NGS theory: Transcriptomics, RNAseq, scRNAseq

Richard H. Scheuermann, PhD
J. Craig Venter Institute

Transcriptome

- The population of mRNAs expressed by a genome at any given time (Abbott, 1999)
- The complete collection of transcribed elements of the genome. (Affymetrix, 2004)
- mRNAs: 35,913 transcripts in human (including alternative spliced variants)
- Non-coding RNAs
    - tRNAs (497 genes)
    - rRNAs (243 genes)
    - snmRNAs (small non-messenger RNAs)
    - microRNAs and siRNAs (small interfering RNAs)
    - snRNAs (small nuclear RNAs)
    - Pseudogenes (~ 2,000)

J Taylor, Oxford Univ.

Transcriptomics

- The study of the characteristics and regulation of the functional RNA transcript population of cells and organisms under specified conditions
  - The population of functional RNA transcripts
  - The mechanisms that regulate their production
  - The dynamics and variability of the transcriptome (time, cell type, genotype, external stimuli)

Transcriptomics and Virology

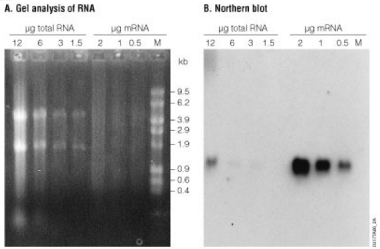
- Study the dynamics of viral gene expression during an infection cycle
- Compare virus gene expression between acute infection, latency, and re-activation
- Get an understanding of the host genes and pathways that respond to viral infection
    - Pathways required for viral replication (candidate drug targets)
    - Host response pathways (possible determinants of virulence)
- Help elucidate the function of unknown genes based on their temporal and spatial patterns (guilt by association)
- Proxy for changes in the proteome and metabolome
- Molecular biomarkers of disease

Outline

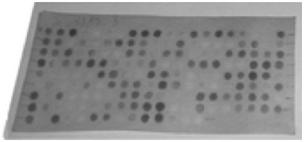
- A very brief history of transcriptomics, including gene expression microarray technologies
- RNA sequencing for transcriptomics analysis
- Single cell RNA sequencing

Early Evolution of Transcriptomics Technologies

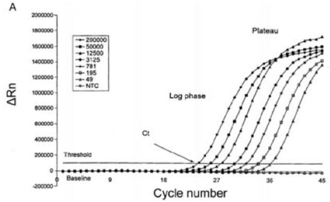
- Northern blot – labor intensive, large amount of material, use of radioactivity, one gene at a time



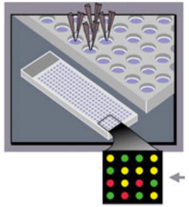
- Macroarrays – more genes, still radioactivity



- qRT-PCR – no radioactivity, but still low throughput



- Microarrays – semi-quantitative, informatics requirements



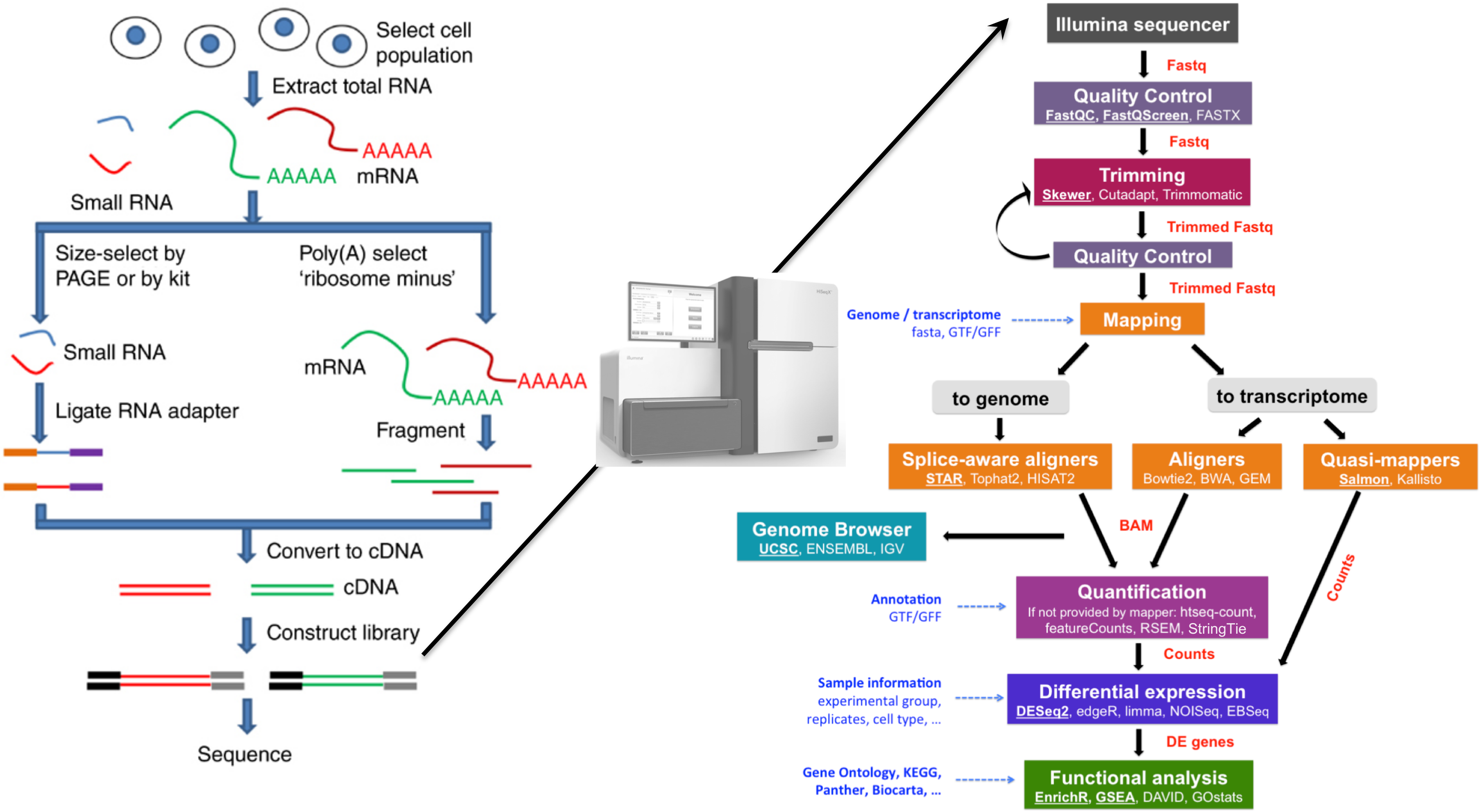
Advantages and Disadvantages of Microarrays

- Advantages
  - Much higher throughput
  - Multiplex transcriptome-level analysis
- Disadvantages
  - Relatively high experimental variability
  - Sensitive to alternative splicing ambiguities
  - Difficult to determine absolute transcript numbers
  - No information about target size is obtained, which can be helpful in recognizing cross-hybridization to non-specific or related targets
  - Only one or two samples can be analyzed at a time
  - Requires prior knowledge of transcript sequences to design probe sets
  - Doesn’t assess allele-specific expression

RNA SEQUENCING FOR TRANSCRIPTOMICS ANALYSIS

RNA-seq Advantages

- Genome-wide gene expression quantification
  - More accurate
  - Unbiased
- Essentially no noise or non-specific signal
- Mapping genes and exon boundaries
  - Single base resolution
  - Alternative splicing detection
- Novel transcripts detected
- But data is voluminous and complex
  - Need scalable, fast and mathematically principled analysis software and LOTS of computing resources



Select cell population

Extract total RNA

AAAAA

AAAAA mRNA

Small RNA

Size-select by PAGE or by kit

Poly(A) select 'ribosome minus'

Small RNA

mRNA

AAAAA

AAAAA

Ligate RNA adapter

Fragment

Convert to cDNA

cDNA

Construct library

Sequence

Illumina sequencer

Fastq

Quality Control
FastQC, FastQScreen, FASTX

Fastq

Trimming
Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome
fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners
STAR, Tophat2, HISAT2

Aligners
Bowtie2, BWA, GEM

Quasi-mappers
Salmon, Kallisto

Genome Browser
UCSC, ENSEMBL, IGV

BAM

Counts

Annotation
GTF/GFF

Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, StringTie

Counts

Sample information
experimental group, replicates, cell type, ...

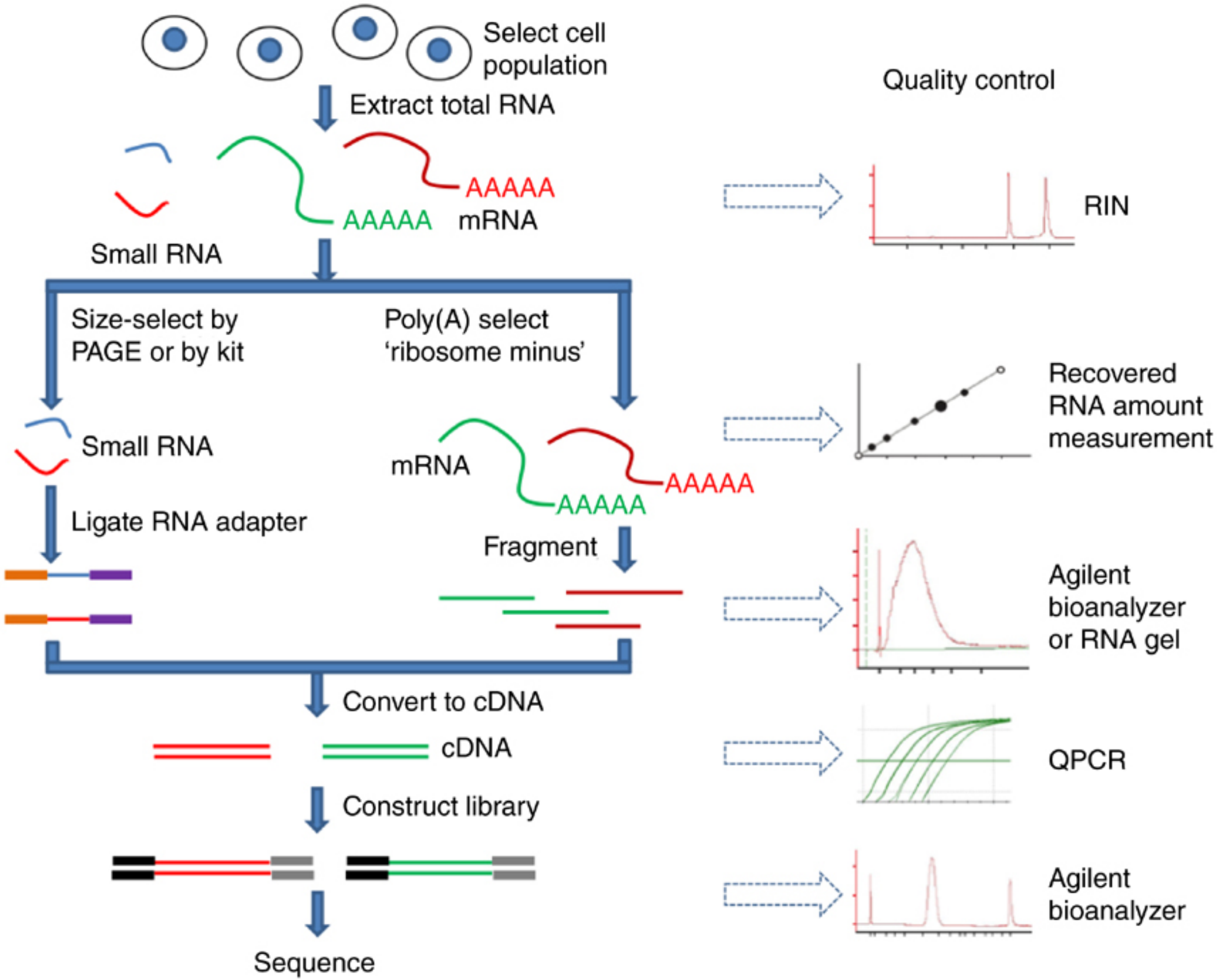
Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq

DE genes

Gene Ontology, KEGG, Panther, Biocarta, ...

Functional analysis
EnrichR, GSEA, DAVID, GOstats

Wet lab



Select cell population

Extract total RNA

Small RNA

AAAAA

AAAAA

mRNA

Size-select by PAGE or by kit

Poly(A) select 'ribosome minus'

Small RNA

mRNA

AAAAA

AAAAA

Ligate RNA adapter

Fragment

Convert to cDNA

cDNA

Construct library

Sequence

Quality control

RIN

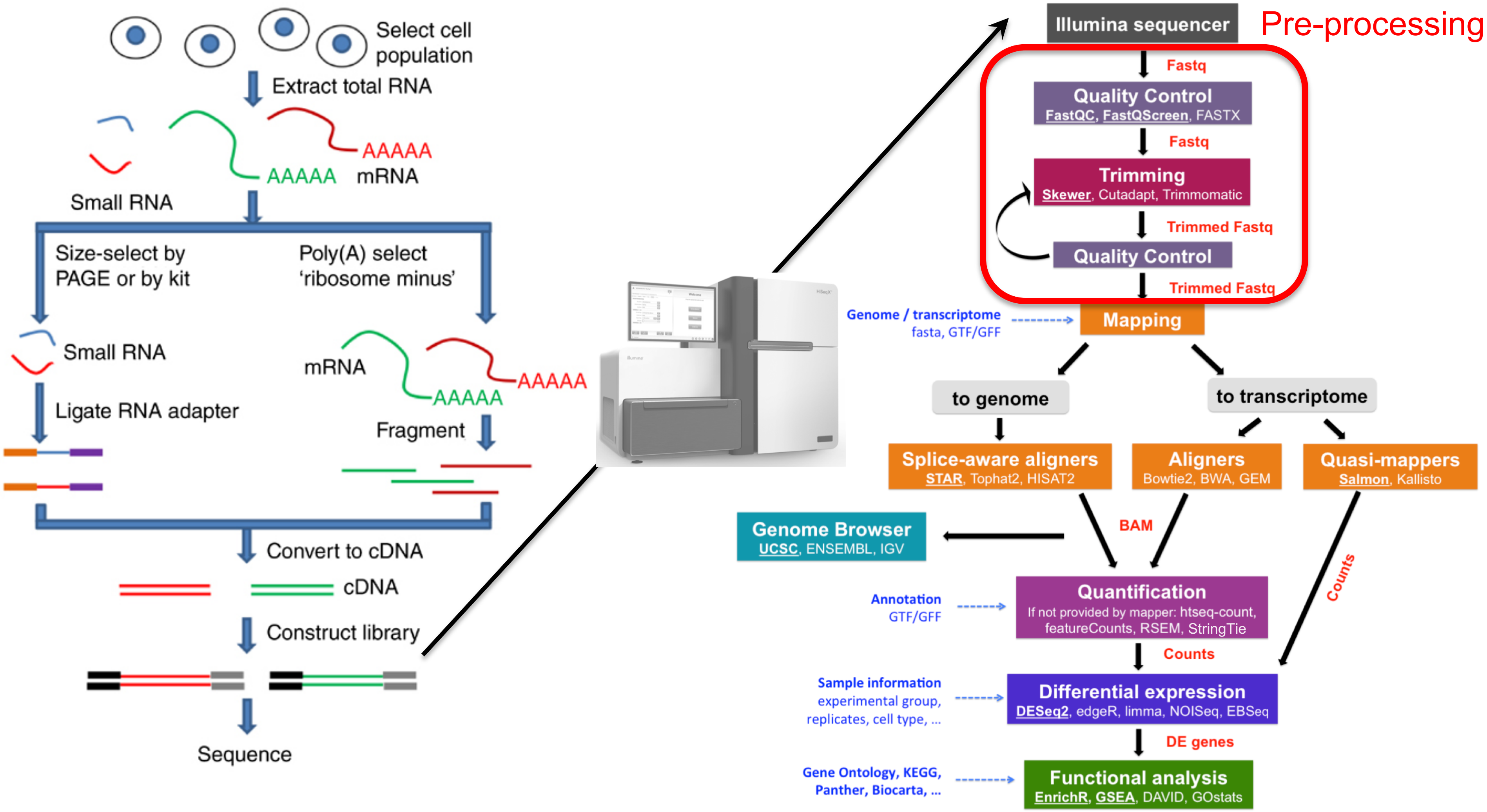
Recovered RNA amount measurement

Agilent bioanalyzer or RNA gel

QPCR

Agilent bioanalyzer

J. Craig Venter™ INSTITUTE



Select cell population

Extract total RNA

Small RNA

AAAAA

AAAAA mRNA

Size-select by PAGE or by kit

Poly(A) select ‘ribosome minus’

