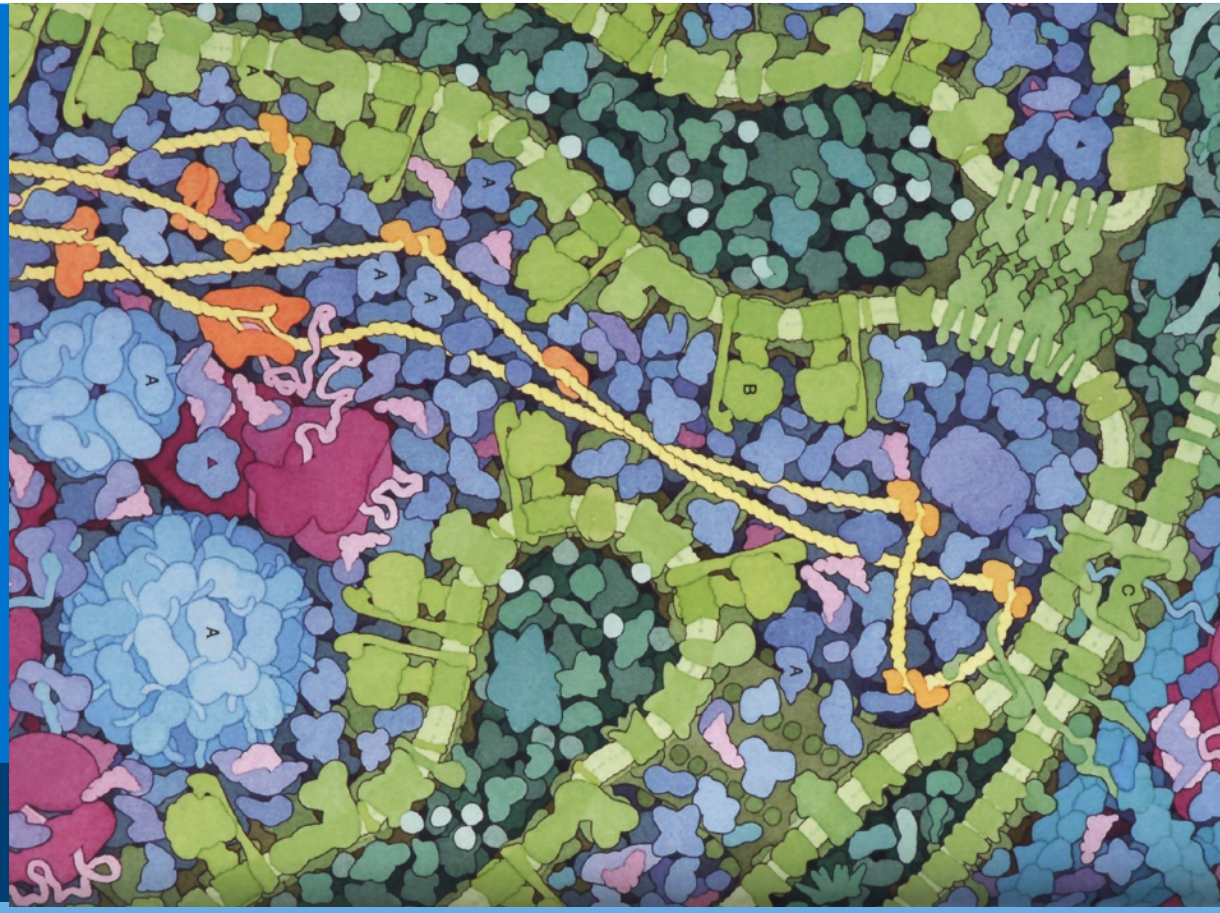
UNIVERSITY OF CAMBRIDGE

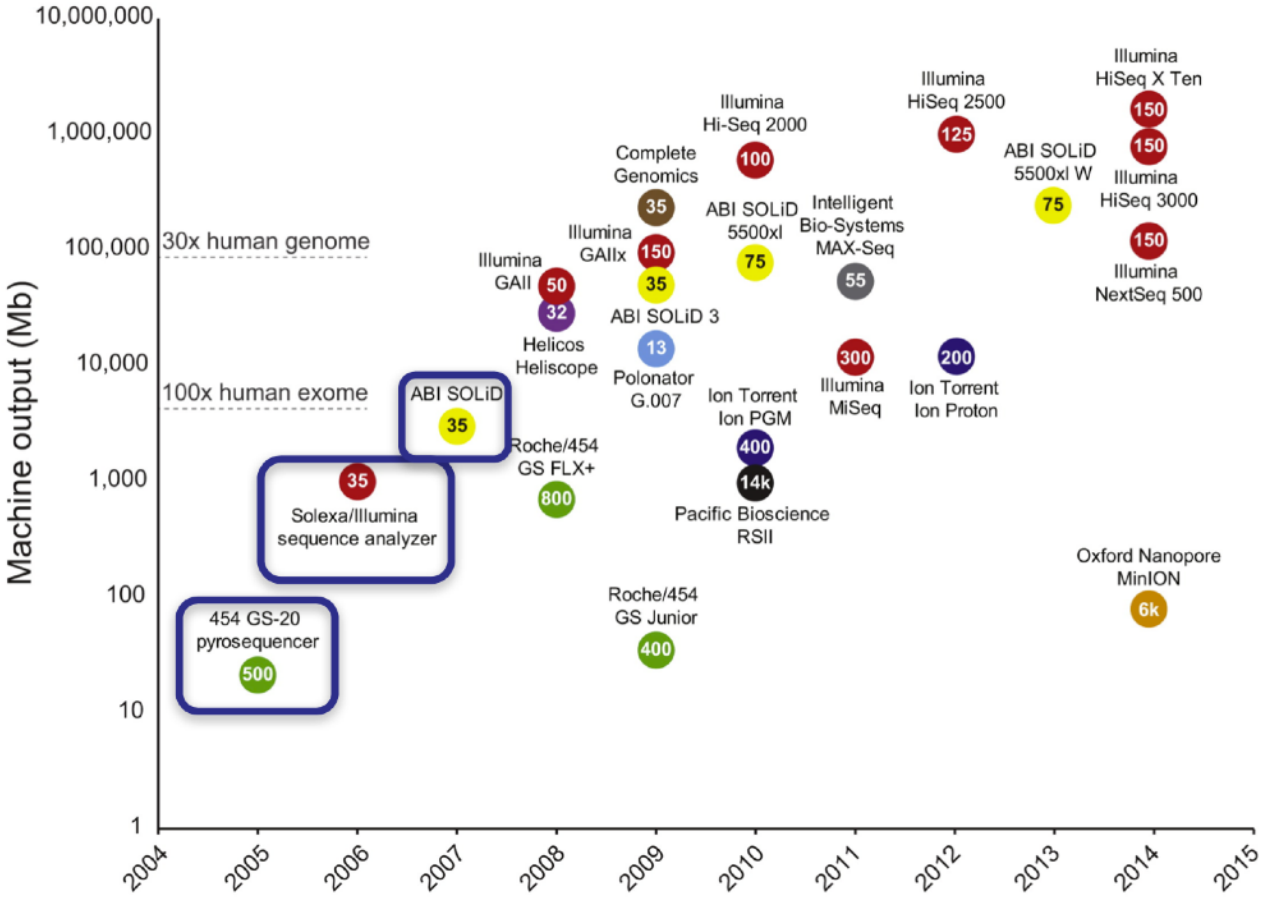
Bioinformatics for PIs - RNA Seq

Dr. Anton Enright
Group Leader,
Academic Lead - Genomics Core,
Dept of Pathology,
University of Cambridge
aje39@cam.ac.uk



Department of Pathology

The Next Generation - High Output, Low Cost



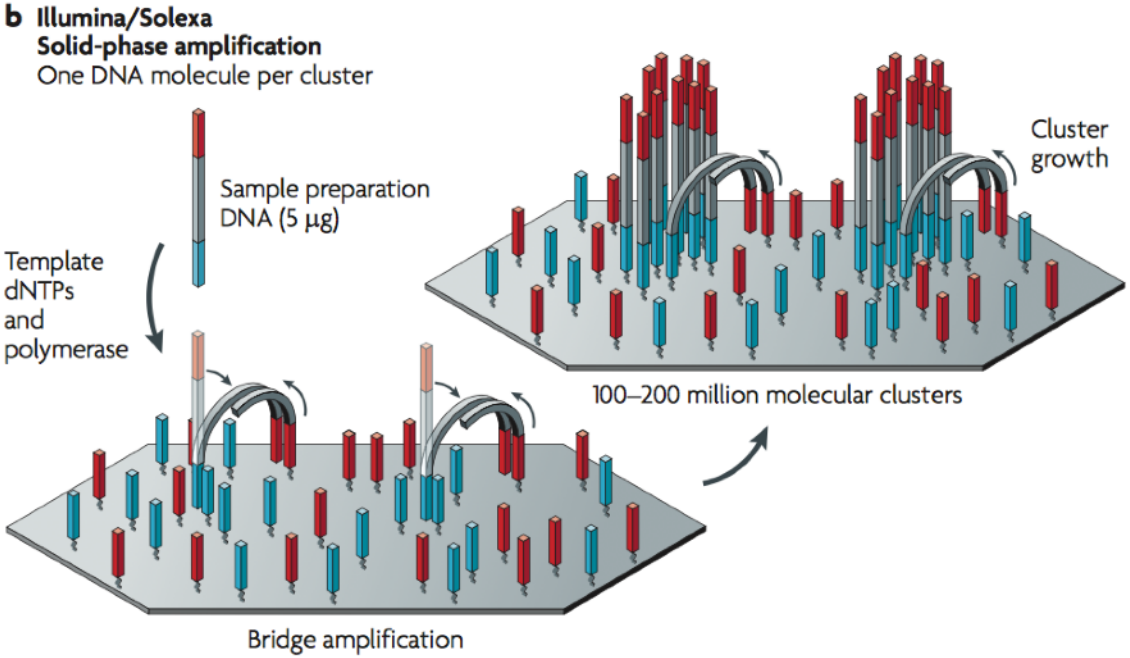
Reuter JA, Spacek DV & Snyder MP, Mol Cell. 2015 May 21;58(4):586-97.

UNIVERSITY OF CAMBRIDGE

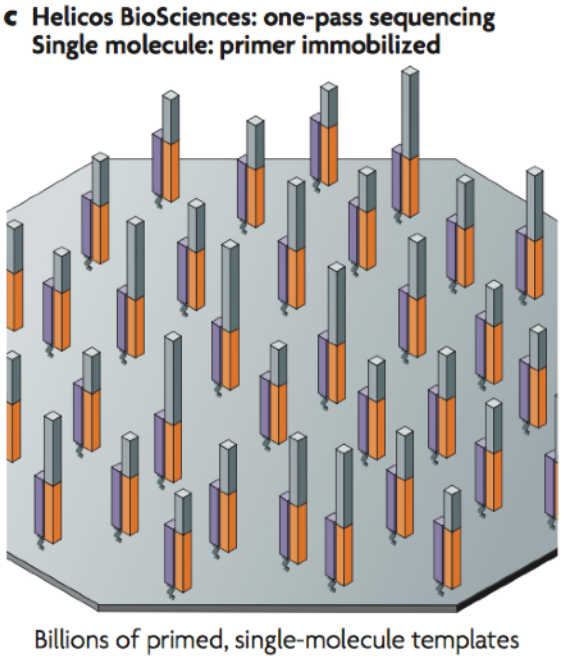
Solid Phase Immobilisation Sequencing (e.g. Illumina)

- Solid Phase methods immobilisation

Images from: Metzker ML, Nat Rev Genet. 2010 Jan;11(1):31-46.



Illumina Sequencing (PCR)

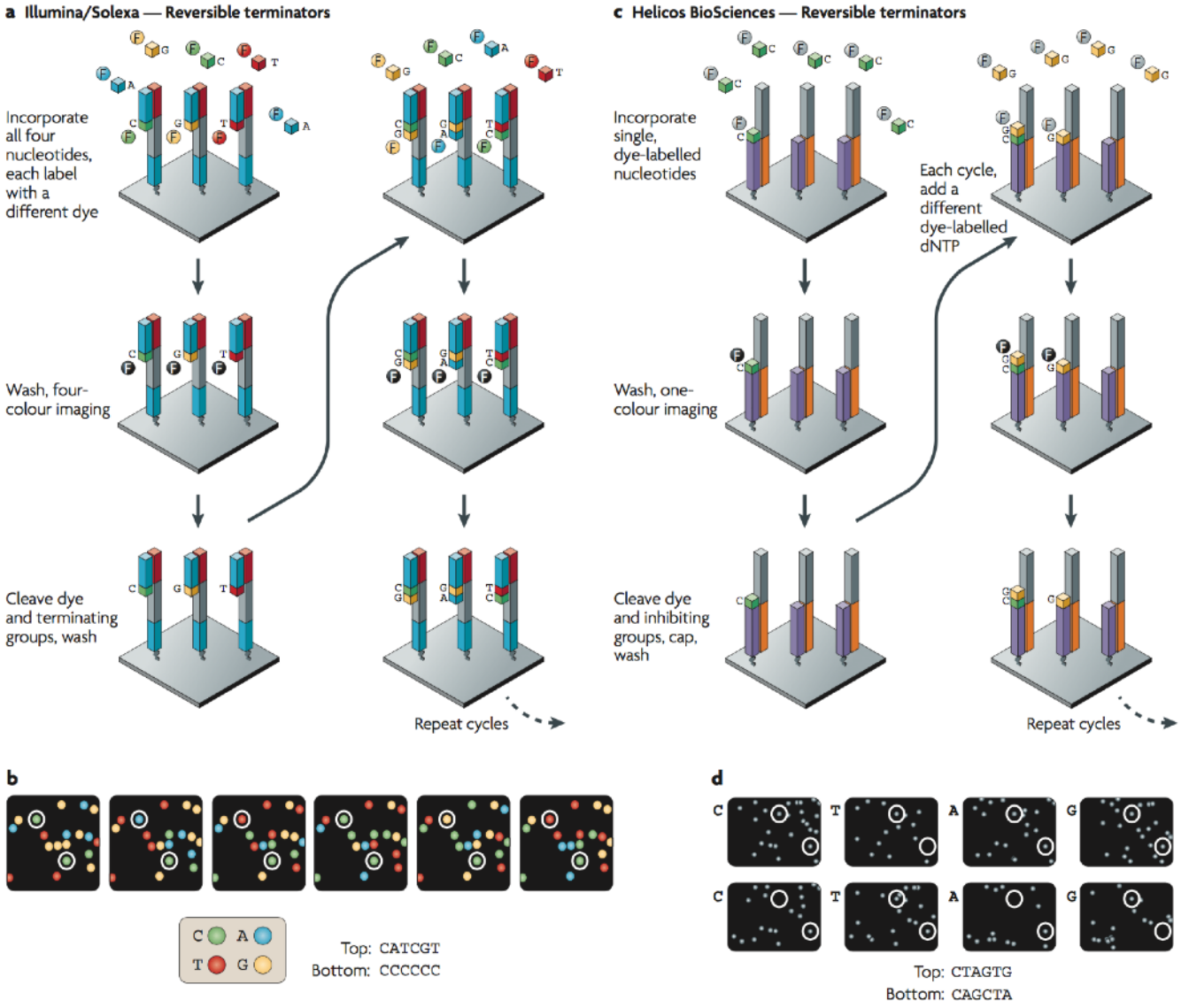


Helicos (Single Molecule)

Illumina - Reversible terminator sequencing

Images from: Metzker ML, Nat Rev Genet. 2010 Jan;11(1):31-46.

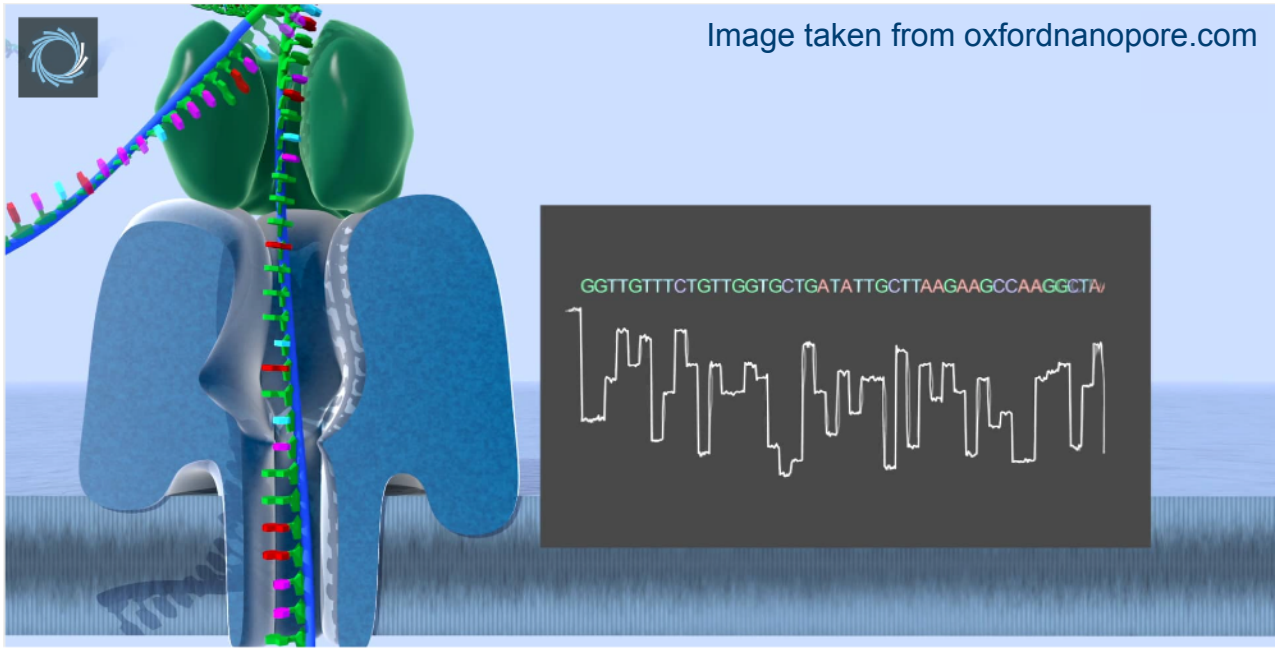
- Nucleotides are flourescently labelled
- They contain terminators attached to the flourophore
- After all four bases are incorporated by a polymerase the clusters are laser imaged
- These terminators are removed after a cycle and the next base is added



4th Gen Approaches to sequencing

Single molecule sequencing with Nanopores

- gDNA fragmented and adapters attached
- gDNA ligated to a molecular motor
- Molecular motor attaches to a protein pore across an artificial membrane
- Voltage across each pore is measured in real-time
- Nucleotides can be called based on current changes



- Very long reads, 50-100kb possible
- Currently relatively high error rates
- Low cost, tiny machine

4th Gen Approaches to sequencing

Single molecule sequencing with Nanopores

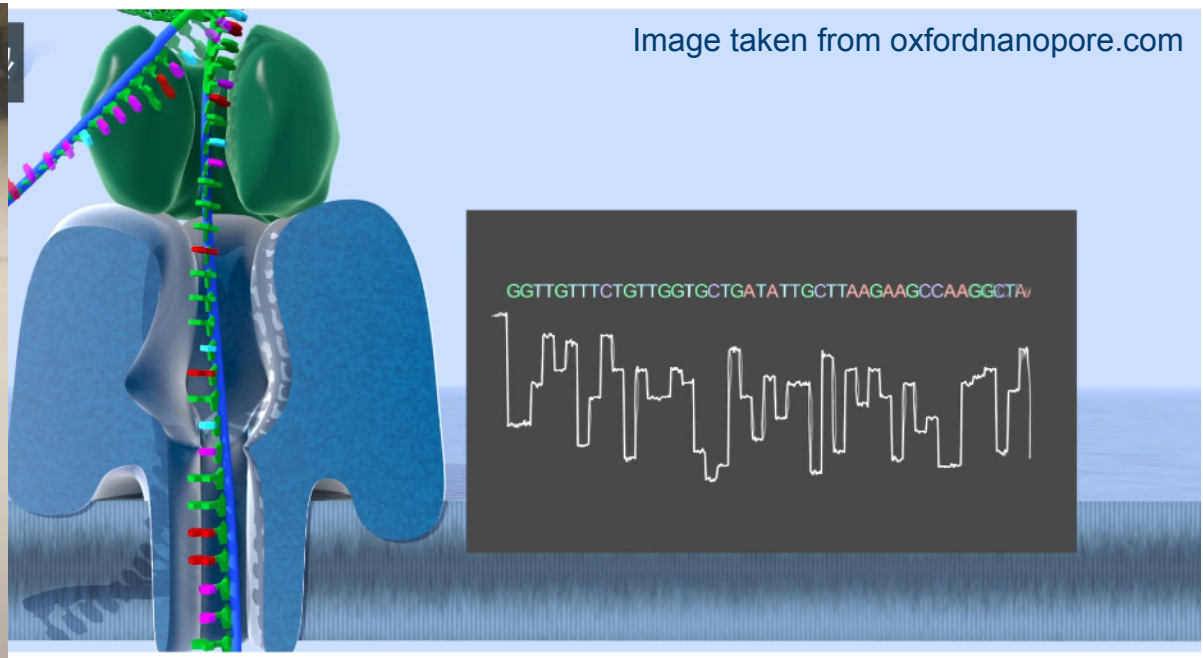
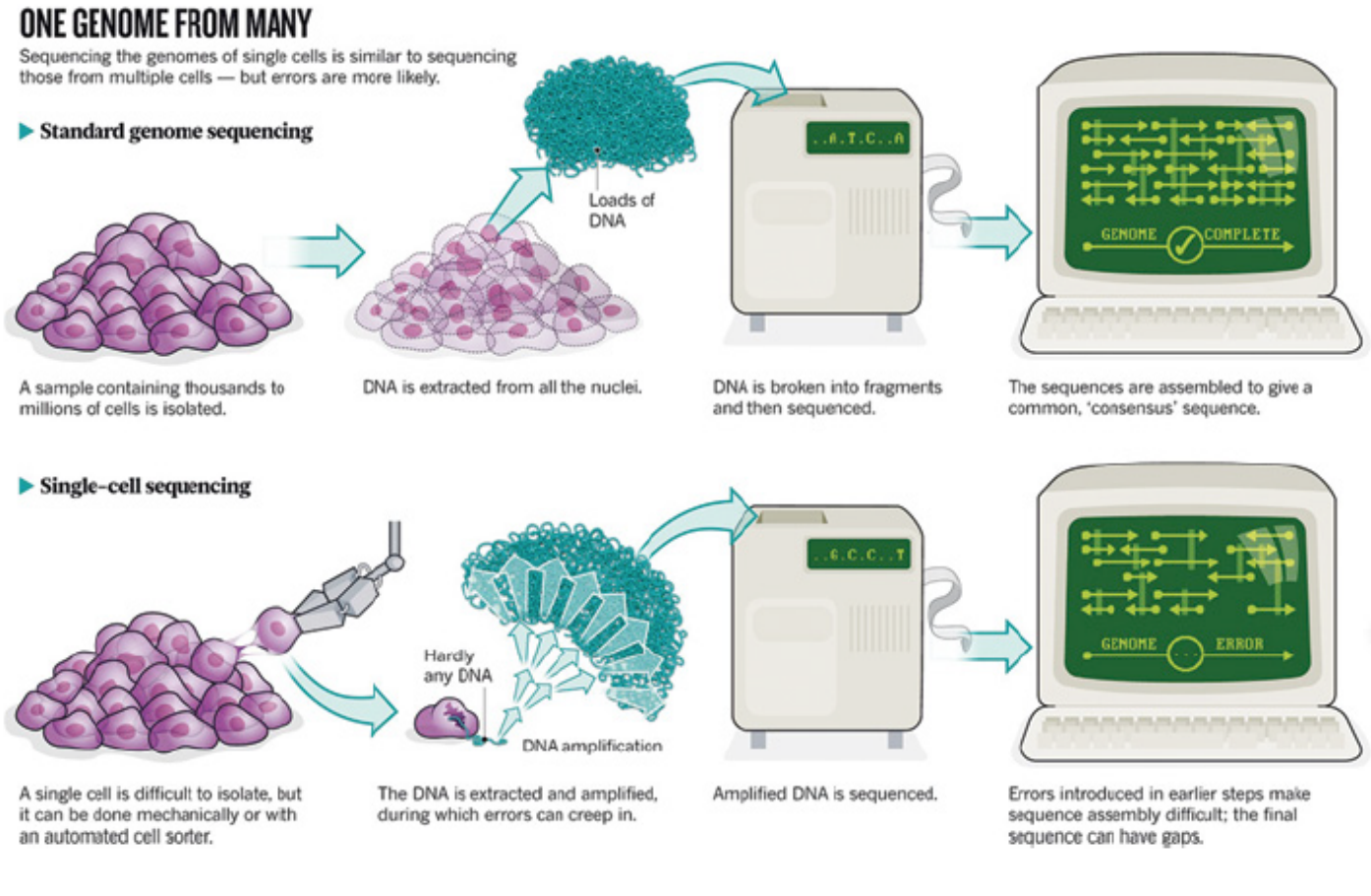
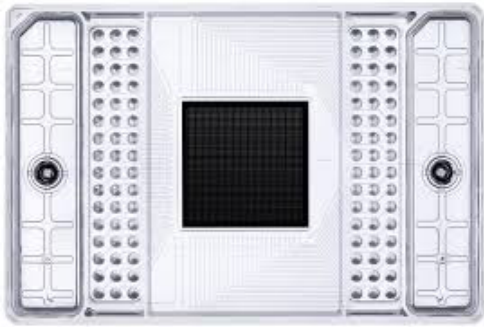


