Differential Gene Expression

Konstantin Riege
FLI Jena

Data Source

PLOS | ONE

RESEARCH ARTICLE

A complex association between DNA methylation and gene expression in human placenta at first and third trimesters

Yen Ching Lim[1], Jie Li[1], Yiyun Ni[1], Qi Liang[1], Junjiao Zhang[1], George S. H. Yeo[2], Jianxin Lyu[1]*, Shengnan Jin[1]*, Chunming Ding[1]*

1 Key Laboratory of Laboratory Medicine, Ministry of Education of China, School of Laboratory Medicine and Life Science, Wenzhou Medical University, Wenzhou, Zhejiang, China, **2** KK Women's and Children's Hospital, Singapore, Singapore

Data Source

- 1st trimester vs. 3rd trimester human RNA-Seq samples (N=4)
  - 1T paired end N1,N2,N3,N4
  - 3T paired end N2,N3
  - 3T single end N1,N4

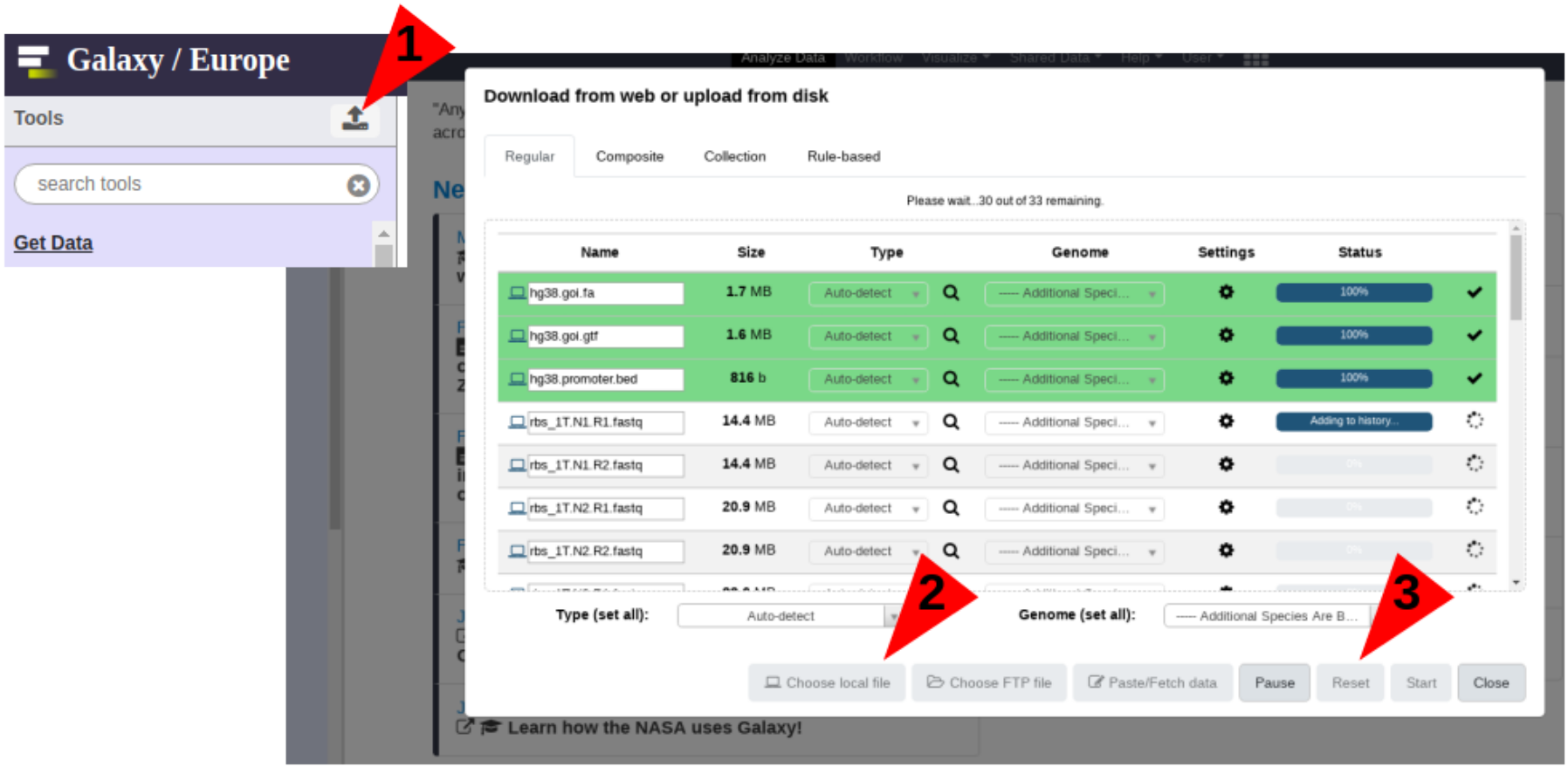
- ~2400 diff. expressed genes
- 21 of them with diff. methylated promoter regions (1kb upstream region)

Data Source

- Downsampling
  - HG38 toy genome composed of 21 genes of interest (GOI) and 4kb intergenic regions (2kb upstream and 2kb downstream)
  - 10% of GOI mapped reads extracted from available RAW sequencing data

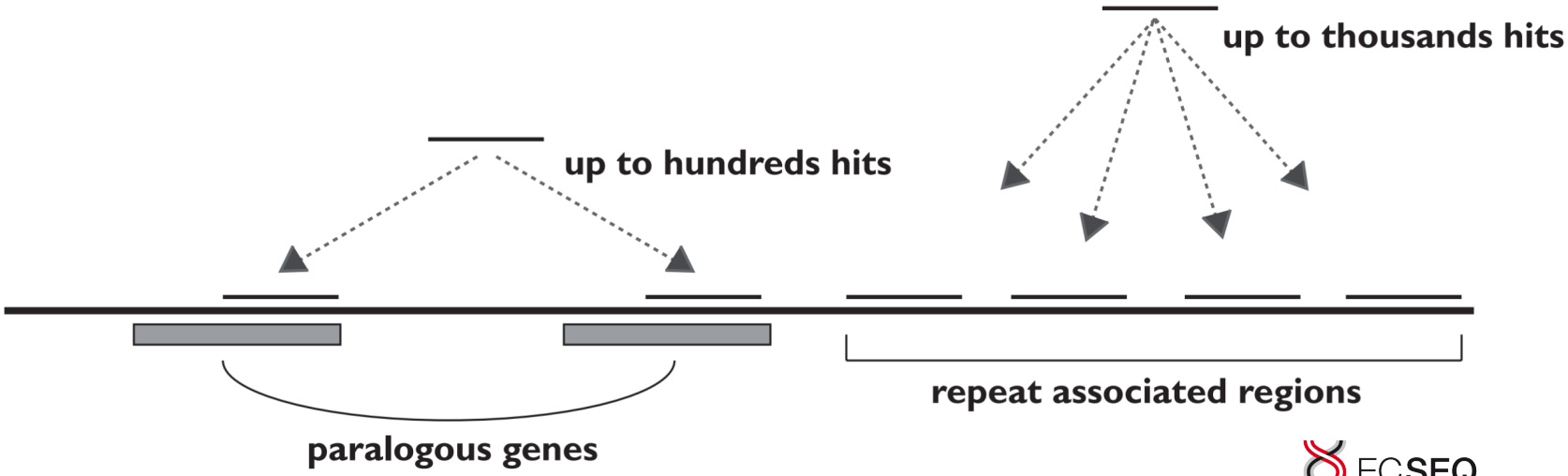
https://github.com/destairdenbi/trainings/tree/master/raw_data/bs_tour

Upload GOI Data



Alignment postprocessing

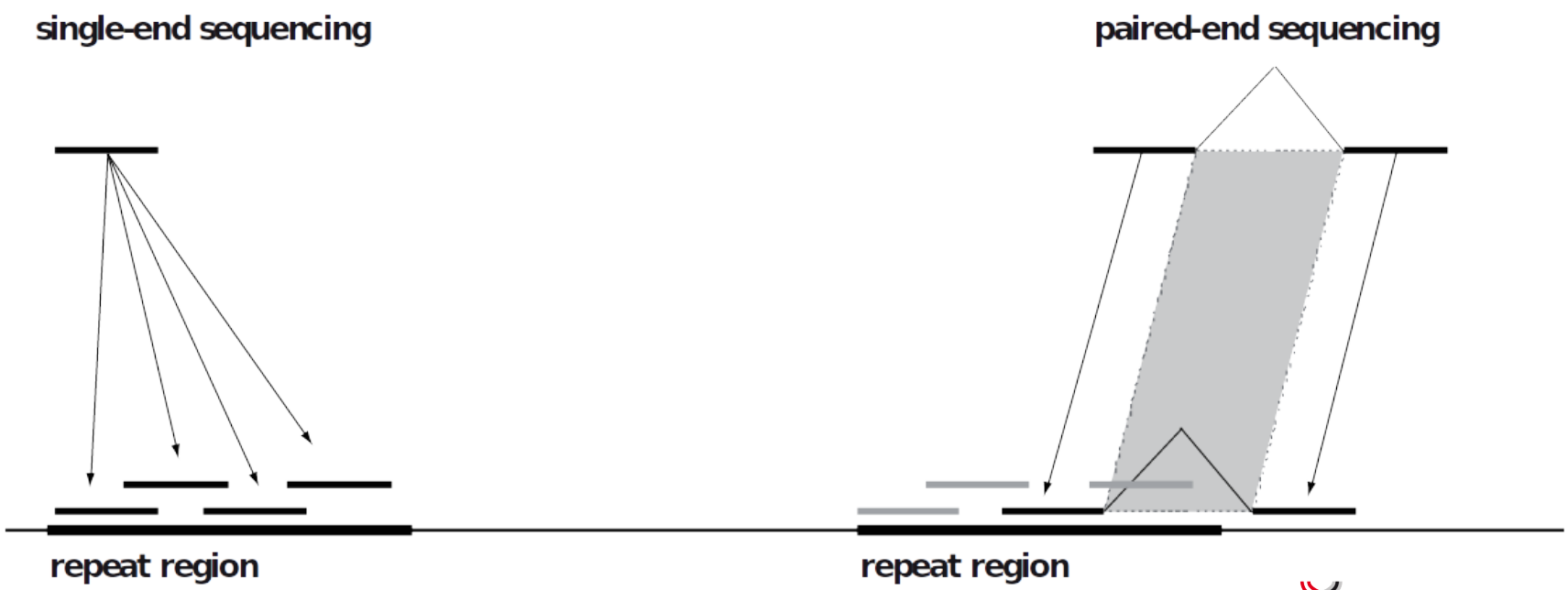
- Extract uniquely/unambigiously aligned reads