Small RNA

mRNA

AAAAA

AAAAA

Ligate RNA adapter

Fragment

Convert to cDNA

cDNA

Construct library

Sequence

Illumina sequencer

Pre-processing

Fastq

Quality Control
FastQC, FastQScreen, FASTX

Fastq

Trimming
Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome
fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners
STAR, Tophat2, HISAT2

Aligners
Bowtie2, BWA, GEM

Quasi-mappers
Salmon, Kallisto

Genome Browser
UCSC, ENSEMBL, IGV

BAM

Counts

Annotation
GTF/GFF

Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, StringTie

Counts

Sample information
experimental group, replicates, cell type, ...

Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq

DE genes

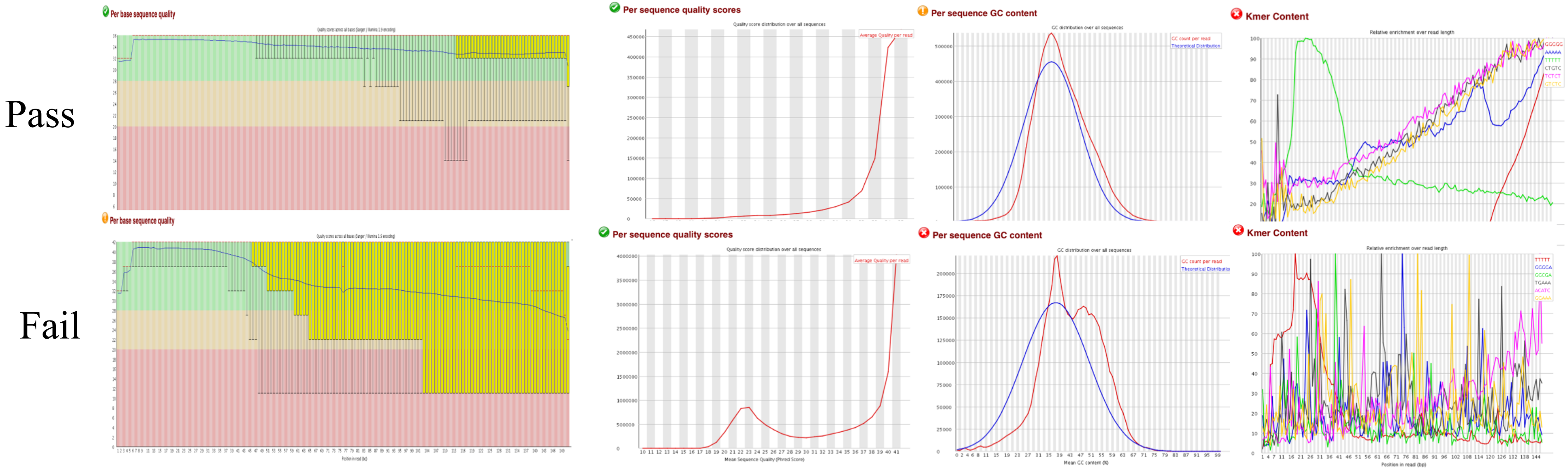
Gene Ontology, KEGG, Panther, Biocarta, ...

Functional analysis
EnrichR, GSEA, DAVID, GOstats

Quality Control

- Quality assessment
  - Evaluate read library quality to identify poor quality samples and contaminants
    - Phred scoring
    - QC content
  - Determine if primer and adapter sequences are present
  - Presence of other over-represented sequences (e.g., rRNA)
  - Software - FastQC, SAMStat, samtools, MISO

fastQC examples and criteria



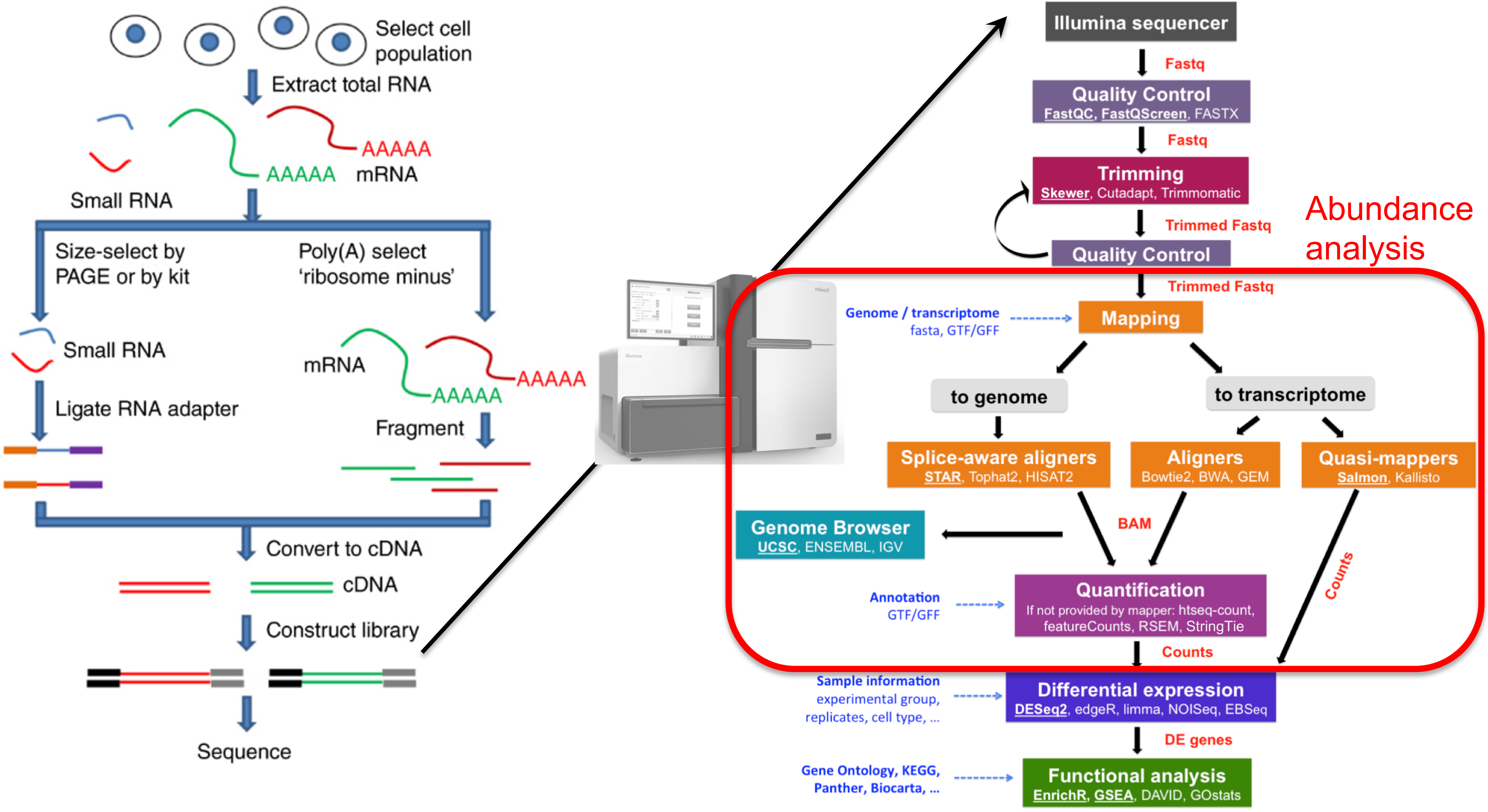
- High quality along the entire read length
- Mean sequence quality curve mostly unimodal above Phred score of 30
- GC count mostly unimodal around 40-43% GC
- Few overrepresented sequences
- Few Kmer sequences with Obs/Exp Overall >10, except polyT early

J. Craig Venter™ INSTITUTE

Adaptor & Quality Trimming

- Trimmomatic
  - Performs both primer/adapter and quality trimming
  - Paired End (PE) aware: writes unpaired reads to separate file
  - User provided adapter/primer .fasta file
- Example parameters for quality trimming
  - Trim leading and trailing bases by phred score (e.g., <3)
  - Quality trimming by user-specified base pair sliding window (e.g., 4 bp) and average phred score (e.g., <12)
  - Remove reads shorter than user-specified length (e.g., 60 bp)

Bolger Anthony, Lohse Marc, Usadel Bjoern. “Trimmomatic: a flexible trimmer for Illumina sequence data” bioinformatics Vol. 30 no. 15 2014, pages 2114–2120 doi:10.1093/bioinformatics/btu170.



Abundance analysis

Select cell population

Extract total RNA

AAAAA

AAAAA

mRNA

Small RNA

Size-select by PAGE or by kit

Poly(A) select 'ribosome minus'

Small RNA

mRNA

AAAAA

AAAAA

Ligate RNA adapter

Fragment

Convert to cDNA

cDNA

Construct library

Sequence

Illumina sequencer

Fastq

Quality Control

FastQC, FastQScreen, FASTX

Fastq

Trimming

Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome

fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners

STAR, Tophat2, HISAT2

Aligners

Bowtie2, BWA, GEM

Quasi-mappers

Salmon, Kallisto

Genome Browser

UCSC, ENSEMBL, IGV

BAM

Counts

Annotation

GTF/GFF

Quantification

If not provided by mapper: htseq-count, featureCounts, RSEM, StringTie

Counts

Sample information

experimental group, replicates, cell type, ...

Differential expression

DESeq2, edgeR, limma, NOISeq, EBSeq

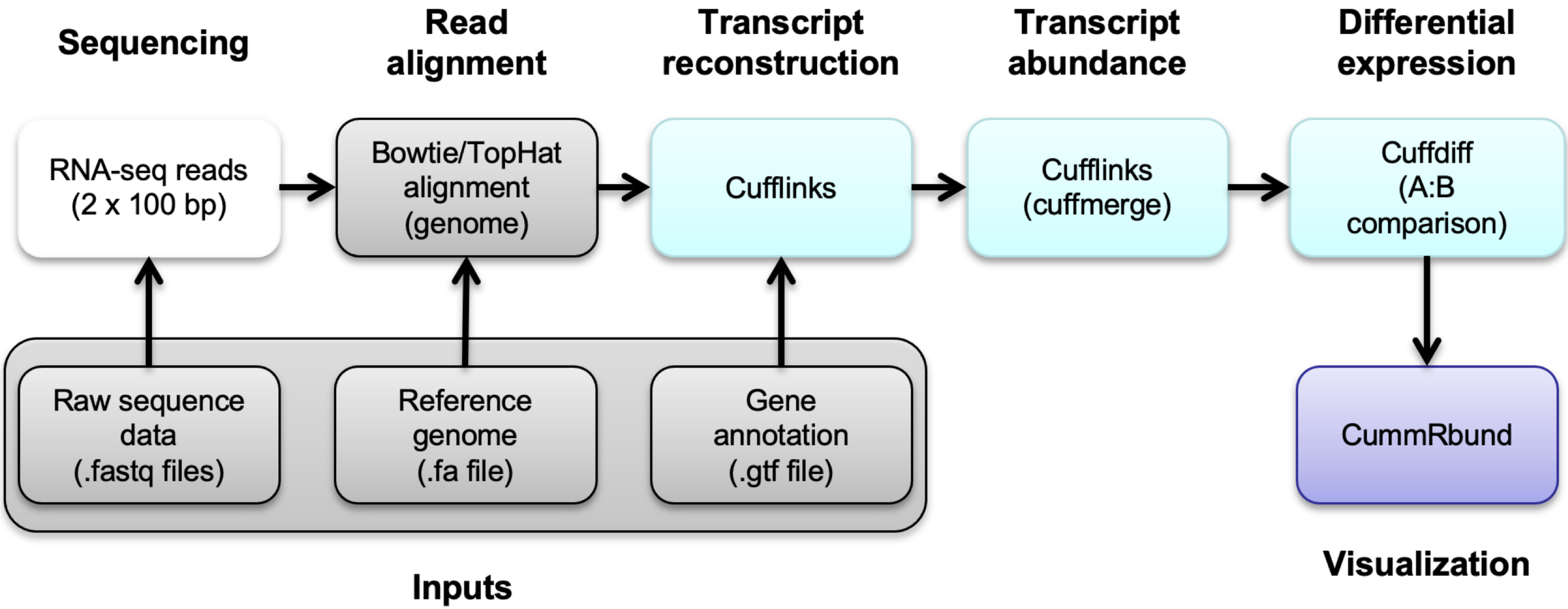
DE genes

Gene Ontology, KEGG, Panther, Biocarta, ...