Image taken from oxfordnanopore.com

- Very long reads, 50-100kb possible
- Currently relatively high error rates
- Low cost, tiny machine

Single Cell Genomics





FLUIDIGM®

10X GENOMICS®

RNA Seq Data

- **Raw Reads (Big - Gigabytes)**
  - FASTQ
- **Mapped Reads (Big - Gigabytes)**
  - SAM Format
  - BAM Format
- **Quantitated Read Counts (Small - Megabytes)**
  - RSEM
  - HTSeq Count
- **Normalised Reads (Small - Megabytes)**
  - EdgeR, Voom
  - DESeq

FASTQ Data Format

- FASTQ Data

```
@HS9_07989:8:1101:1349:2153#10/1
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAGAC
+
A:D<DG@@D?DAH4FGAD?DBBDC>:A?B)GGD@CEEC6E->8DAE.>=>8D9HCF>FEC@FB:?CE;A/?(/:+
@HS9_07989:8:1101:1275:2171#10/1
TCGCTNGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
B@DGA!DD>CFFFHFGEGIDF8JHFHFGDEGGHG=HECE?HFDDGDEFDGCHDFGFFFB>EEFFFHFHCGGFH:F
@HS9_07989:8:1101:1436:2228#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DGEGFFFFFEFEEFJGIDGIJGEGFFDCGIFIEHGCGEGGGDGDHEHGCHDFAFEFE;HEGFHHDECGGHH@D
@HS9_07989:8:1101:1500:2229#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DCGGDFFFDEHDGADGIGBFDFEGGGDIEGGHEH@?EBHGFG=FADGGCDDFCFE8F8DEBBDDFECGB?H@?
```

FASTQ Data Format

- FASTQ Data

```
@HS9_07989:8:1101:1349:2153#10/1
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAGAC
+
A:D<DG@@D?DAH4FGAD?DBBDC>:A?B)GGD@CEEC6E->8DAE.>=>8D9HCF>FEC@FB:?CE;A/?(/:+
@HS9_07989:8:1101:1275:2171#10/1
TCGCTNGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
B@DGA!DD>CFFFHFGEGIDF8JHFHFGDEGGHG=HECE?HFDDGDEFDGCHDFGFFFB>EEFFFHFHCGGFH:F
@HS9_07989:8:1101:1436:2228#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DGEGFFFFFEFEEFJGIDGIJGEGFFDCGIFIEHGCGEGGGDGDHEHGCHDFAFEFE;HEGFHHDECGGHH@D
@HS9_07989:8:1101:1500:2229#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DCGGDFFFDEHDGADGIGBFDFEGGGDIEGGHEH@?EBHGFG=FADGGCDDFCFE8F8DEBBDDFECGB?H@?
```

FASTQ Data Format

- FASTQ Data

Header Line (@)

```
@HS9_07989:8:1101:1349:2153#10/1
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAGAC
+
A:D<DG@@D?DAH4FGAD?DBBDC>:A?B)GGD@CEEC6E->8DAE.>=>8D9HCF>FEC@FB:?CE;A/?(/:+
@HS9_07989:8:1101:1275:2171#10/1
TCGCTNGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
B@DGA!DD>CFFFHFGEGIDF8JHFHFGDEGGHG=HECE?HFDDGDEFDGCHDFGFFFB>EEFFFHFHCGGFH:F
@HS9_07989:8:1101:1436:2228#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DGEGFFFFFEFEEFJGIDGIJGEGFFDCGIFIEHGCGEGGGDGDHEHGCHDFAFEFE;HEGFHHDECGGHH@D
@HS9_07989:8:1101:1500:2229#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DCGGDFFFDEHDGADGIGBFDFEGGGDIEGGHEH@?EBHGFG=FADGGCDDFCFE8F8DEBBDDFECGB?H@?
```

FASTQ Data Format

- FASTQ Data

Header Line (@)
Sequence Line

```
@HS9_07989:8:1101:1349:2153#10/1
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAGAC
+
A:D<DG@@D?DAH4FGAD?DBBDC>:A?B)GGD@CEEC6E->8DAE.>=>8D9HCF>FEC@FB:?CE;A/?(/:+
@HS9_07989:8:1101:1275:2171#10/1
TCGCTNGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
B@DGA!DD>CFFFHFGEGIDF8JHFHFGDEGGHG=HECE?HFDDGDEFDGCHDFGFFFB>EEFFFHFHCGGFH:F
@HS9_07989:8:1101:1436:2228#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DGEGFFFFFEFEEFJGIDGIJGEGFFDCGIFIEHGCGEGGGDGDHEHGCHDFAFEFE;HEGFHHDECGGHH@D
@HS9_07989:8:1101:1500:2229#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DCGGDFFFDEHDGADGIGBFDFEGGGDIEGGHEH@?EBHGFG=FADGGCDDFCFE8F8DEBBDDFECGB?H@?
```

FASTQ Data Format

- FASTQ Data

Header Line (@)
Sequence Line
Description Line

```
@HS9_07989:8:1101:1349:2153#10/1
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAGAC
+
A:D<DG@@D?DAH4FGAD?DBBDC>:A?B)GGD@CEEC6E->8DAE.>=>8D9HCF>FEC@FB:?CE;A/?(/:+
@HS9_07989:8:1101:1275:2171#10/1
TCGCTNGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
B@DGA!DD>CFFFHFGEGIDF8JHFHFGDEGGHG=HECE?HFDDGDEFDGCHDFGFFFB>EEFFFHFHCGGFH:F
@HS9_07989:8:1101:1436:2228#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DGEGFFFFFEFEEFJGIDGIJGEGFFDCGIFIEHGCGEGGGDGDHEHGCHDFAFEFE;HEGFHHDECGGHH@D
@HS9_07989:8:1101:1500:2229#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DCGGDFFFDEHDGADGIGBFDFEGGGDIEGGHEH@?EBHGFG=FADGGCDDFCFE8F8DEBBDDFECGB?H@?
```

FASTQ Data Format

- FASTQ Data

Header Line (@)
Sequence Line
Description Line
Quality Score Line

```
@HS9_07989:8:1101:1349:2153#10/1
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAGAC
+
A:D<DG@@D?DAH4FGAD?DBBDC>:A?B)GGD@CEEC6E->8DAE.>=>8D9HCF>FEC@FB:?CE;A/?(/:+
@HS9_07989:8:1101:1275:2171#10/1
TCGCTNGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
B@DGA!DD>CFFFHFGEGIDF8JHFHFGDEGGHG=HECE?HFDDGDEFDGCHDFGFFFB>EEFFFHFHCGGFH:F
@HS9_07989:8:1101:1436:2228#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DGEGFFFFFEFEEFJGIDGIJGEGFFDCGIFIEHGCGEGGGDGDHEHGCHDFAFEFE;HEGFHHDECGGHH@D
@HS9_07989:8:1101:1500:2229#10/1
TCGCTTGGTGCAGATCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTAGCTTATCTCGTATGCCGTC
+
A@DCGGDFFFDEHDGADGIGBFDFEGGGDIEGGHEH@?EBHGFG=FADGGCDDFCFE8F8DEBBDDFECGB?H@?
```

FASTQ Quality Score

Phred score	Probability that the base is incorrect	Precision of the base
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %

Phil Green (author of Phred)

0	nul	1	soh	2	stx	3	etx	4	eot	5	enq	6	ack	7	bel
8	bs	9	ht	10	nl	11	vt	12	np	13	cr	14	so	15	si
16	dle	17	dc1	18	dc2	19	dc3	20	dc4	21	nak	22	syn	23	etb
24	can	25	em	26	sub	27	esc	28	fs	29	gs	30	rs	31	us
32	sp	33	!	34	"	35	#	36	$	37	%	38	&	39	'
40	(	41	)	42	*	43	+	44	,	45	-	46	.	47	/
48	0	49	1	50	2	51	3	52	4	53	5	54	6	55	7
56	8	57	9	58	:	59	;	60	<	61	=	62	>	63	?
64	@	65	A	66	B	67	C	68	D	69	E	70	F	71	G
72	H	73	I	74	J	75	K	76	L	77	M	78	N	79	O
80	P	81	Q	82	R	83	S	84	T	85	U	86	V	87	W
88	X	89	Y	90	Z	91	[	92	\	93	]	94	^	95	_
96	`	97	a	98	b	99	c	100	d	101	e	102	f	103	g
104	h	105	i	106	j	107	k	108	l	109	m	110	n	111	o
112	p	113	q	114	r	115	s	116	t	117	u	118	v	119	w
120	x	121	y	122	z										

The standard ASCII Table

$Q=-\log_{10}p$

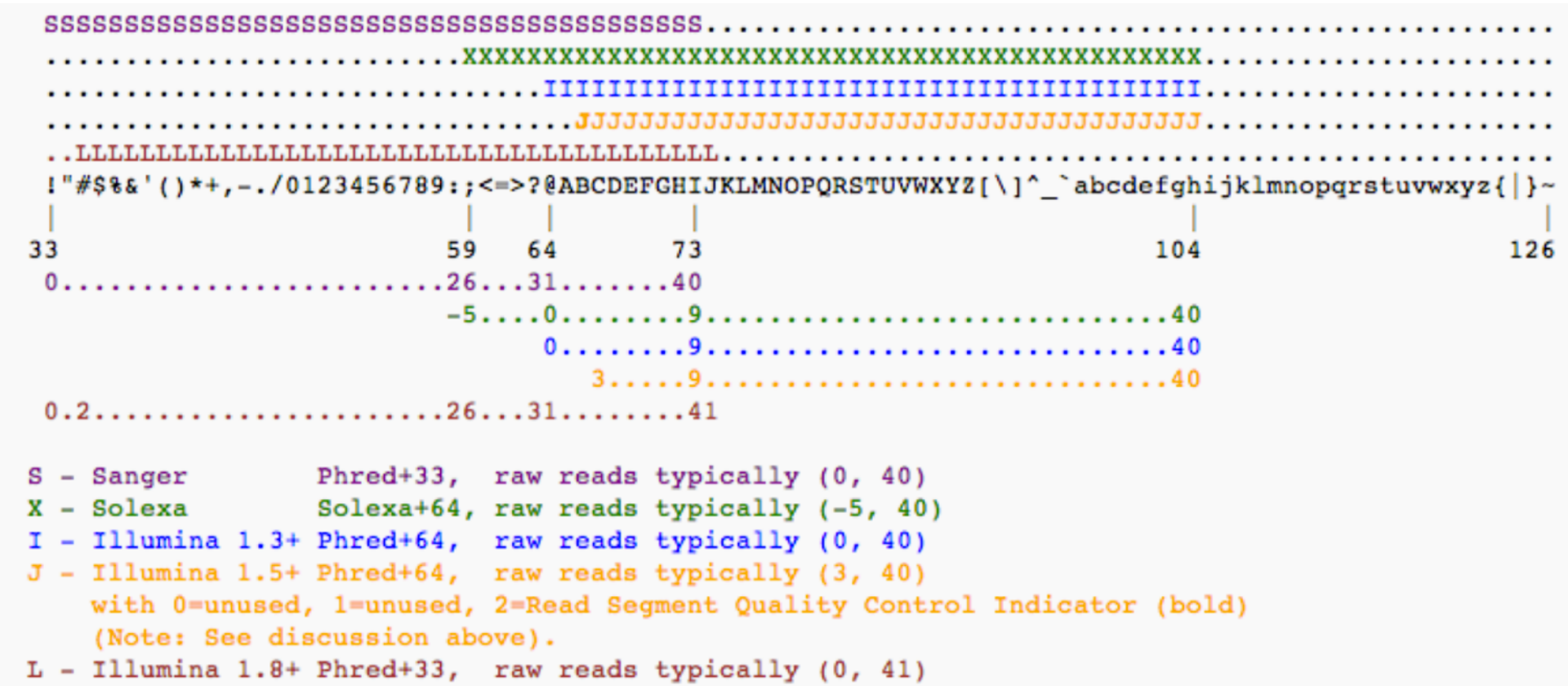
Q is quality score

p is probability that the base call is wrong

		0	!	1	"	2	#	3	$	4	%	5	&	6	'
7	(	8	)	9	*	10	+	11	,	12	-	13	.	14	/
15	0	16	1	17	2	18	3	19	4	20	5	21	6	22	7
23	8	24	9	25	:	26	;	27	<	28	=	29	>	30	?
31	@	32	A	33	B	34	C	35	D	36	E	37	F	38	G
39	H	40	I												

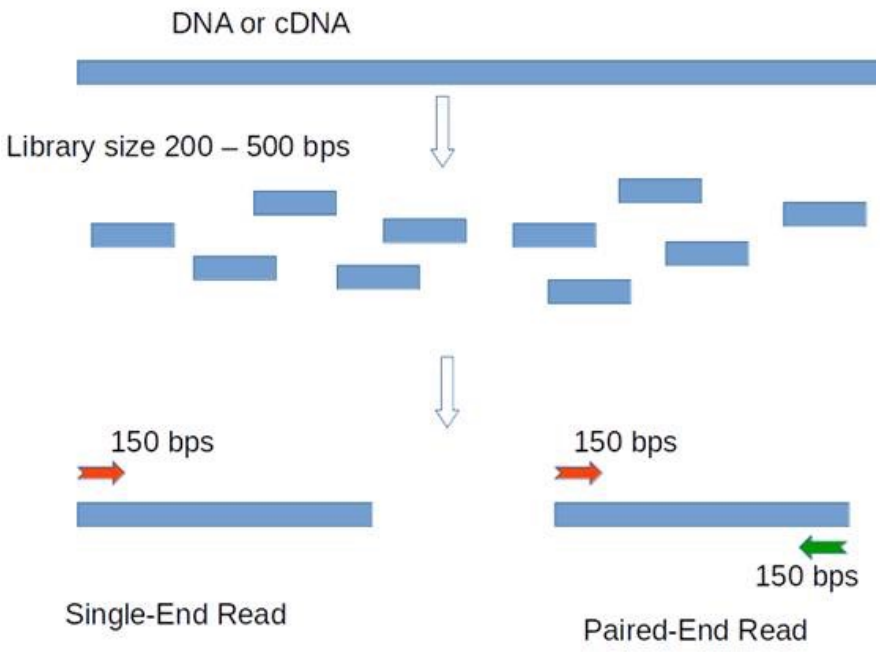
Some ascii characters are unprintable so the entire table is shifted by 33 giving a final lookup table as follows where each symbol represents a unique Phred score.

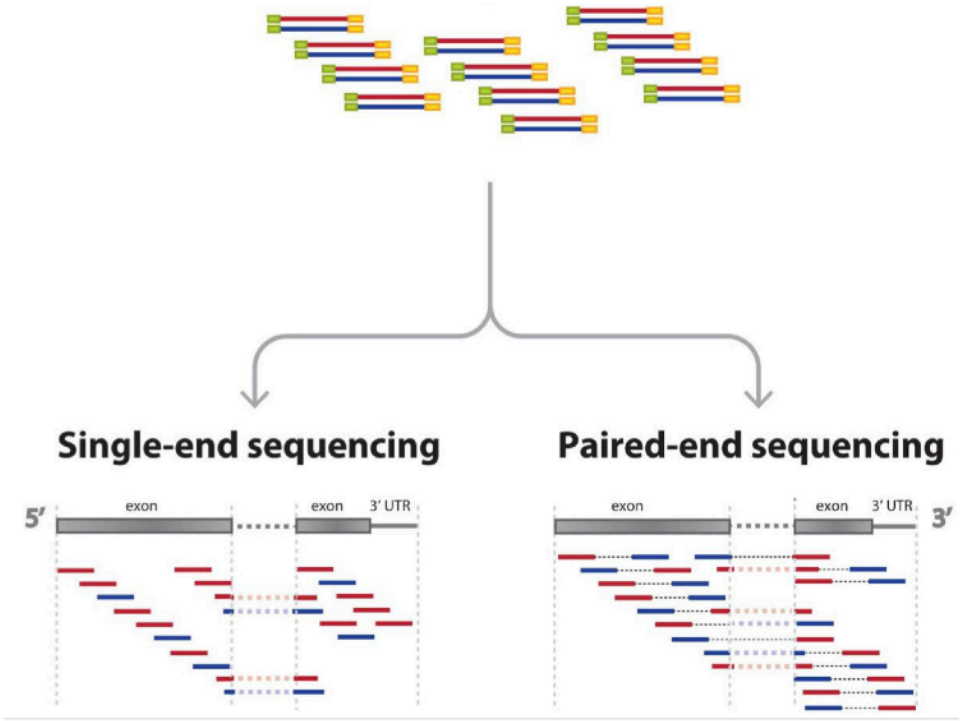
Different FASTQ encodings



https://en.wikipedia.org/wiki/FASTQ_format

Single End / Paired End





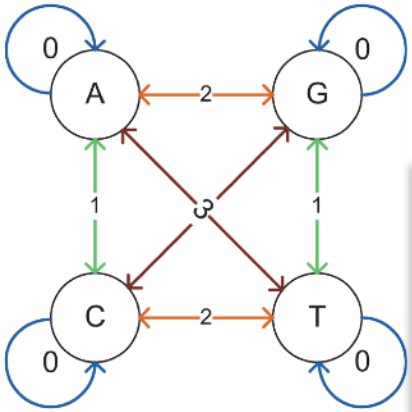
Some Paired end approaches are stranded, either RF or FR, best to check with sequencing facility - Mapping tools need to know this.

Other Formats - ABI SoLiD ColorSpace

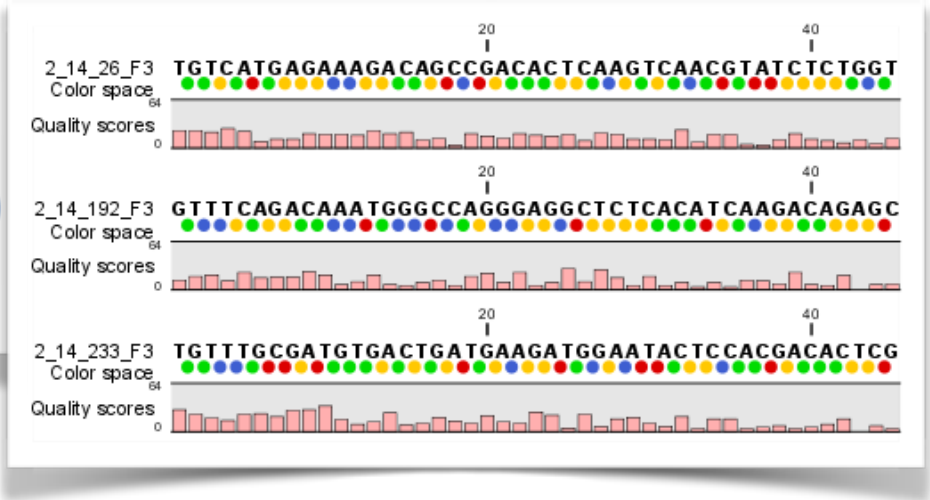
- ColorSpace, e.g. ABI SoLiD a sequence of colour calls
  - SOLID Machine calls pairs of bases with a coloured tag
- E.G. Green is either AC,CA,GT or TG
- Blue is AA, TT, GG or CC

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

A



B

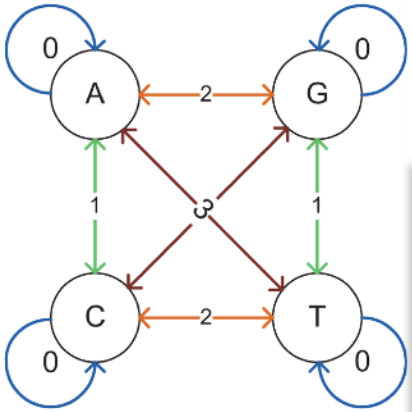


Other Formats - ABI SoLiD ColorSpace

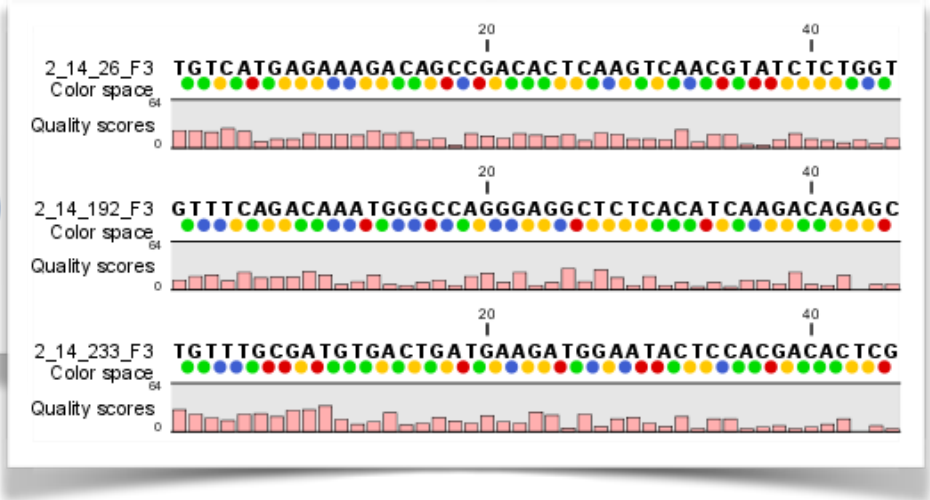
- ColorSpace, e.g. ABI SoLiD a sequence of colour calls
  - SOLID Machine calls pairs of bases with a coloured tag
- E.G. Green is either AC,CA,GT or TG
- Blue is AA, TT, GG or CC

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

A



B



RNA Seq - Mapping

Read Mapping - RNA Seq

- Most methods make use of the Burrows-Wheeler Transformation
  - BWA, Bowtie, Hisat, STAR
  - Input genome is loaded and indexed to a file
  - Location of all *k-mers* of a fixed size can be rapidly detected
  - Allows rapid scanning of short reads with minimal differences to the reference
  - Significantly faster than BLAST, FASTA or Smith-Waterman for short high-quality reads against a reference

RNA Seq - Mapping

- Taking individual reads and mapping to the genome
  - Single End
  - Paired End
- Usually > 50nt enough to unambiguously map a read in human
- Reads are usually shattered into fragments and mapped independently
  - Allows detection of splice junctions

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, STAR

Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

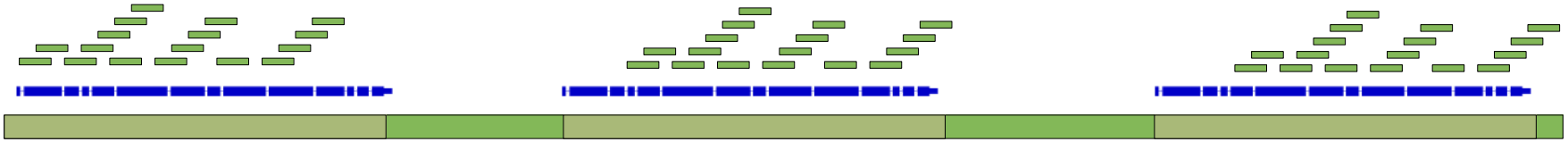
Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, STAR

Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, STAR

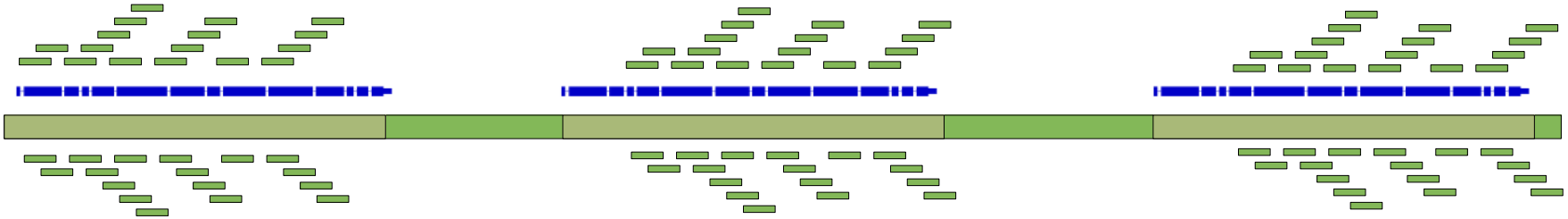


Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, STAR

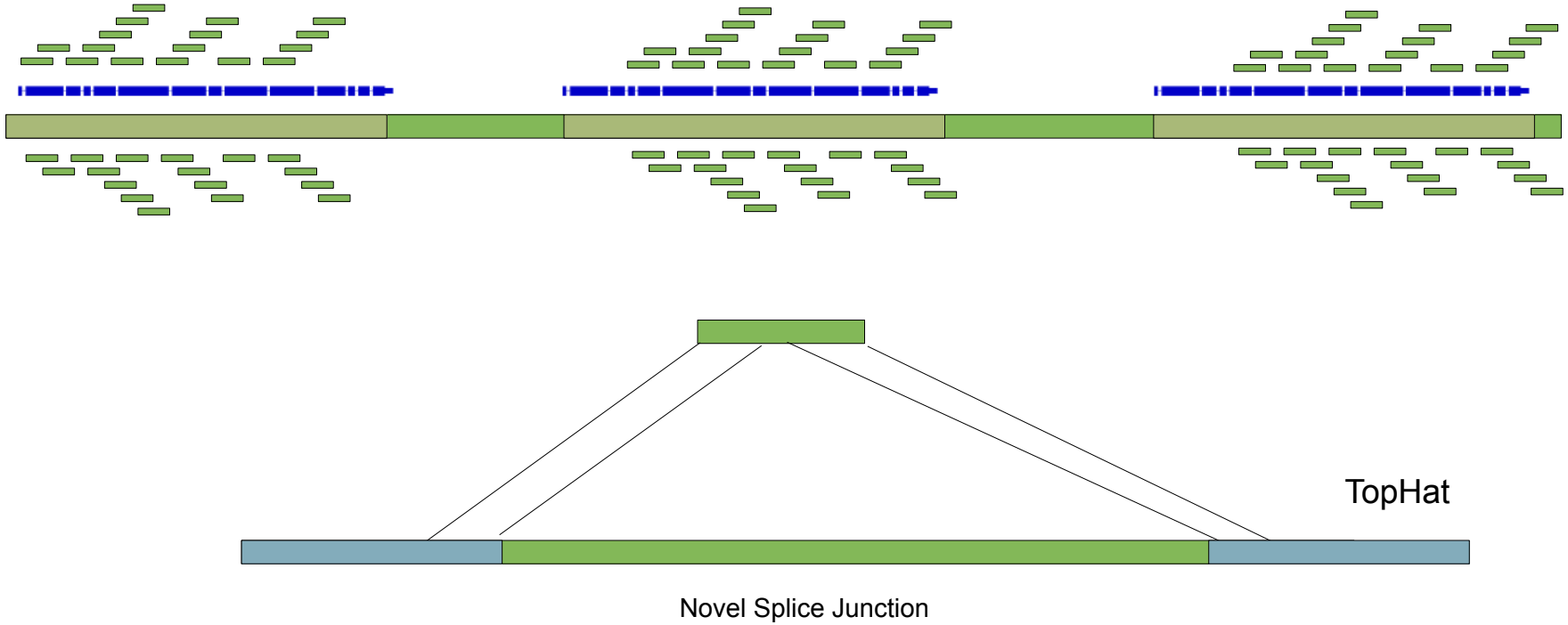


Results in count data and FPKMs

mRNA Seq: Analysis Types

Reference Transcriptome

Call expression levels of known genes, predict splice variants
e.g. HiSat2, TopHat2, STAR



Results in count data and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information

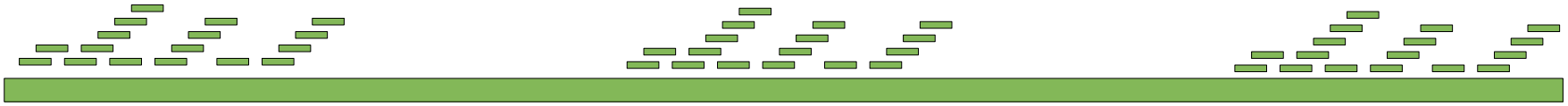
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



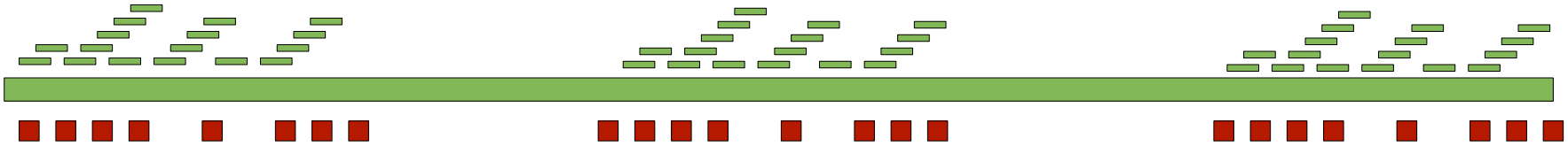
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



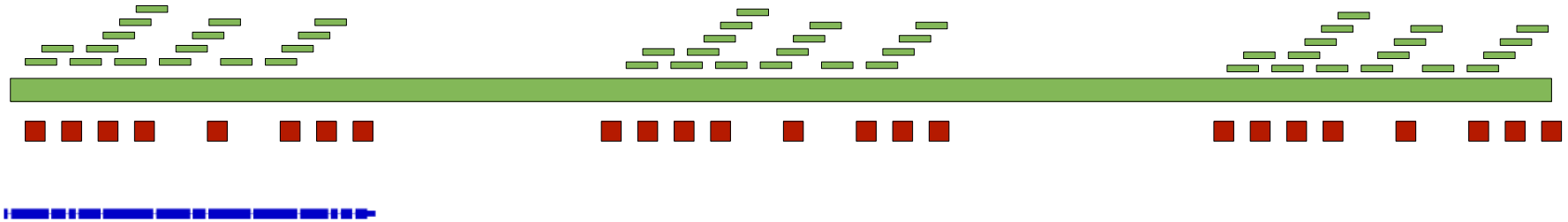
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



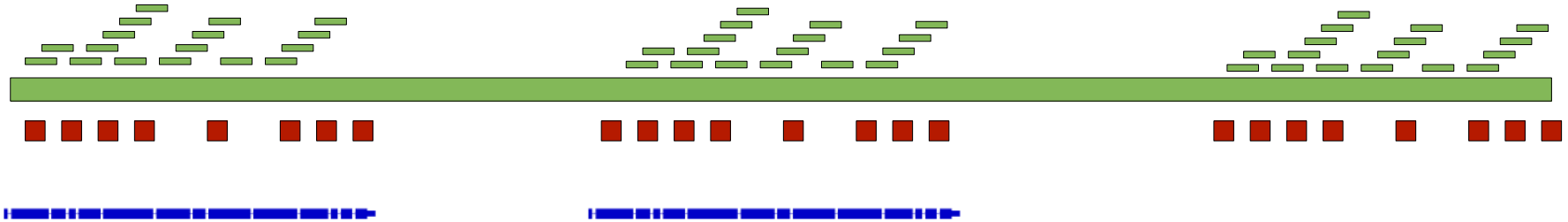
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



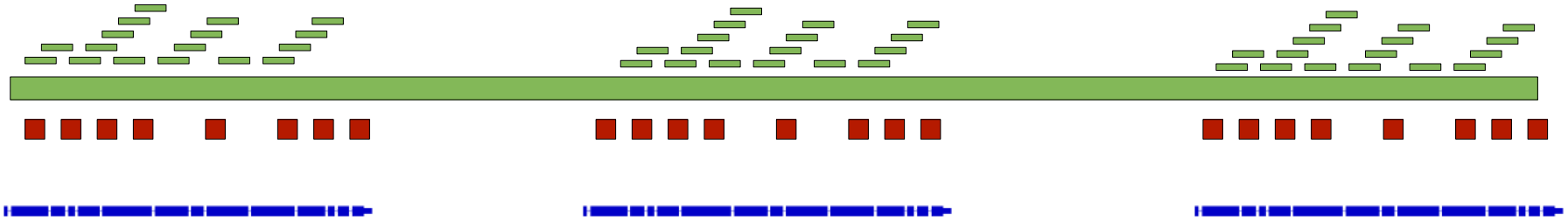
Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*de novo*)

Use the genome to find splice-junctions, assemble transcripts using this information



Cufflinks, Scripture etc.

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome

Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome

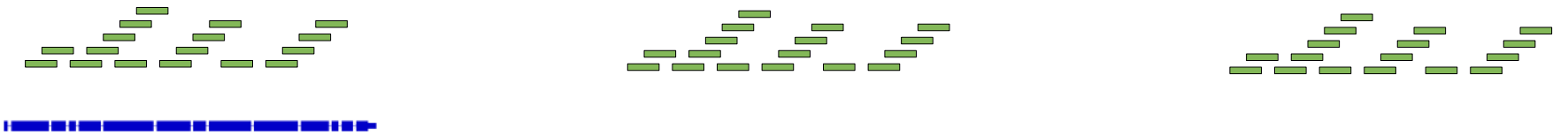
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



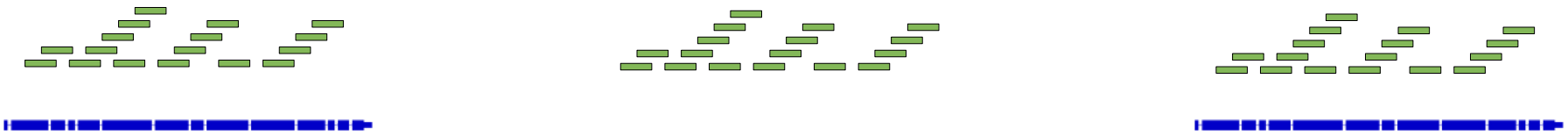
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



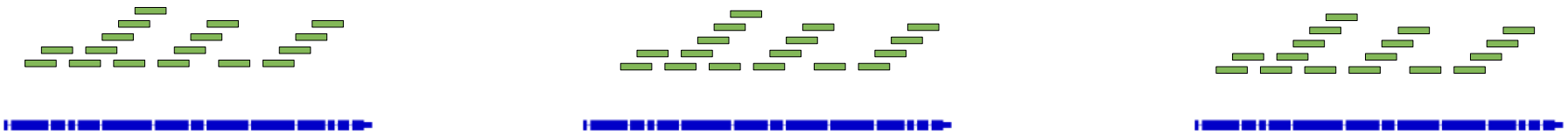
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



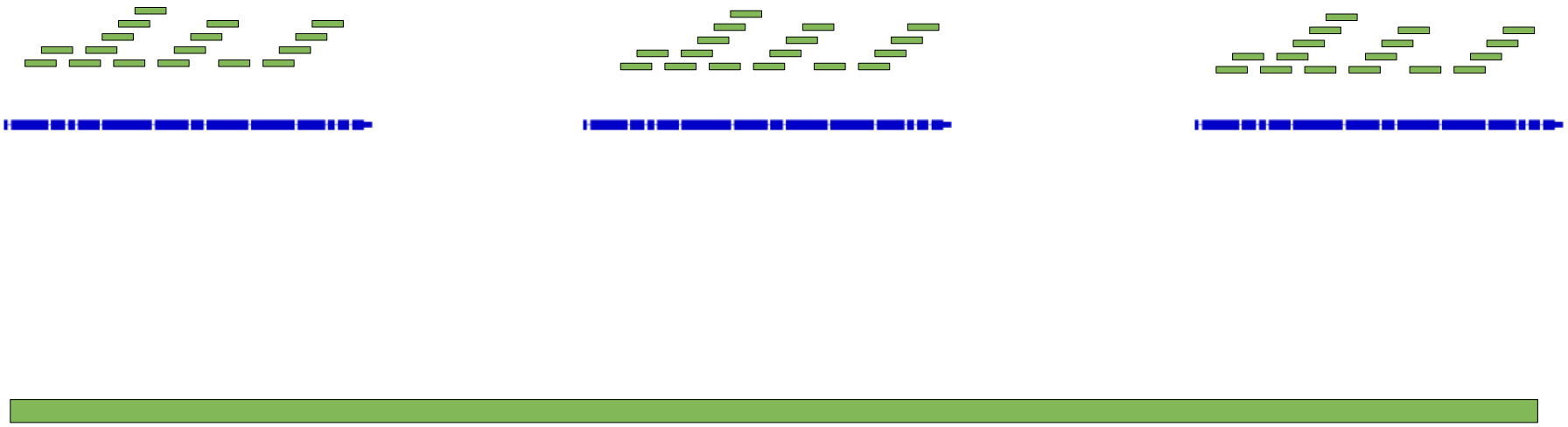
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



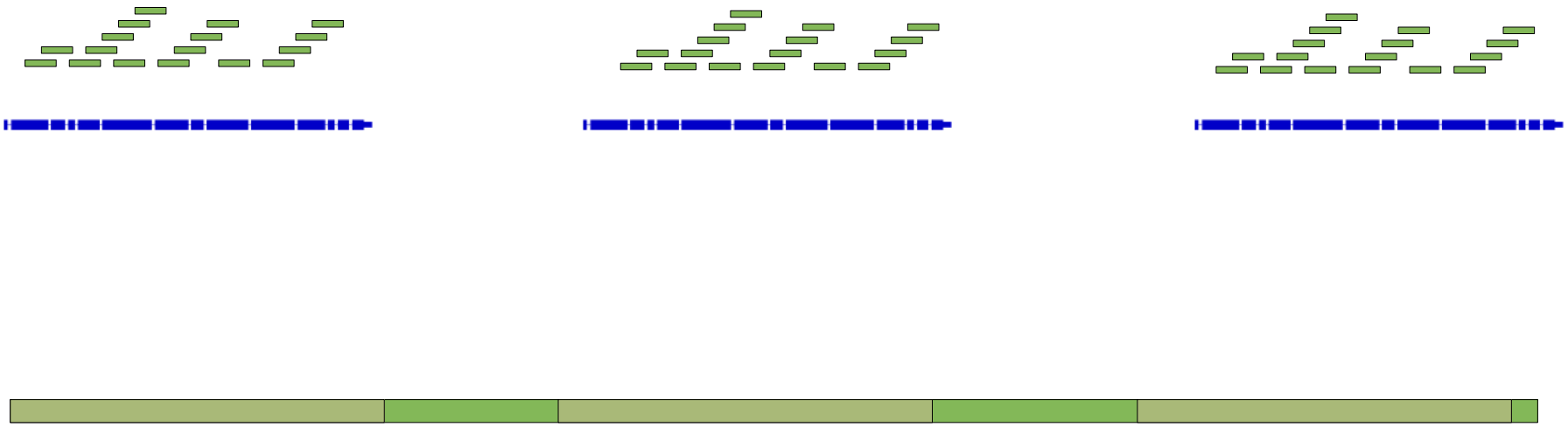
Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

mRNA Seq: Analysis Types

Transcriptome Assembly (*ab initio*)

Assemble the transcriptome without a reference genome



Trinity, TransAbyss, Velvet etc

Results in transcripts isoforms, counts and FPKMs

Data Types encountered

- **FASTQ** - Raw Read Data

- **SAM** - Text Alignment Format
- **BAM** - Binary Alignment Format
- **ebwt** - Bowtie Genome Index File

- **GFF** - Gene Models on a reference
- **GTF** - Gene Models on a reference
- **BED** - Gene Models on a reference

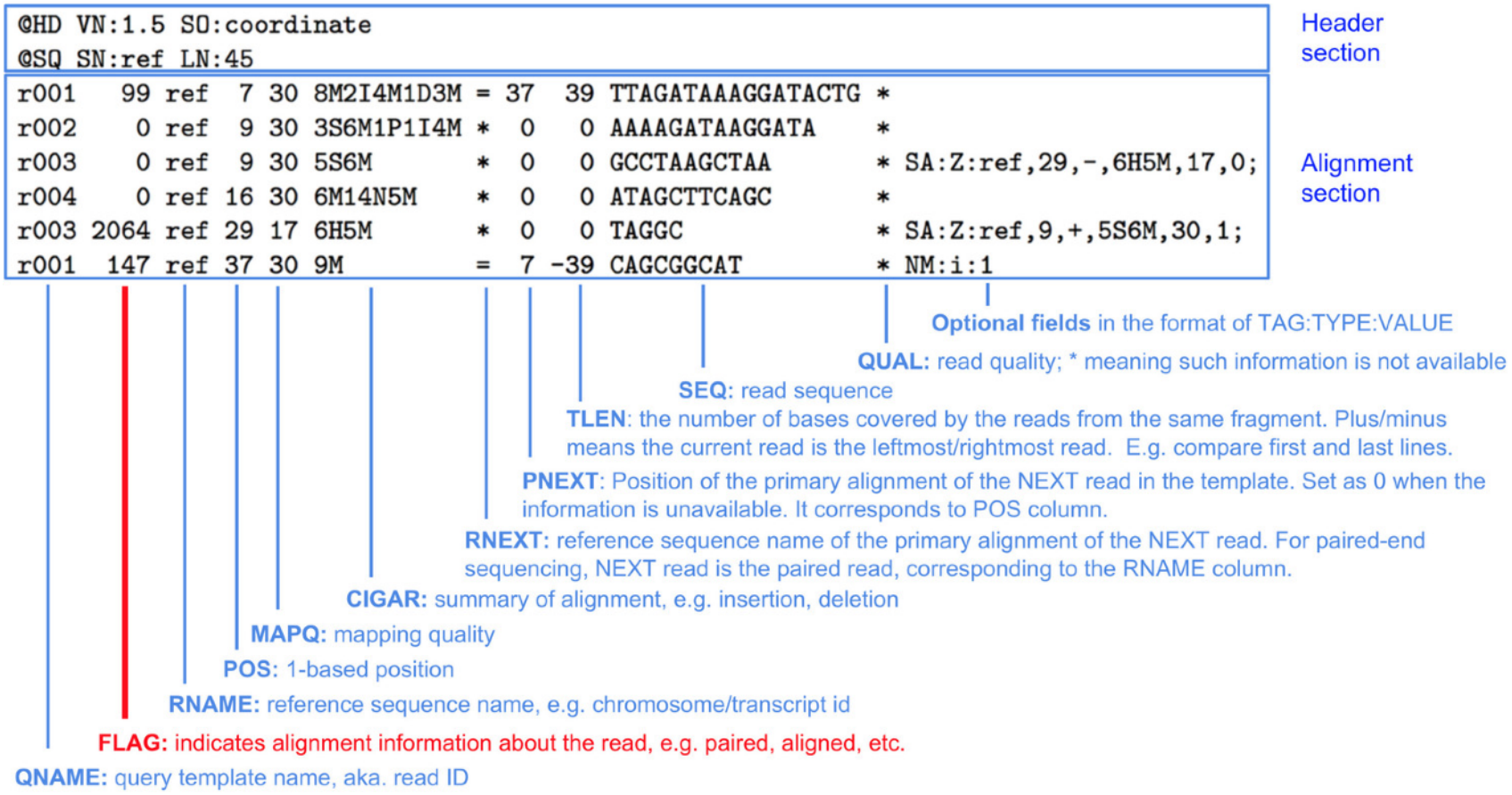
- **TDF** - Coverage File (IGV)

- **SAMTOOLS** is a package for converting between different formats

SAM/BAM Format: e.g. SAMTOOLS

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

SAM/BAM Format: e.g. SAMTOOLS



RNA Seq - Reference Data

Reference Genome = DNA Sequence
Reference Transcriptome = Location and Annotation of each gene/exon/transcript

Results in count data and FPKMs

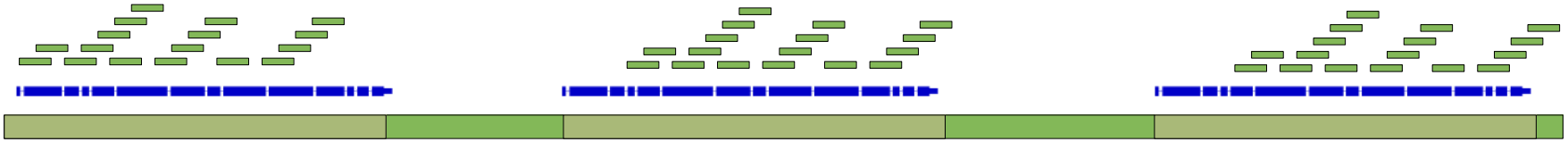
RNA Seq - Reference Data

Reference Genome = DNA Sequence
Reference Transcriptome = Location and Annotation of each gene/exon/transcript

Results in count data and FPKMs

RNA Seq - Reference Data

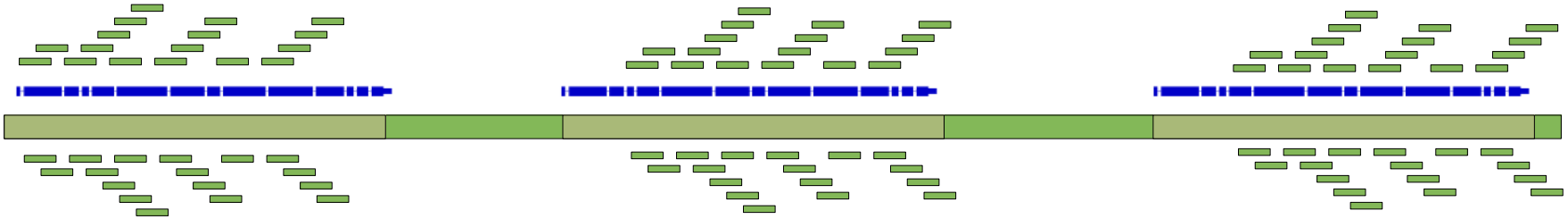
Reference Genome = DNA Sequence
Reference Transcriptome = Location and Annotation of each gene/exon/transcript



Results in count data and FPKMs

RNA Seq - Reference Data

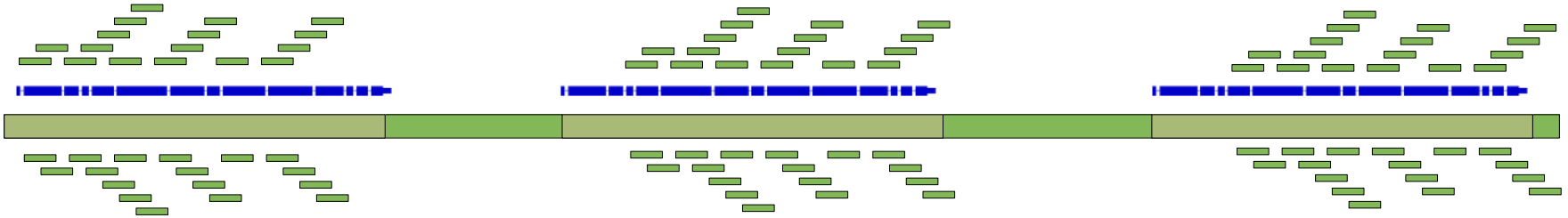
Reference Genome = DNA Sequence
Reference Transcriptome = Location and Annotation of each gene/exon/transcript

