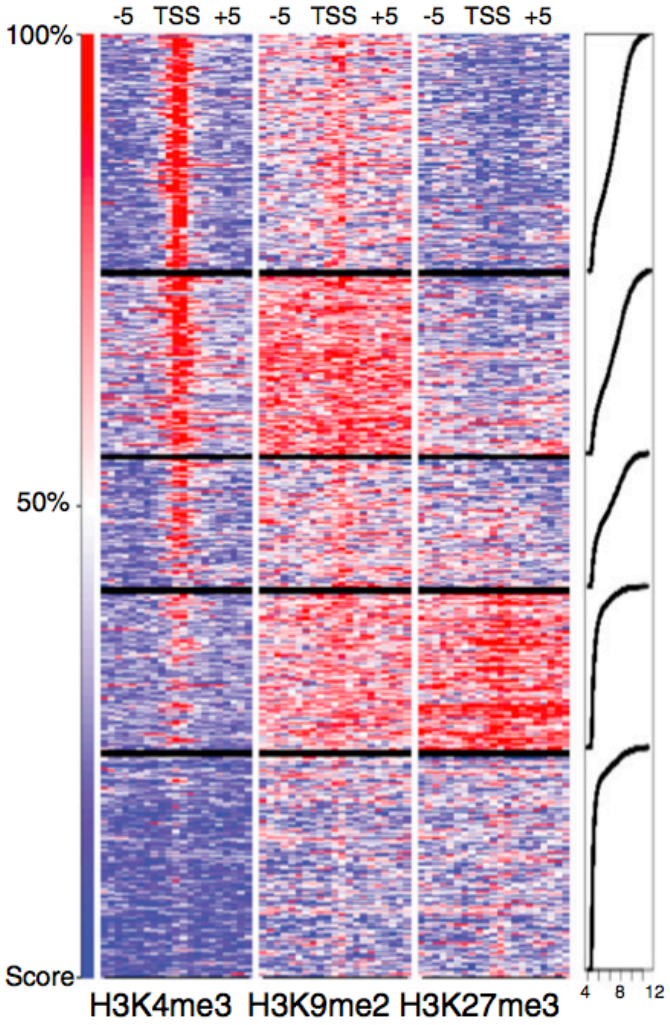
Methods	# reads	Time (Mins)	Memory (GB)
featureCounts	7,122,463	1.8	1.0
summarizeOverlaps	7,122,463	10 days	3.2
htseq-count	7,114,419	19.6	3.4

Liao, Smyth and Shi, submitted

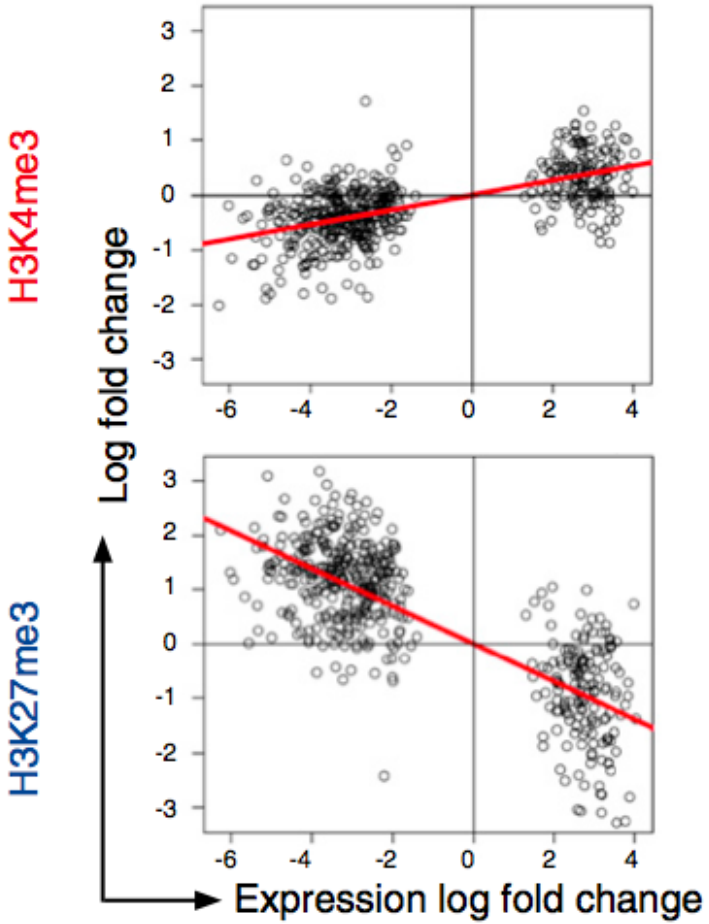
Walter+Eliza Hall
Institute of Medical Research

Summarizing histone ChIP-seq reads

Mammary stem cells



Mammary stem cells VS
luminal progenitors



Pal et al. Cell Reports, 2013

Summarizing H3K27me3 reads to regions including gene bodies and promoters, for each gene. Multi-overlapping is allowed.

