"IBS 574 - Computational Biology & Bioinformatics"
Spring 2018, Thursday (03/22) 2.00-4.00PM

Gene Expression: II

RNA-seq data analysis

Quality Control, Mapping, Gene Count & Differential Expression

Ashok R. Dinasarapu Ph.D

Scientist, Bioinformatics
Dept. of Human Genetics, Emory University, Atlanta

RNA-Seq Exercise:

raw reads to differential expression

https://bitbucket.org/adinasarapu/ibs_class/src

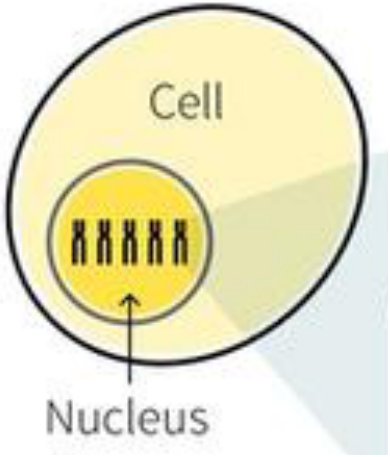
03/27/2018

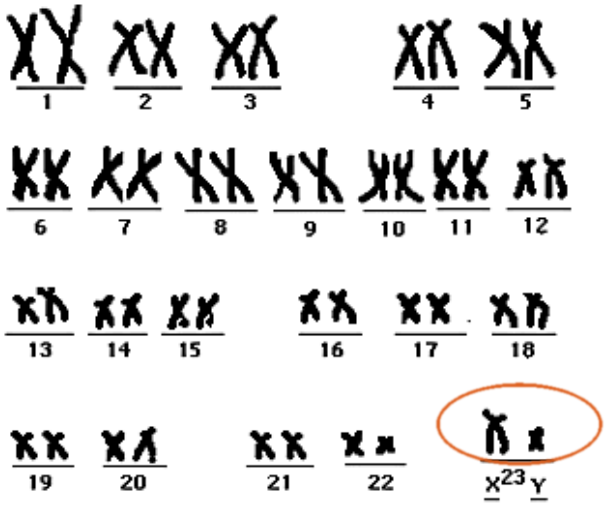
1. Quality Control (QC)
2. Mapping (Alignment)