ECSEQ BIOINFORMATICS

Alignment postprocessing

- Select only alignments with both read mates mapped properly paired



ECSEQ BIOINFORMATICS

Alignment postprocessing

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

Alignment postprocessing

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

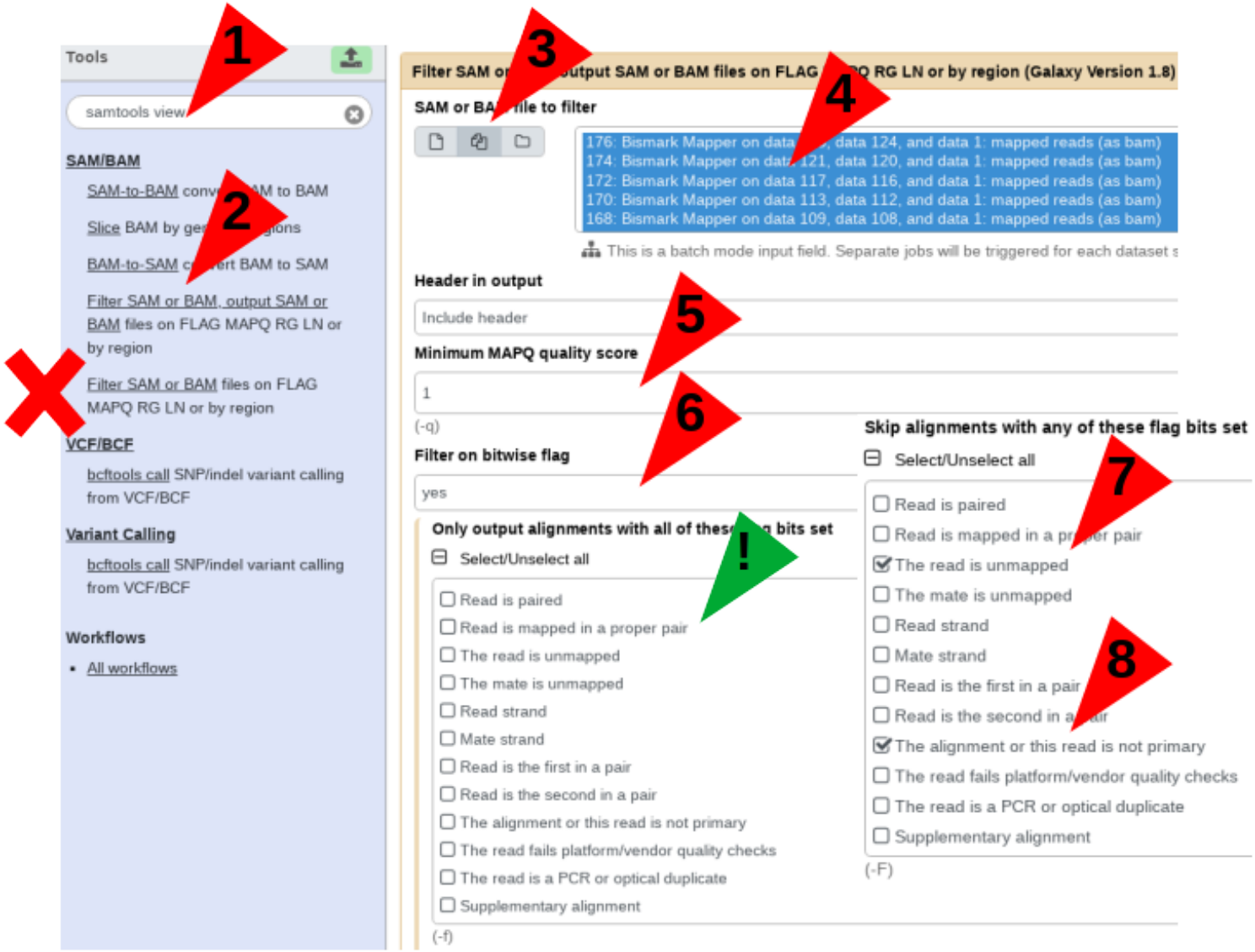
Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Alignment postprocessing

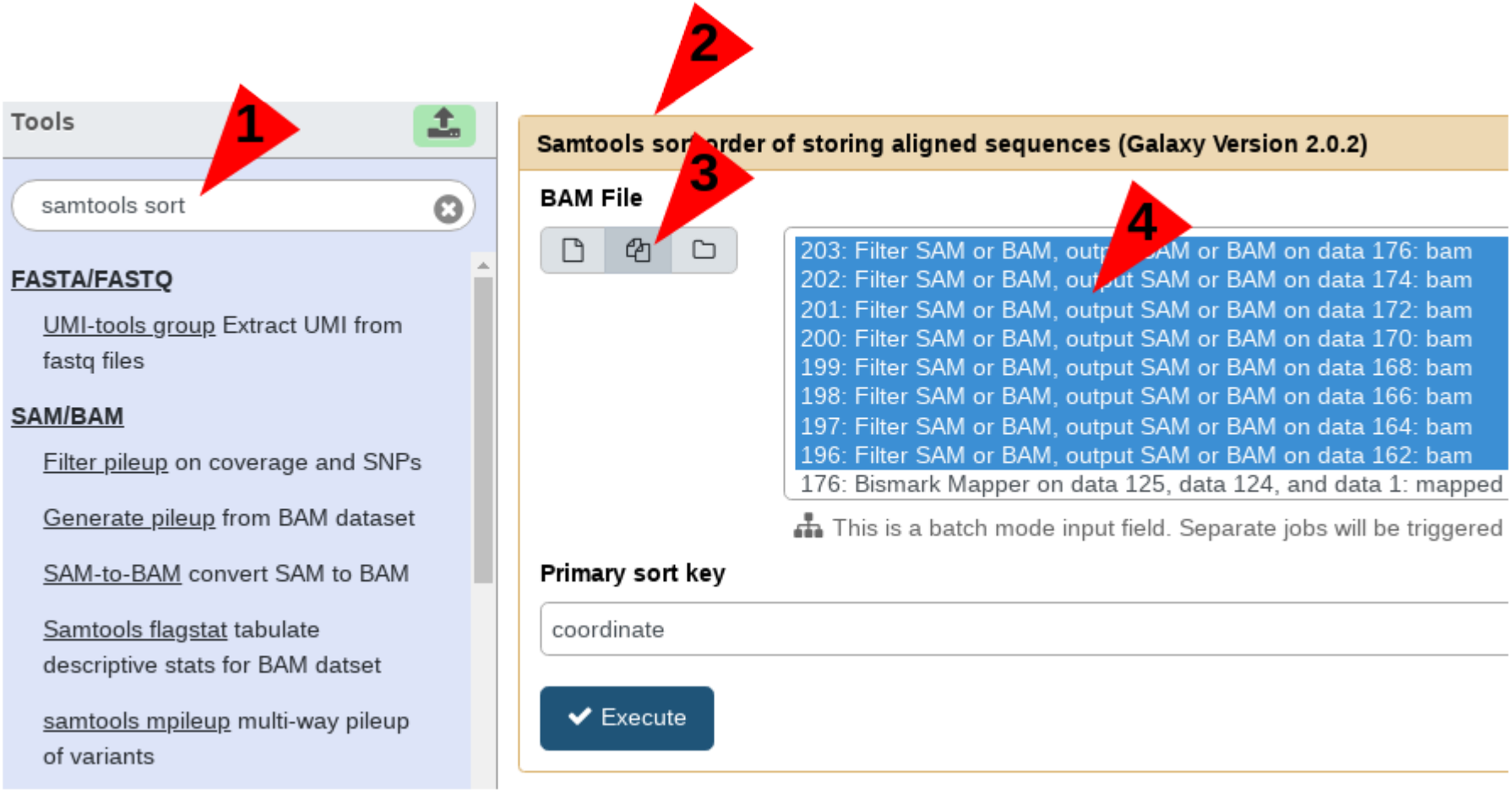
- Filter BAM by
  - alignment flags - pair information: 2
  - alignment tags - mapping count: NH:i:*
  - mapping quality - often an indicator for unique alignments
    - STAR: =255
    - HISAT2: >0
- Sort BAM