Methods	# fragments	Time (Mins)	Memory (MB)
featureCounts	5,392,155	1.0	6
countOverlaps	5,392,155	24.4 (36.6)	7000 (783)
htseq-count (union)	4,978,050	36.0	31
htseq-count (intersection-nonempty)	4,993,644	35.7	31

Pal et al. Cell Reports, 2013

Features of featureCounts

- Perform precise and accurate read assignments by taking care of indels, junctions and fusions in the reads
- Support multi-threaded running
- Support both SAM and BAM format input
- Support strand-specific read summarization
- Allows users to specify if multi-overlapping is allowed
- Allow users to check if both ends are mapped and/or if the paired-end distances satisfy the distance criteria
- Allow users to specify whether chimeric fragments should be counted
- Allow users to use mapping quality scores to filter out reads.

featureCounts output

A count table that can be readily used for differential expression analysis.

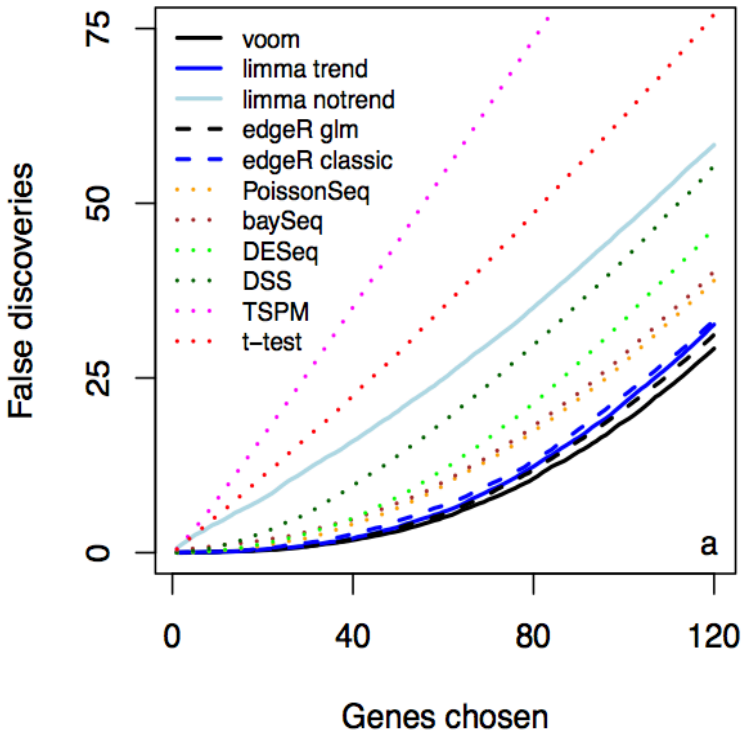
Differential expression analysis with limma/voom

- voom is an acronym for mean-variance modelling at the observational level
- Limma, a very popular package for analyzing microarray data, can now be used to analyze RNA-seq data