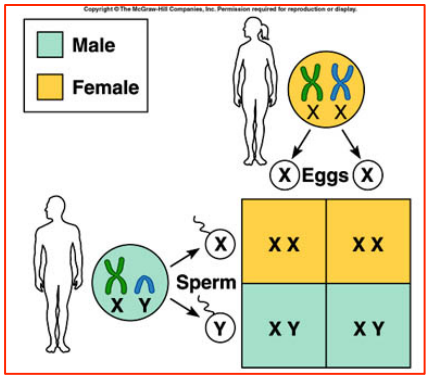
03/29/2018

1. Gene count & normalization
2. Differential expression

Genome – the haploid set of chromosomes

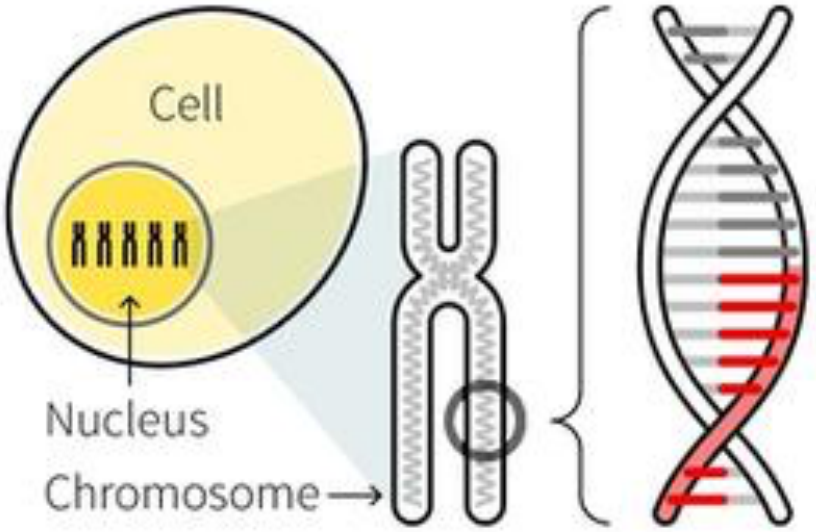


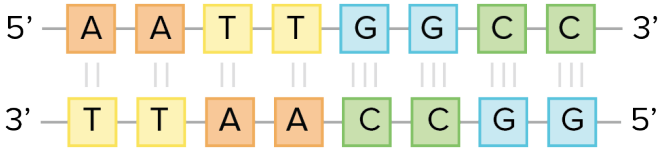




DNA, the genetic material, is made up of molecules called **nucleotides**. Each nucleotide contains a phosphate group, a sugar group and a **nitrogen base**. Nitrogen bases Adenine (**A**) and Guanine (**G**) are called purines, while Thymine (**T**) and Cytosine (**C**) are called pyrimidines. The **human genome** contains ~ 3 billion of these **base pairs**, which reside in the 23 pairs of chromosomes within the nucleus of all our cells.

Genome – the haploid set of chromosomes



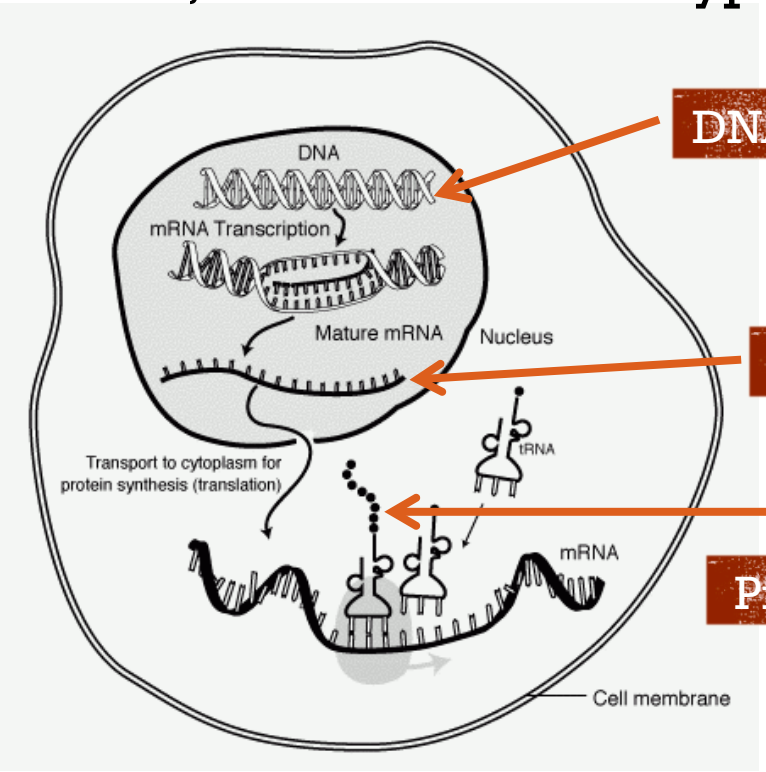


Human genome length = 3.3×10^9 *bp*

DNA, the genetic material, is made up of molecules called **nucleotides**. Each nucleotide contains a phosphate group, a sugar group and a **nitrogen base**. Nitrogen bases Adenine (**A**) and Guanine (**G**) are called purines, while Thymine (**T**) and Cytosine (**C**) are called pyrimidines. The **human genome** contains ~ 3 billion of these **base pairs**, which reside in the 23 pairs of chromosomes within the nucleus of all our cells.