Functional analysis

EnrichR, GSEA, DAVID, GOstats

Tuxedo RNA-seq Pipeline



Sequencing

Read alignment

Transcript reconstruction

Transcript abundance

Differential expression

RNA-seq reads (2 x 100 bp)

Bowtie/TopHat alignment (genome)

Cufflinks

Cufflinks (cuffmerge)

Cuffdiff (A:B comparison)

Raw sequence data (.fastq files)

Reference genome (.fa file)

Gene annotation (.gtf file)

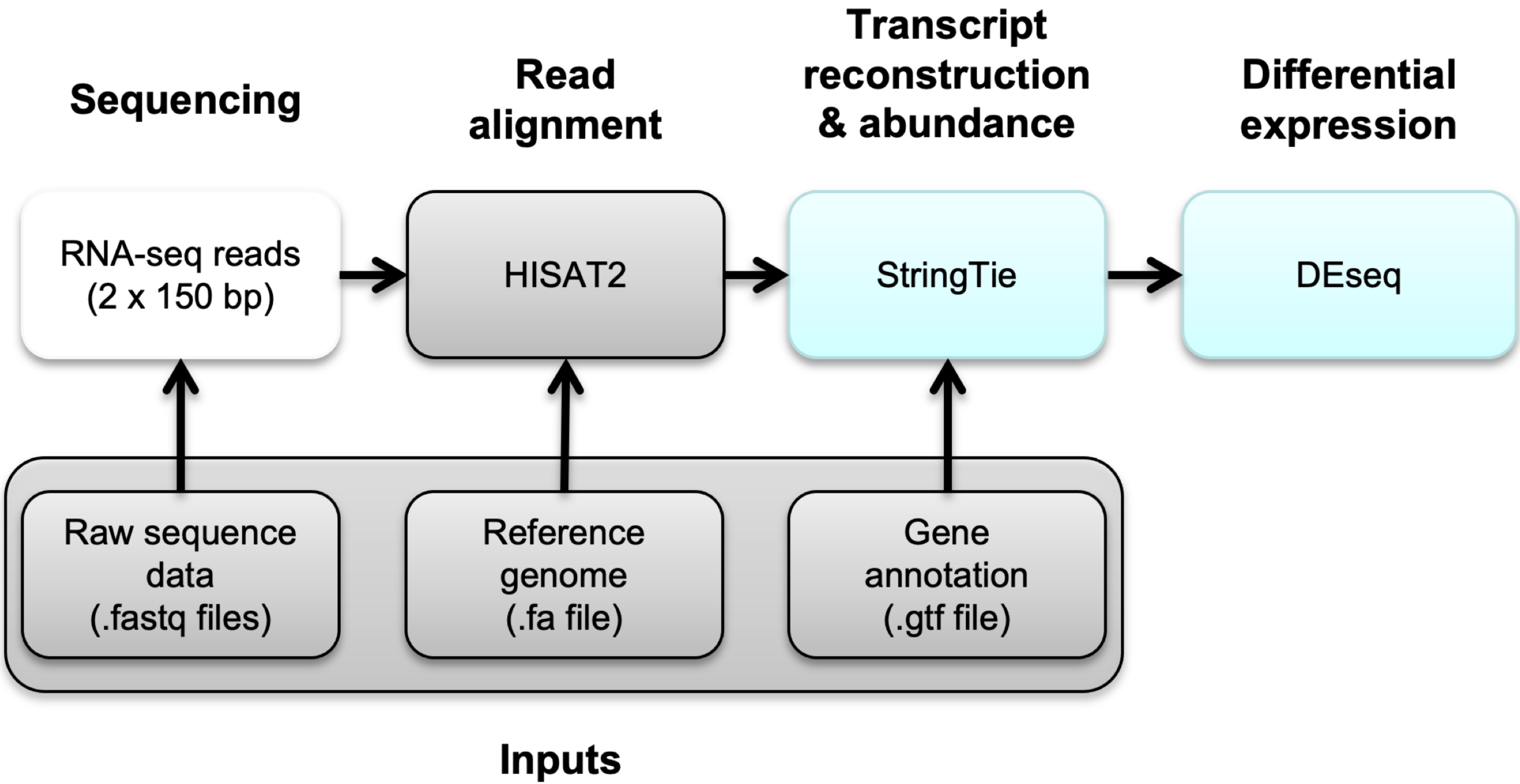
CummRbund

Inputs

Visualization

J. Craig Venter™ INSTITUTE

Current “Tuxedo” RNA-seq Pipeline



Sequencing

Read alignment

Transcript reconstruction & abundance

Differential expression

RNA-seq reads (2 x 150 bp)

HISAT2

StringTie

DEseq

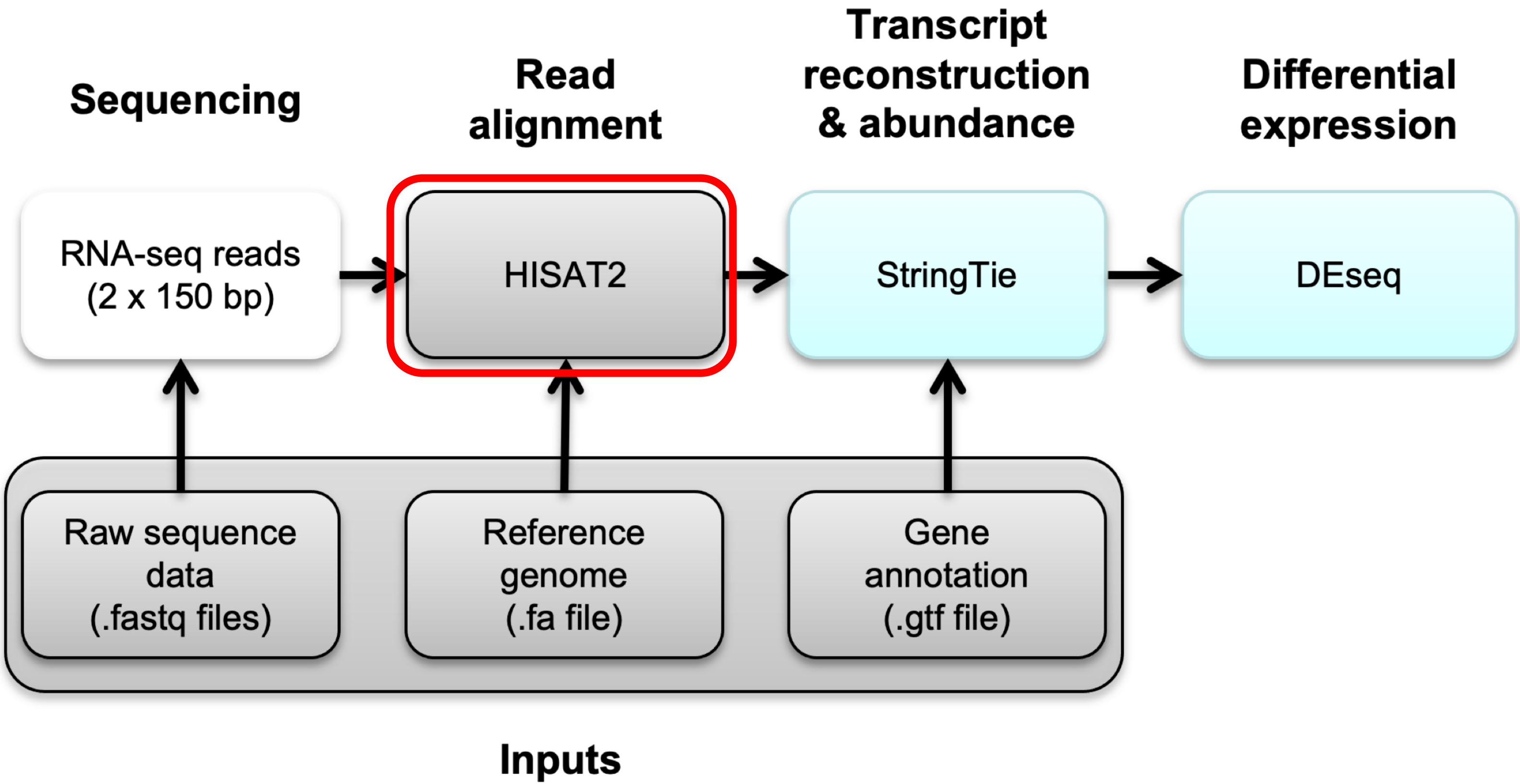
Raw sequence data (.fastq files)

Reference genome (.fa file)

Gene annotation (.gtf file)

Inputs

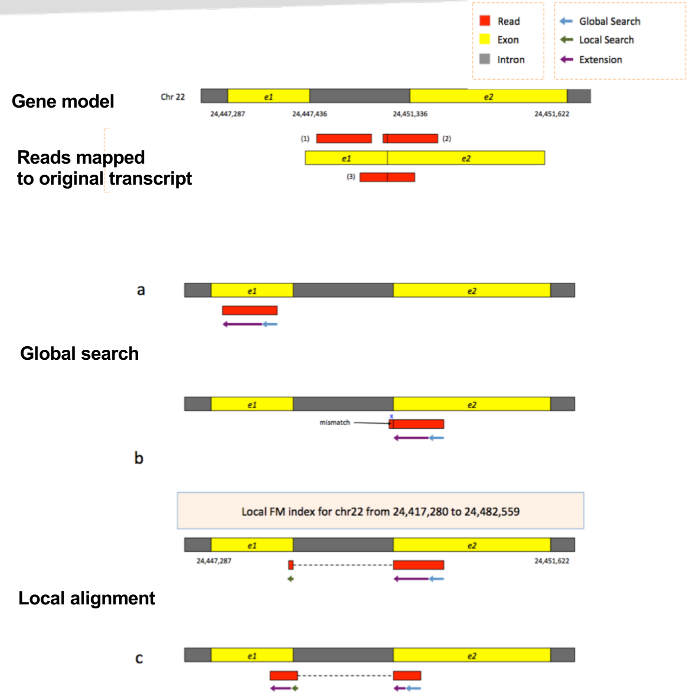
Current “Tuxedo” RNA-seq Pipeline


Sequencing
Read alignment
Transcript reconstruction & abundance
Differential expression
RNA-seq reads (2 x 150 bp)
HISAT2
StringTie
DEseq
Raw sequence data (.fastq files)
Reference genome (.fa file)
Gene annotation (.gtf file)
Inputs

Venter TUTE

HISAT2

- ***Strategy:*** HISAT2 uses a genome indexing scheme in order to make the alignment process more efficient and more accurate
  - A genome index is a type of preprocessing that compresses the size of the text and makes queries fast: "*Like the index at the end of a book, an index of a large DNA sequence allows one to rapidly find shorter sequences embedded within it.*"
  - HISAT2 is based on a Hierarchical Graph FM * index (HGFM). It generates one global GFM index but also many local indexes (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover human genome). At an index size of 56 Kbp, over 90% of introns are contained within the same index.
  - The algorithm first searches the global index for a given read to find a region of interest, then loads the local index for that region and aligns the read. This gives significant efficiency boost, but also increases accuracy as the alignment process only attempts to align the potentially spliced reads within the context of the small index as opposed to the whole genome.



Supplementary Figure 8

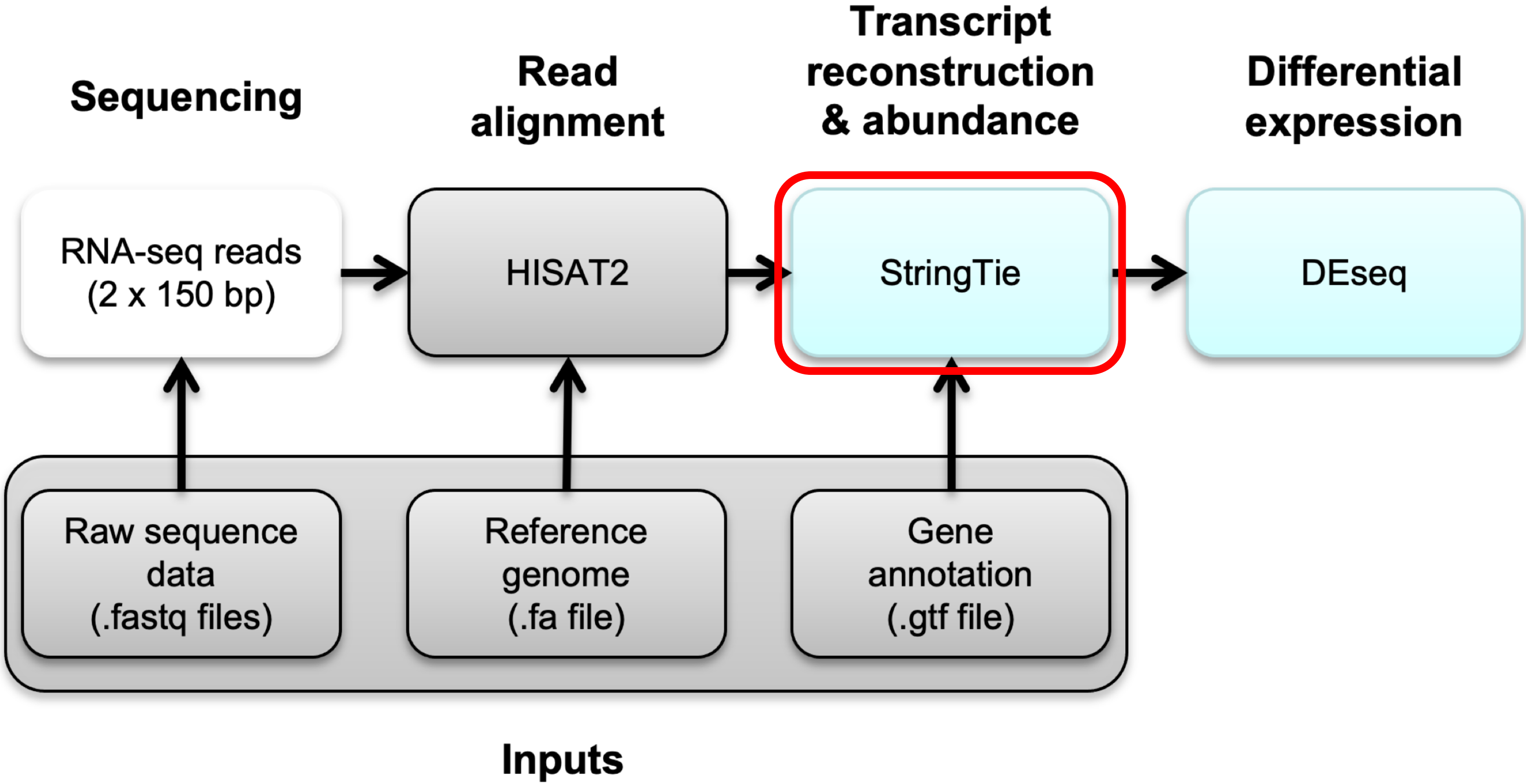
Three working examples demonstrating how HISAT applies its hierarchical indexing for fast and sensitive alignment.

The examples include alignment of one exonic read and two junction reads (one an intermediate-anchored read and the other a long-anchored read). Reads are error-free and 100-bp long.