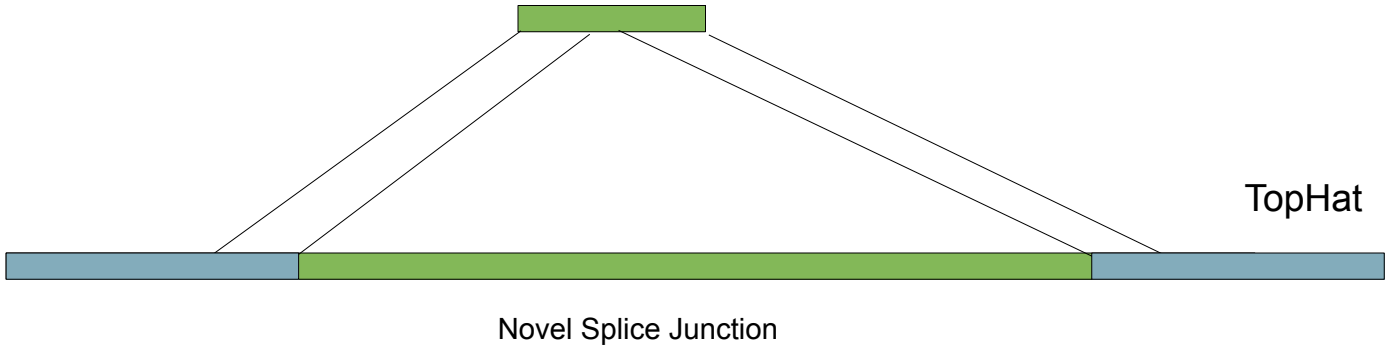


Results in count data and FPKMs

RNA Seq - Reference Data

Reference Genome = DNA Sequence
Reference Transcriptome = Location and Annotation of each gene/exon/transcript

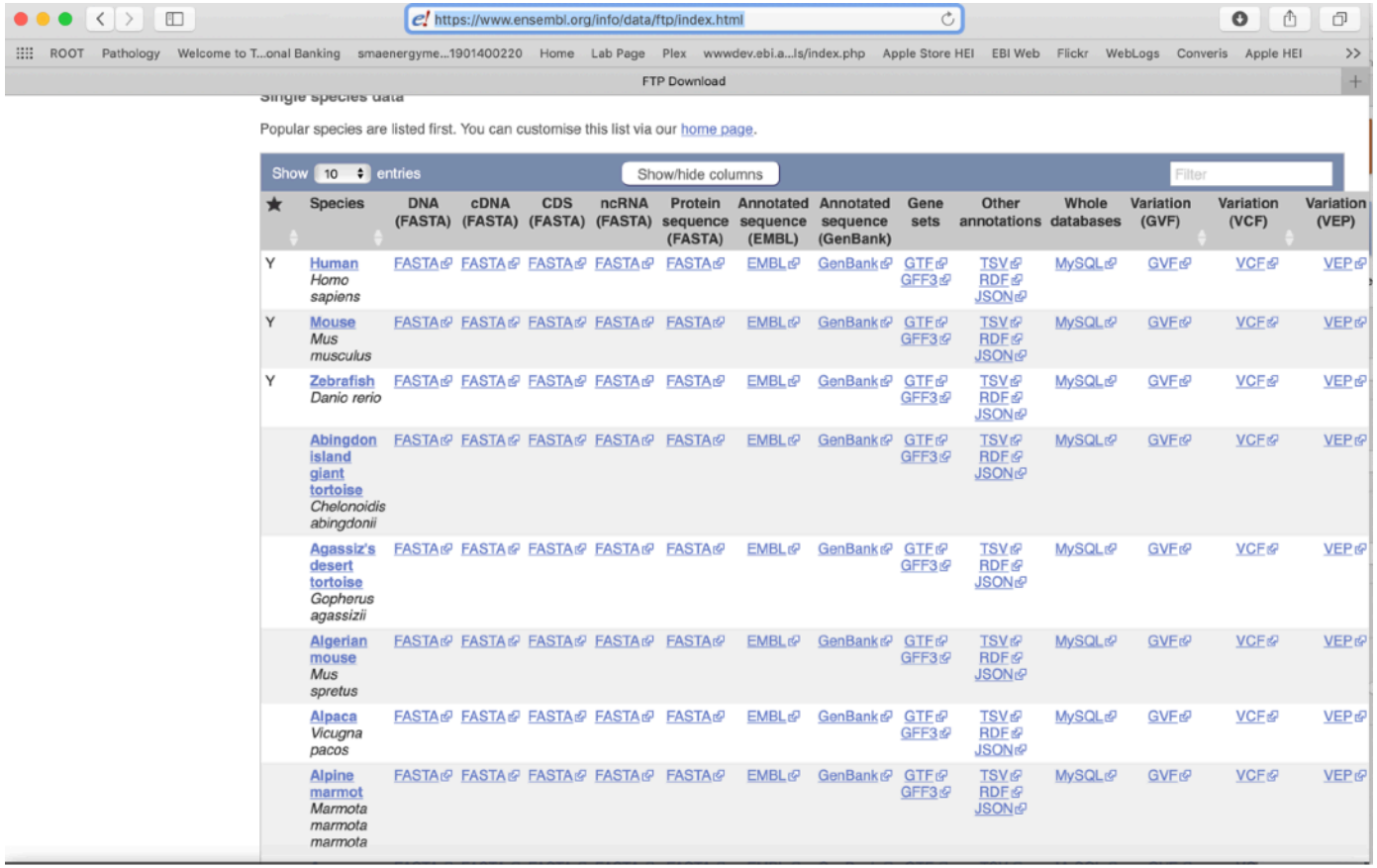




Results in count data and FPKMs

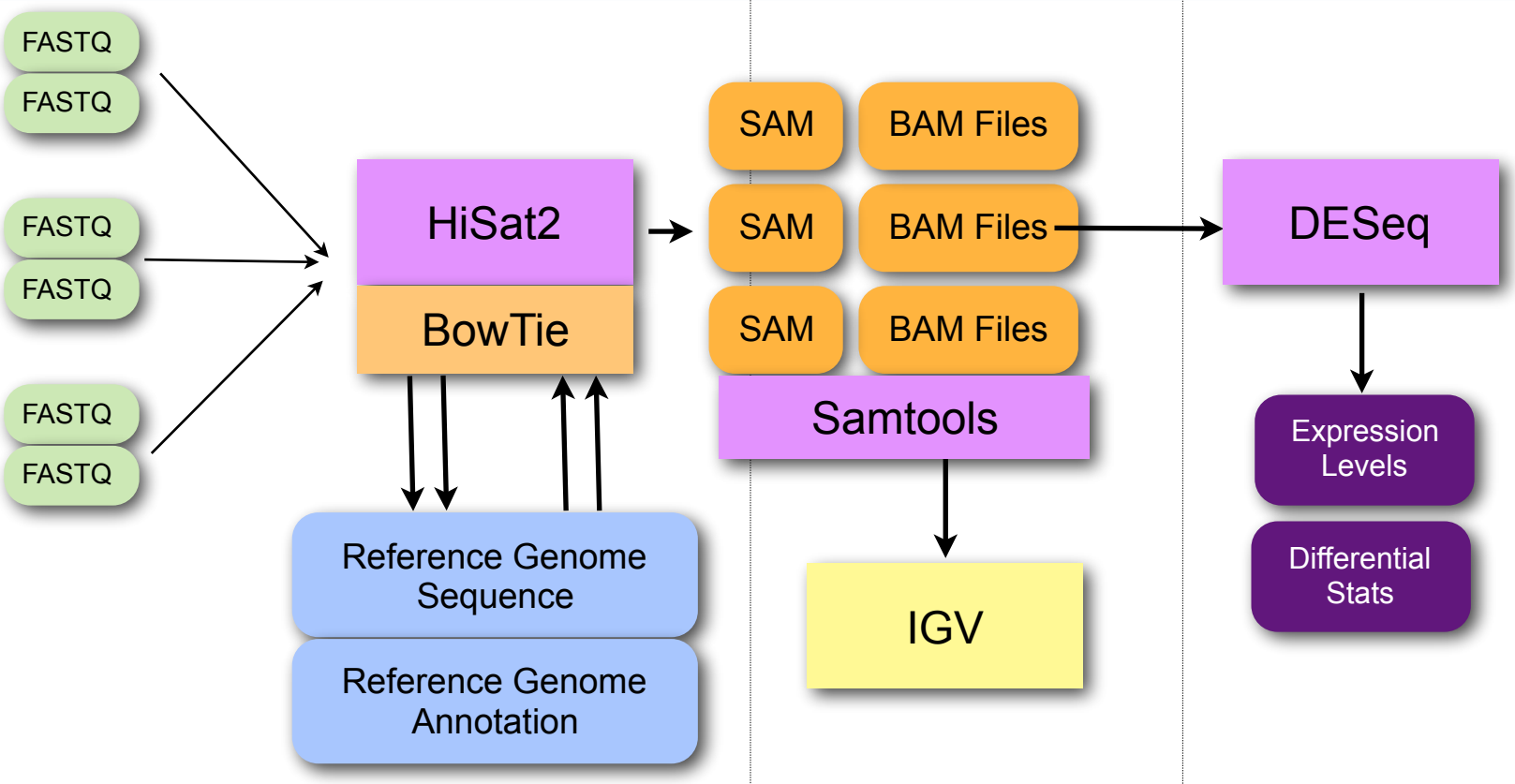
Where to get genome sequence and annotation ?

https://www.ensembl.org/info/data/ftp/index.html

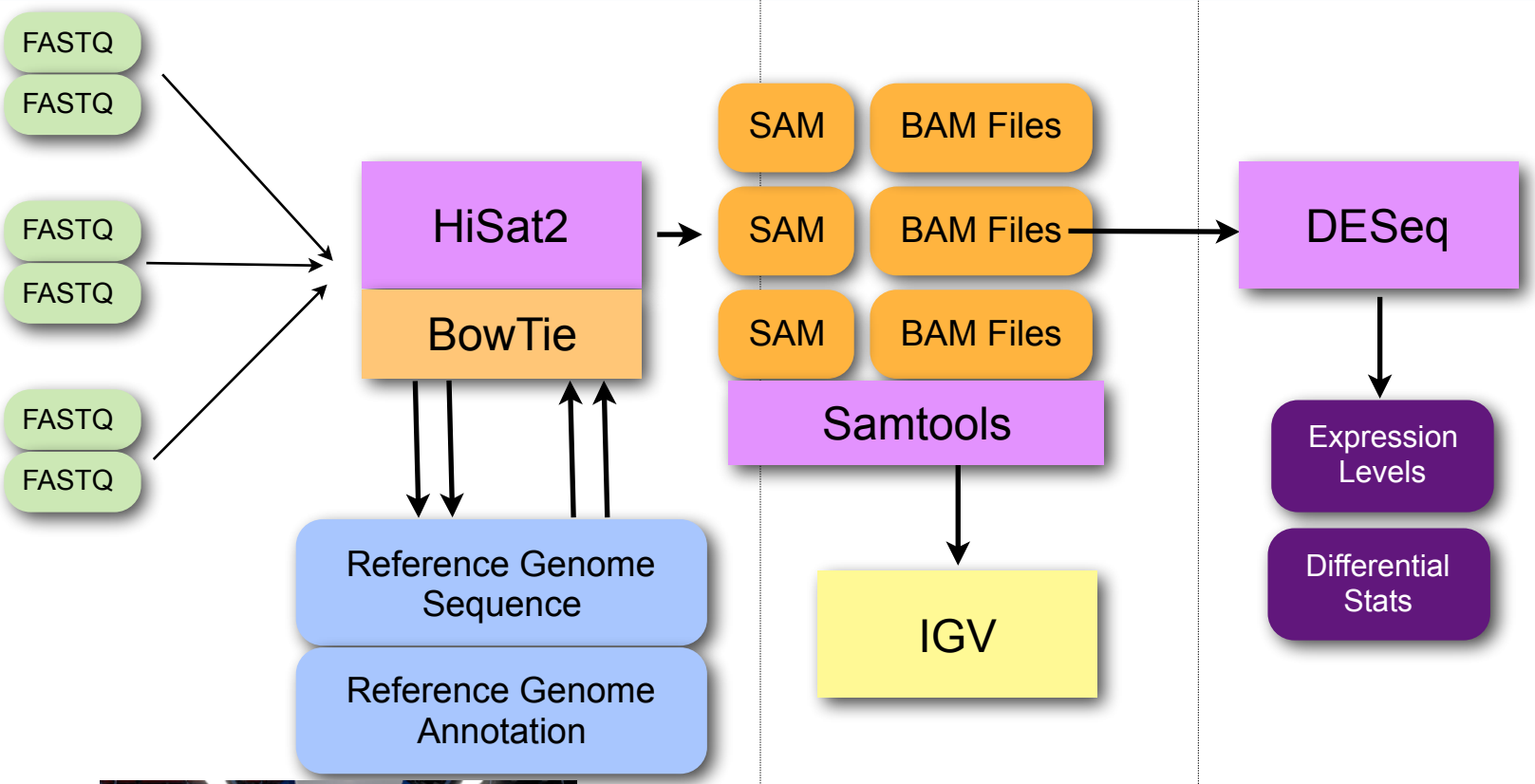


★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)
Y	Human *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Zebrafish *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
	Abingdon island giant tortoise *Chelonoidis abingdonii*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
	Agassiz's desert tortoise *Gopherus agassizii*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
	Algerian mouse *Mus spretus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
	Alpaca *Vicugna pacos*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
	Alpine marmot *Marmota marmota marmota*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP

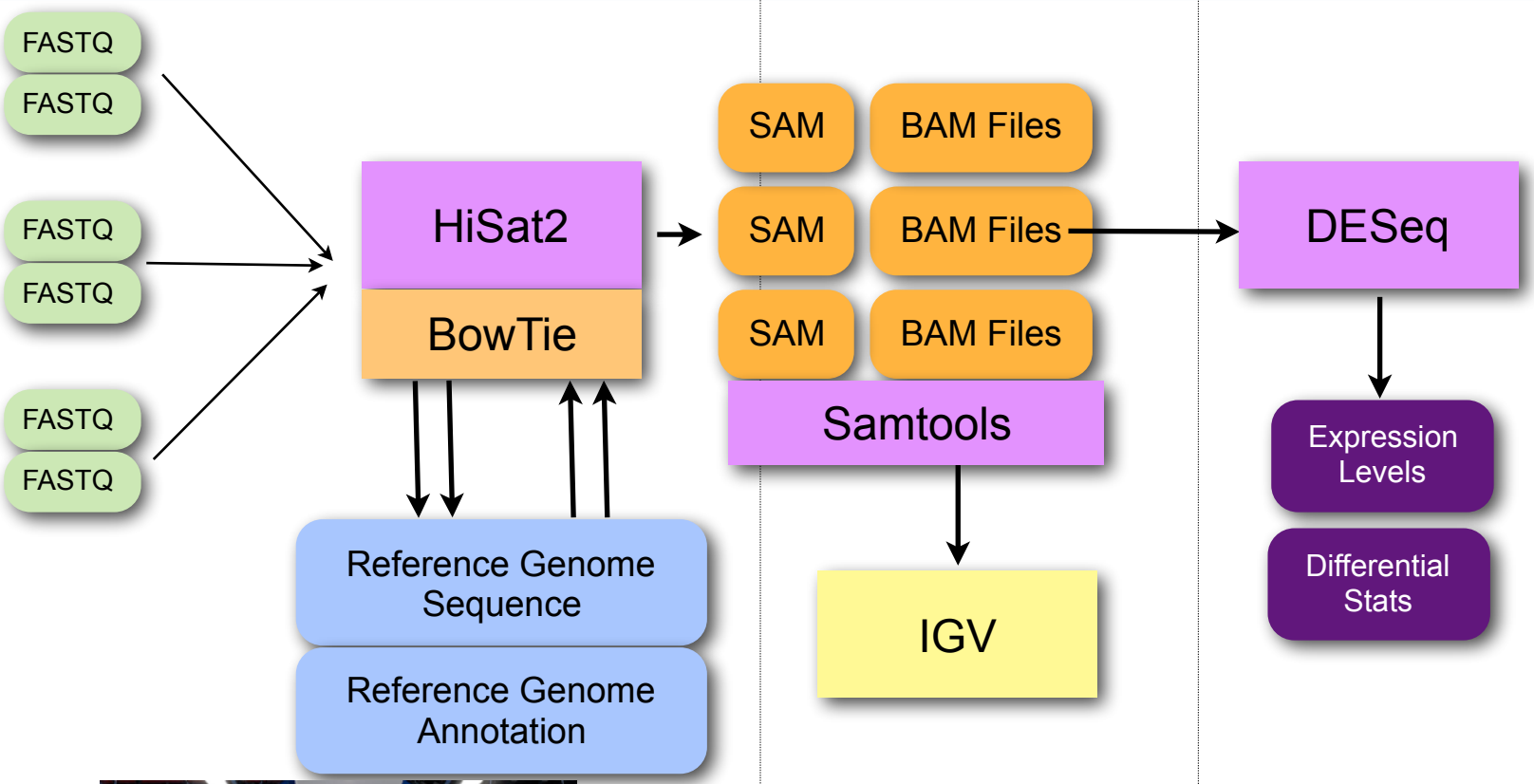
mRNA Seq Workflow



mRNA Seq Workflow

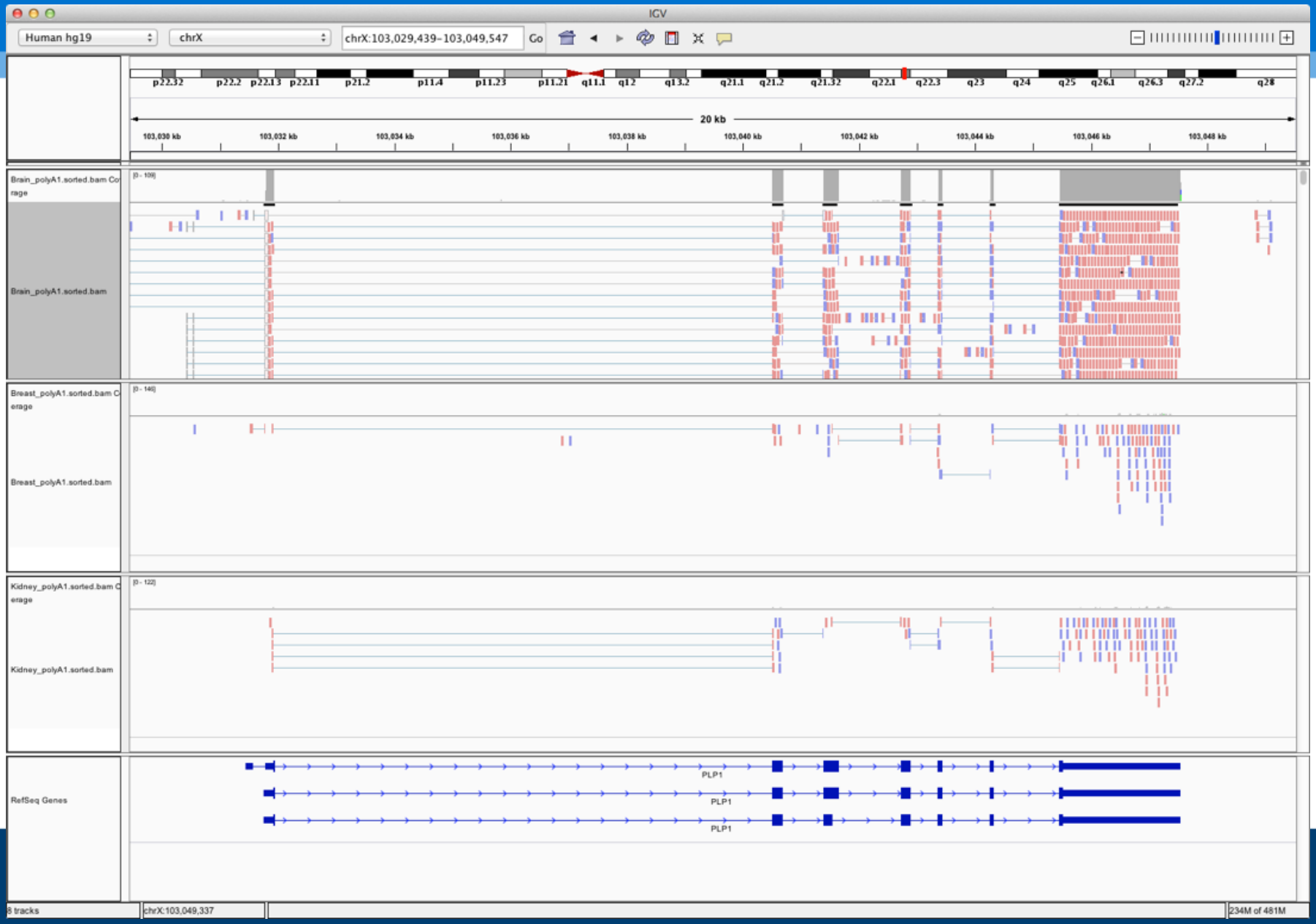


mRNA Seq Workflow



IGV

http://www.broadinstitute.org/igv/

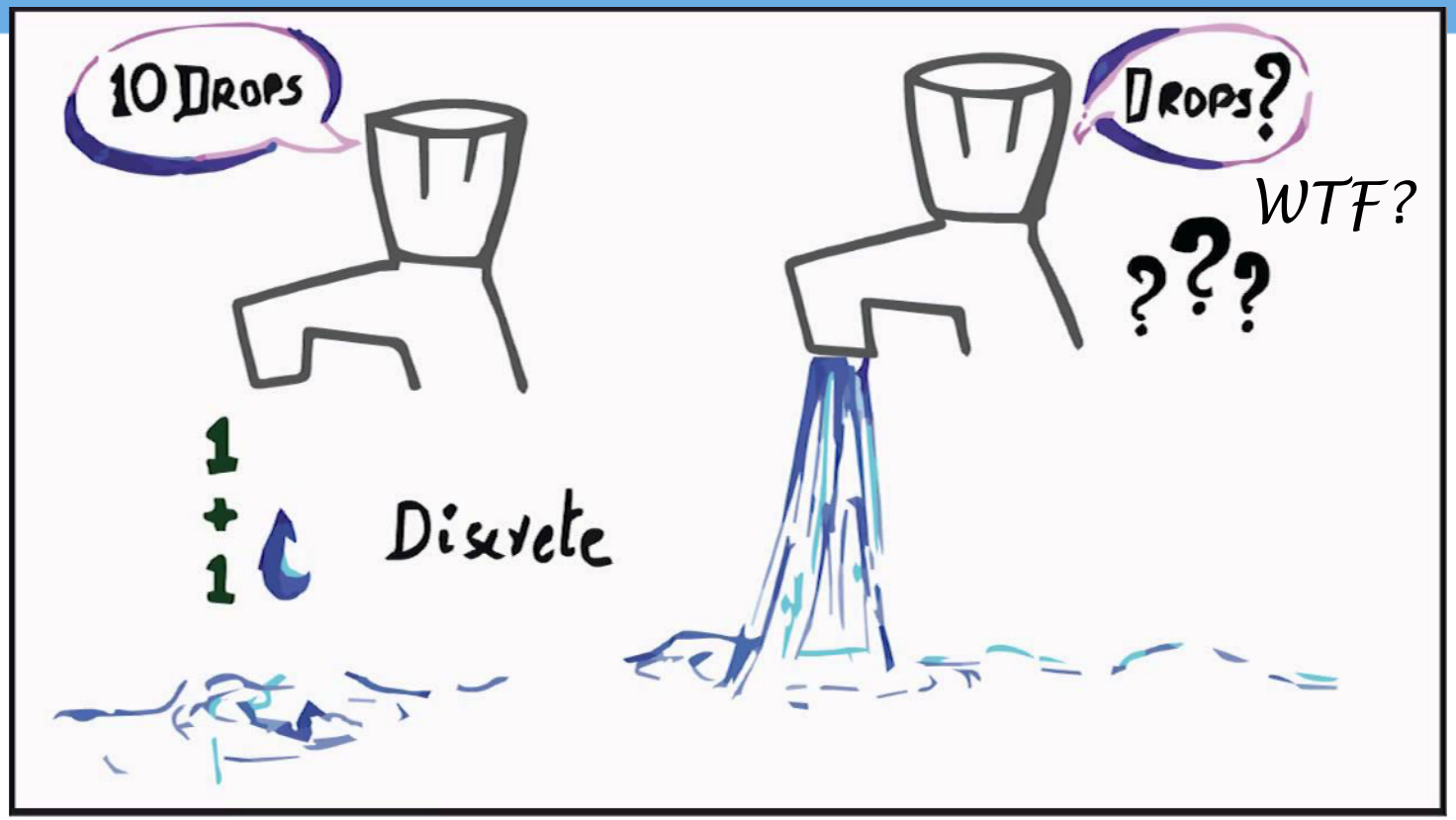


Data Normalisation and Statistics - RNASeq

Count Data is Different! - NGS Analysis

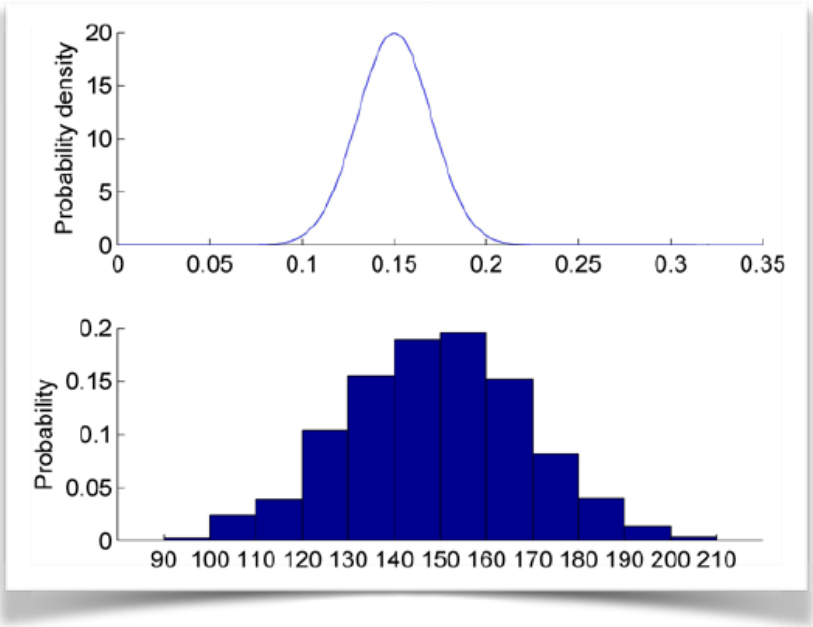
Count Data is Different! - NGS Analysis

Count Data is Different! - NGS Analysis



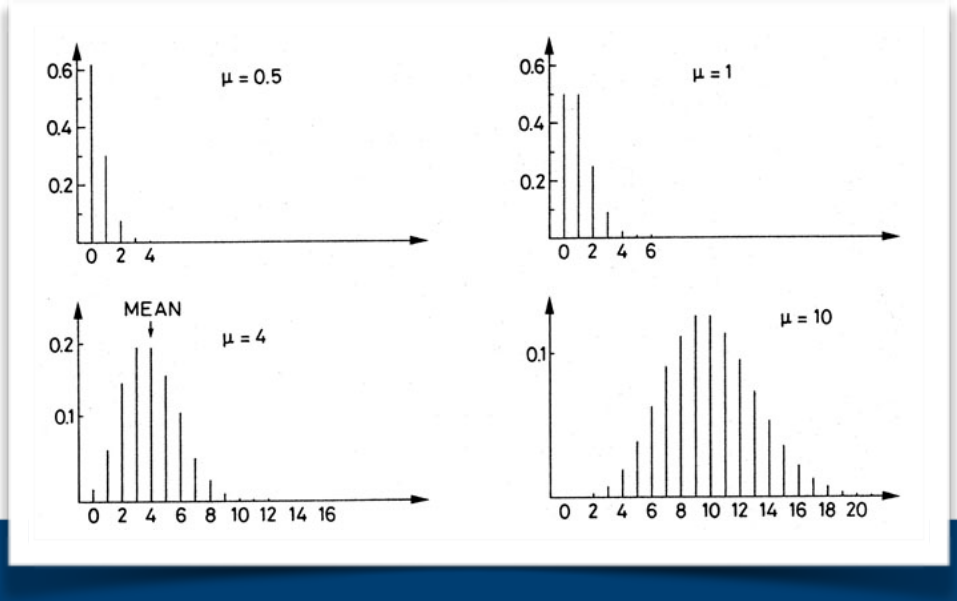
Sequencing Data Statistics

- **Sequencing data produces counts not measurements**
  - Forms a discrete distribution
  - Very different from continuous distributions (e.g. intensities from a microarray probe)
  - Positive values
  - Skewed
  - Heteroscedastic
  - Massive dynamic range
  - Large differences sample to sample



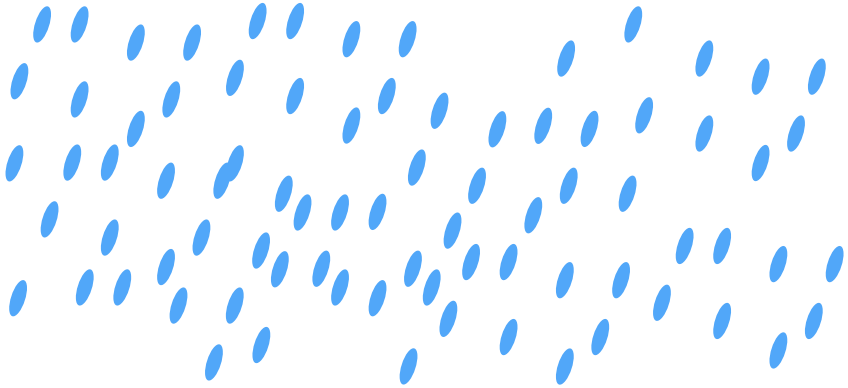
Discrete Count Data - How to model it

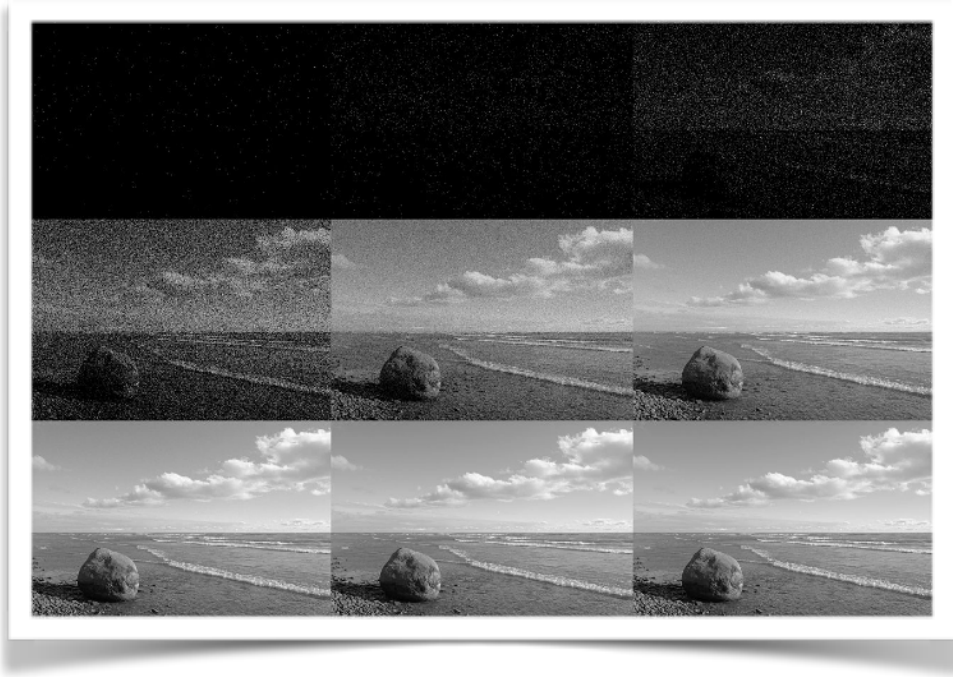
- **The Poisson Distribution**
- Example: A short, light rain shower with ***r*** drops per meter^2.
- What is the probability to find ***k*** drops on a paving stone of size 1 m^2?
- For poisson, the variance should equal the mean



Discrete Count Data

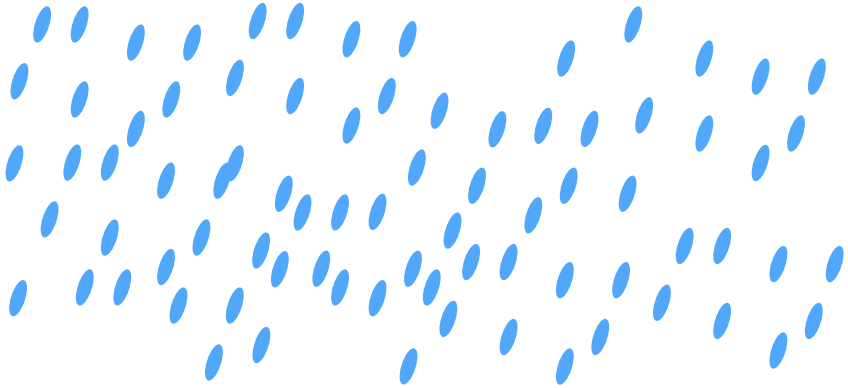
- Shot Noise

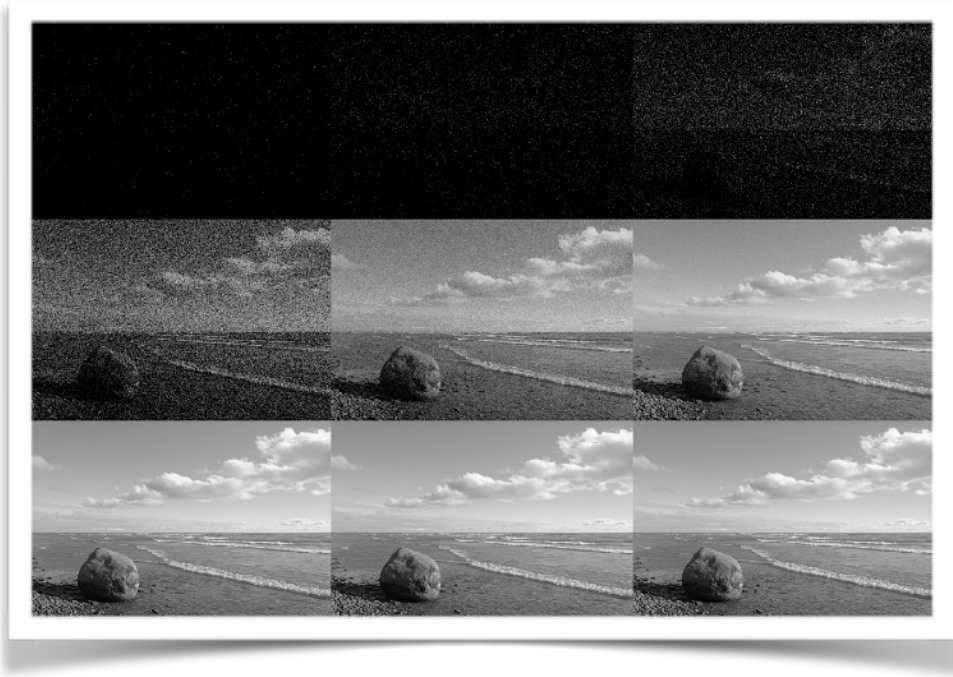




Discrete Count Data

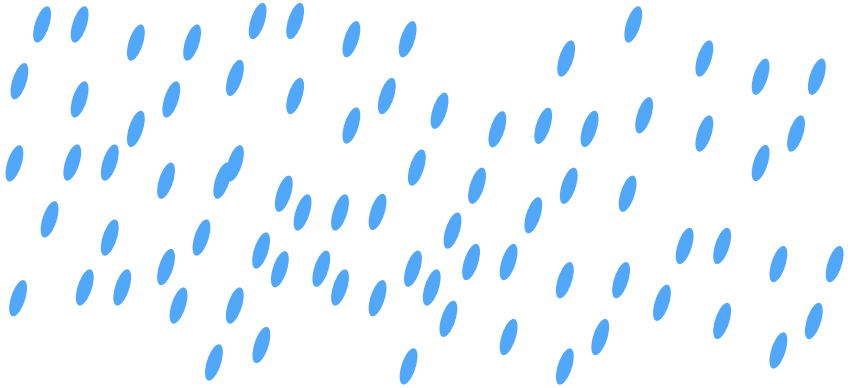
- Shot Noise

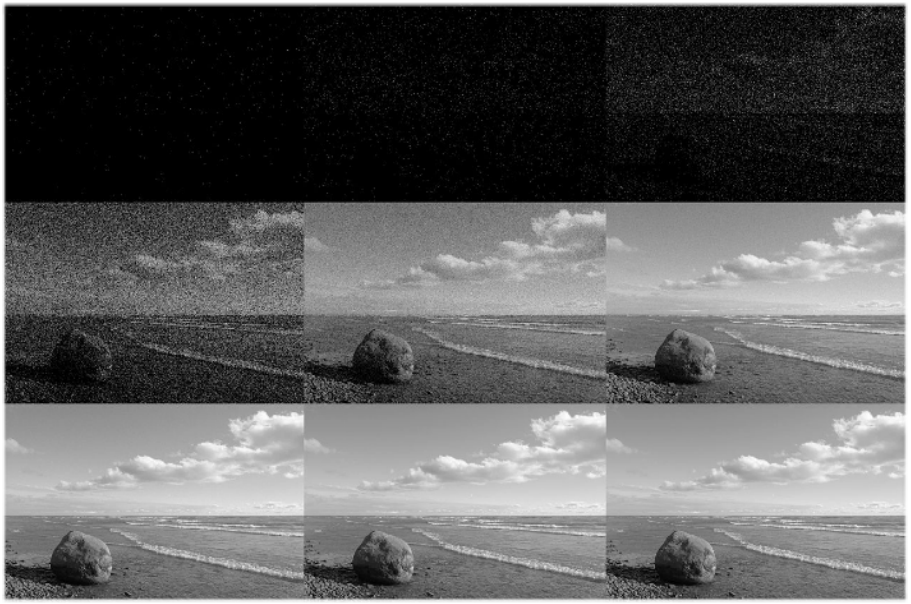




Discrete Count Data

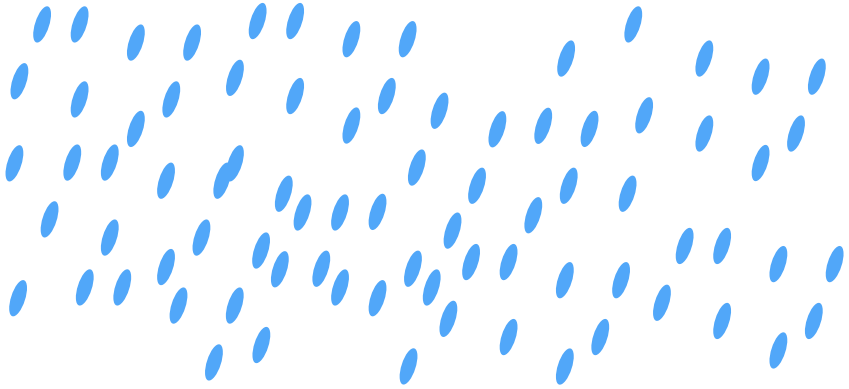
- Shot Noise

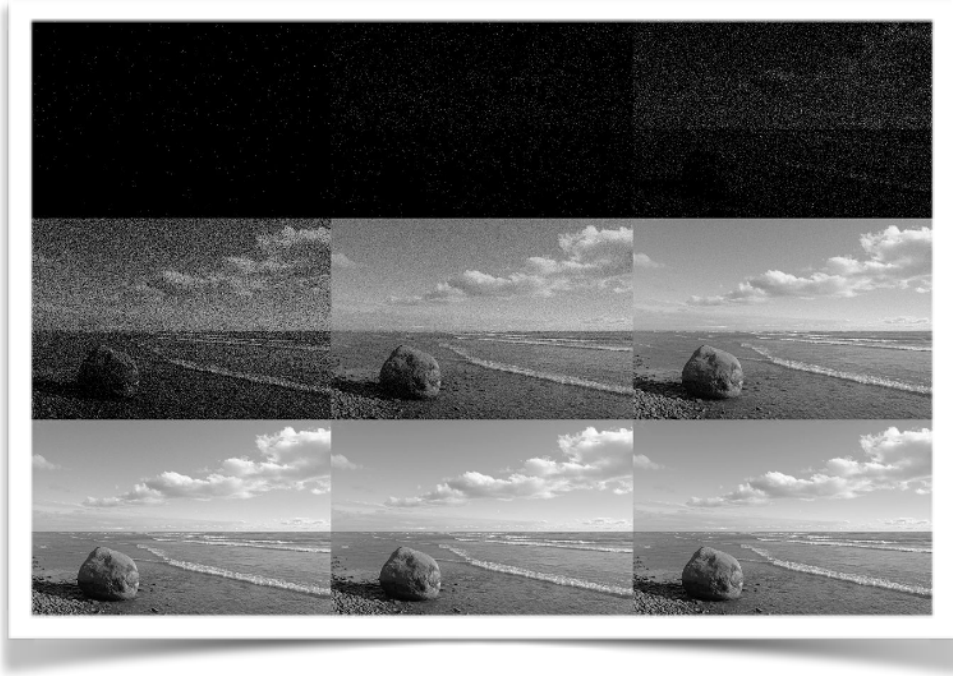




Discrete Count Data

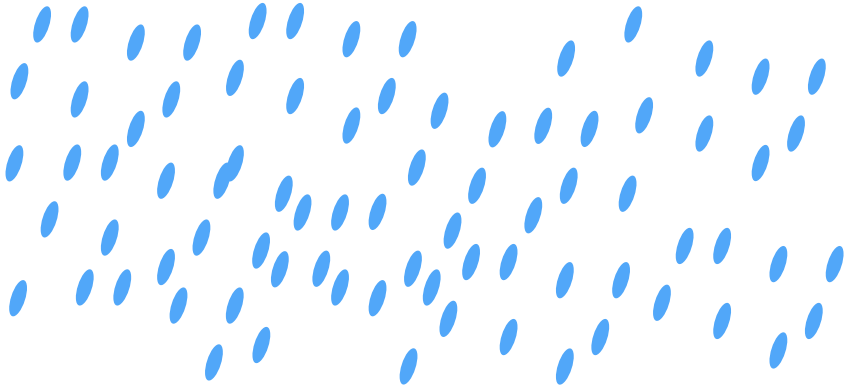
- Shot Noise

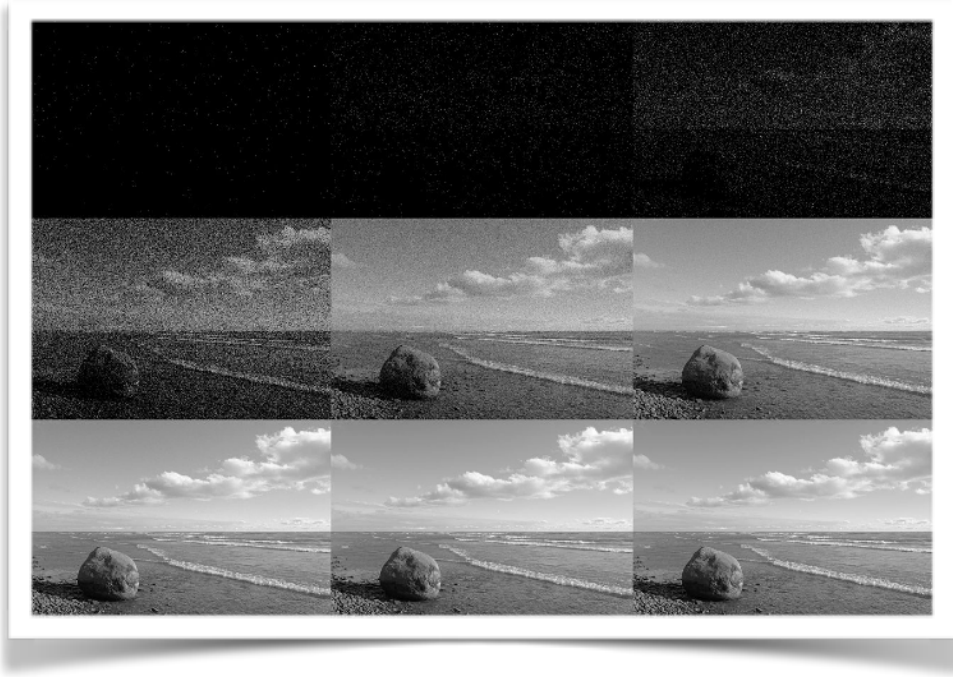




Discrete Count Data

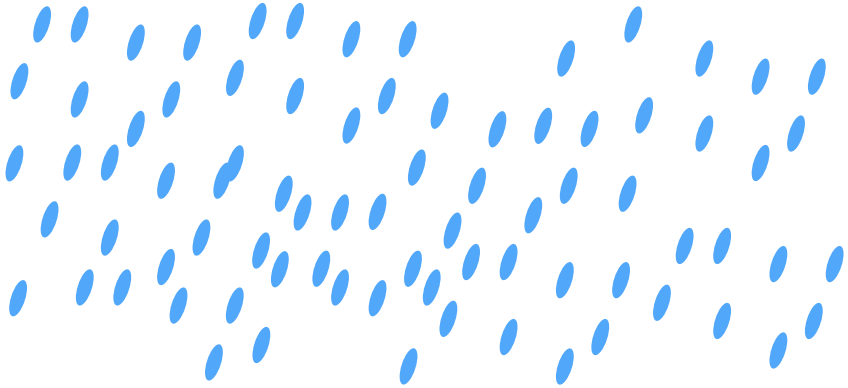
- Shot Noise

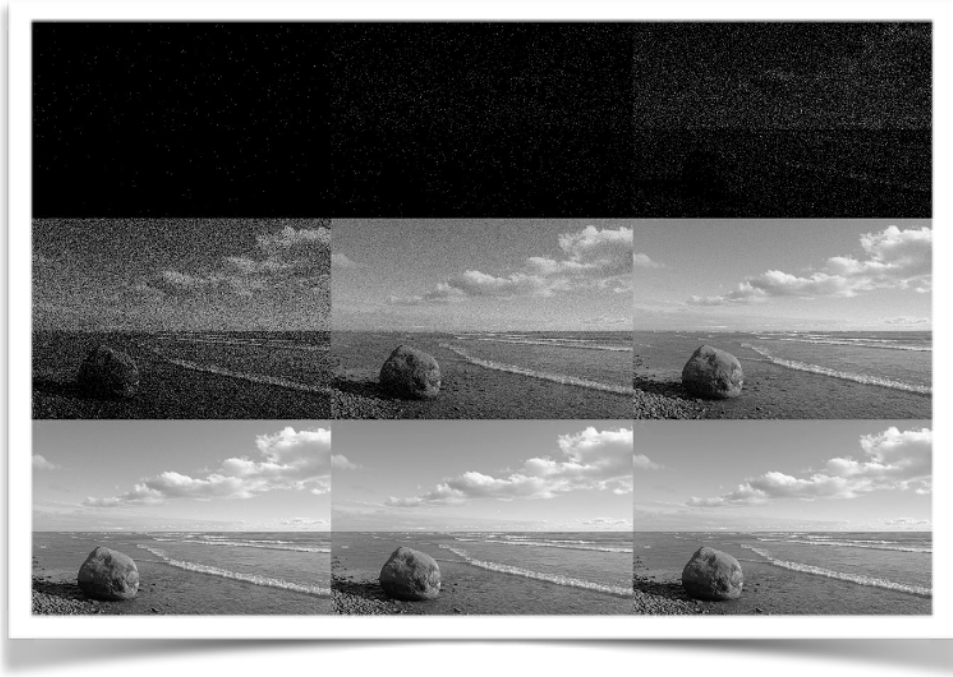




Discrete Count Data

- Shot Noise

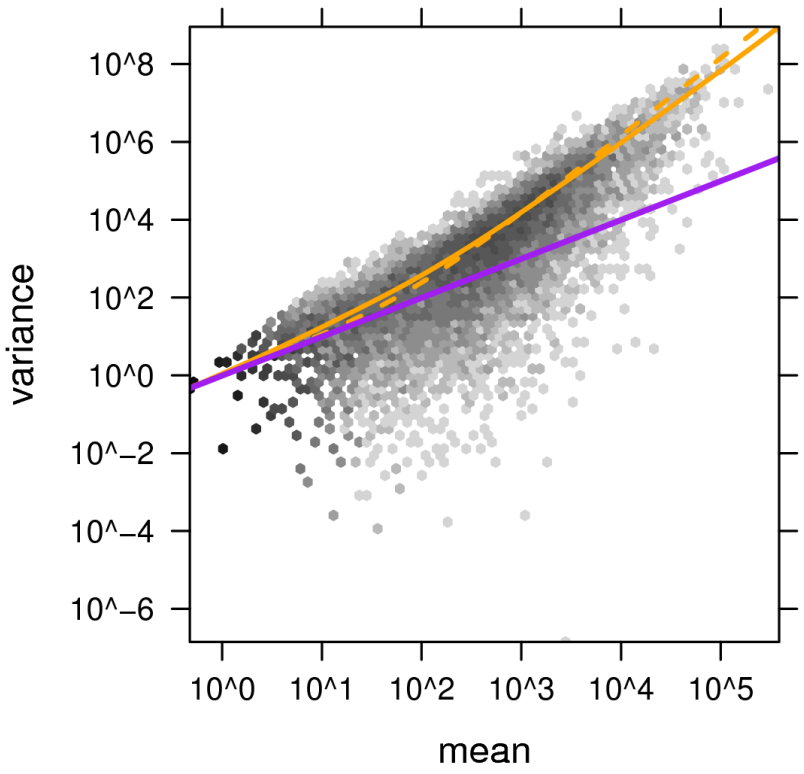


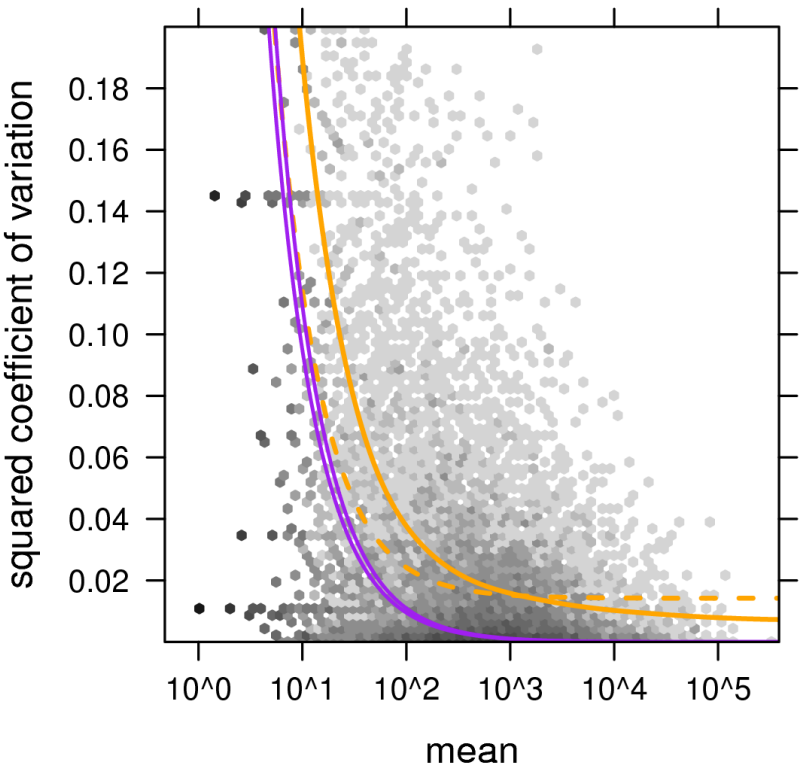


It's Raining Genes… Hallelujah!