1. Samtools view
2. Samtools sort

Alignment postprocessing



Alignment postprocessing

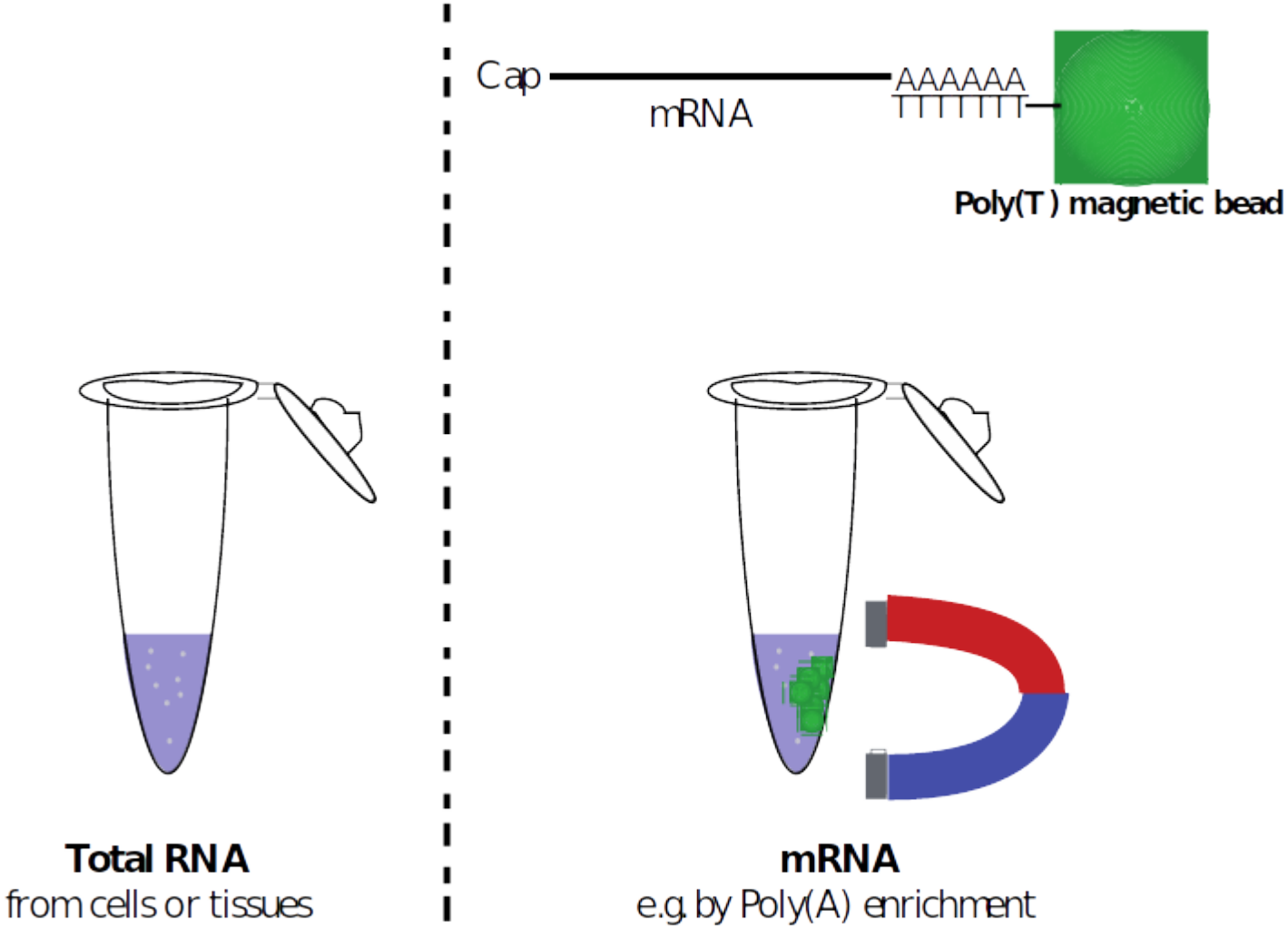


Quantification

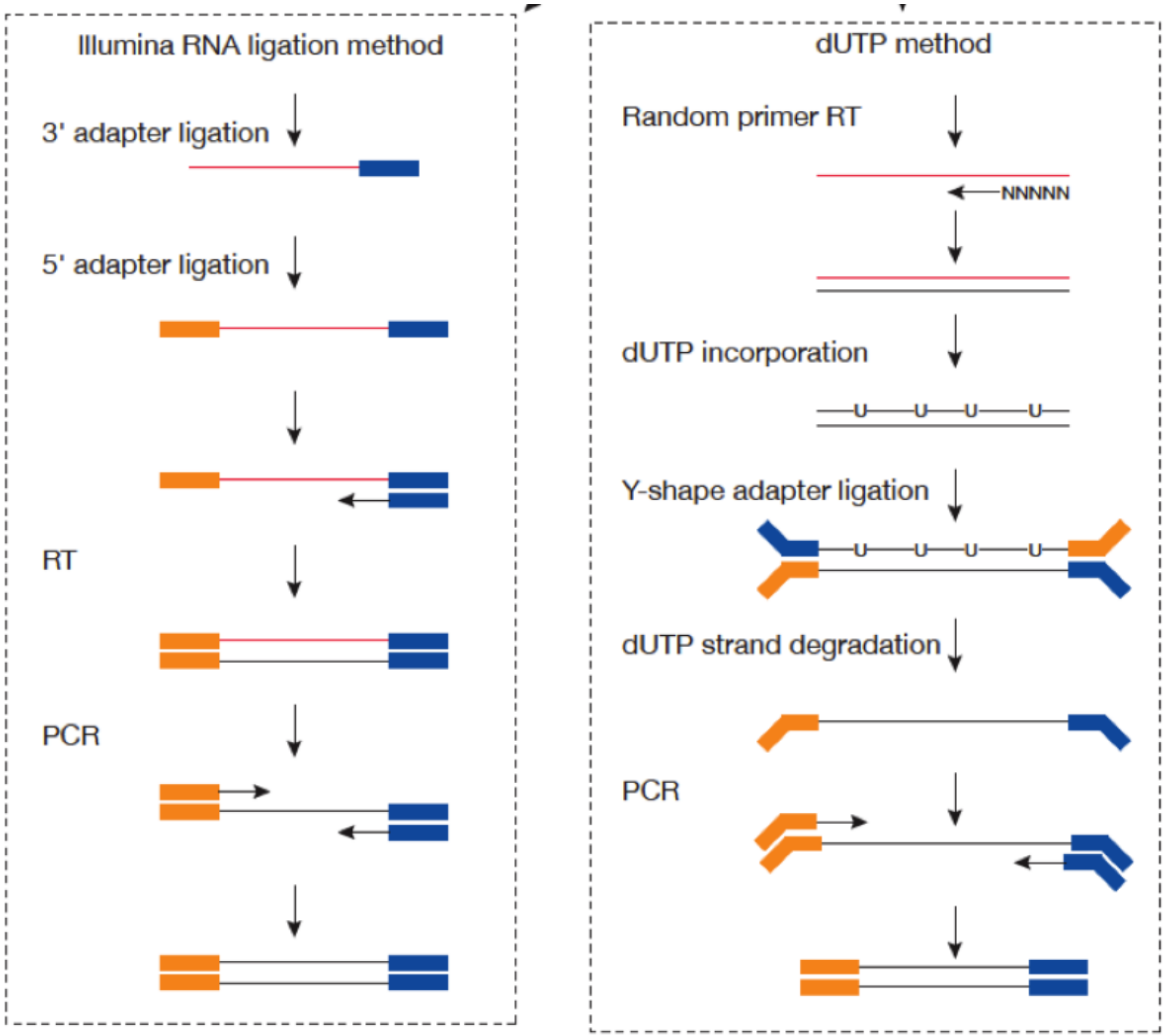
- Requieremets:
  - Annotation
  - sequencing strandness information