*Burrows-Wheeler transform and the Ferragina-Manzini (FM) index - https://en.wikipedia.org/wiki/FM-index

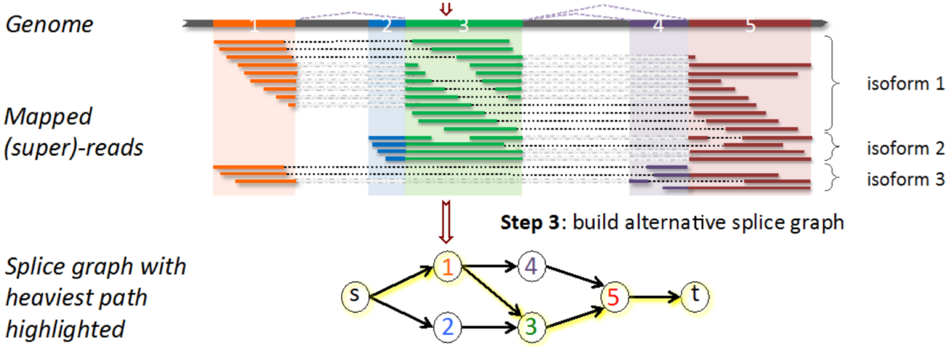
** Figure adapted from "Kim D, Langmead B and Salzberg SL. **HISAT: a fast spliced aligner with low memory requirements**. *Nature Methods* 2015" supplemental

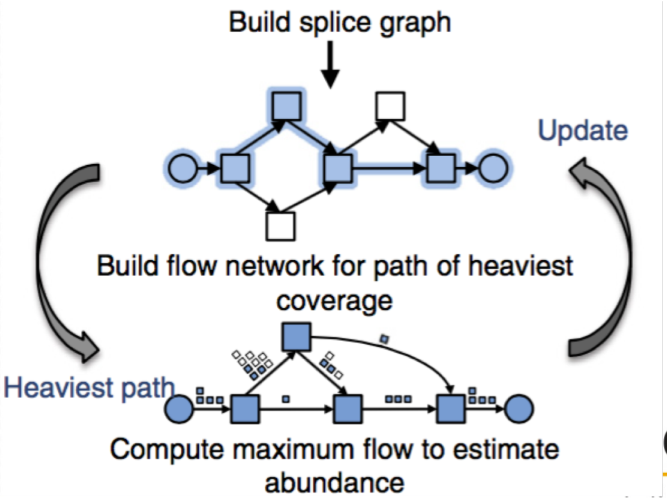
Current “Tuxedo” RNA-seq Pipeline


Sequencing
Read alignment
Transcript reconstruction & abundance
Differential expression
RNA-seq reads (2 x 150 bp)
HISAT2
StringTie
DEseq
Raw sequence data (.fastq files)
Reference genome (.fa file)
Gene annotation (.gtf file)
Inputs

StringTie

- ***Strategy:*** StringTie uses a graph-based approach called network flow
  - 1. pulls in a cluster of reads for a region given by the alignment.
  - 2. Builds a splice graph of all isoforms for a given gene based on the annotation provided.
  - 3. Estimates heaviest flow using reads aligned to exons (nodes) and for that transcript a flow network is built.
  - 4. From the flow network, the abundance of that transcript is then estimated by maximal flow. These assembled reads are then removed.
  - 5. Process iterates until all reads are assigned to a transcript.

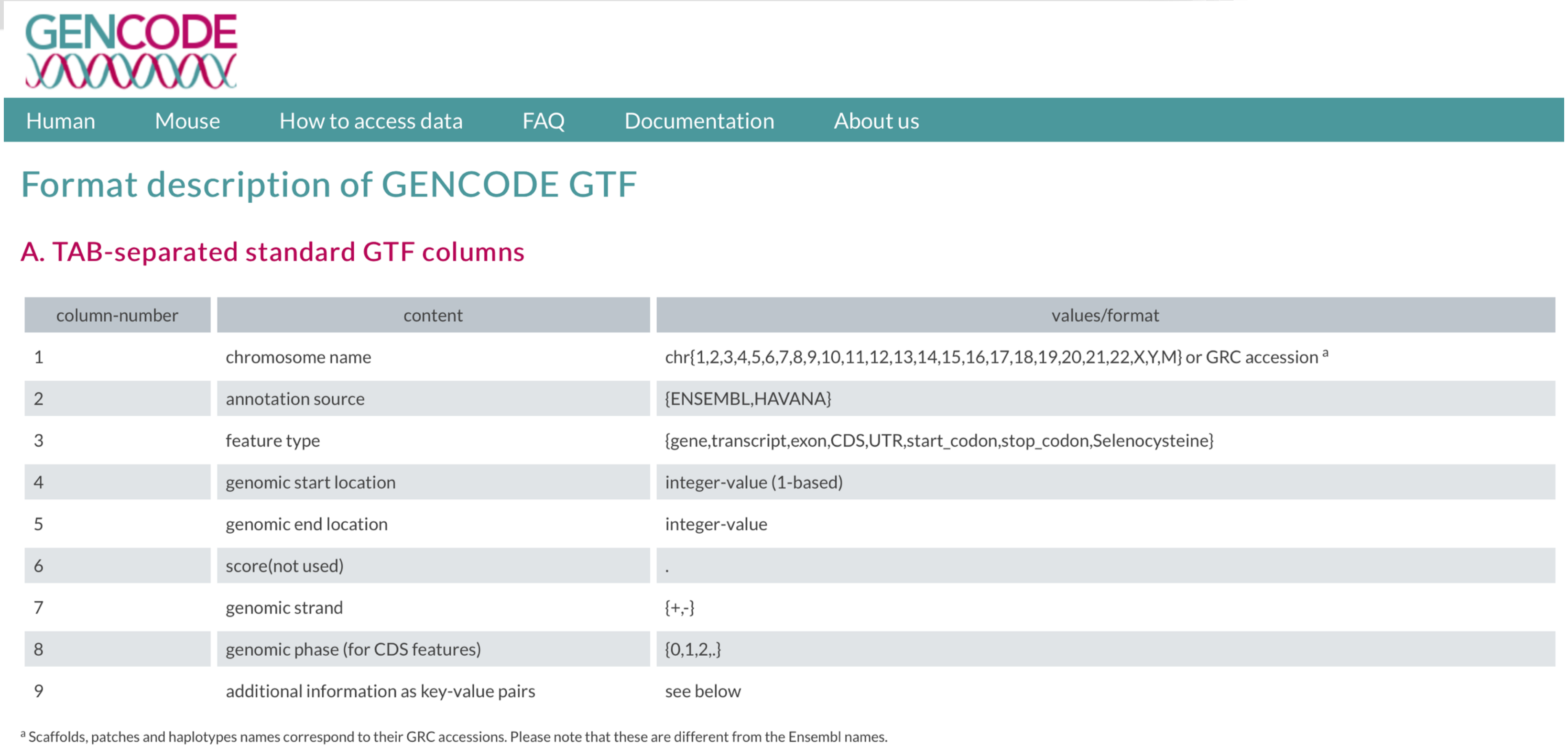

Genome
1
2
3
4
5
Mapped (super)-reads
isoform 1
isoform 2
isoform 3
Step 3: build alternative splice graph
Splice graph with heaviest path highlighted
s
1
4
2
3
5
t


Build splice graph
Update
Build flow network for path of heaviest coverage
Heaviest path
Compute maximum flow to estimate abundance

Pertea, M., Pertea, G., Antonescu, C. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).

Craig Venter™
INSTITUTE

StringTie optional input - GTF



Format description of GENCODE GTF

A. TAB-separated standard GTF columns

column-number	content	values/format
1	chromosome name	chr{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y,M} or GRC accession [a]
2	annotation source	{ENSEMBL,HAVANA}
3	feature type	{gene,transcript,exon,CDS,UTR,start_codon,stop_codon,Selenocysteine}
4	genomic start location	integer-value (1-based)
5	genomic end location	integer-value
6	score(not used)	.
7	genomic strand	{+,-}
8	genomic phase (for CDS features)	{0,1,2,.}
9	additional information as key-value pairs	see below

[a] Scaffolds, patches and haplotypes names correspond to their GRC accessions. Please note that these are different from the Ensembl names.

https://www.gencodegenes.org/pages/data_format.html

J. Craig Venter™ INSTITUTE

Additional information

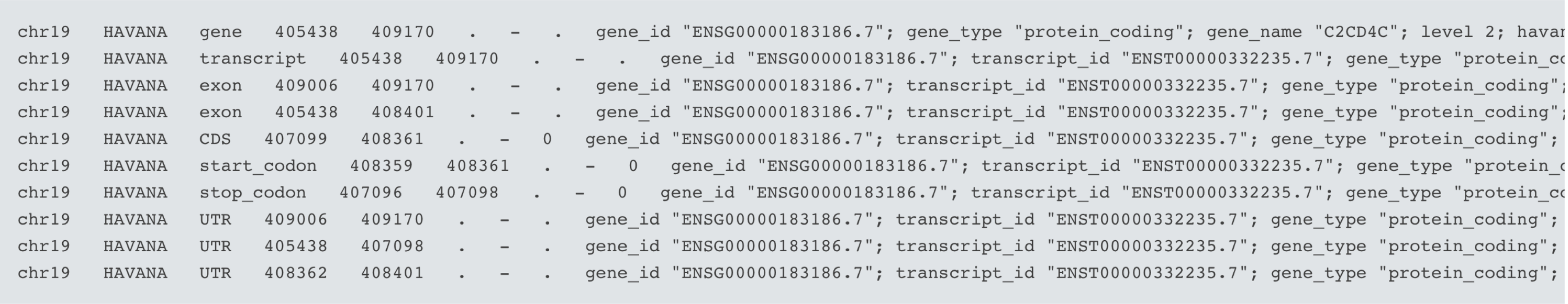
B. Key-value pairs in 9th column (format: key "value";)

B.1. Mandatory fields

key name	feature type(s)	value format	release
gene_id	all	ENSGXXXXXXXXXXX.X [b,c] _X[g]	all
transcript_id [d]	all except gene	ENSTXXXXXXXXXXX.X [b,c] _X[g]	all
gene_type	all	**list of biotypes**	all
gene_status [e]	all	{KNOWN, NOVEL, PUTATIVE}	until 25 and M11
gene_name	all	string	all
transcript_type [d]	all except gene	**list of biotypes**	all
transcript_status[d,e]	all except gene	{KNOWN, NOVEL, PUTATIVE}	until 25 and M11
transcript_name [d]	all except gene	string	all
exon_number [f]	all except gene/transcript/Selenocysteine	integer (exon position in the transcript from its 5' end)	all
exon_id [f]	all except gene/transcript/Selenocysteine	ENSEXXXXXXXXXXX.X [b] _X[g]	all
level	all	1 (verified loci), 2 (manually annotated loci), 3 (automatically annotated loci)	all

Example GTF

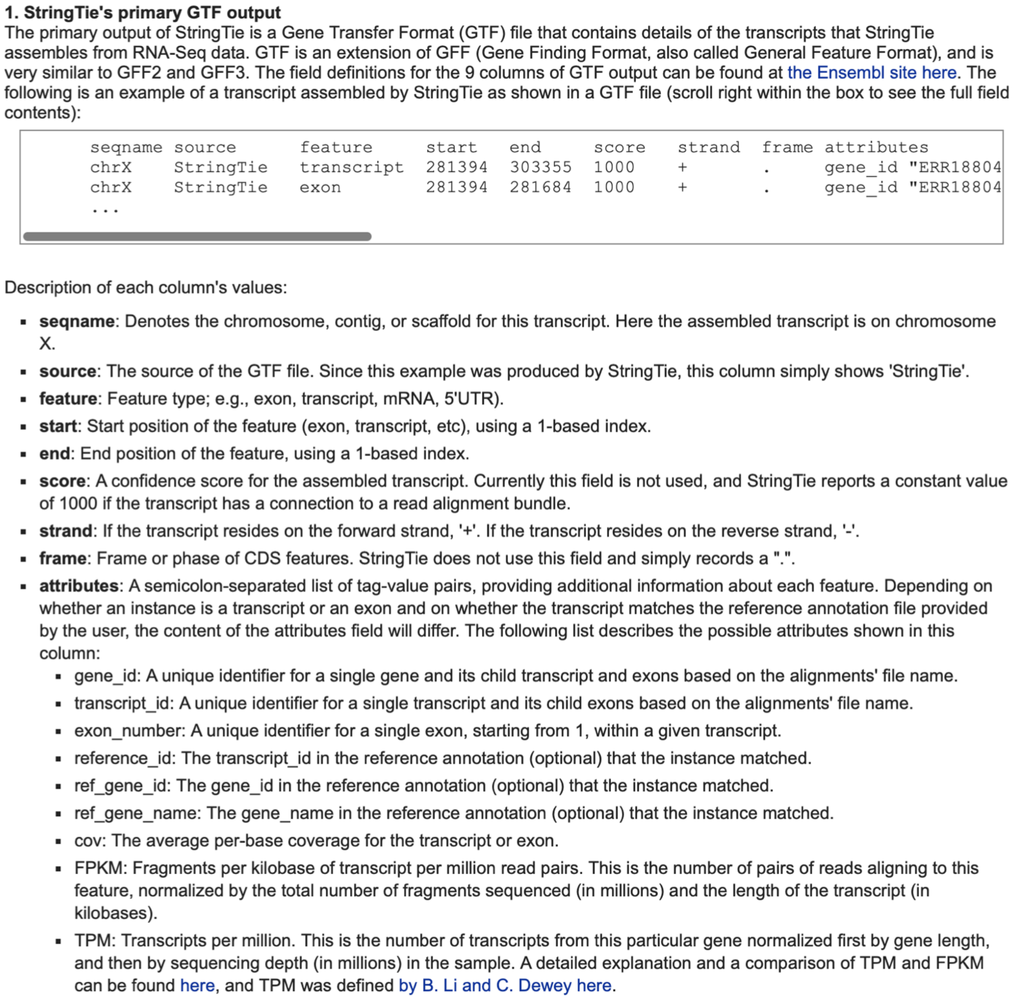
Example GTF lines:



```
chr19   HAVANA   gene   405438   409170   .   -   .   gene_id "ENSG00000183186.7"; gene_type "protein_coding"; gene_name "C2CD4C"; level 2; havar
chr19   HAVANA   transcript   405438   409170   .   -   .   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_co
chr19   HAVANA   exon   409006   409170   .   -   .   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_coding";
chr19   HAVANA   exon   405438   408401   .   -   .   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_coding";
chr19   HAVANA   CDS   407099   408361   .   -   0   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_coding";
chr19   HAVANA   start_codon   408359   408361   .   -   0   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_c
chr19   HAVANA   stop_codon   407096   407098   .   -   0   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_co
chr19   HAVANA   UTR   409006   409170   .   -   .   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_coding";
chr19   HAVANA   UTR   405438   407098   .   -   .   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_coding";
chr19   HAVANA   UTR   408362   408401   .   -   .   gene_id "ENSG00000183186.7"; transcript_id "ENST00000332235.7"; gene_type "protein_coding";
```