- Shot Noise is very different for individual genes

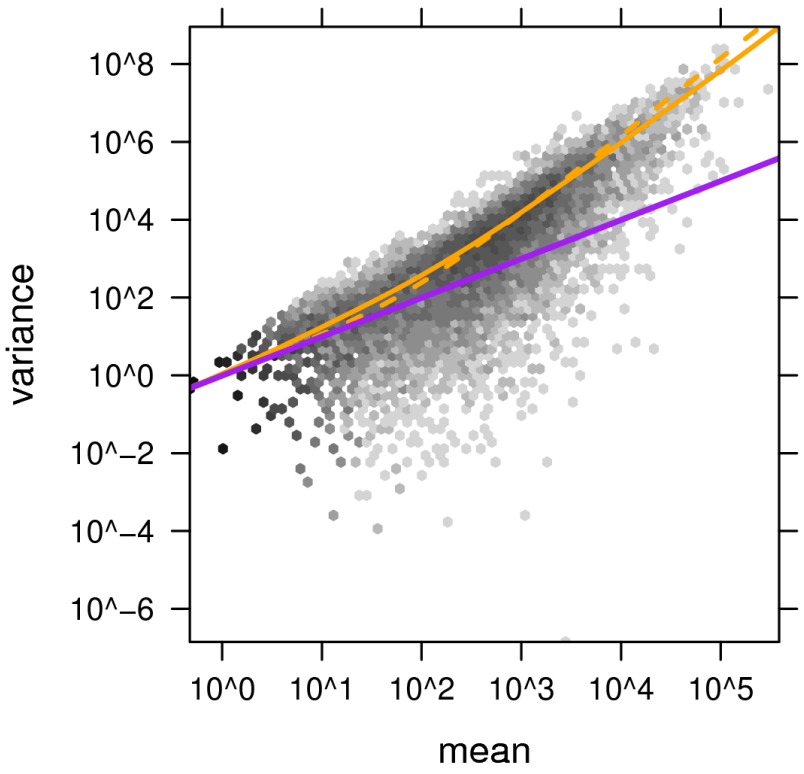
Sequencing Data - Variance - 2 replicates

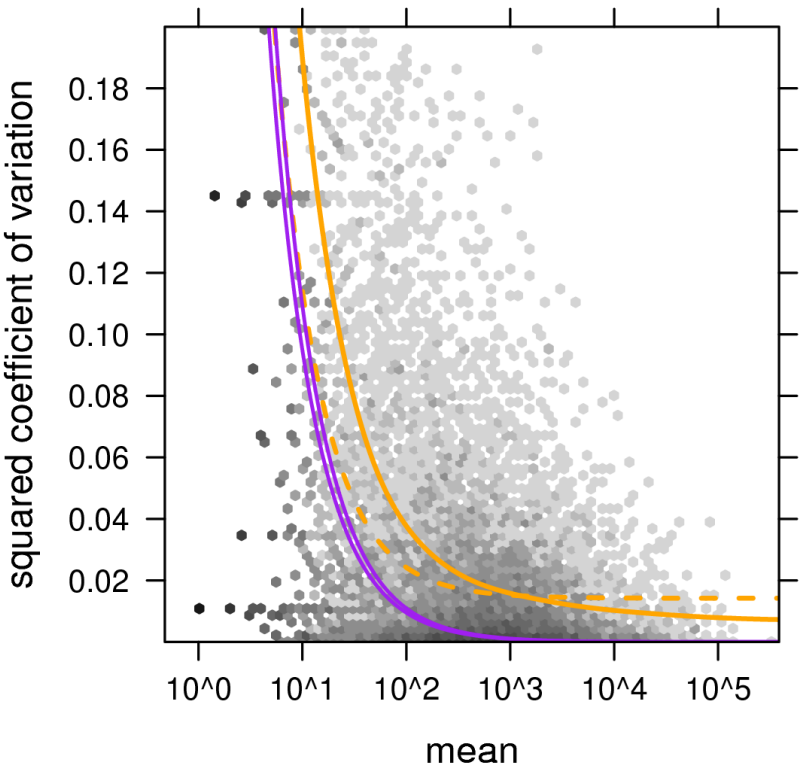




Poisson	$v = \mu$	———
Poisson + constant CV	$v = \mu + \alpha \mu^2$	- - - - -
Poisson + local regression	$v = \mu + f(\mu^2)$	———

Sequencing Data - Variance - 2 replicates



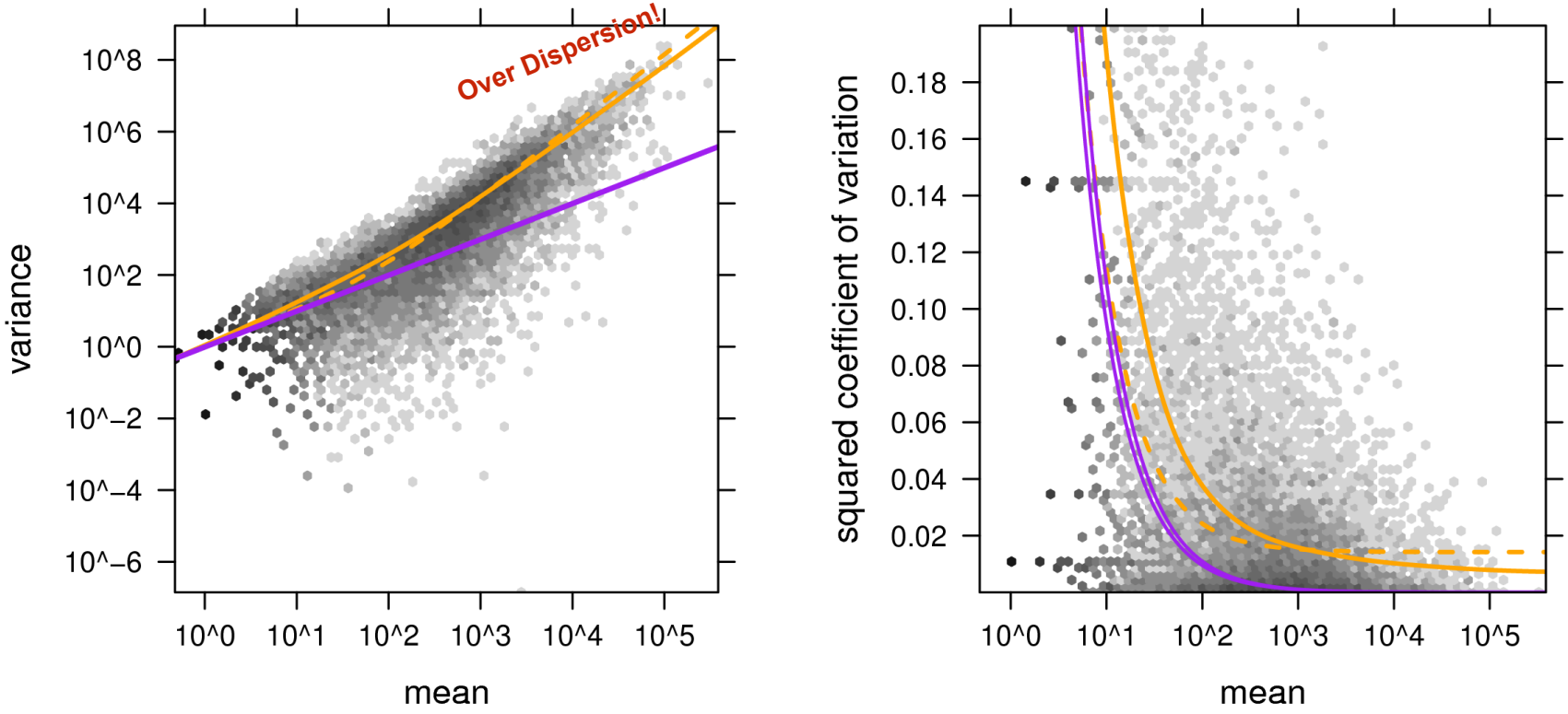


Poisson	$v = \mu$
Poisson + constant CV	$v = \mu + \alpha \mu^2$
Poisson + local regression	$v = \mu + f(\mu^2)$

Poisson Invalidated for genes with high expression as their variance exceeds the model

The variance should be approximately equal to the mean

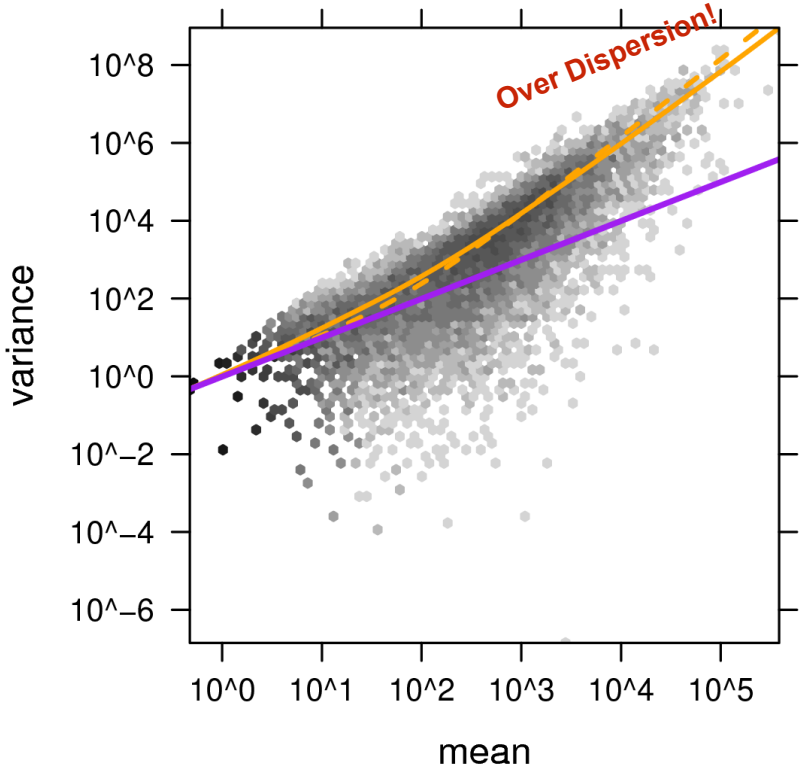
Sequencing Data - Variance - 2 replicates

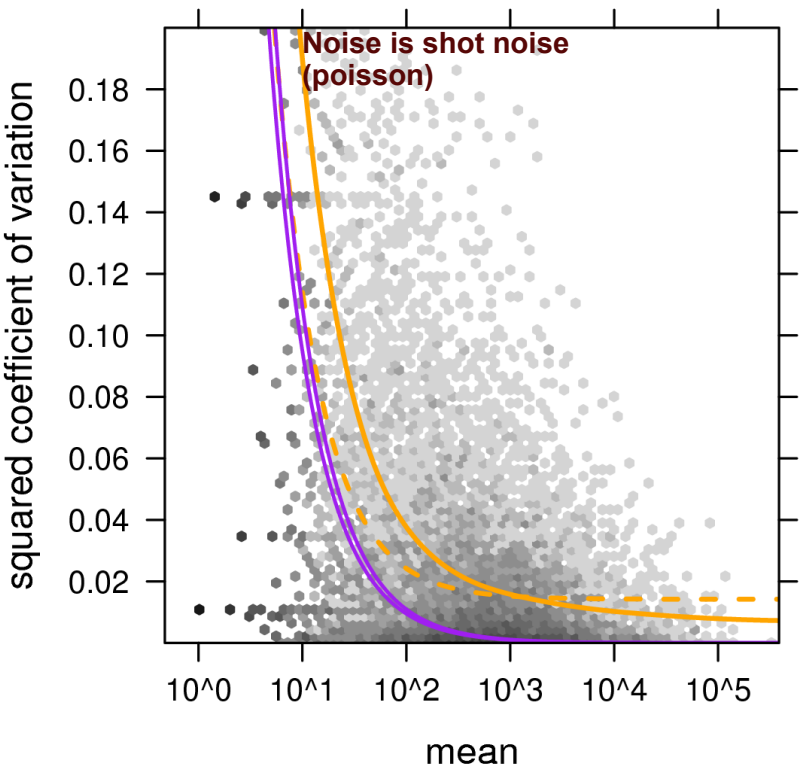


Poisson Invalidated for genes with high expression as their variance exceeds the model

The variance should be approximately equal to the mean

Sequencing Data - Variance - 2 replicates



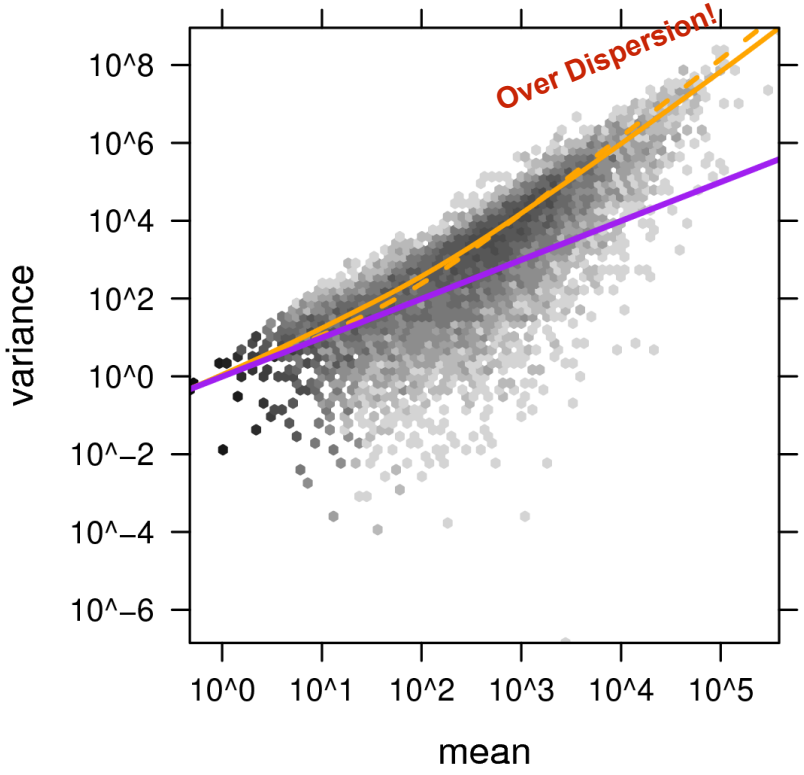


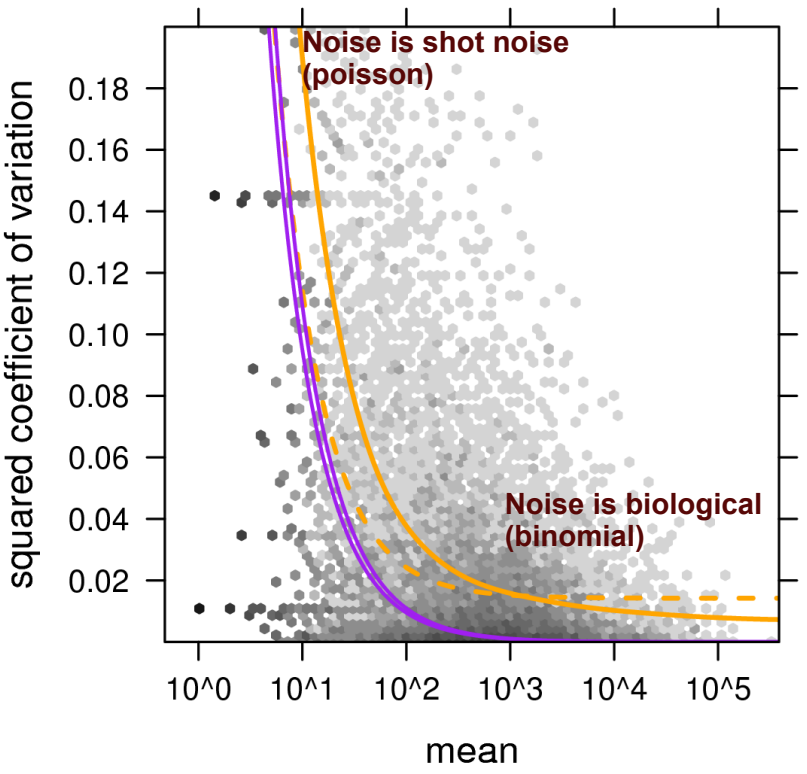
Poisson	$v = \mu$
Poisson + constant CV	$v = \mu + \alpha \mu^2$
Poisson + local regression	$v = \mu + f(\mu^2)$

Poisson Invalidated for genes with high expression as their variance exceeds the model

The variance should be approximately equal to the mean

Sequencing Data - Variance - 2 replicates





Poisson	$v = \mu$
Poisson + constant CV	$v = \mu + \alpha \mu^2$
Poisson + local regression	$v = \mu + f(\mu^2)$

Poisson Invalidated for genes with high expression as their variance exceeds the model

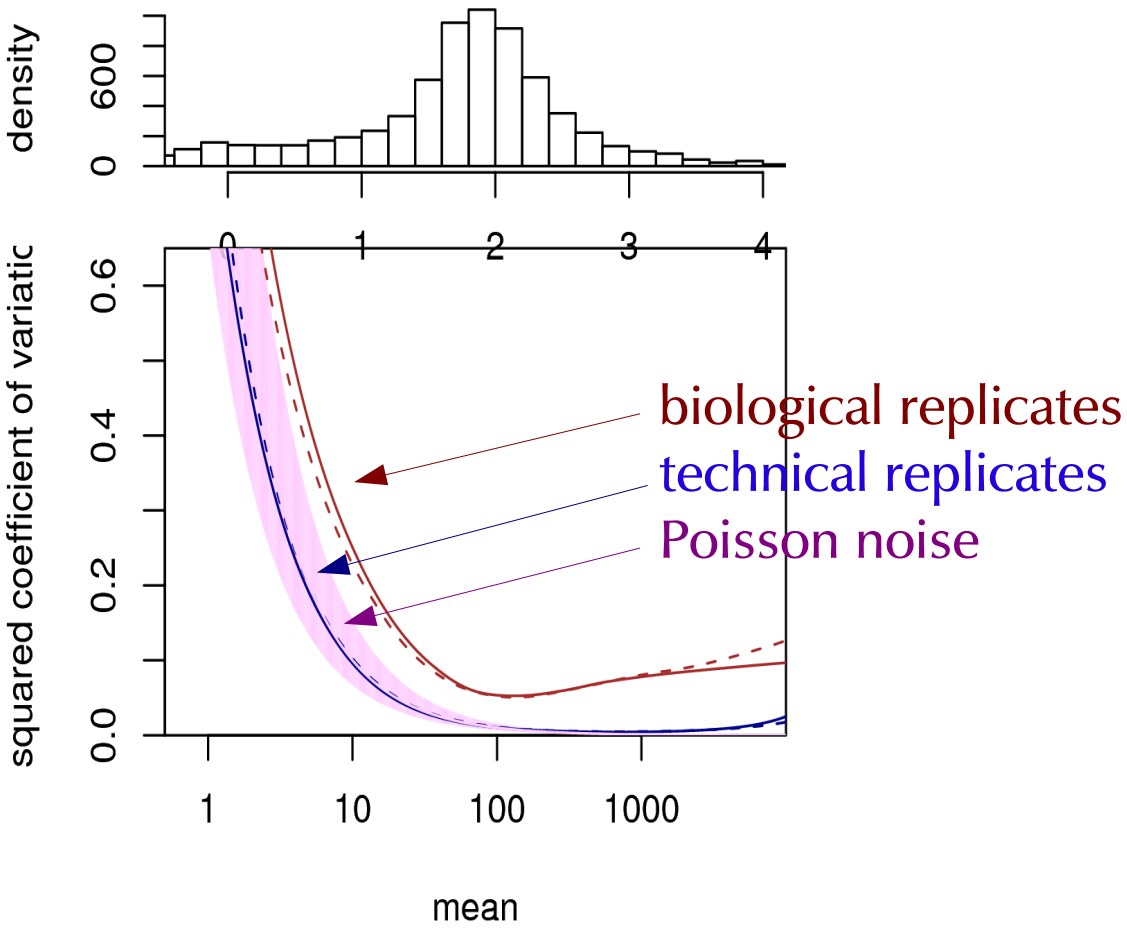
The variance should be approximately equal to the mean

Sequencing Data - Summary

- **Shot Noise**
  - unavoidable, appears even with perfect replication
  - dominant noise for weakly expressed genes
- **Technical Noise**
  - from sample preparation and sequencing
  - negligible (if all goes well)
- **Biological Noise**
  - unaccounted-for differenced between samples
  - Dominant noise for strongly expressed genes

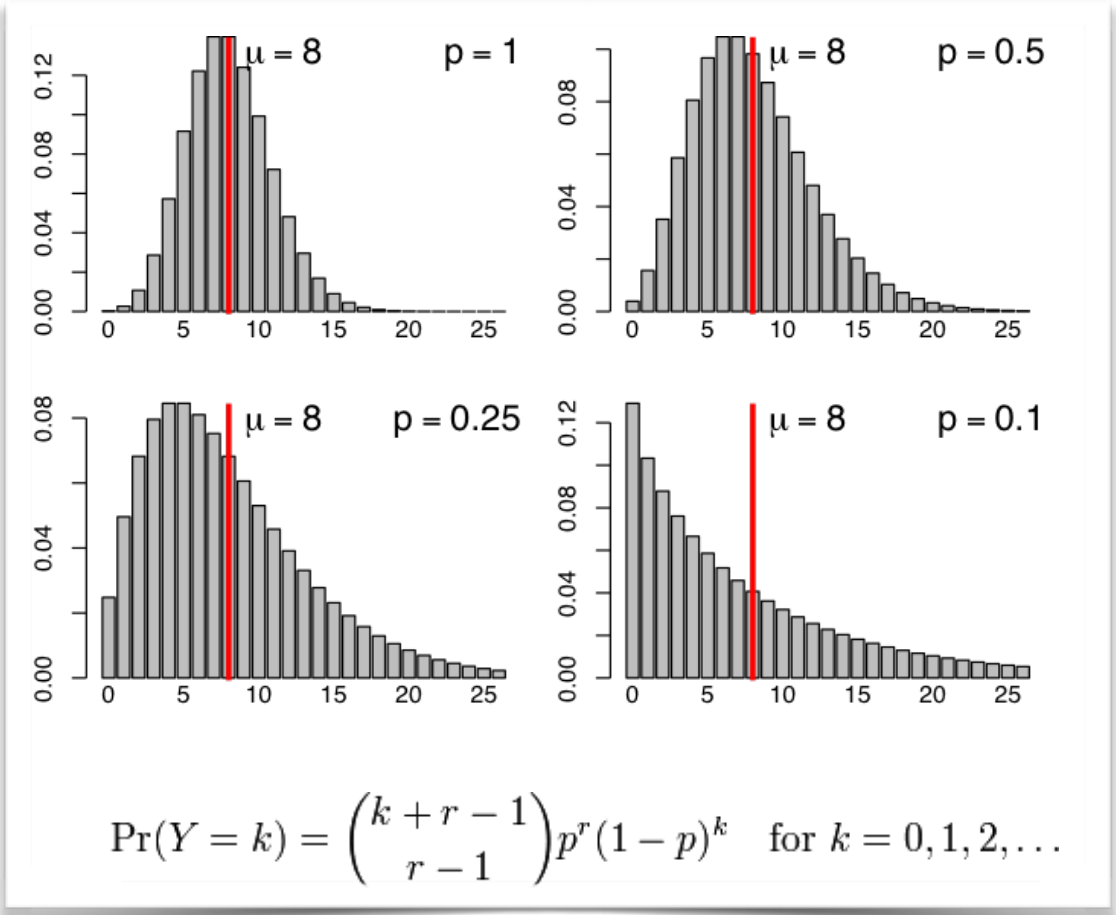
Sequencing Data - Summary

- **Shot No**
  - unavoid
  - domina
- **Technic**
  - from sa
  - negligib
- **Biologic**
  - unaccou
  - Domina



How to resolve this issue ?