The central dogma of molecular biology

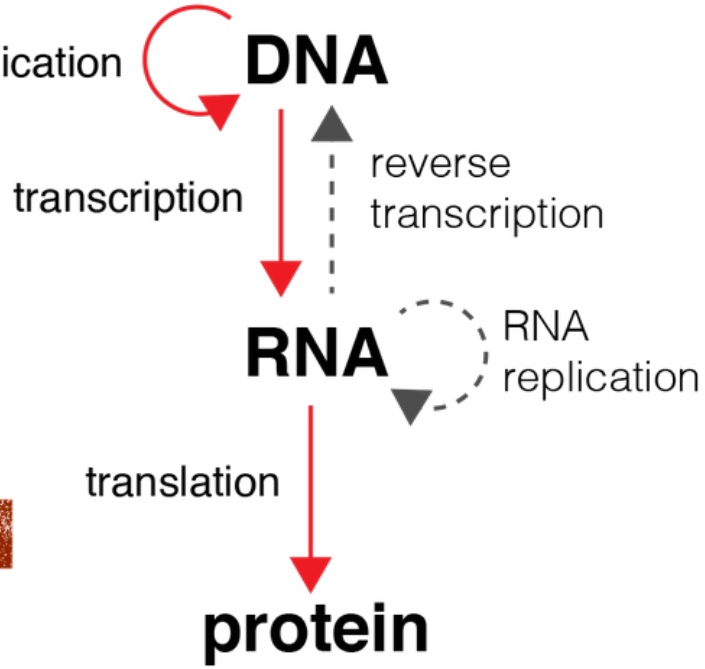
In multicellular organisms, nearly all cells have the same DNA, but different cell types express distinct proteins.