J. Craig Venter™
INSTITUTE

StringTie output - GTF



1. StringTie's primary GTF output

The primary output of StringTie is a Gene Transfer Format (GTF) file that contains details of the transcripts that StringTie assembles from RNA-Seq data. GTF is an extension of GFF (Gene Finding Format, also called General Feature Format), and is very similar to GFF2 and GFF3. The field definitions for the 9 columns of GTF output can be found at the Ensembl site here. The following is an example of a transcript assembled by StringTie as shown in a GTF file (scroll right within the box to see the full field contents):

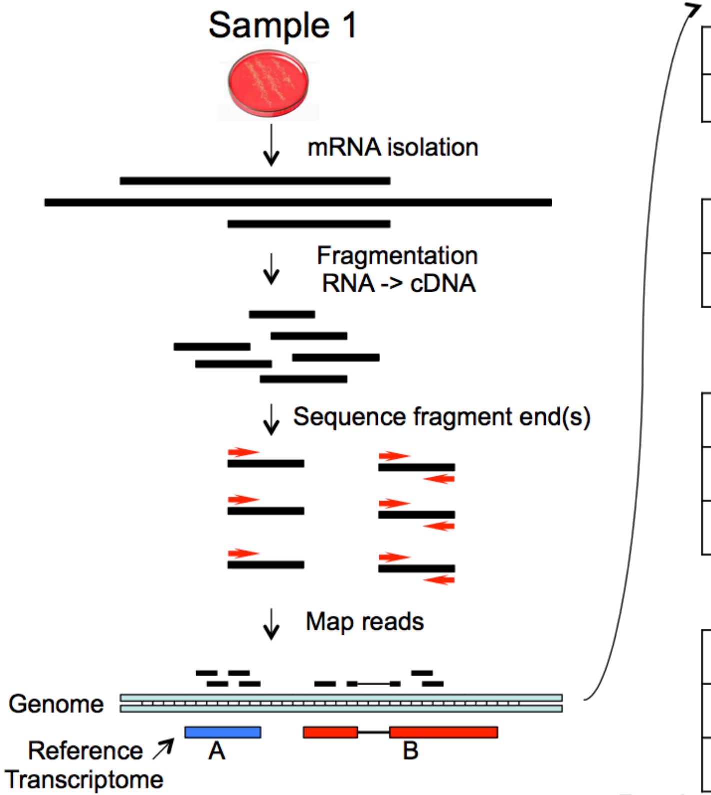
```
    seqname source      feature     start   end     score   strand  frame attributes
    chrX    StringTie   transcript  281394  303355  1000    +       .     gene_id "ERR18804
    chrX    StringTie   exon        281394  281684  1000    +       .     gene_id "ERR18804
    ...
```

Description of each column's values:

- **seqname**: Denotes the chromosome, contig, or scaffold for this transcript. Here the assembled transcript is on chromosome X.
- **source**: The source of the GTF file. Since this example was produced by StringTie, this column simply shows 'StringTie'.
- **feature**: Feature type; e.g., exon, transcript, mRNA, 5'UTR).
- **start**: Start position of the feature (exon, transcript, etc), using a 1-based index.
- **end**: End position of the feature, using a 1-based index.
- **score**: A confidence score for the assembled transcript. Currently this field is not used, and StringTie reports a constant value of 1000 if the transcript has a connection to a read alignment bundle.
- **strand**: If the transcript resides on the forward strand, '+'. If the transcript resides on the reverse strand, '-'.
- **frame**: Frame or phase of CDS features. StringTie does not use this field and simply records a ".".
- **attributes**: A semicolon-separated list of tag-value pairs, providing additional information about each feature. Depending on whether an instance is a transcript or an exon and on whether the transcript matches the reference annotation file provided by the user, the content of the attributes field will differ. The following list describes the possible attributes shown in this column:
  - gene_id: A unique identifier for a single gene and its child transcript and exons based on the alignments' file name.
  - transcript_id: A unique identifier for a single transcript and its child exons based on the alignments' file name.
  - exon_number: A unique identifier for a single exon, starting from 1, within a given transcript.
  - reference_id: The transcript_id in the reference annotation (optional) that the instance matched.
  - ref_gene_id: The gene_id in the reference annotation (optional) that the instance matched.
  - ref_gene_name: The gene_name in the reference annotation (optional) that the instance matched.
  - cov: The average per-base coverage for the transcript or exon.
  - FPKM: Fragments per kilobase of transcript per million read pairs. This is the number of pairs of reads aligning to this feature, normalized by the total number of fragments sequenced (in millions) and the length of the transcript (in kilobases).
  - TPM: Transcripts per million. This is the number of transcripts from this particular gene normalized first by gene length, and then by sequencing depth (in millions) in the sample. A detailed explanation and a comparison of TPM and FPKM can be found here, and TPM was defined by B. Li and C. Dewey here.

http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual

RPKM Normalization



Calculate transcript abundance

	Gene A	Gene B
Sample 1	4	4

# of Reads

	Gene A	Gene B
Sample 1	4	2

Reads per kilobase of exon

	Gene A	Gene B	Total
Sample 1	4	2	6
Sample 2	7	5	12

Reads per kilobase of exon

	Gene A	Gene B	Total
Sample 1	.7	.3	6
Sample 2	.6	.4	12

Reads per kilobase of exon per million mapped reads

RPKM

J. Craig Venter™ INSTITUTE

Alternative Normalized Counts

RPKM

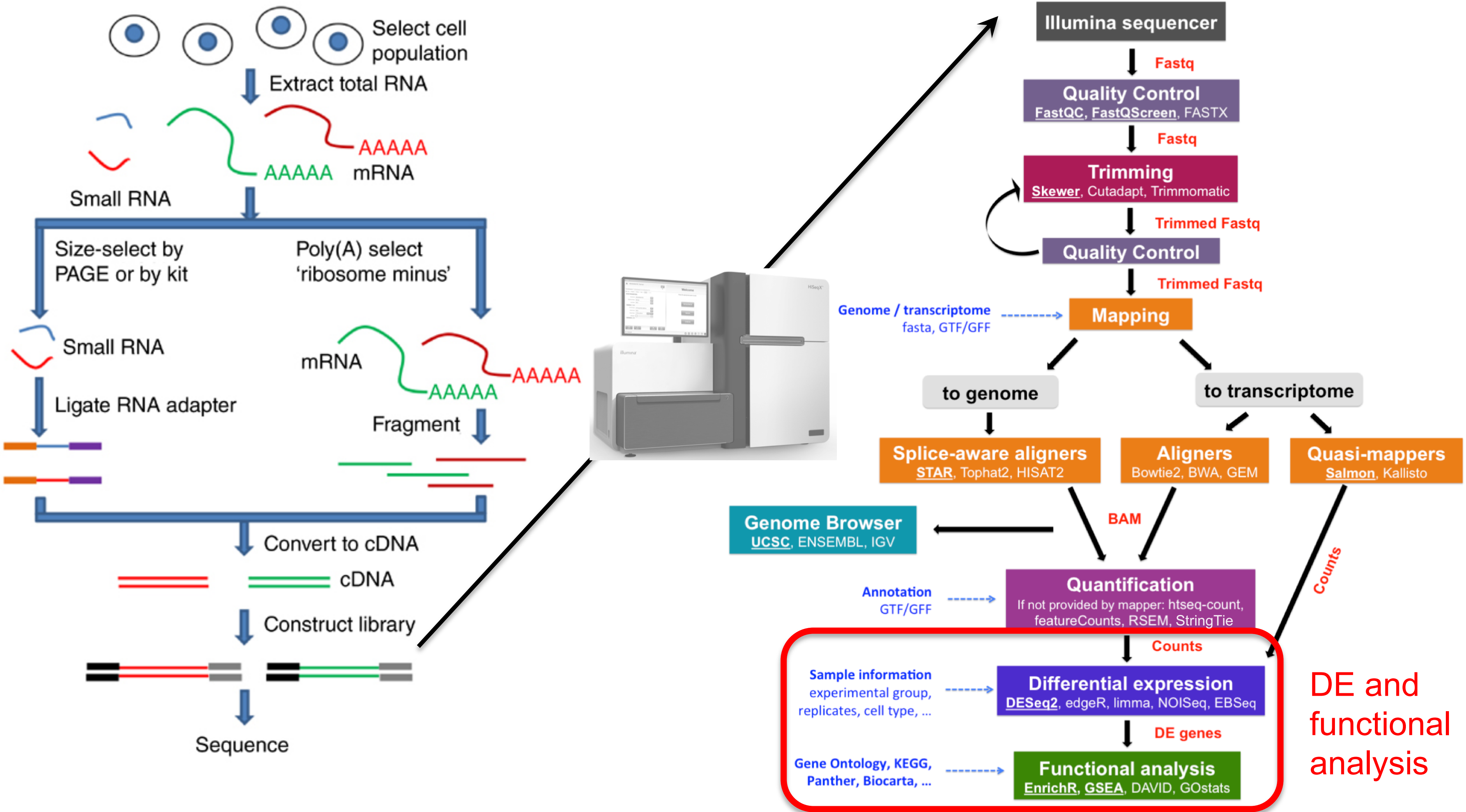
- Reads per kilobase per million normalizes the raw count by transcript length and sequencing depth.
- RPKM = (CDS read count * 10^9) / (CDS length * total mapped read count)

FPKM

- Same as RPKM except if the data is paired then only one of the mates is counted, i.e., fragments are counted rather than reads

TPM

- Transcripts per million (as proposed by Wagner et al 2012) is a modification of RPKM designed to be consistent across samples. It is normalized by total transcript count instead of read count in addition to average read length.
- TPM = (CDS read count * mean read length * 10^6) / (CDS length * total transcript count)



Select cell population

Extract total RNA

AAAAA

AAAAA mRNA

Small RNA

Size-select by PAGE or by kit

Poly(A) select ‘ribosome minus’

Small RNA

mRNA

AAAAA

AAAAA

Ligate RNA adapter

Fragment

Convert to cDNA

cDNA

Construct library

Sequence

Illumina sequencer

Fastq

Quality Control
FastQC, FastQScreen, FASTX

Fastq

Trimming
Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome
fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners
STAR, Tophat2, HISAT2

Aligners
Bowtie2, BWA, GEM

Quasi-mappers
Salmon, Kallisto

Genome Browser
UCSC, ENSEMBL, IGV

BAM

Counts

Annotation
GTF/GFF

Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, StringTie

Counts

Sample information
experimental group, replicates, cell type, ...

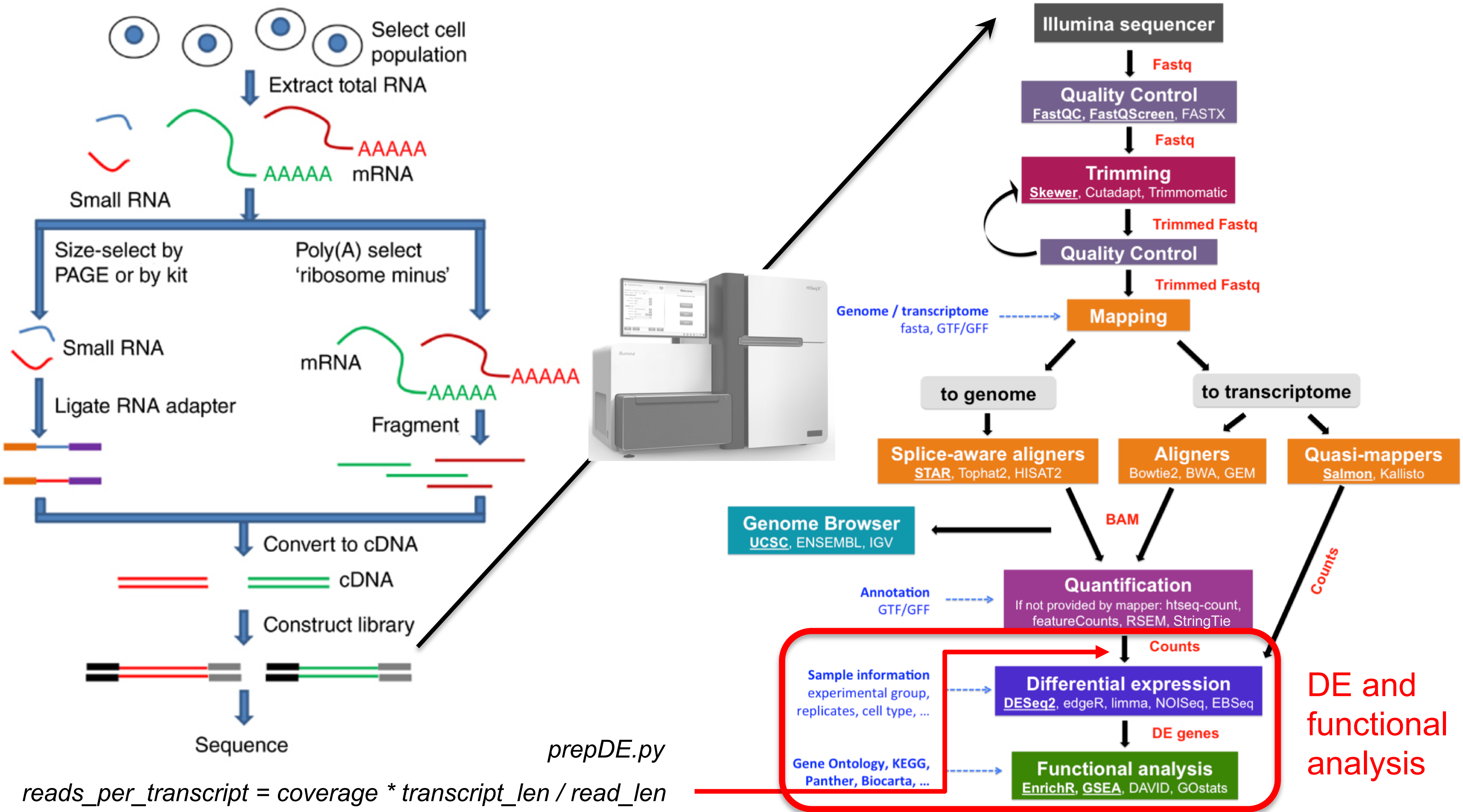
Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq

DE genes

Gene Ontology, KEGG, Panther, Biocarta, ...

Functional analysis
EnrichR, GSEA, DAVID, GOstats

DE and functional analysis



Select cell population

Extract total RNA

AAAAA

AAAAA mRNA

Small RNA

Size-select by PAGE or by kit

Poly(A) select 'ribosome minus'

Small RNA

mRNA

AAAAA

AAAAA

Ligate RNA adapter

Fragment

Convert to cDNA

cDNA

Construct library

Sequence

Illumina sequencer

Fastq

Quality Control
FastQC, FastQScreen, FASTX

Fastq

Trimming
Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome
fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners
STAR, Tophat2, HISAT2

Aligners
Bowtie2, BWA, GEM

Quasi-mappers
Salmon, Kallisto

Genome Browser
UCSC, ENSEMBL, IGV

BAM

Counts

Annotation
GTF/GFF

Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, StringTie

Counts

Sample information
experimental group, replicates, cell type, ...

Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq

DE genes

Gene Ontology, KEGG, Panther, Biocarta, ...

Functional analysis
EnrichR, GSEA, DAVID, GOstats

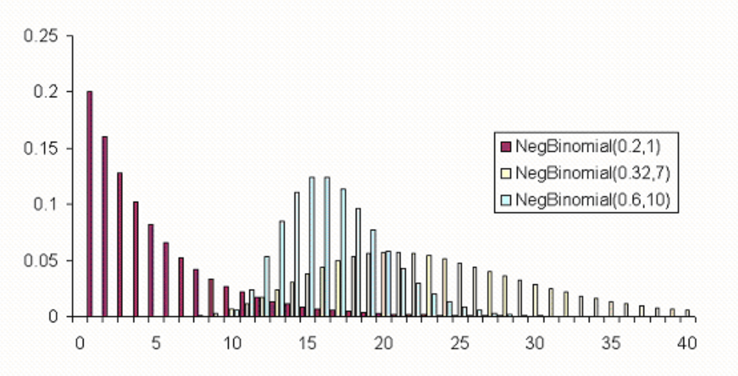
DE and functional analysis

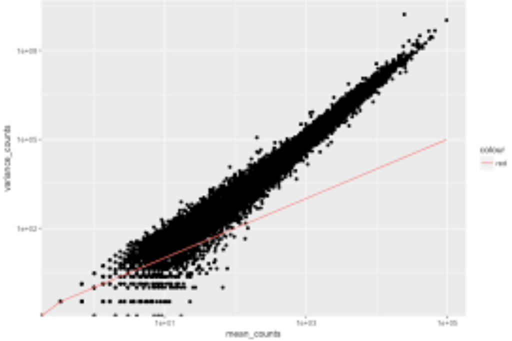
prepDE.py

$\mathit{reads\_per\_transcript} = \mathit{coverage} * \mathit{transcript\_len} / \mathit{read\_len}$

DEseq2

- Characteristics of RNAseq data
  - Non-normal distribution of expression values
  - Discrete rather than continuous
  - Dependence of variance on the mean (overdispersion)
  - Small sample sizes
- DEseq2
  - Model expression as a negative binomial distribution
  - Corrects dispersion estimates that are too low through modeling of the dependence of the dispersion on the average expression strength over all samples



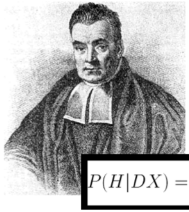


Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

Functional enrichment analyses

- GO Enrichment
- GO-BAYES
- GSEA
- Enrichr

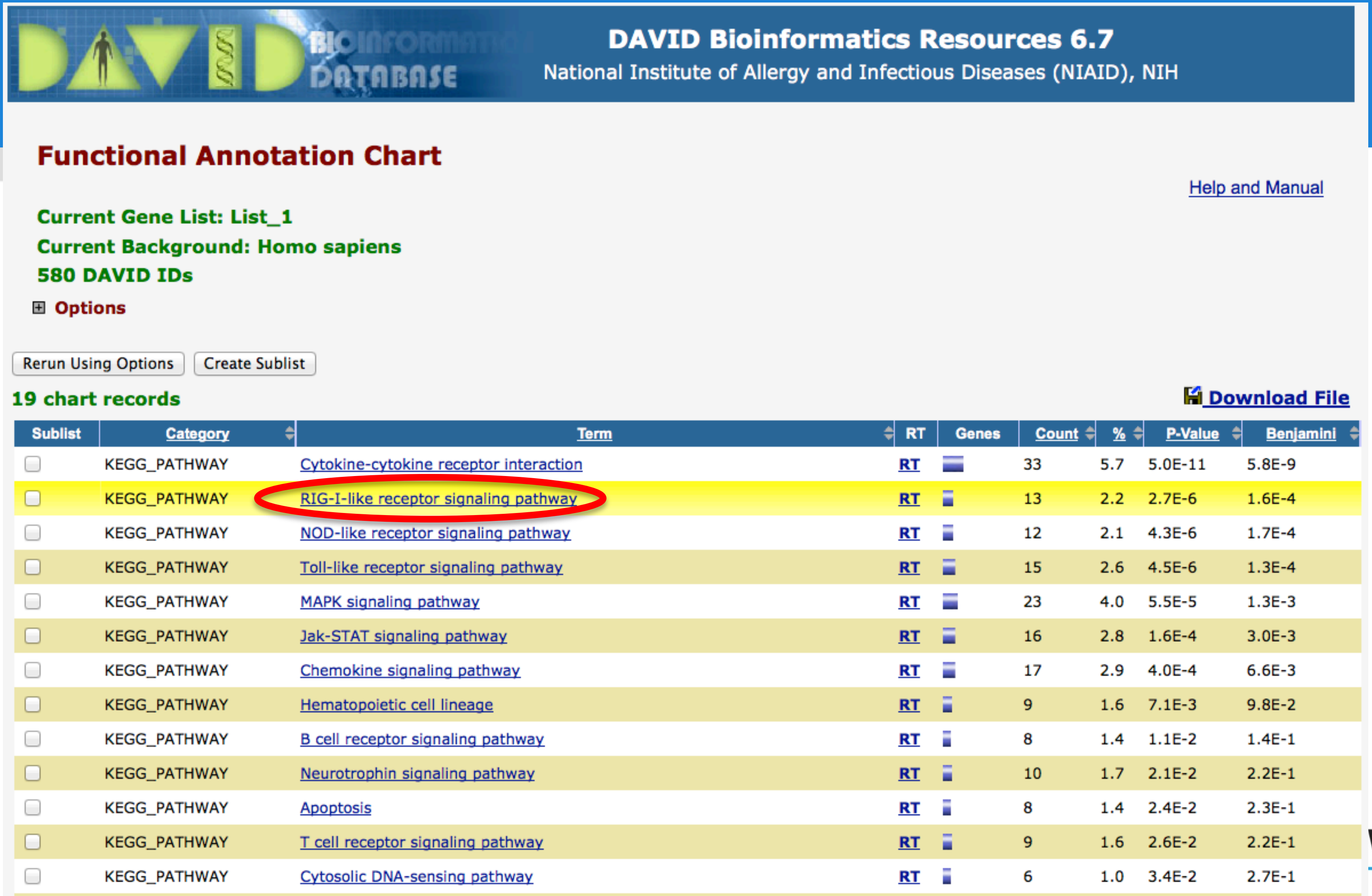
the Gene Ontology



$$P(H|DX) = \frac{P(H|X) \times P(D|HX)}{P(D|X)}$$

GSEA
Gene Set Enrichment Analysis

Enrichr



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Chart

Help and Manual

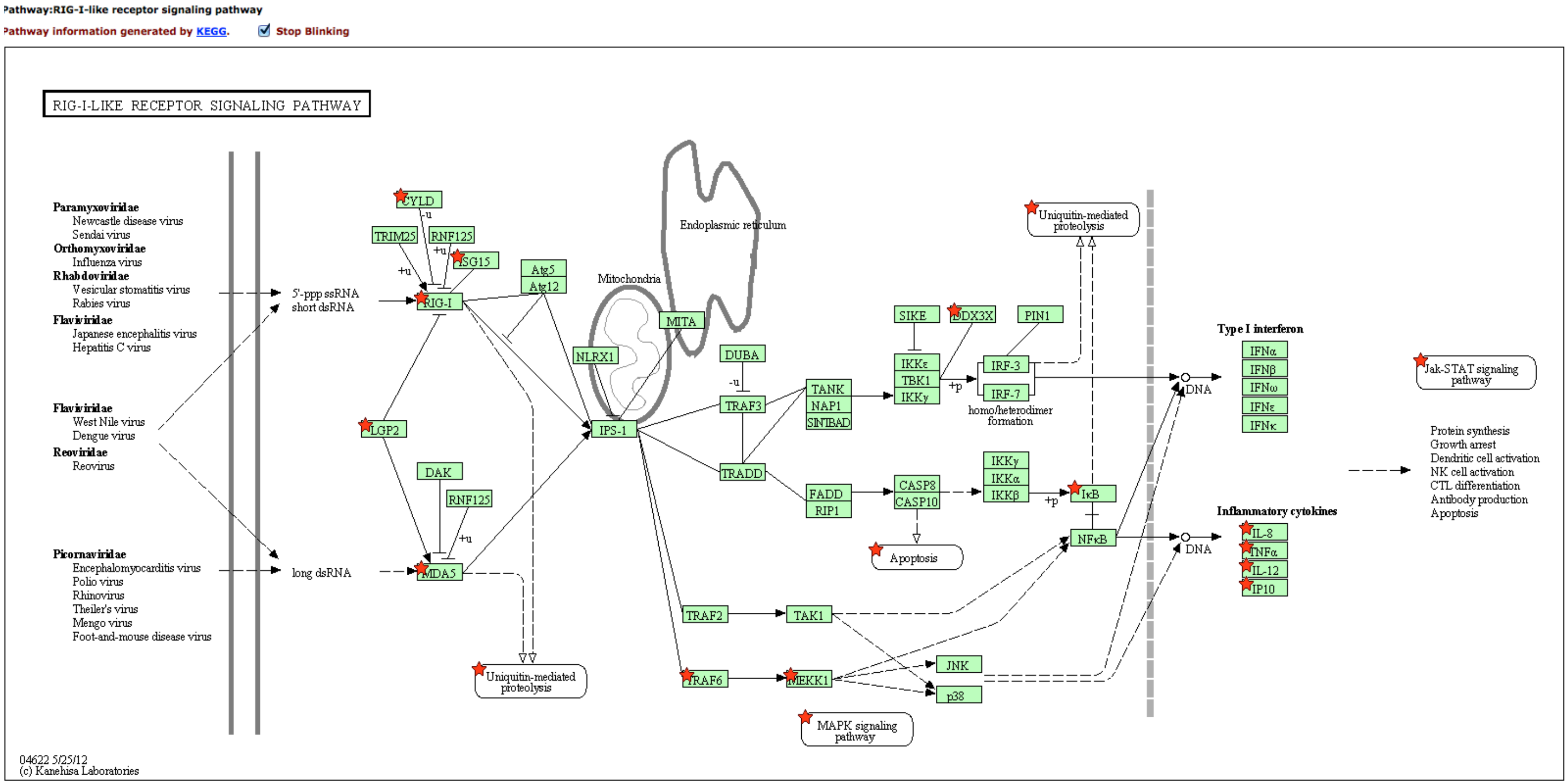
Current Gene List: List_1
Current Background: Homo sapiens
580 DAVID IDs

⊞ **Options**

19 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT	▬	33	5.7	5.0E-11	5.8E-9
□	KEGG_PATHWAY	RIG-I-like receptor signaling pathway	RT	▬	13	2.2	2.7E-6	1.6E-4
□	KEGG_PATHWAY	NOD-like receptor signaling pathway	RT	▬	12	2.1	4.3E-6	1.7E-4
□	KEGG_PATHWAY	Toll-like receptor signaling pathway	RT	▬	15	2.6	4.5E-6	1.3E-4
□	KEGG_PATHWAY	MAPK signaling pathway	RT	▬	23	4.0	5.5E-5	1.3E-3
□	KEGG_PATHWAY	Jak-STAT signaling pathway	RT	▬	16	2.8	1.6E-4	3.0E-3
□	KEGG_PATHWAY	Chemokine signaling pathway	RT	▬	17	2.9	4.0E-4	6.6E-3
□	KEGG_PATHWAY	Hematopoietic cell lineage	RT	▬	9	1.6	7.1E-3	9.8E-2
□	KEGG_PATHWAY	B cell receptor signaling pathway	RT	▬	8	1.4	1.1E-2	1.4E-1
□	KEGG_PATHWAY	Neurotrophin signaling pathway	RT	▬	10	1.7	2.1E-2	2.2E-1
□	KEGG_PATHWAY	Apoptosis	RT	▬	8	1.4	2.4E-2	2.3E-1
□	KEGG_PATHWAY	T cell receptor signaling pathway	RT	▬	9	1.6	2.6E-2	2.2E-1
□	KEGG_PATHWAY	Cytosolic DNA-sensing pathway	RT	▬	6	1.0	3.4E-2	2.7E-1

Highlight factors in KEGG Pathways



Pathway:RIG-I-like receptor signaling pathway

Pathway information generated by KEGG. ☑ Stop Blinking

Virology and Transcriptomics

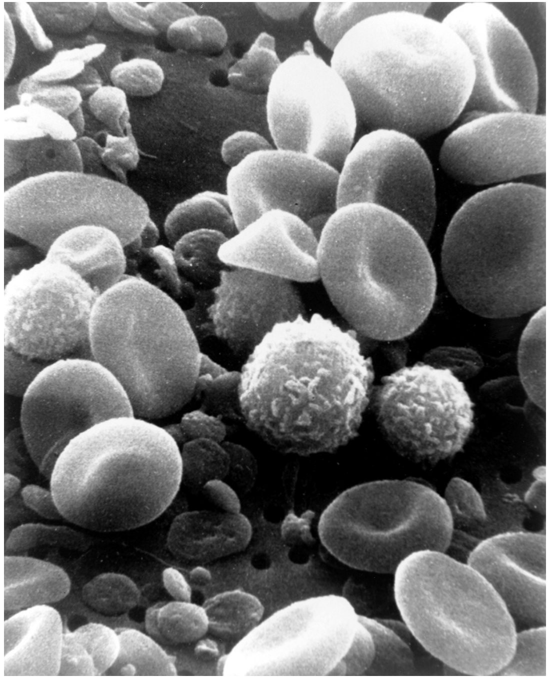
- Study the dynamics of viral gene expression during and infection cycle
- Compare virus gene expression between acute infection, latency, and re-activation
- Get an understanding of the genes and pathways that respond to viral infection
  - Pathways required for viral replication (candidate drug targets)
  - Host response pathways (possible determinants of virulence)
- Help elucidate the function of unknown genes based on their temporal and spatial patterns (guilt by association)
- Proxy for changes in the proteome and metabolome
- Molecular biomarkers of disease