- DETOUR: infer experimental setup

Infer experiment strandness



Infer experiment strandness

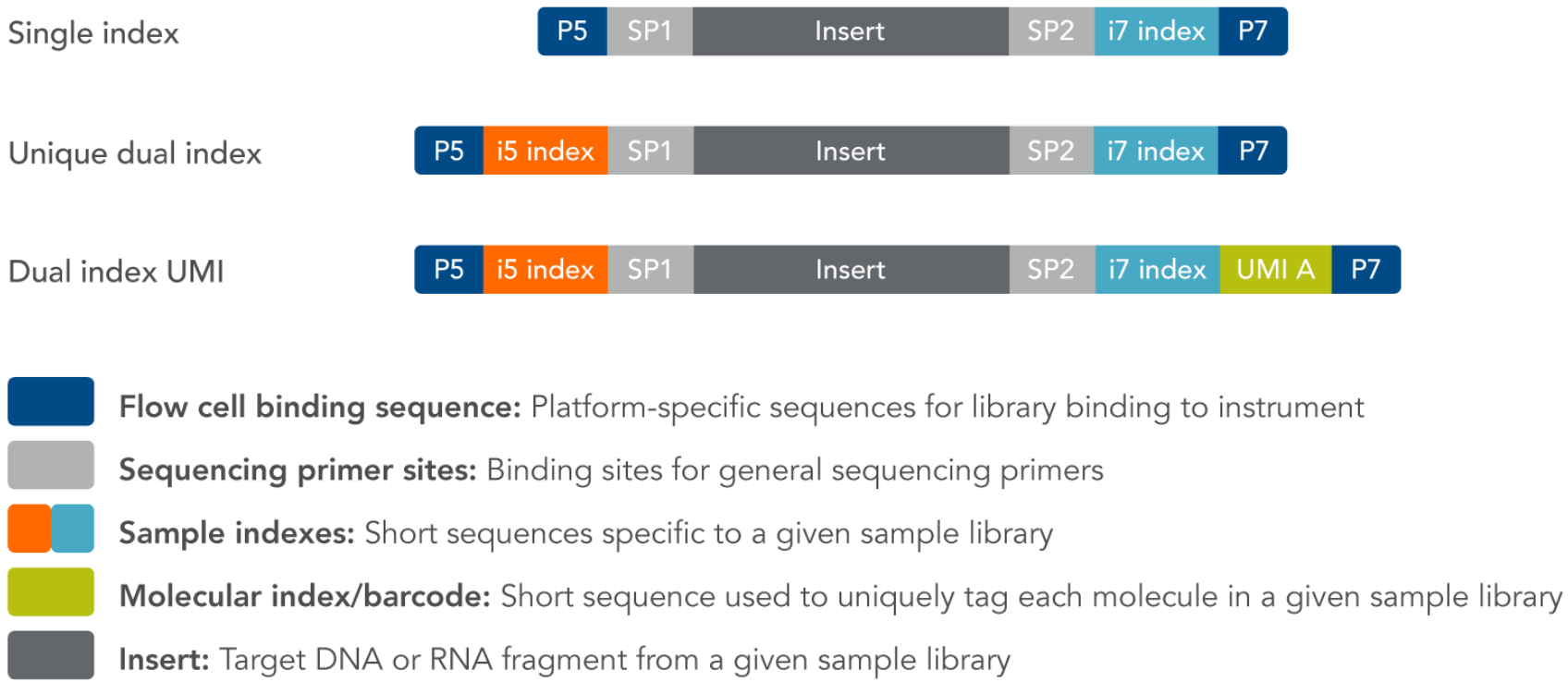


fragment RNA + cDNA

RNA from cDNA using Deoxyuridine Triphosphate + cRNA

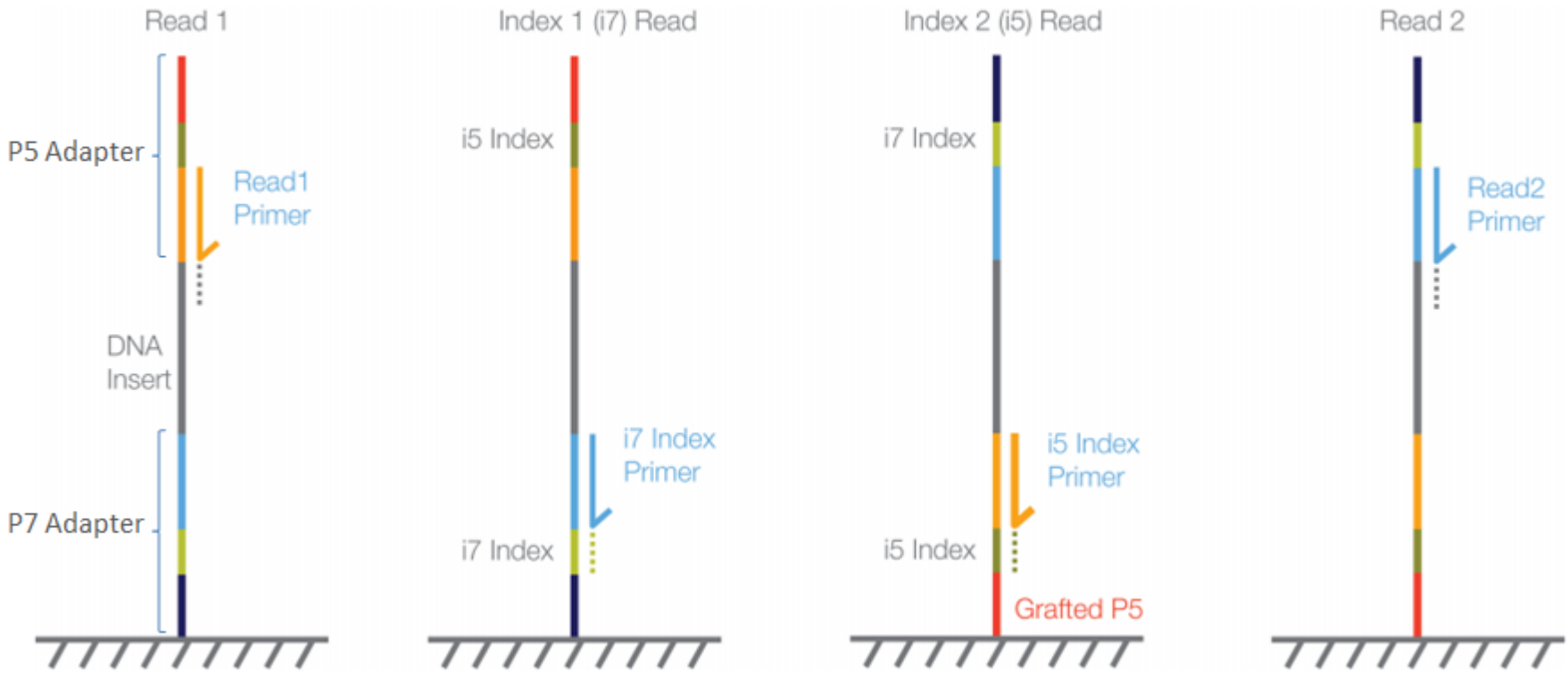
Image Credit: Zhao Zhang

Infer experiment strandness



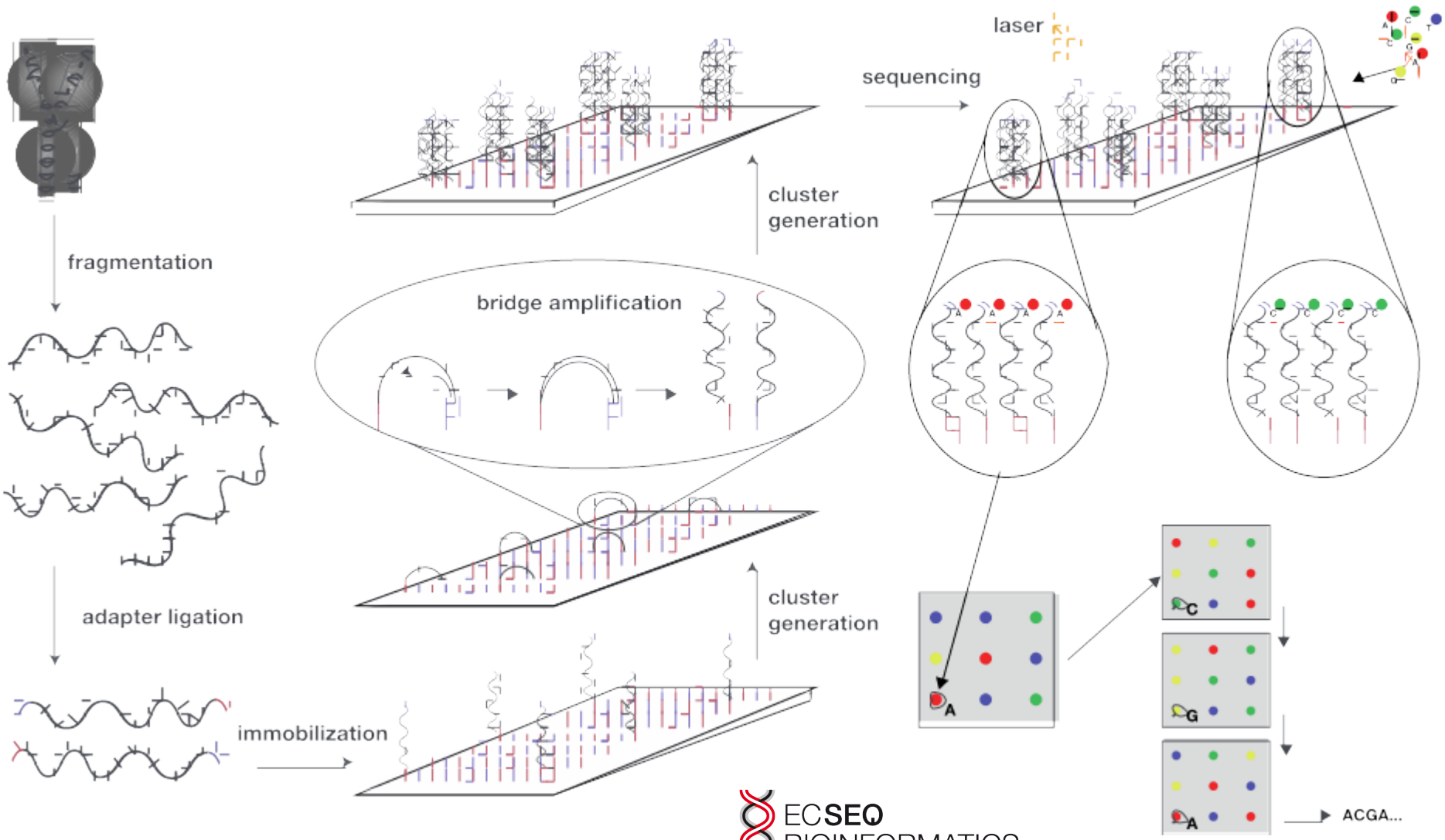
https://eu.idtdna.com/pages/products/next-generation-sequencing/adapters

Infer experiment strandness



https://support.illumina.com/bulletins/2016/04/adapter-trimming-why-are-adapter-sequences-trimmed-from-only-the--ends-of-reads.html