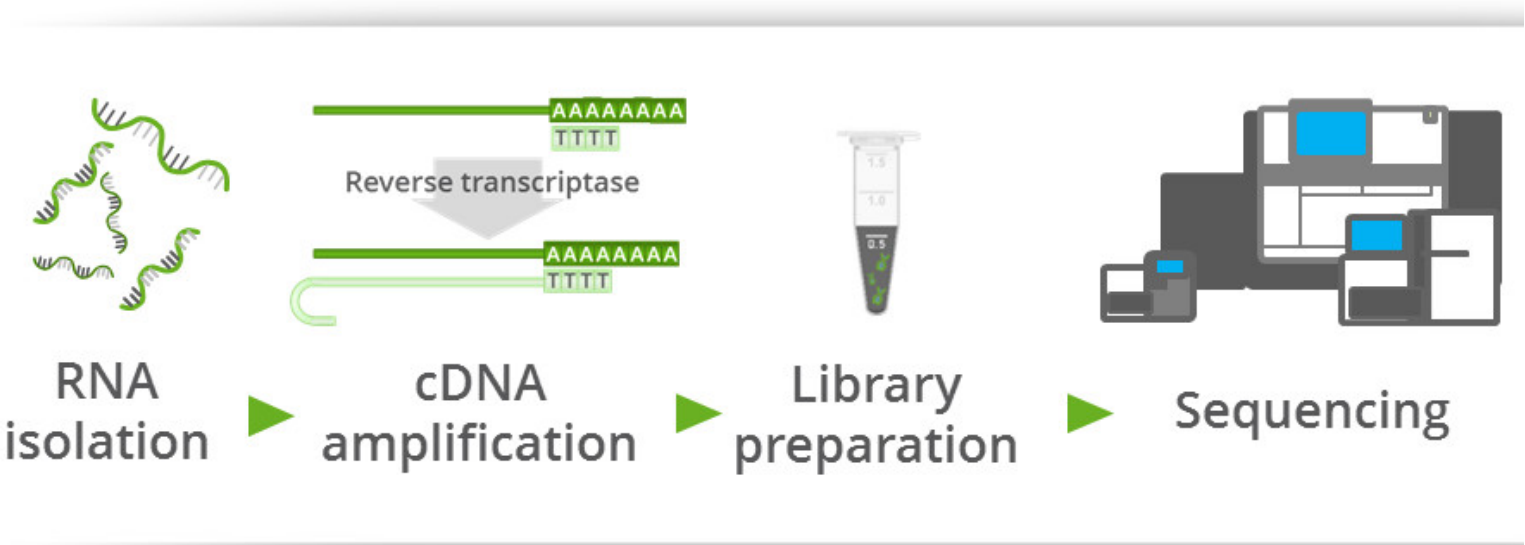
DNA

RNA

Protein



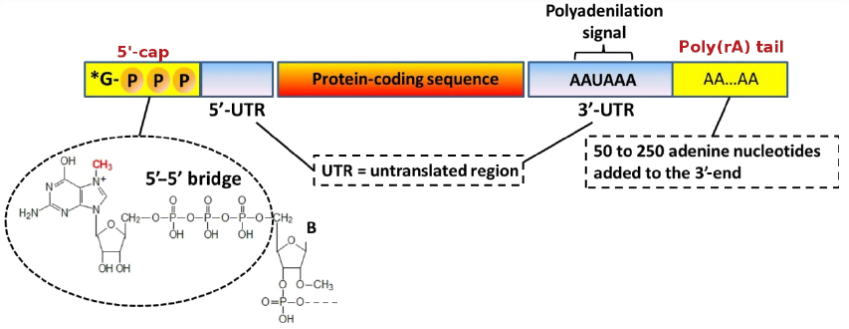
RNA-seq: a tool for transcriptomics

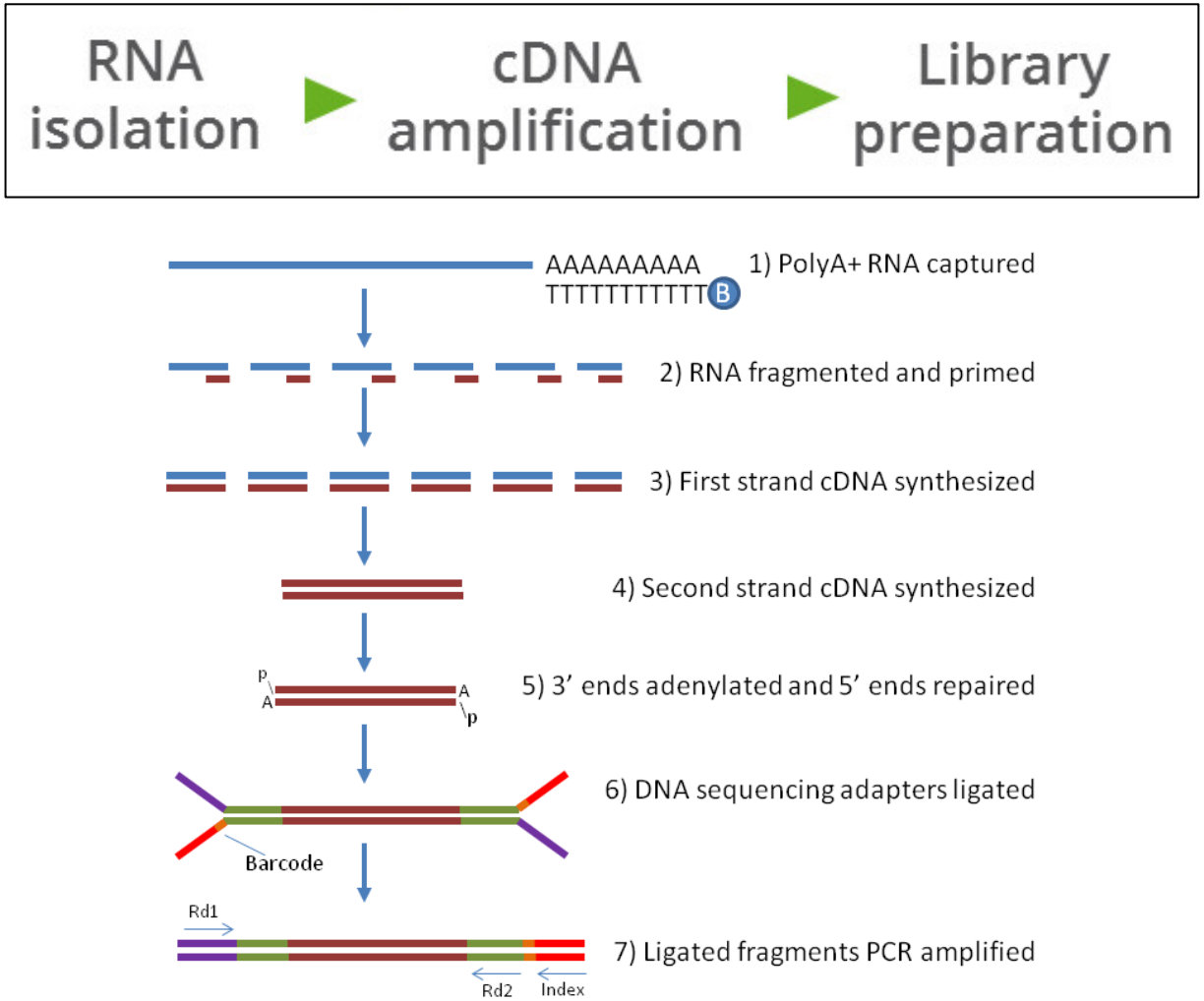


https://www.gatc-biotech.com