SINGLE CELL RNA-SEQ

J. Craig Venter™ INSTITUTE

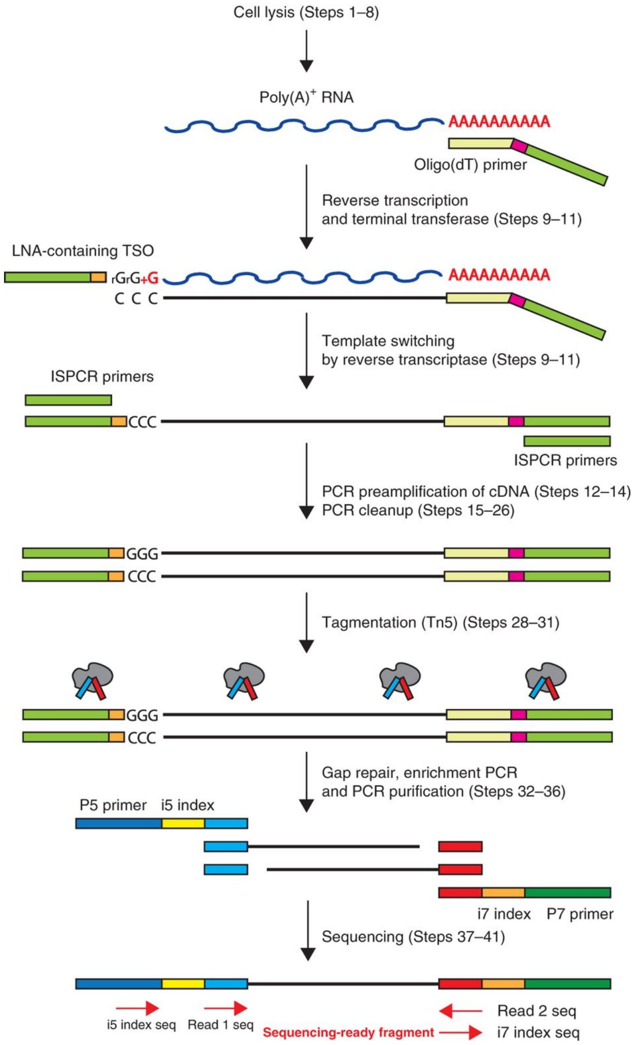
Single Cell Profiling

- Cells are the fundamental functional units of multicellular organisms
- Different cell types play different physiological roles in the body
- Cell identity and function (phenotype) is dictated by the subset of genes/proteins expressed
- Abnormalities in the expressed genome (disorders) form the physical basis of disease
- Understanding normal and abnormal cellular phenotypes is key for diagnosing disease and for identifying therapeutic targets



Bruce Wetzel & Harry Schaefer, National Cancer Institute
http://en.wikipedia.org/wiki/Image:SEM_blood_cells.jpg

- Transcriptional profiling of bulk samples obscures the cellular complexity of tissues
- Single cell RNA sequencing allows us to quantify cellular phenotypes in an unbiased fashion, enabling the evaluation of both known and novel cell subsets in tissue samples
- Explainable Artificial Intelligence has emerged as a valuable tool to characterize this complexity



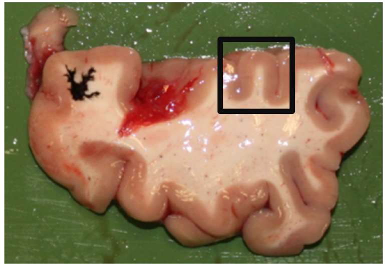
Smart-seq2

- Poly-A hybridization with 30nt polyT and 25nt 5' anchor sequence
- RT adding untemplated C
- Template switching with TSO
- Locked nucleic acid binds to untemplated C
- RT switches template
- Preamplification / cleanup
- DNA fragmentation and adapter ligation together
- Gap repair, enrich, purify

Picelli S, (2013) **Nat Methods** 10:1096-8.

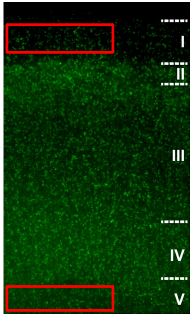
Smart-Seq2

Frozen cortical tissue slab

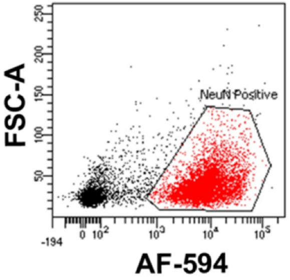


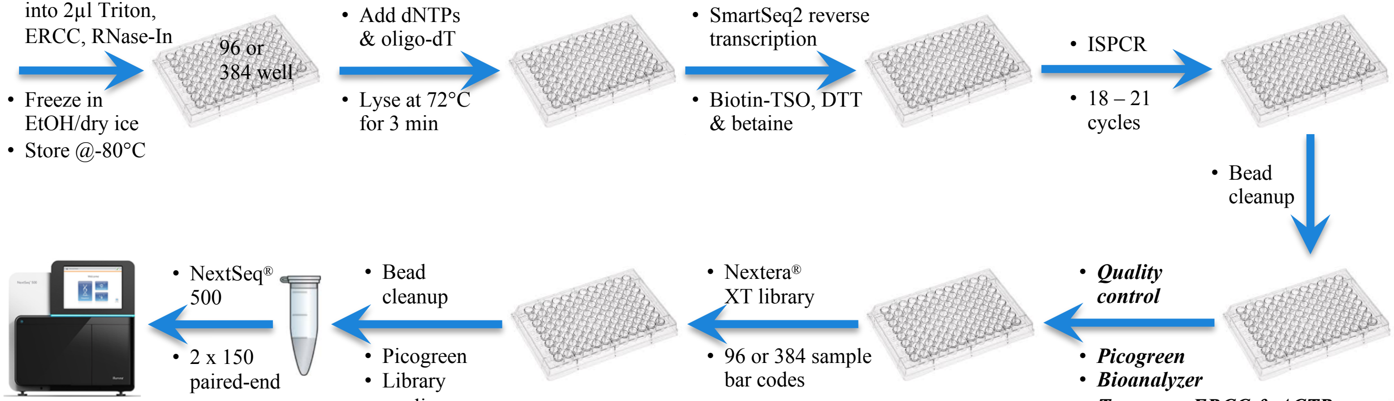
350 µm sections

Nissl stained



Sort





- Deposit single nuclei into 2µl Triton, ERCC, RNase-In
- Freeze in EtOH/dry ice
- Store @-80°C

- Add dNTPs & oligo-dT
- Lyse at 72°C for 3 min

- SmartSeq2 reverse transcription
- Biotin-TSO, DTT & betaine

- ISPCR
- 18 – 21 cycles

- Bead cleanup

- ***Quality control***
- ***Picogreen***
- ***Bioanalyzer***
- ***Taqman – ERCC & ACTB***

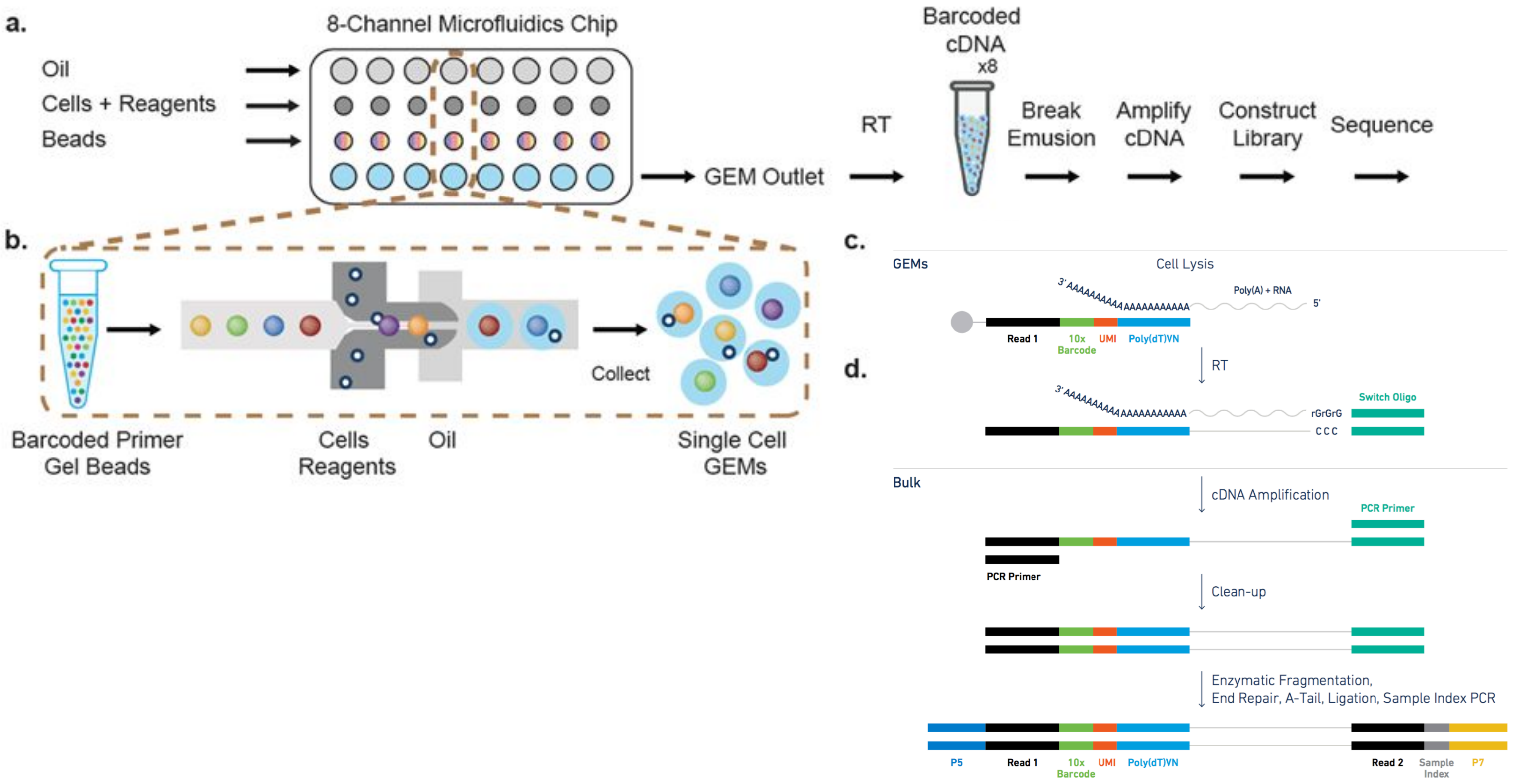
- Nextera® XT library
- 96 or 384 sample bar codes

- Bead cleanup
- Picogreen
- Library pooling

- NextSeq® 500
- 2 x 150 paired-end

Craig Venter™ INSTITUTE

10X Genomics

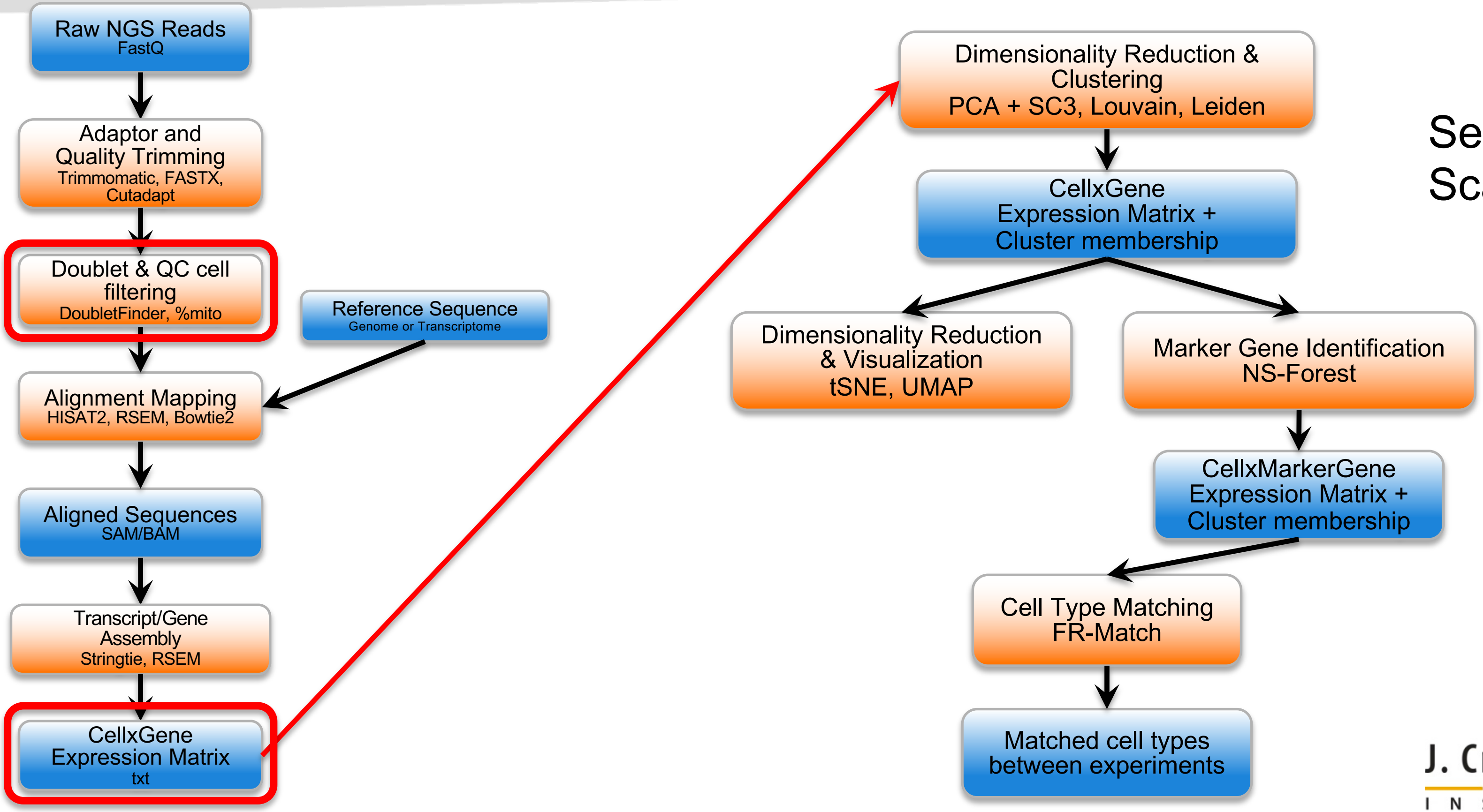


Smart-seq vs 10X

characteristics	Smart-seq	10X
Genes detected	4000 - 6000	1000 - 2000
Transcript structure	Full length transcripts	3’ or 5’ end only
Alternative splicing	Yes	No
PCR amplification bias	Yes	No (UMI)
Throughput	100-1000’s	10,000
Labor intensive	Yes	No
Cost	~$30/cell	~$1/cell