Infer experiment strandness



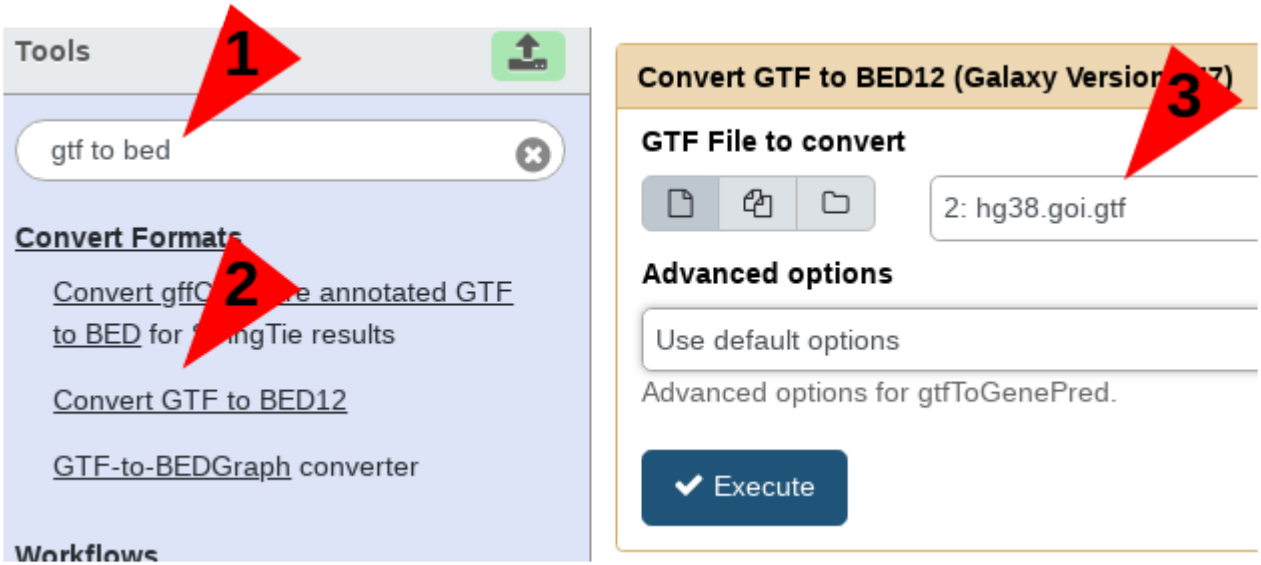
ECSEQ BIOINFORMATICS

Infer experiment strandness

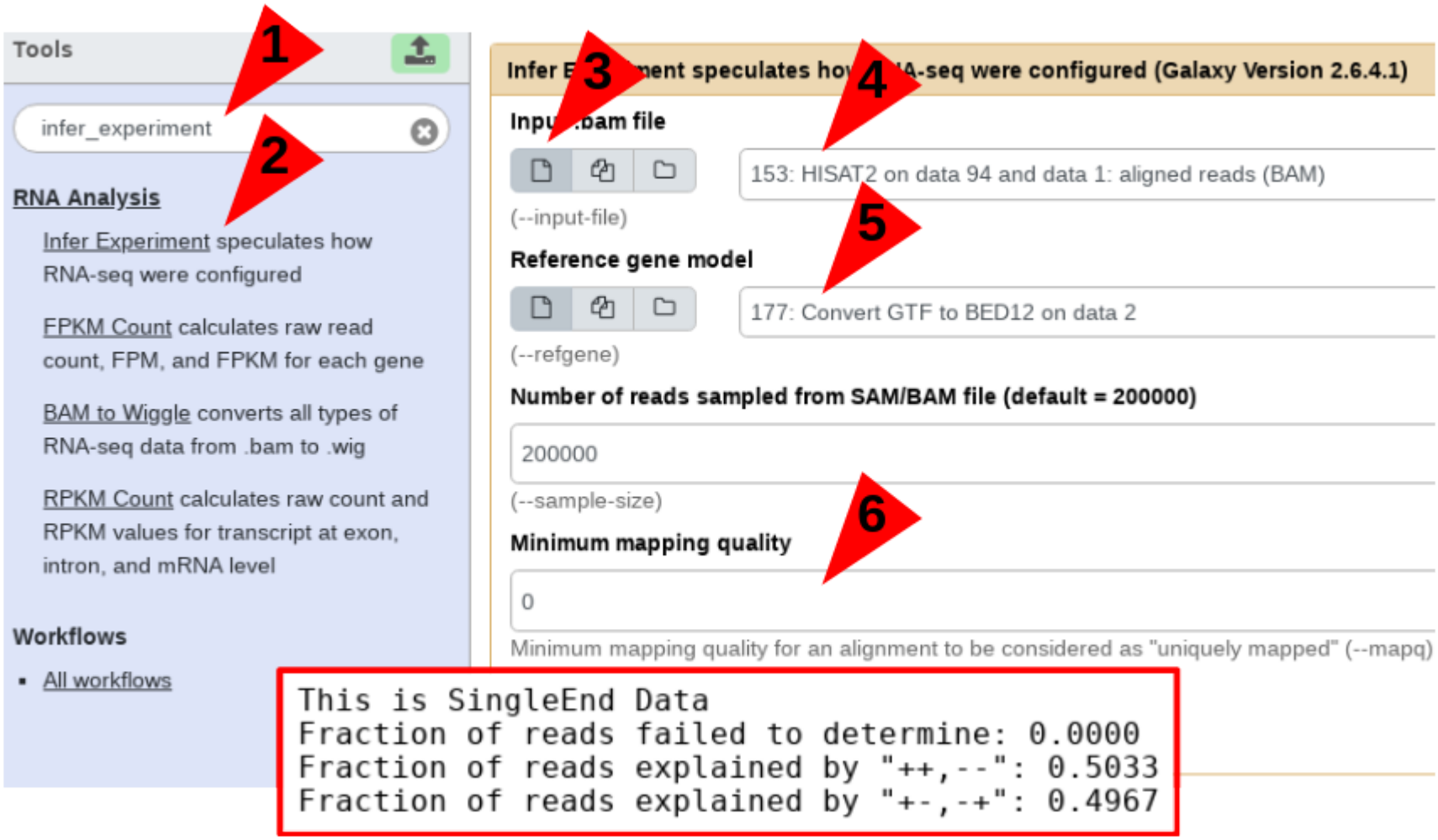
- Tool box: RSeQC
- Requierements:
  - Annotation in BED format