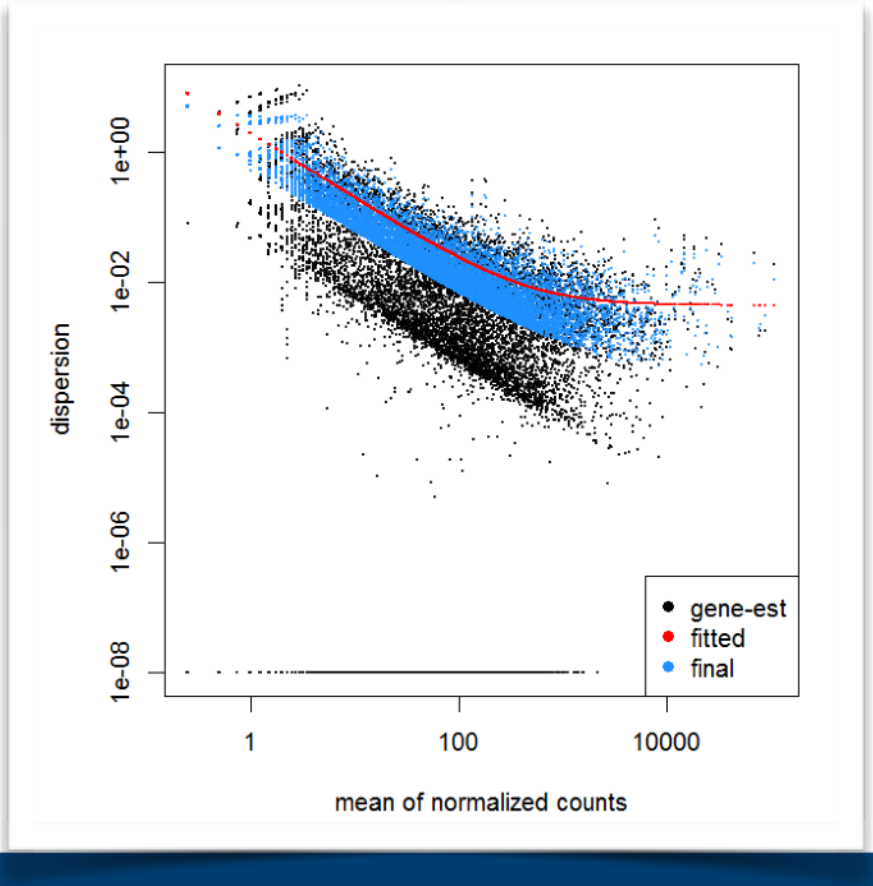
- The negative binomial distribution



$$\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

Negative Binomial Approaches

- DESeq2 - Simon Anders & Wolfgang Huber
- EdgeR - Robinson & Smyth



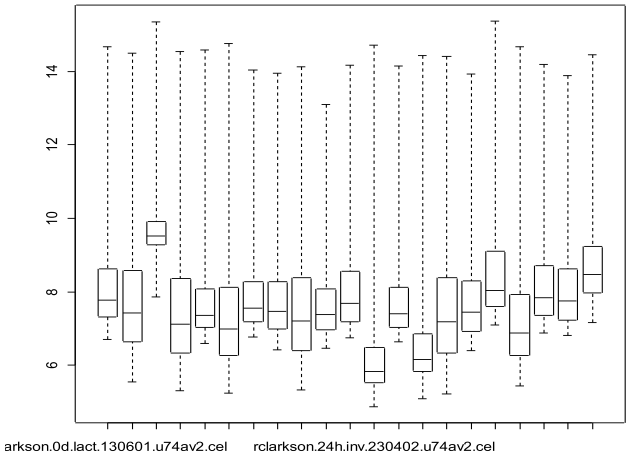
RNASeq Quantitation and Normalisation

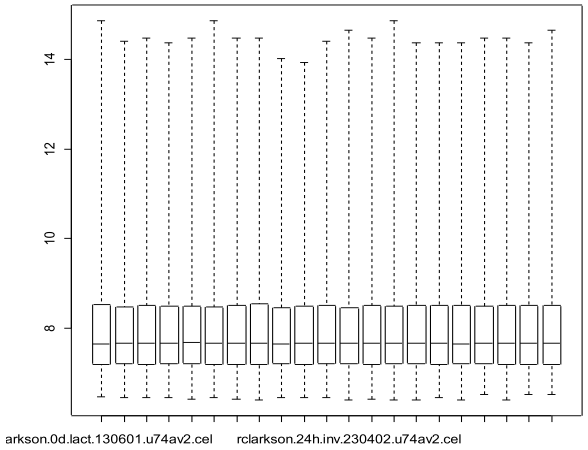
- Genes which are longer, produce more reads, they are not necessarily more expressed
- The expression of the majority of genes should be invariant between samples
- Any observed variation in their expression levels is due to technical and biological sample variation

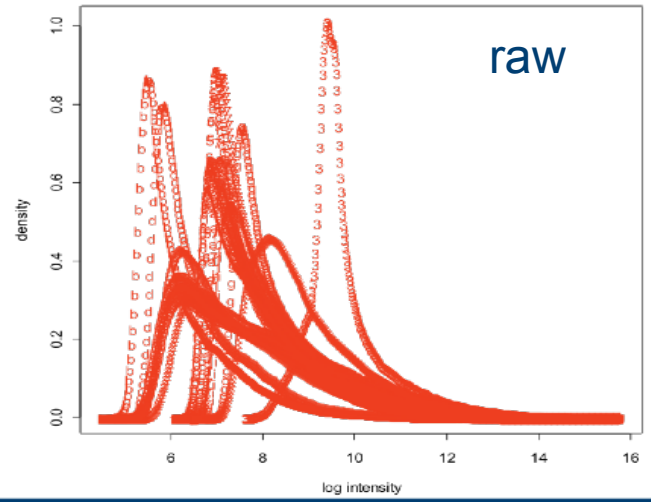
Why Normalise ?

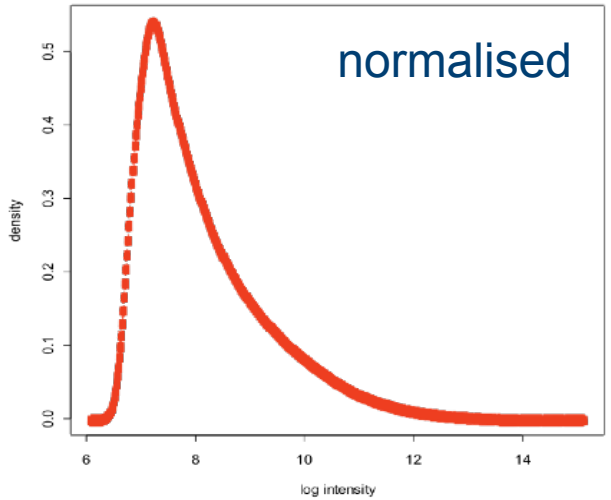
- The experimental goal is to identify biological variation (expression changes between samples)
- Technical variation can hide the real data
- Unavoidable systematic bias should be recognised and corrected
- Normalisation is necessary to effectively make comparisons between samples
- There are different methods of normalisation the assumptions of where variation exist will determine the normalisation techniques used.
- Always look at data before and after normalisation

Normalisation



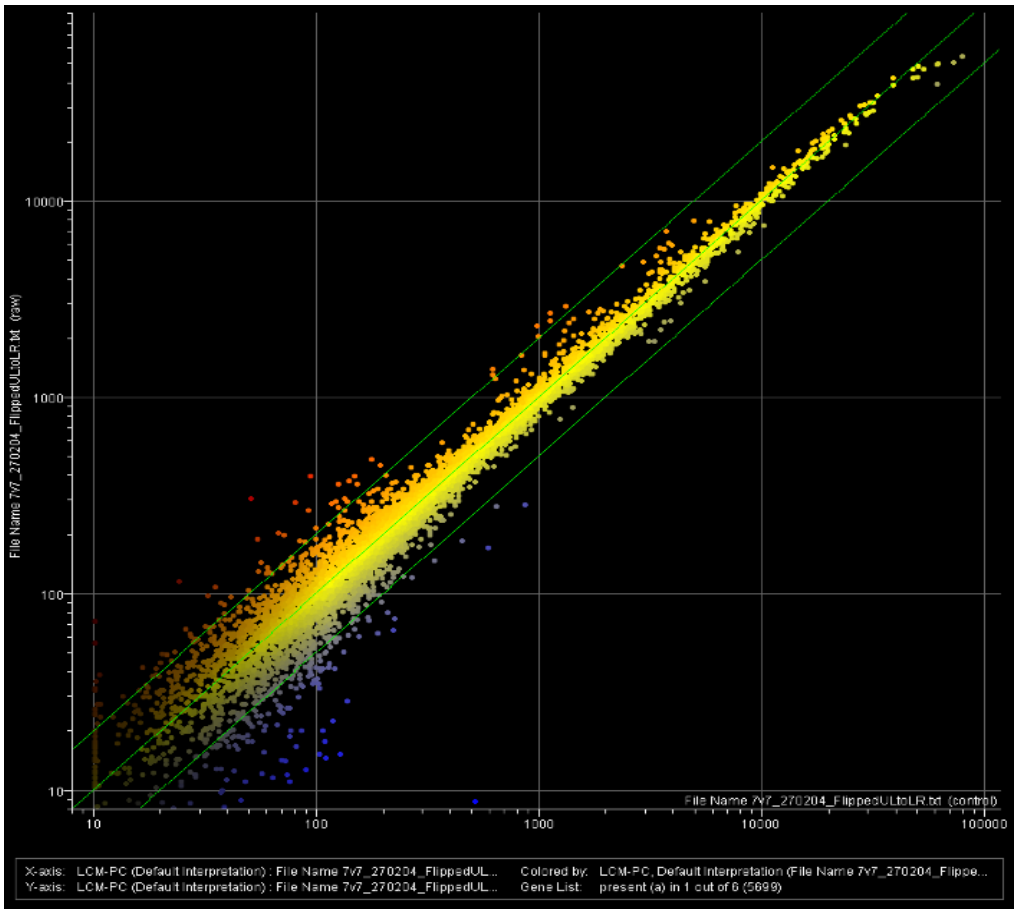






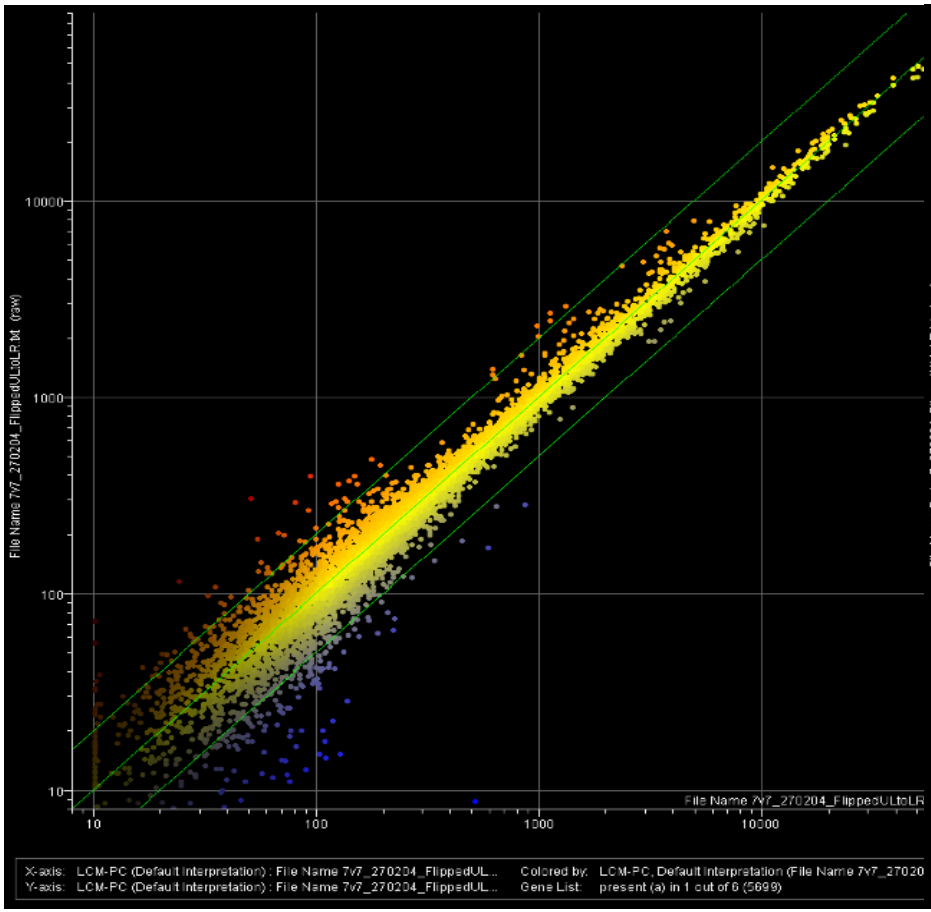
Normalisation has limits

Liver slice

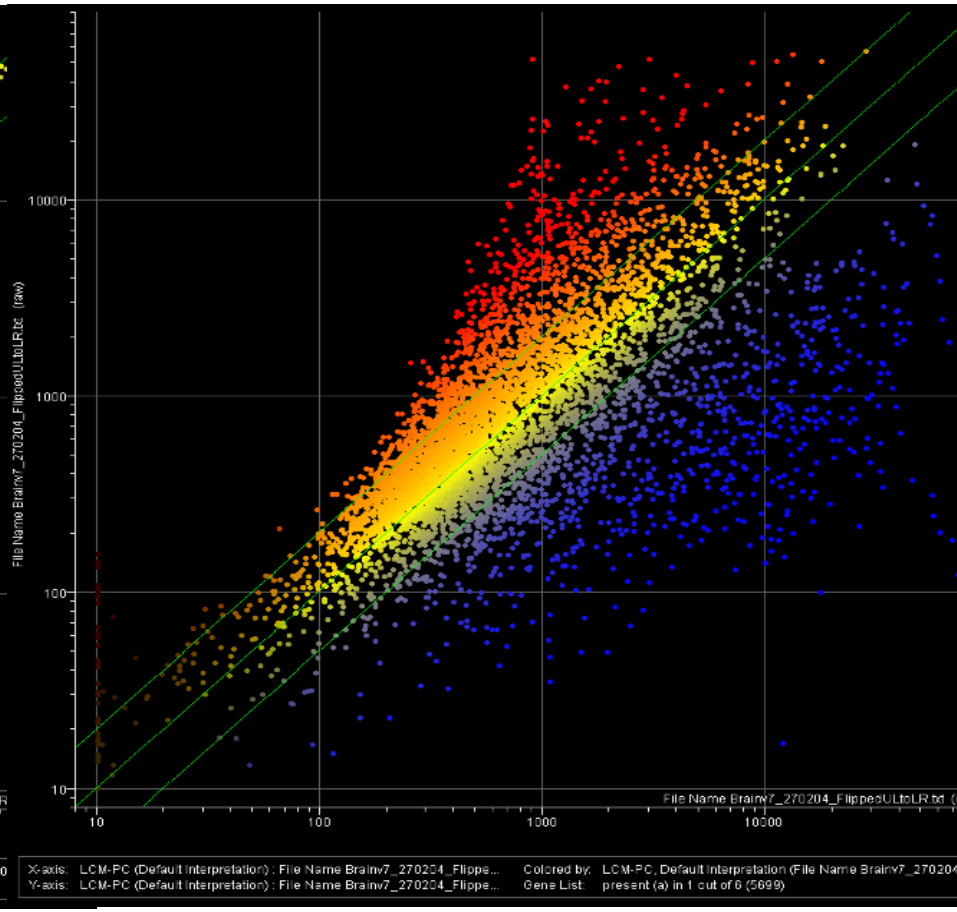


Normalisation has limits

Liver slice



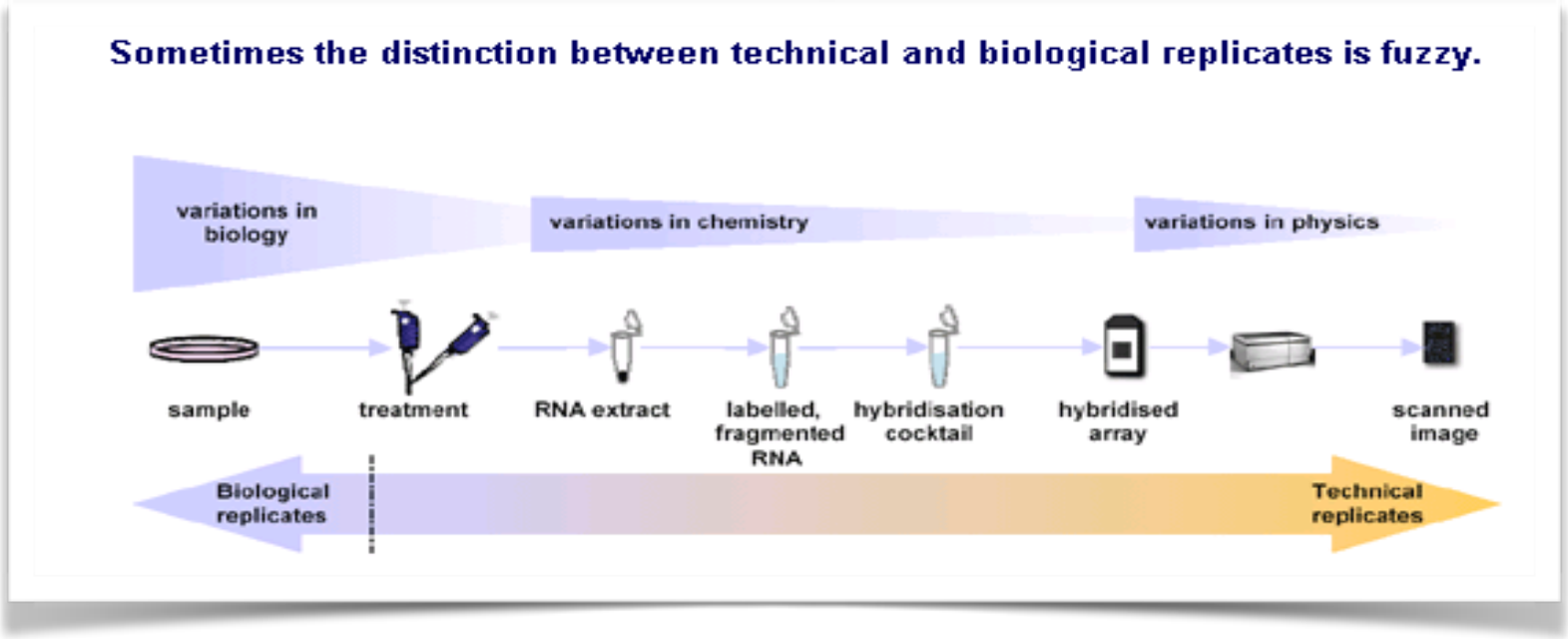
Brain



Experimental Design

- **Know what you want to achieve**
  - Is RNASeq Appropriate ? Could a microarray be better ?
- **Samples will fail**
  - RNA Degradation, rRNA contamination
  - Not enough input
  - Other Issues
- **Sample QC**
  - Nanodrop, Qubit, Bioanalyser, Tapestation
- Statistical Analysis will only work if enough replicates are present
  - **At least 3 biological replicates per condition**

Technical vs. Biological Replicates



RNASeq has very low technical variation - technical replicates NOT required

RNASeq - Read Depth

- For straight-forward differential expression
  - **10M reads per sample is plenty**
  - **5M reads for small RNA Sequencing**

- For alternative splicing
  - **20-50M reads**

- For *de novo* discovery of novel transcripts (e.g. lncRNAs)
  - **50-100M reads per sample.**

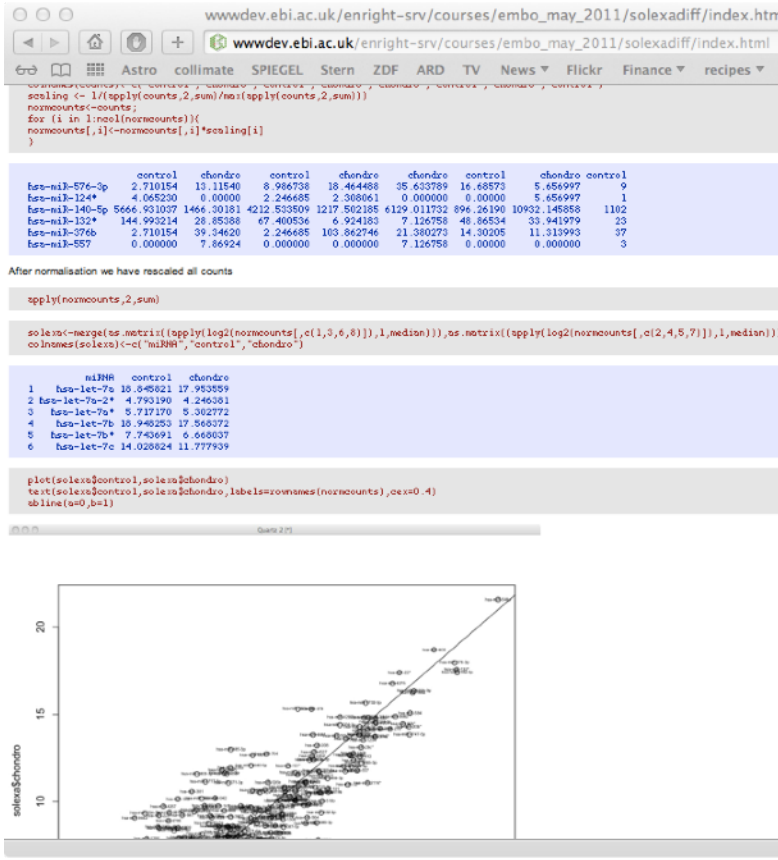
mRNA Seq - Example Dataset

Breast Cancer Samples (MCF7)