J. Craig Venter™ INSTITUTE

scRNA-seq Processing & Analysis Workflow



Raw NGS Reads
FastQ

Adaptor and Quality Trimming
Trimmomatic, FASTX, Cutadapt

Doublet & QC cell filtering
DoubletFinder, %mito

Reference Sequence
Genome or Transcriptome

Alignment Mapping
HISAT2, RSEM, Bowtie2

Aligned Sequences
SAM/BAM

Transcript/Gene Assembly
Stringtie, RSEM

CellxGene Expression Matrix
txt

Dimensionality Reduction & Clustering
PCA + SC3, Louvain, Leiden

CellxGene Expression Matrix + Cluster membership

Dimensionality Reduction & Visualization
tSNE, UMAP

Marker Gene Identification
NS-Forest

CellxMarkerGene Expression Matrix + Cluster membership

Cell Type Matching
FR-Match

Matched cell types between experiments

Seurat
Scanpy

J. Craig Venter™ INSTITUTE

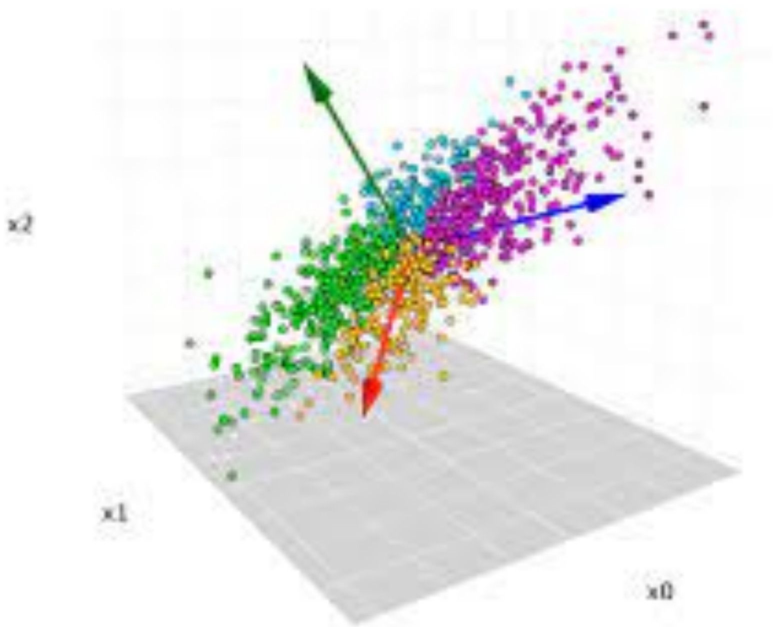
Dimensionality Reduction, Unsupervised Clustering and Visualization

- ***Strategy:*** using cell by gene expression values, perform PCA, unsupervised clustering and visualization in projected space
    - Unsupervised clustering
        - Louvain – graph-based community detection algorithm; Vincent Blondel, University of Louvain; J. Stat. Mech. (2008) P10008
        - Leiden – improvement to Louvain to ensure that all communities are guaranteed to be connected; Vincent Traag, Leiden University; https://www.nature.com/articles/s41598-019-41695-z
        - SC3 - unsupervised consensus clustering using multiple distance metrics, and dimensionality reduction methods for a user defined range of k (clusters); Martin Hemberg, Wellcome Trust Sanger Institute; https://doi.org/10.1038/nmeth.4236.
    - Visualization in projected space
        - tSNE - van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE" (PDF). Journal of Machine Learning Research. 9: 2579–2605.
        - UMAP - McInnes, Leland; Healy, John; Melville, James (2018-12-07). "Uniform manifold approximation and projection for dimension reduction". arXiv:1802.03426.
    - Platforms
        - Seurat - Rahul Satija, New York Genome Center; http://satijalab.org/seurat/
        - Scanpy – Fabian Theis, Helmholtz University; https://scanpy.readthedocs.io/en/stable/

J. Craig Venter™ INSTITUTE

Principal Component Analysis (PCA)

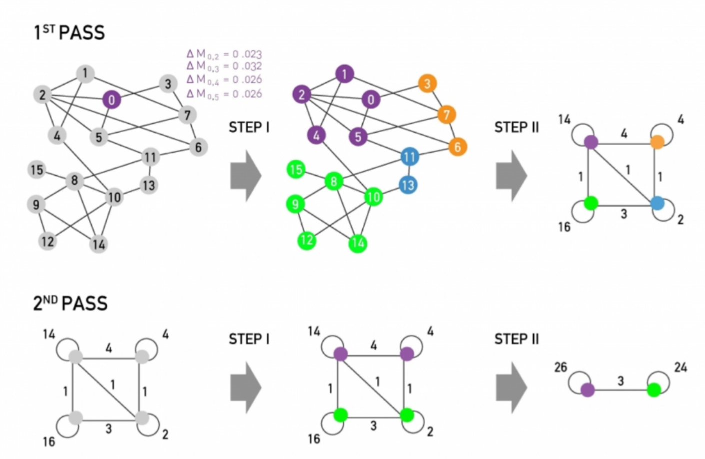
- Linear transformation for combining difference between data objects across all N original dimensions
- Convert originally correlated variables into linearly uncorrelated variables (PC: principal components) by:
  - Eigenvalue decomposition of data covariance matrix, or
  - Single value decomposition of data matrix
- The goal is to use a subset of transformed dimensions to represent the difference across all original dimensions for dimensionality reduction
- Generated dimensions (PCs) lose the meaning of the original variables
- May not be able to identify small data clusters, depending on the relative scaling of the original variables



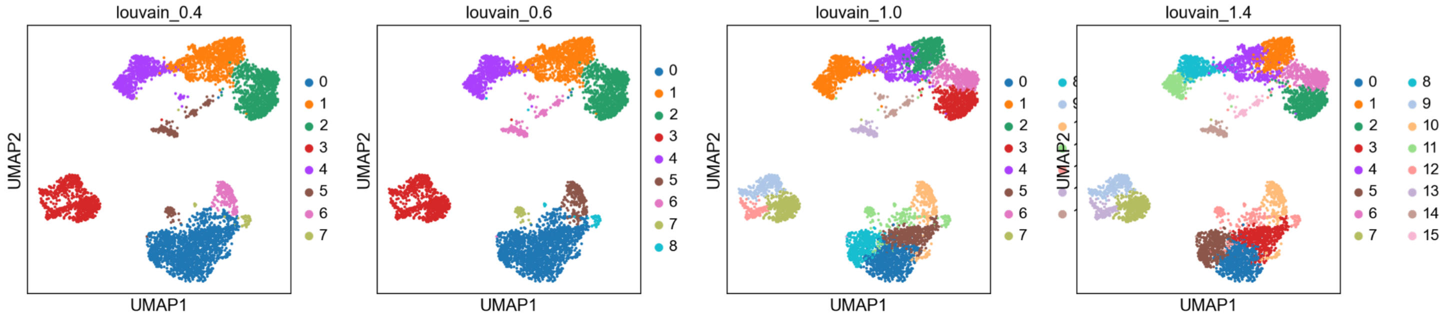
J. Craig Venter™ INSTITUTE

Clustering using Louvain

1. Build an unweighted k nearest neighbor (KNN) graph
2. Add weights, and obtain a shared nearest neighbor (SNN) graph
3. Iterate to optimize modularity



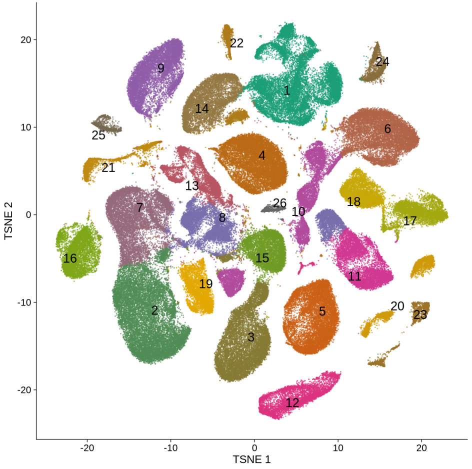
Effect of resolution parameter

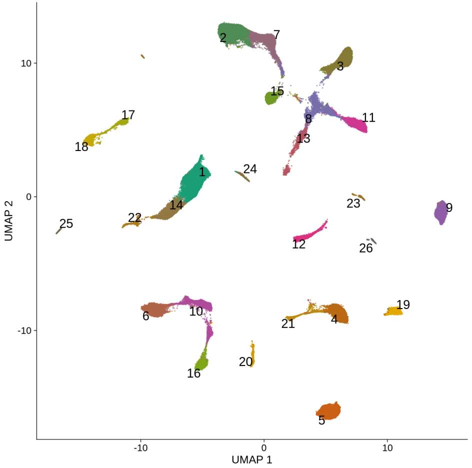


J. Craig Venter
INSTITUTE

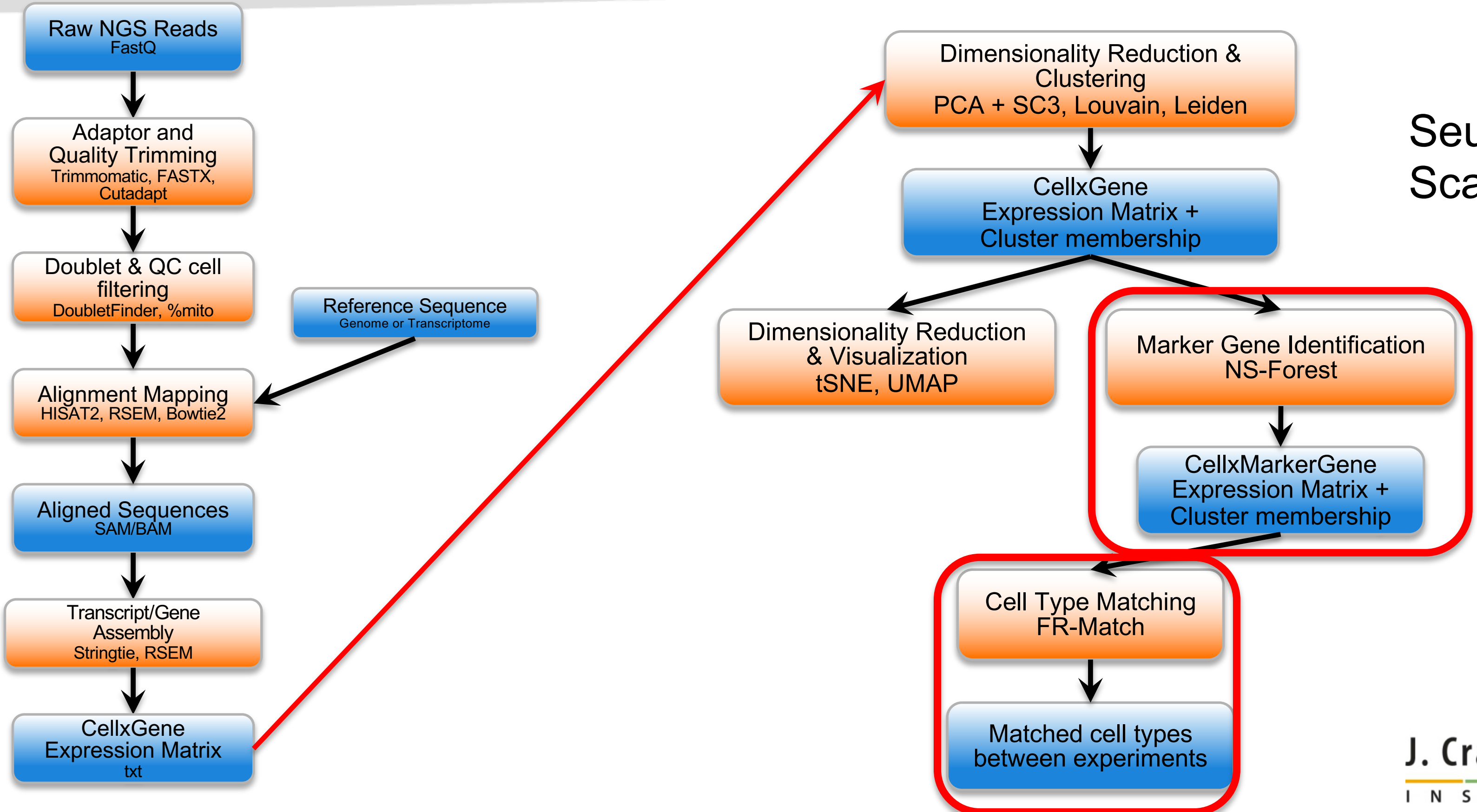
t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Nonlinear transformation
- Goals is to plot similar data objects nearby, and dissimilar objects distant by:
  - Constructing t-distributed probability model over pairs of high-dimensional data objects
  - Minimizing Kullback-Leibler (KL) distance between the high-D distribution and the low-D distribution
- Generated dimensions lose the meaning of the original variables
- Distance between objects on transformed low-D space does not correspond to (in a ratio to) original variance in high-D space
- Small data clusters can be identified, depending on their similarity with the other data objects
- Requires setting at least two algorithm parameters that usually change the 2D layout significantly





scRNA-seq Processing & Analysis Workflow



Raw NGS Reads
FastQ

Adaptor and Quality Trimming
Trimmomatic, FASTX, Cutadapt

Doublet & QC cell filtering
DoubletFinder, %mito

Reference Sequence
Genome or Transcriptome

Alignment Mapping
HISAT2, RSEM, Bowtie2

Aligned Sequences
SAM/BAM

Transcript/Gene Assembly
Stringtie, RSEM

CellxGene Expression Matrix
txt

Dimensionality Reduction & Clustering
PCA + SC3, Louvain, Leiden

CellxGene Expression Matrix + Cluster membership

Dimensionality Reduction & Visualization
tSNE, UMAP

Marker Gene Identification
NS-Forest

CellxMarkerGene Expression Matrix + Cluster membership

Cell Type Matching
FR-Match

Matched cell types between experiments

Seurat
Scanpy

J. Craig Venter™
INSTITUTE

Resources

- Publications
  - NS-Forest v1.0 - Aevermann B, et al. (2018) ***Human Molecular Genetics***, 27(R1):R40-R47. PMID: 29590361
  - NS-Forest v2.0 - Aevermann B, et al. (2021) ***Genome Research***, 31:1767-1780. PMID: 34088715
  - FR-Match v1.0 - Zhang Y, et al. (2021) ***Briefings in Bioinformatics***, 22:bbaa339. PMID: 33249453
  - FR-Match v2.0 - https://www.biorxiv.org/content/10.1101/2021.10.17.464718v2
  - Cortical layer 1 cell types - Boldog E, et al. (2018) ***Nature Neuroscience***, 21: 1185-1195. PMID: 30150662
  - MTG human cell types - Hodge RD, et al. (2019) ***Nature***, 573:61-68. PMID: 31435019
  - M1 human, mouse, marmoset – Bakken T, et al. (2021) ***Nature***, 598:111-119. PMID: 34616062
- Source Code
  - NS-Forest source code is available at https://github.com/JCVenterInstitute/NSForest
  - FR-Match source code is available at https://github.com/JCVenterInstitute/FRmatch
- Protocols
  - NS-Forest protocol is available at https://www.protocols.io/view/ns-forest-version-2-un7evhn
  - FR-Match protocol is available at https://www.protocols.io/view/fr-match-cell-type-matching-for-scrnaseq-data-bmyfk7tn
- Ontology
  - PCL is available through the BioPortal - https://bioportal.bioontology.org/ontologies/PCL