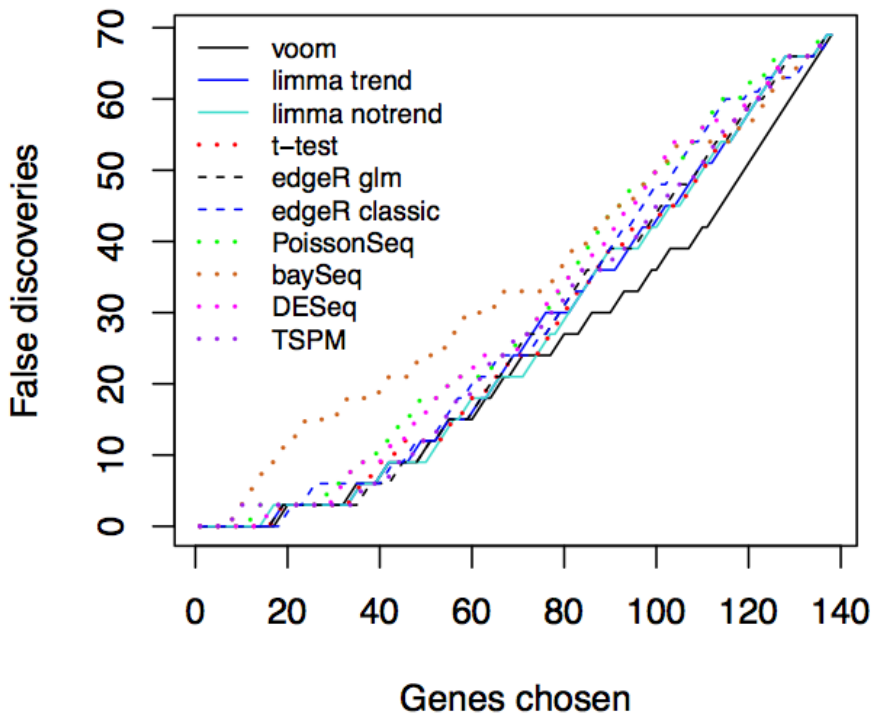
Walter+Eliza Hall
Institute of Medical Research