1. GTF to BED conversion
2. infer_experiment.py

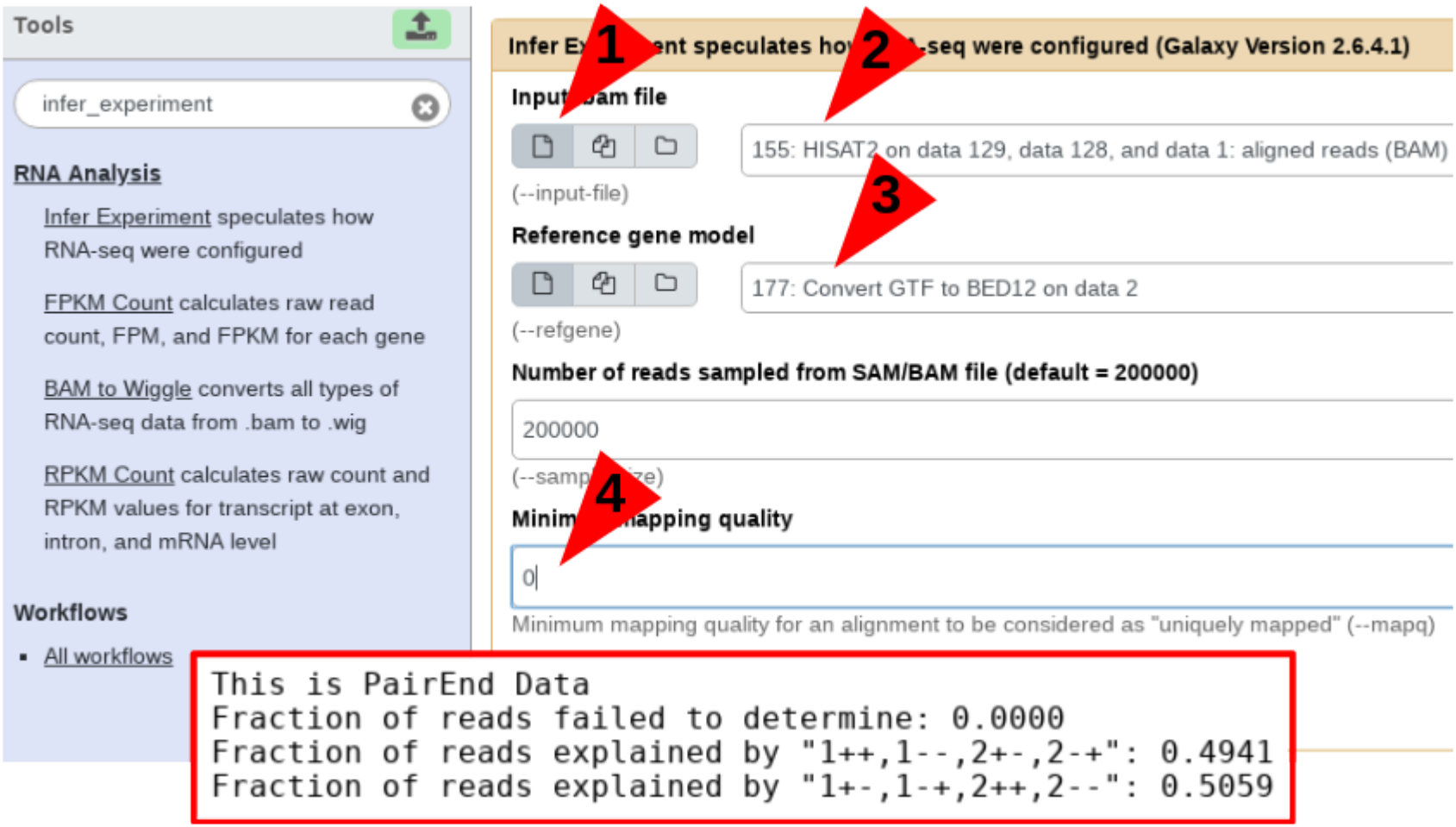
Infer experiment strandness



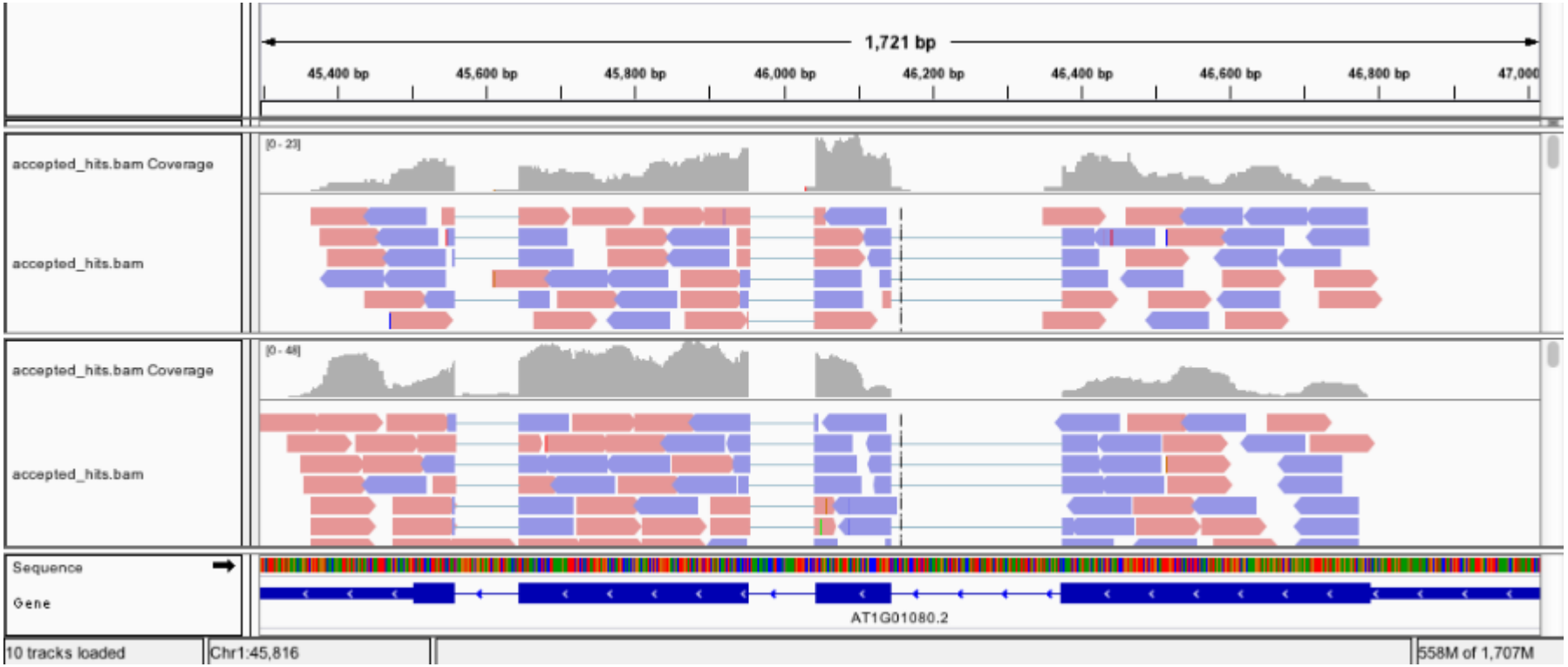
Infer experiment strandness: SE



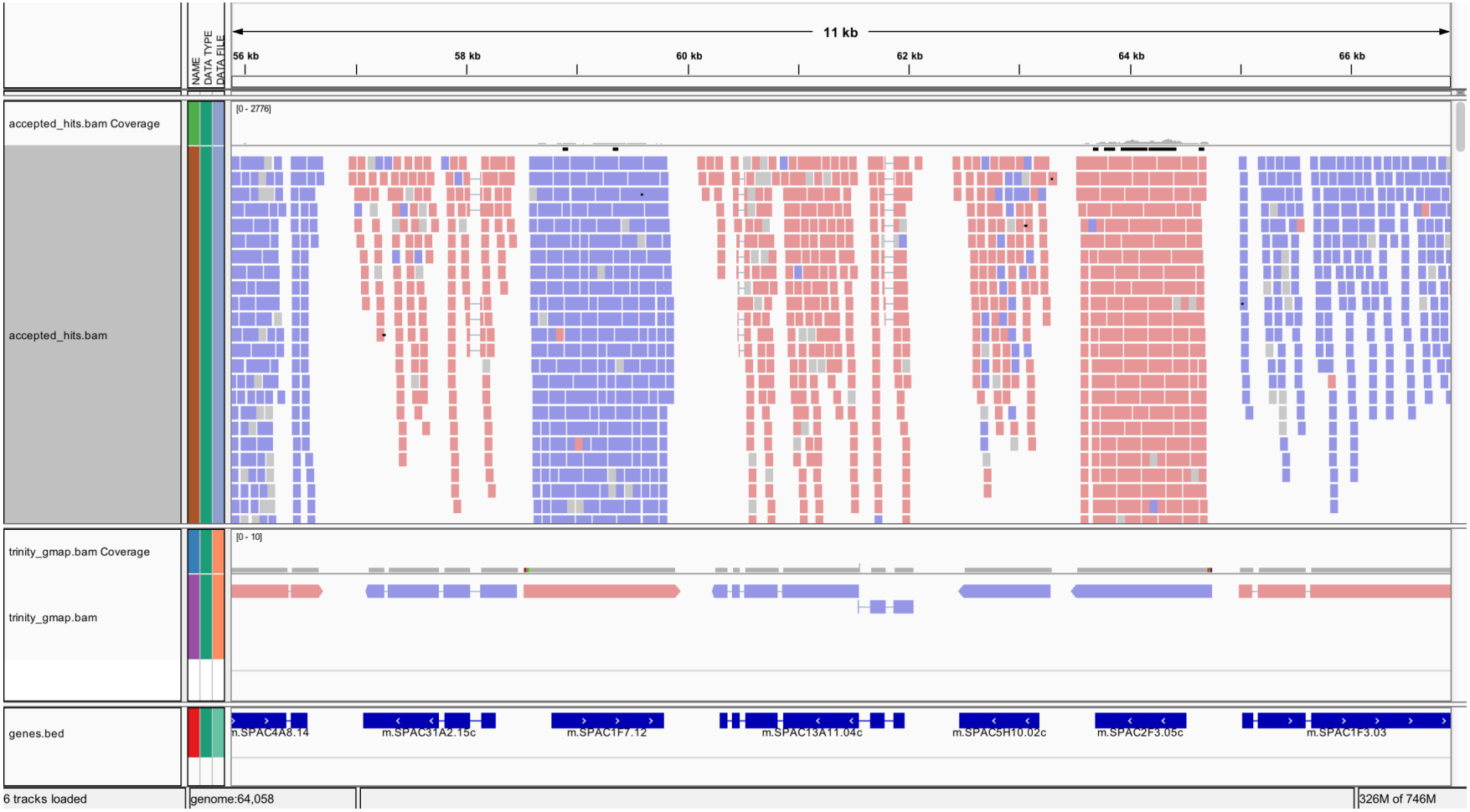
Infer experiment strandness: PE



Infer experiment strandness



Infer experiment strandness

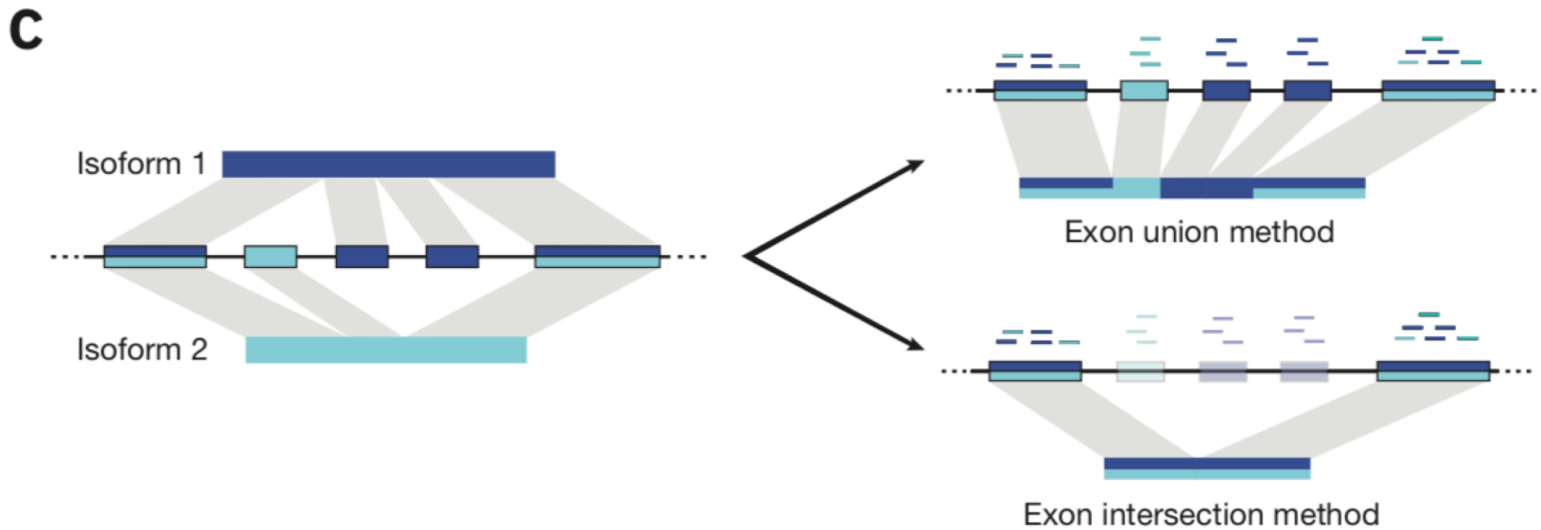


Quantification

- Count uniquely mapped reads only
- What to count?
  - Fragments
  - Reads
- Where to count?
  - Gene body
  - Exons