Jiannis Raggousis, McGill University, Montreal

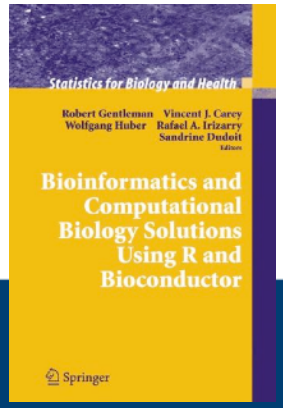
MCF7 Line
3 Scrambled Controls
3 miR-210 knockdowns

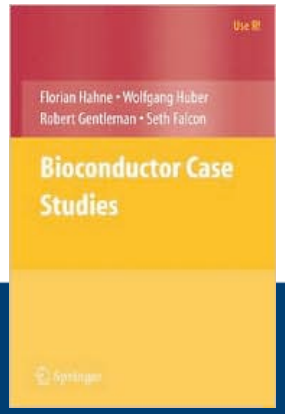
Sequenced on Illumina
MiSeq
Wellcome Trust Advanced Course 2016



R & BioConductor

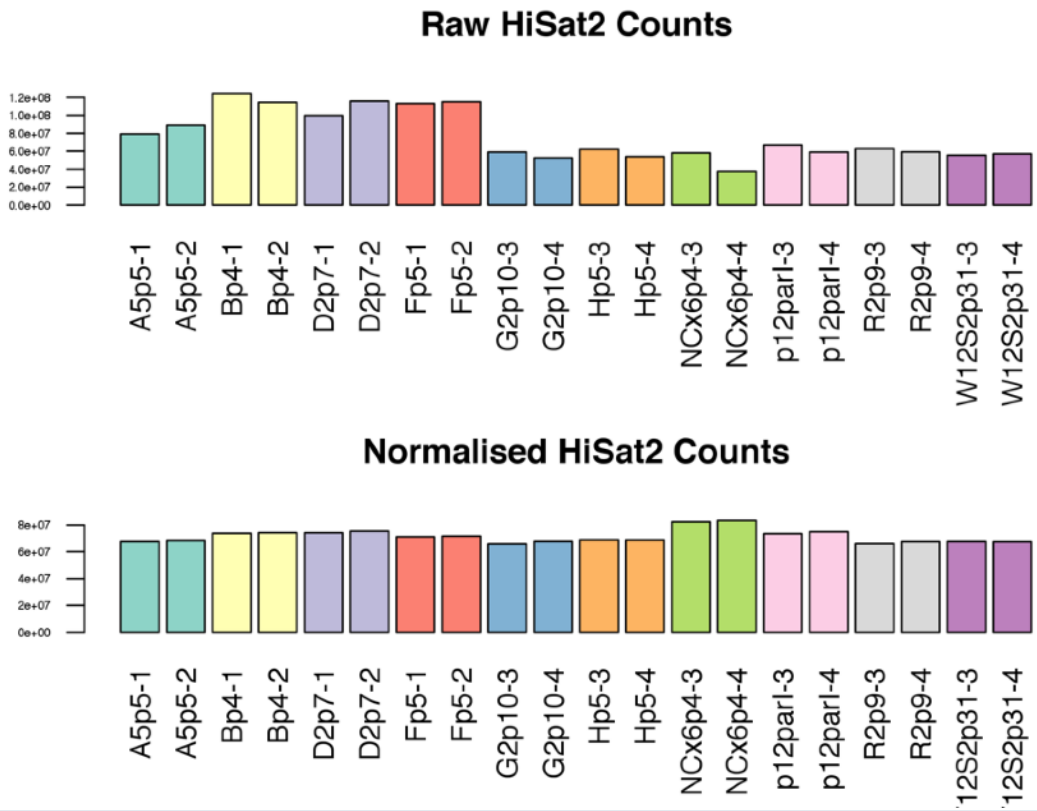
- Freely available
- Updated constantly
- Available here:
  - http://www.r-project.org/
  - http://www.bioconductor.org/
- LIMMA user guide and the Springer book series are well worth investigating
- DESeq2 Vignette and Manual Pages excellent also
- Full R courses available



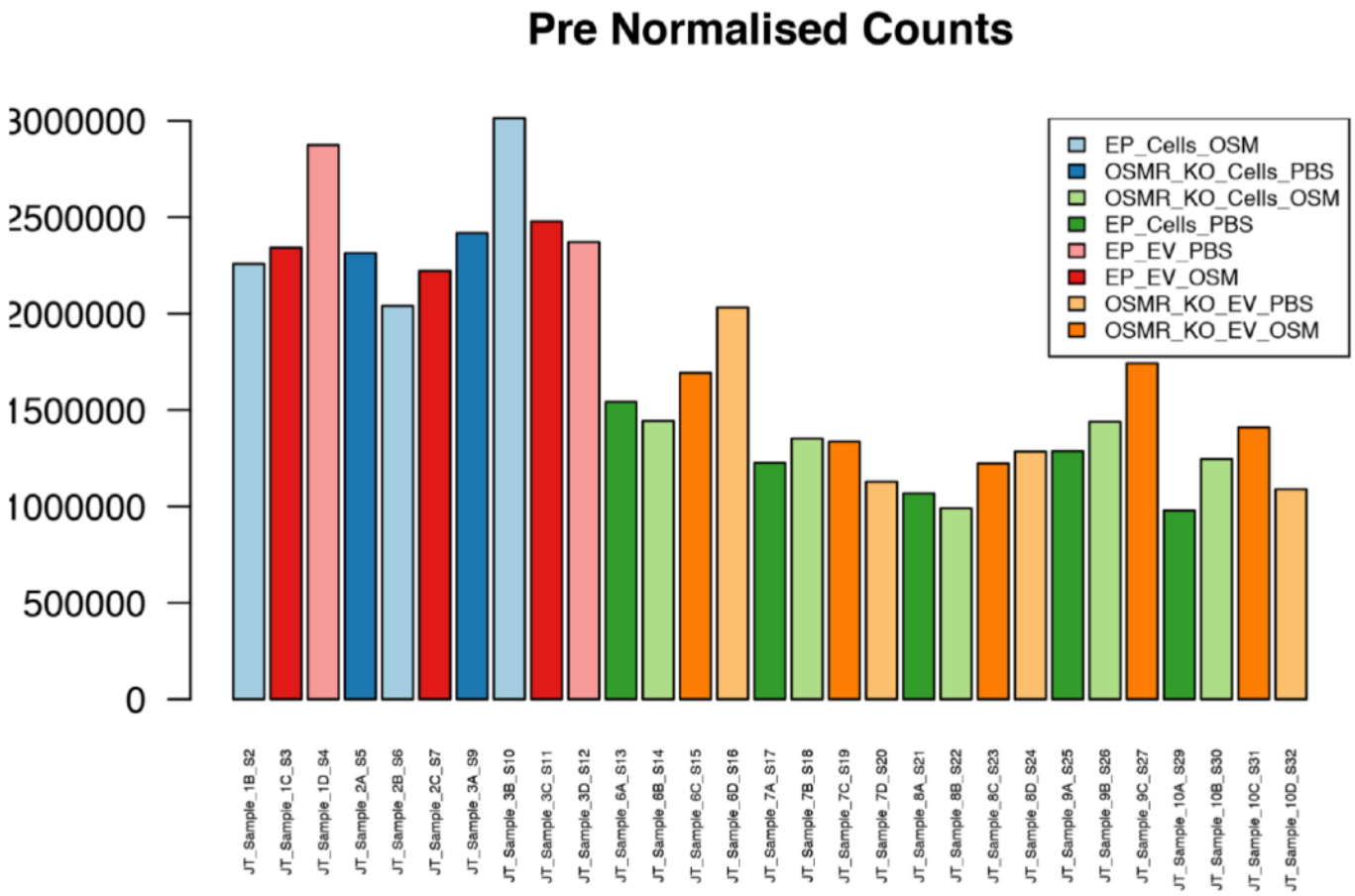


Post Sequencing QC

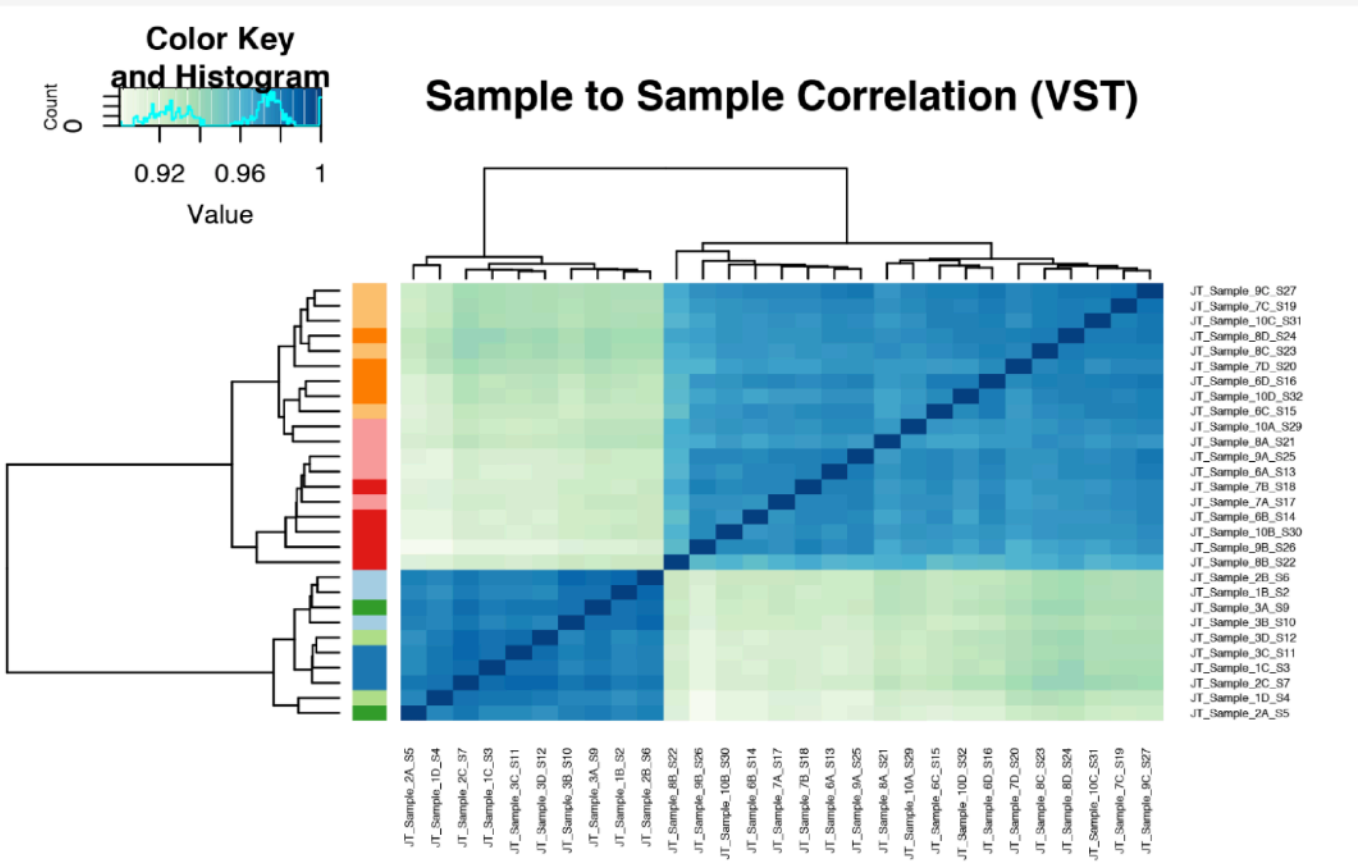
- Raw counts of mapped reads



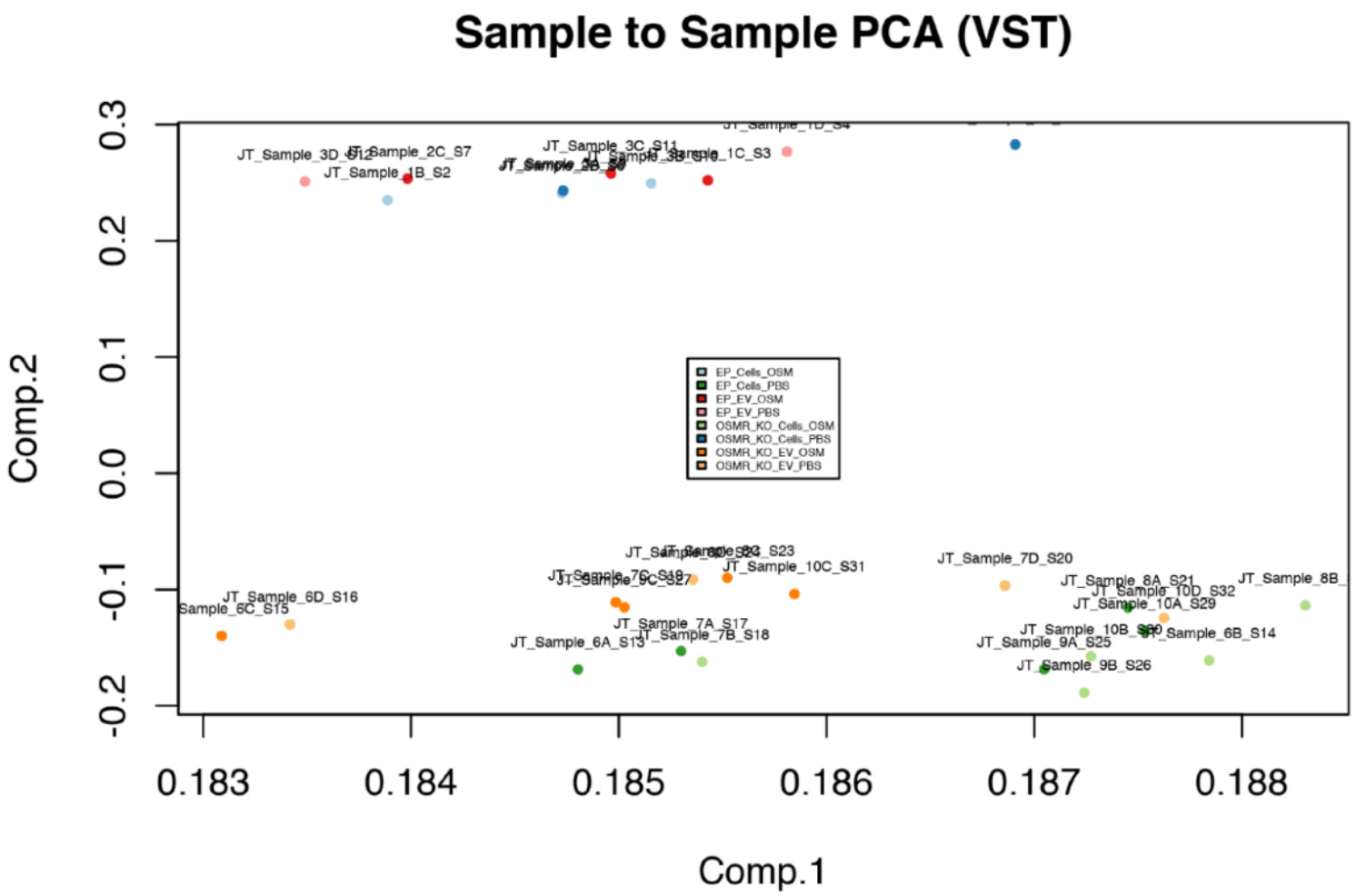
Another example



Sample QC - Sample Correlation

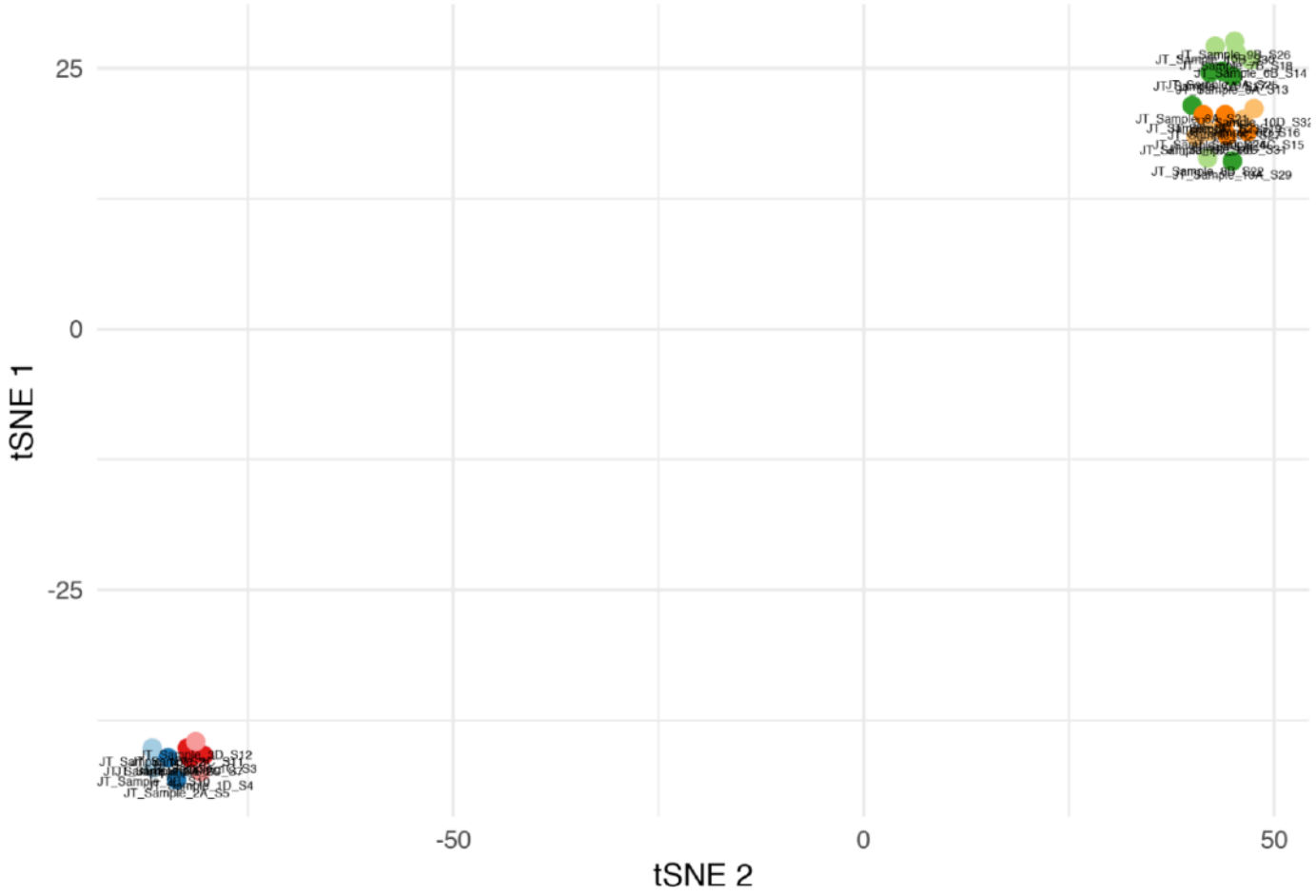


PCA or tSNE - Sample Analysis



UNIVERSITY OF CAMBRIDGE

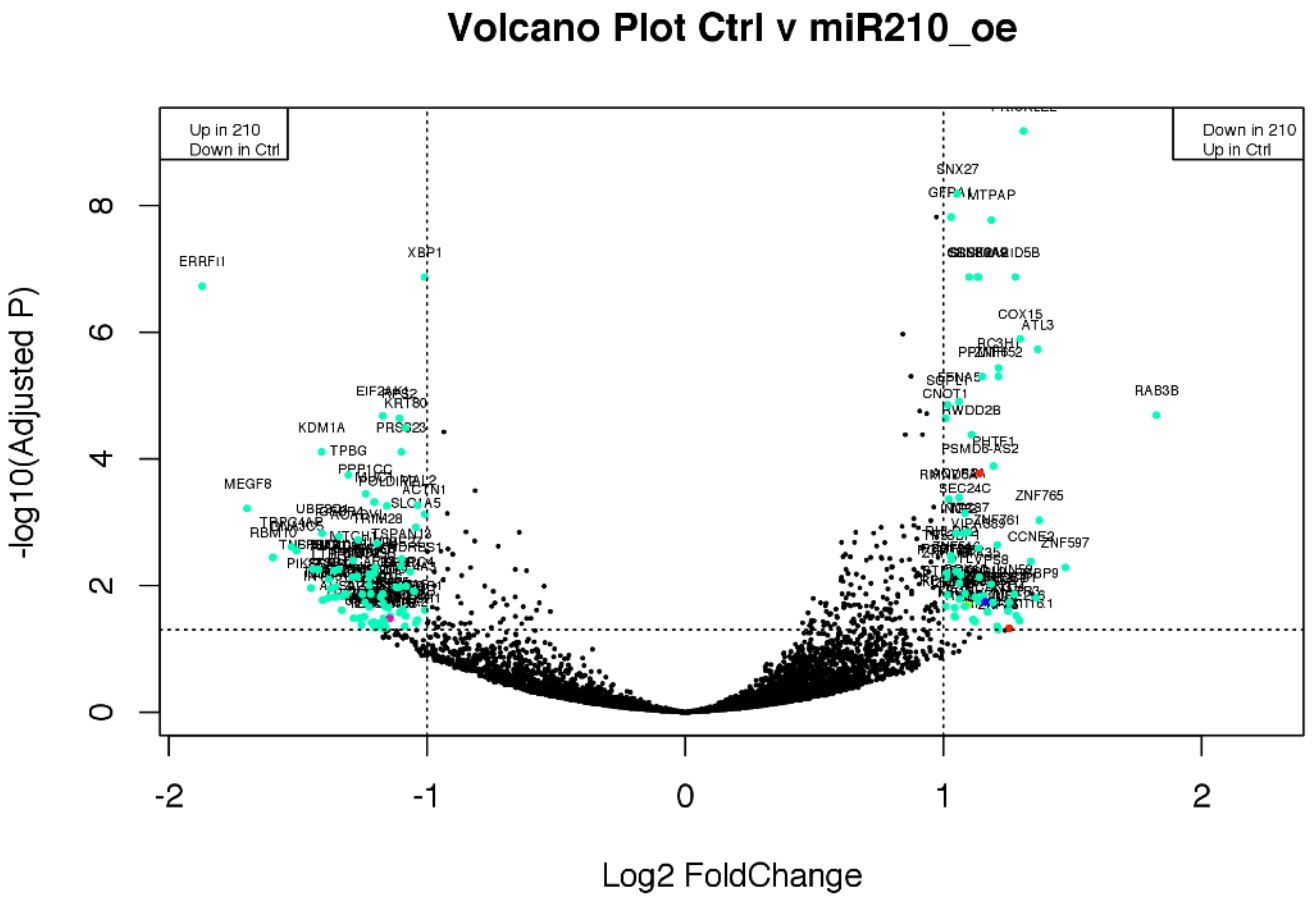
PCA or tSNE - Sample Analysis



Differential Expression Analysis

- **Data has passed QC**
  - ***Failed samples have been removed***
- **Data has been normalised**
- **The experimental design is correct and has power (e.g. enough replicates)**
- Define a 'contrast', i.e. a particular test.
    - Wildtype versus Knockout
    - 0hr versus 10 hour
    - Disease versus Normal
  - Complicated contrasts can be constructed
  - A statistical test (e.g. t-test) is computed to test for significant changes for the contrast
- **A Log Fold-Change (LFC) and Adjusted P-Value (adj.P) will be produced for each gene**

Analysis of Differential Expression - Volcano Plot



Downstream Analysis of Differential Expression

- Confirm key findings (e.g. PCR)
- Gene Ontology Analysis
- Pathway / Network Analysis
- Detection of Motifs
  - TF binding motifs upstream of DE genes,
  - microRNA binding motifs in 3'UTRs
- Gene Set Enrichment Analysis