Voom is a powerful and accurate method

simulation



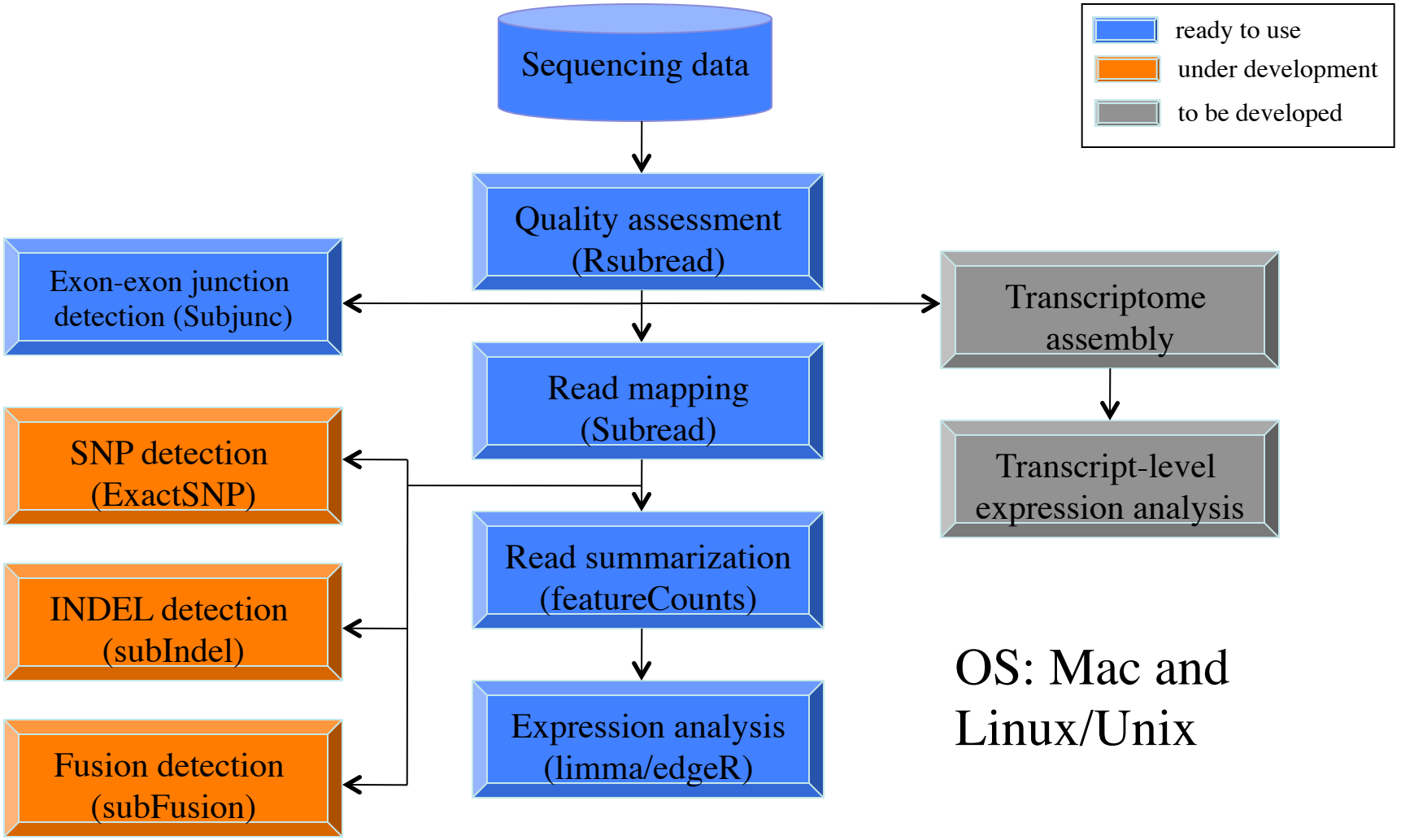
SEQC spike-in data



Law, Chen, Shi and Smyth, submitted

Walter+Eliza Hall
Institute of Medical Research

An Biocondutor R pipeline for analyzing RNA-seq data



OS: Mac and Linux/Unix

Walter+Eliza Hall
Institute of Medical Research

Essential R commands for an RNA-seq analysis

```
buildindex(basename="hg19",reference="hg19.fa")
align(index="hg19",readfile1="reads.fq",output_file="reads.sam")
fcounts <- featureCounts(files="reads.sam",genome="hg")
design <- model.matrix(~celltype)
y <- voom(fcounts$counts[isexpr,],design)
fit <- eBayes(lmFit(y,design))
topTable(fit,coef=2)
```

Data used in the Workshop

- Reference sequence
  - Human chromosome 1 (GRCh37)

- Read data (SEQC pilot data)
  - Two libraries from sequencing of Universal Human Reference RNA sample: A_1 and A_2
  - Two libraries from sequencing of Human Brain Reference RNA sample: B_1 and B_2
  - Reads mapped to Chr 1 were chosen (according to Subread) for this workshop

- Annotation
  - NCBI human RefSeq annotation (build 37.2)

Index building

- ***buildindex*** command in Rsubread
  - Input and output
    - It takes FASTA format reference sequences as input and saves generated index files to the hard disk
  - Configurable memory use
    - Divide genome into blocks and save memory usage when performing read mapping
  - Build base-space or color-space genome index
  - Removing uninformative subreads
    - Improve mapping accuracy and reduce running cost

Alignment

- ***align*** command in Rsubread
  - Input and output
    - It takes FASTQ/FASTA read files and reference index as input and outputs mapping results to the hard disk
  - Number of subreads selected and consensus threshold
    - Control mapping sensitivity and specificity
  - Output uniquely mapped reads
    - Break ties using Edit distance and mapping quality scores
  - Ouput multiple best mapping locations
  - Detect indels of up to 16bp long
    - Next release will support the detection of long indels (>200bp)

- ***Subjunc*** command in Rsubread
  - Designed for detecting exon-exon junctions

Subread (*align*) and Subjunc

- Both can be used for mapping RNA-seq reads for the purpose of expression analysis
  - Subjunc is slightly more accurate but less sensitive
  - Subread is faster
  - We use Subread in this Workshop

- Subjunc should be used instead of Subread for
  - Exon-exon junction detection
  - Genomic variant detection (SNPs, indels …) in RNA-seq data

- Subread can also be used for mapping gDNA-seq reads

Read summarization

- ***featureCounts*** command in Rsubread
  - Input and output
    - It takes as input a list of SAM/BAM format files and annotation data and returns an *R List* object including a read count table
    - It supports GFF/GTF format and SAF (Simplified Annotation Format) annotations
    - A SAF format *R data frame*:

GeneID	*Chr*	*Start*	*End*	*Strand*
497097	chr1	3204563	3207049	-
497097	chr1	3411783	3411982	-
497097	chr1	3660633	3661579	-
100503874	chr1	3637390	3640590	-
100503874	chr1	3648928	3648985	-
100038431	chr1	3670236	3671869	-
…				

Expression analysis

- ***Voom*** in limma package
  - Input and output
    - It takes as input a numeric *matrix* containing raw counts or a *DGEList* object and returns an *EList* object containing normalized expression values and inverse variance weights

- *lmFit* and *eBayes* in limma package
  - *lmFit* takes *voom* output as input and fit linear models to genes
  - *eBayes* takes *lmFit* output as input and assesses differential expression

All materials included in this Workshop can be found at:

http://bioinf.wehi.edu.au/RNAseqCaseStudy

Walter+Eliza Hall
Institute of Medical Research

WEHI Bioinformatics