RNA-seq: a tool for transcriptomics

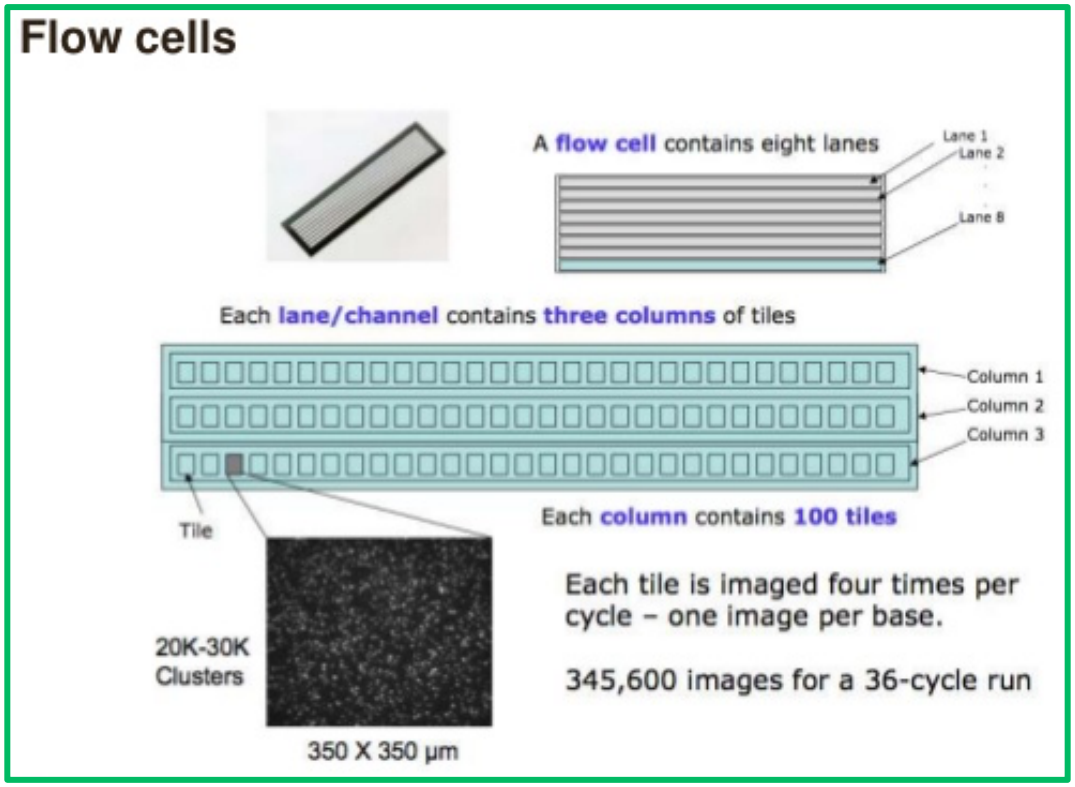
Polyadenylation is part of the RNA processing pathway that leads to the production of mature **mRNA** molecules