Quantification

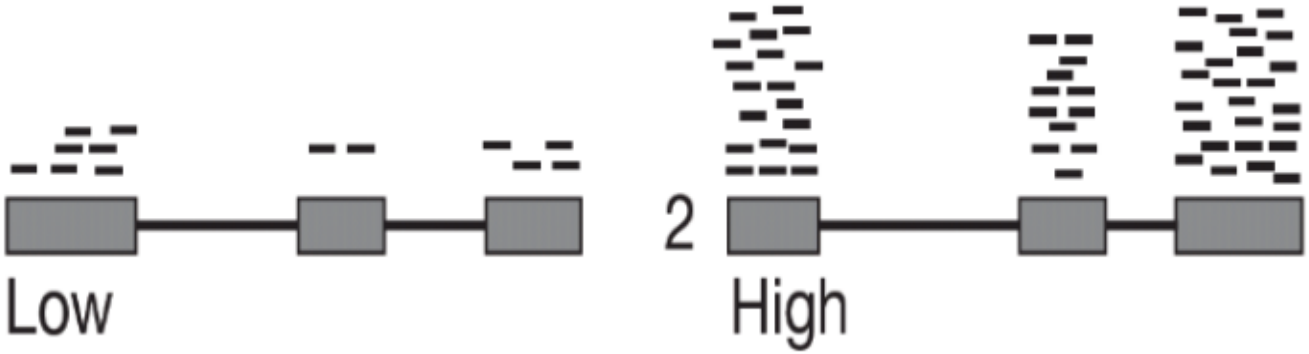
- Count fragments over all exons



Quantification

- For comparison, normalize counts by
  - Transcript length

  - Sequencing depth

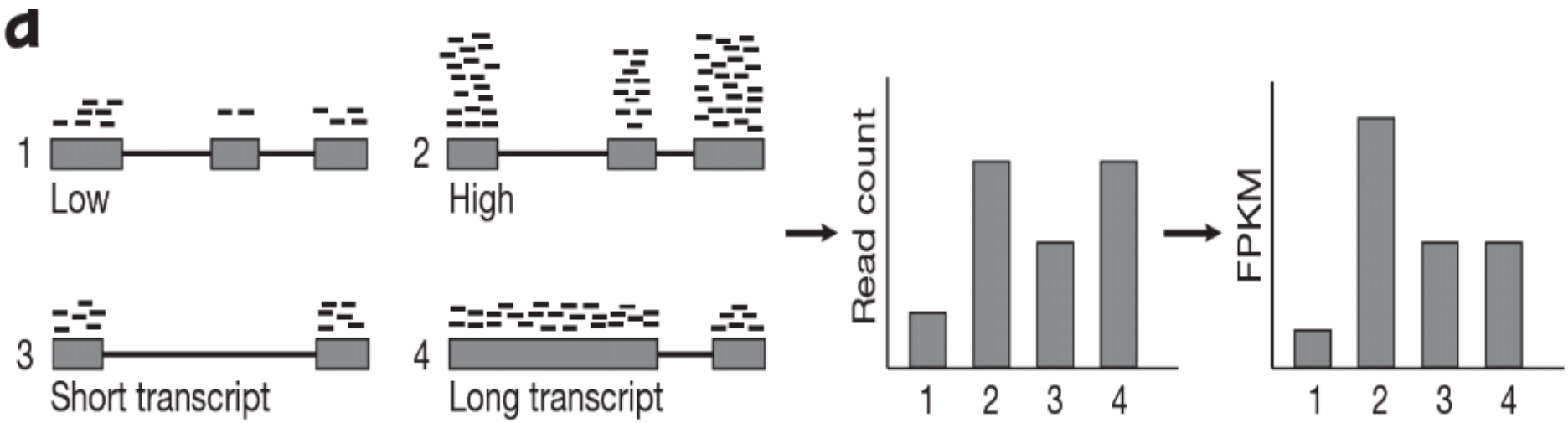


Quantification

- Fragments Per Kilobase Million (FPKM)
  - Total reads of a sample (w), divided by 1Mio
  - Divide the exon read counts (x) by the “per million” scaling factor
  - Divide this values by the length (l) of the gene (sum of exon lengths), devided by 1000 (kilobase scaling)

$$FPKM = 10^9 \times \frac{x}{wl}$$

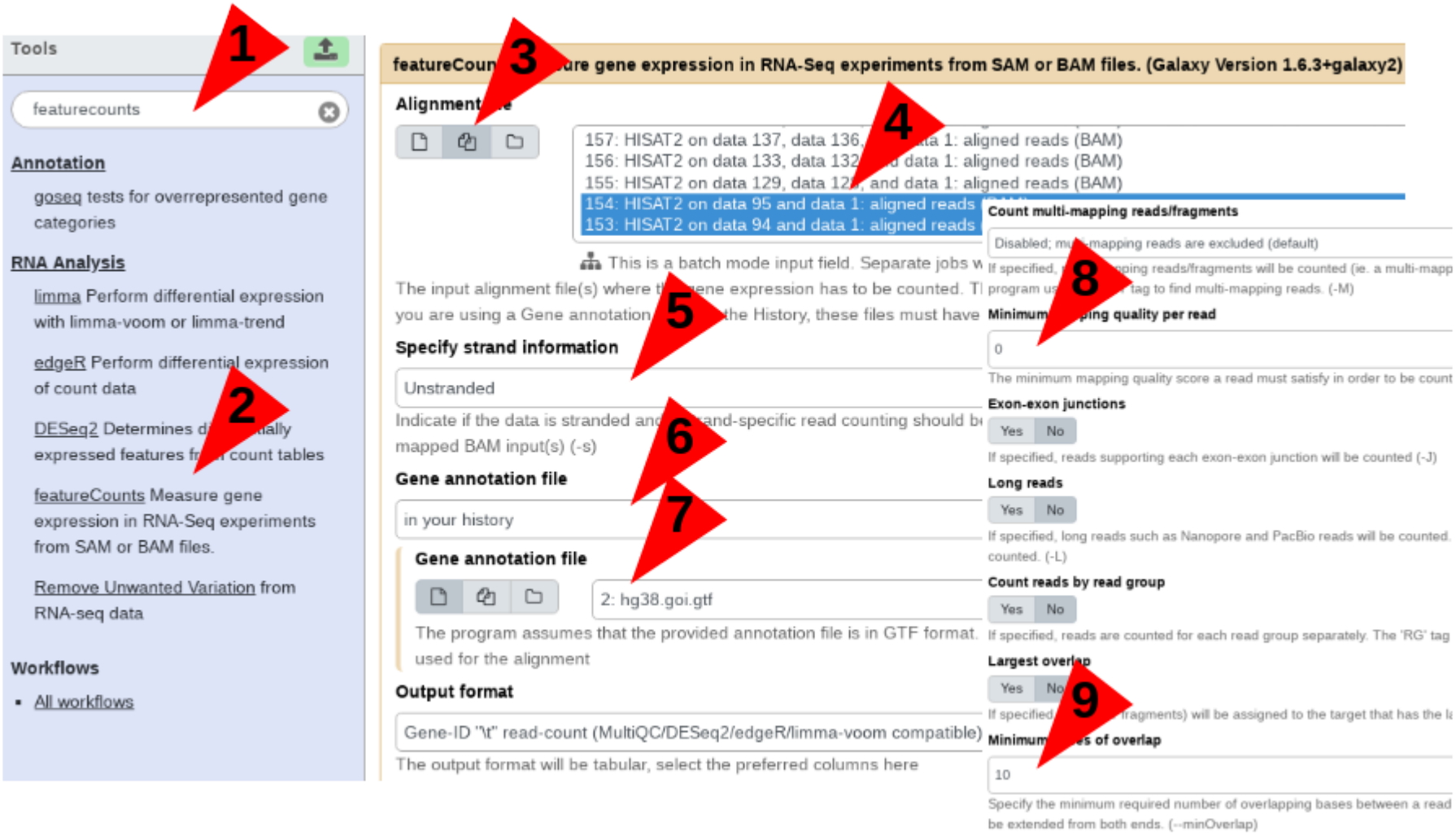
Quantification



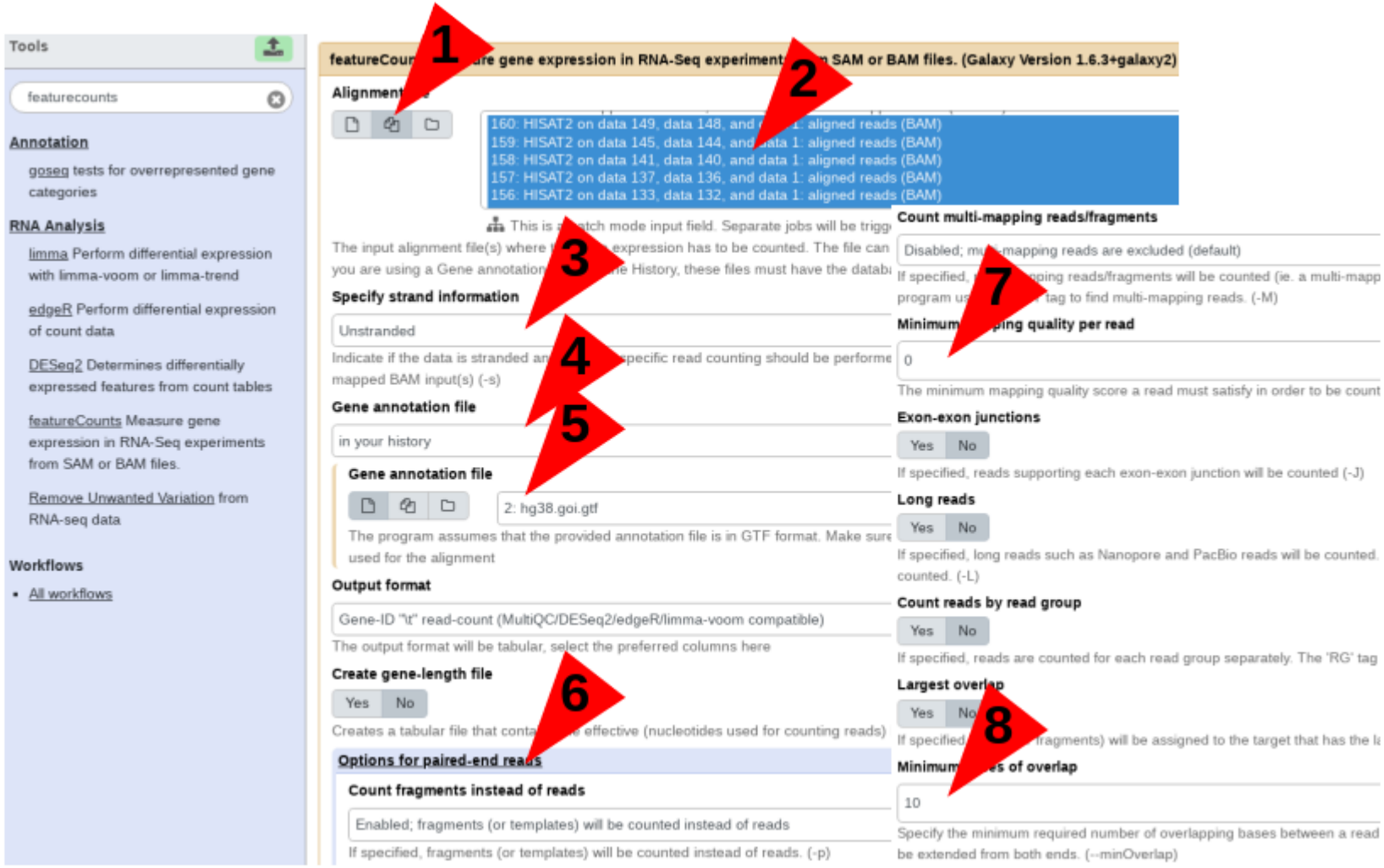
Quantification

- Tool box: R Subread - featureCounts
- Requierements:
  - BAM
  - GFF/GTF
  - Sequencing strandness information