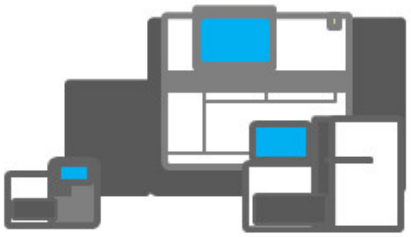




Illumina RNA library preparation. PolyA+ RNA is enriched using oligo(dT) beads followed by fragmentation and reverse transcription. The 5' and 3' ends of cDNA fragments are next prepared to allow efficient ligation of "Y" adapters containing a unique barcode and primer binding sites. Finally, ligated cDNAs are PCR-amplified and ready for cluster generation and sequencing.

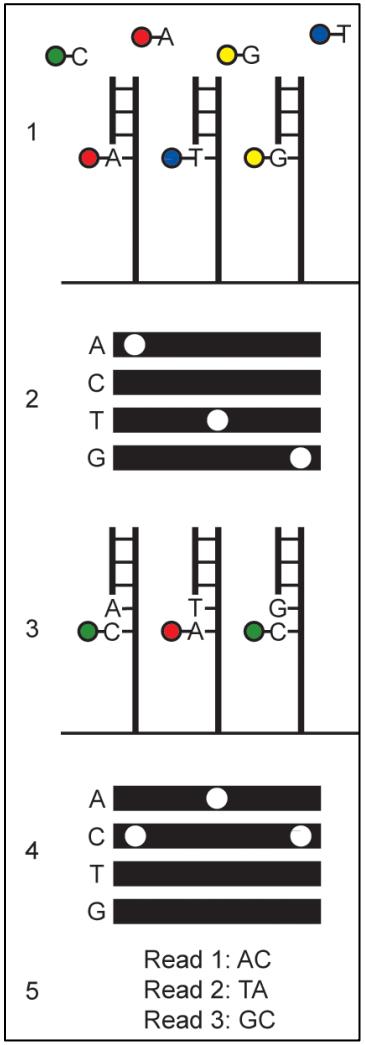