Quantification: SE

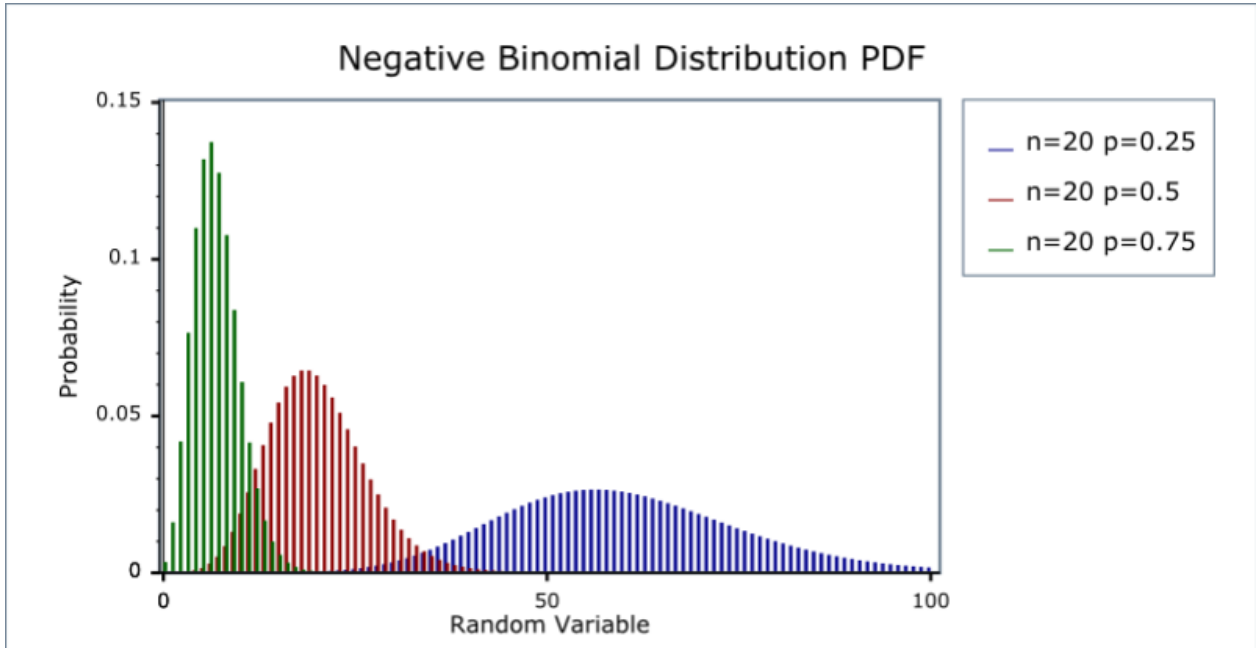


Quantification: PE



Differential Gene Expression

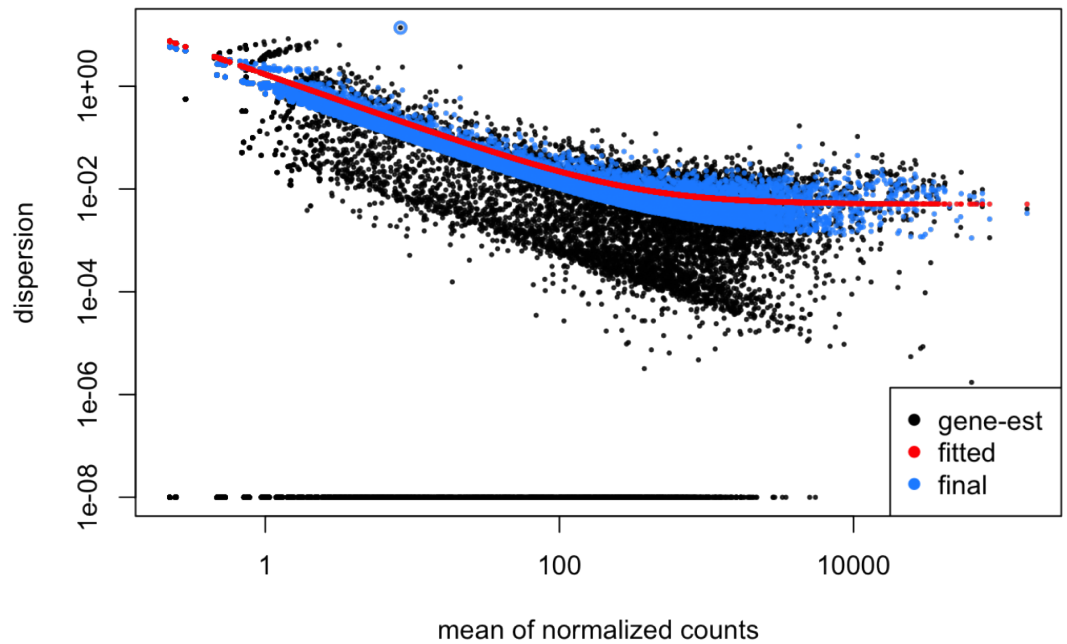
- Sophisticated read count normalization
  - per gene estimation of dispersion for Negative Binomial read count distribution (NB can model over-dispersion due to biological noise)



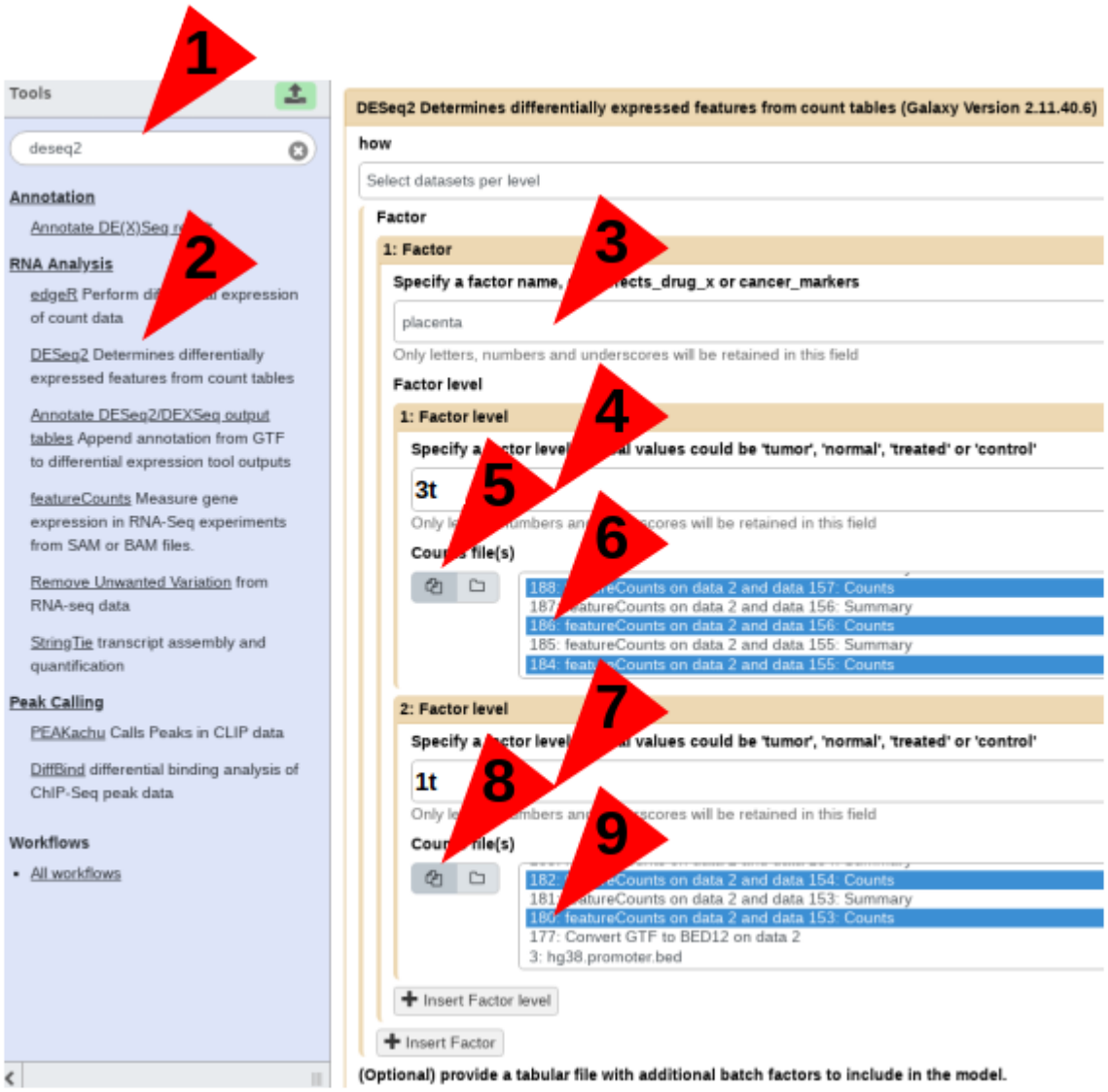
https://www.statisticshowto.datasciencecentral.com/negative-binomial-experiment/

Differential Gene Expression

- Tool box: DESeq2
  - Dispersion shrinkage to regression fit
  - Hypothesis test
  - FDR correction: Benjamini Hochberg rank sum test



Differential Gene Expression



Differential Gene Expression