RNA-seq: a tool for transcriptomics





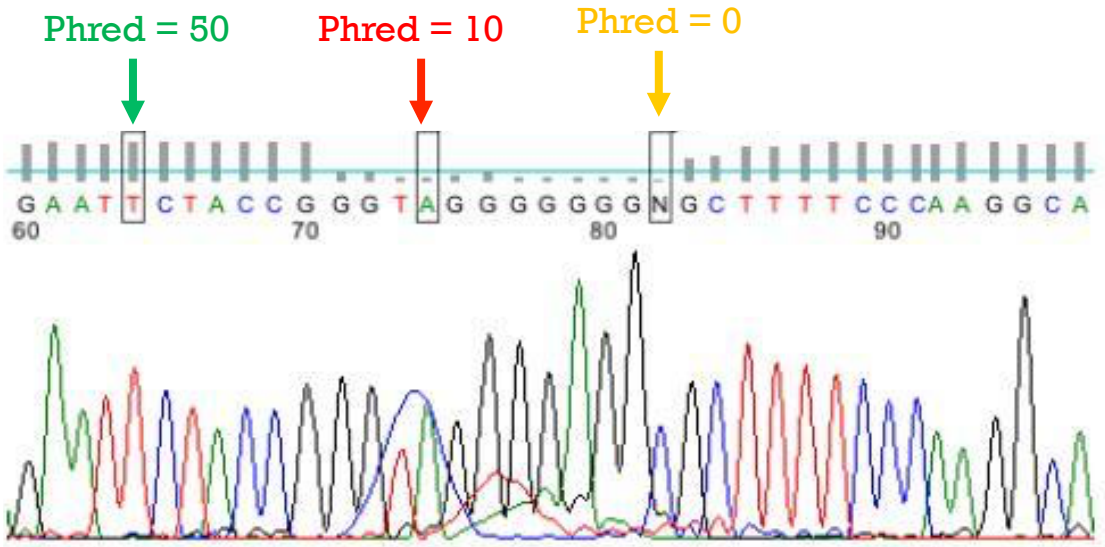
- Sequencing

Sequence detection method of illumina



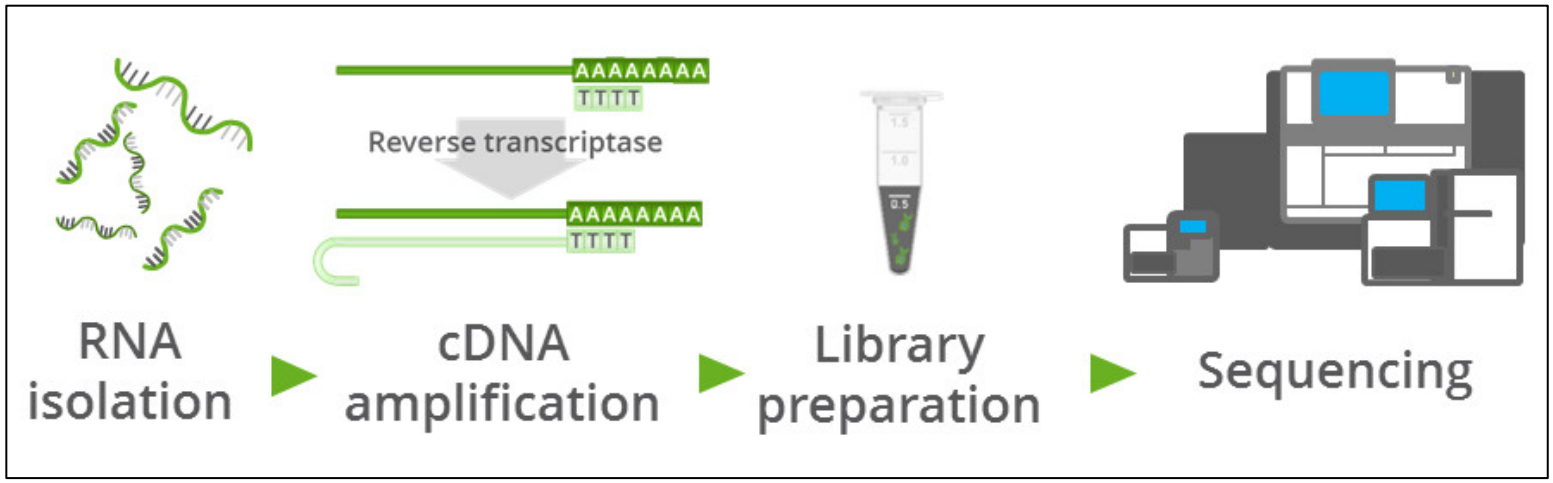
A) Illumina detection is fluorescence-based using reversible terminator dNTPs, resulting in one nucleotide incorporation per cycle (Materials Methods, 2013)

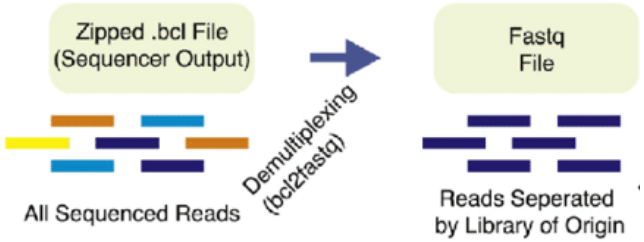
B) DNA sequence tracing and Phred score corresponding to each colored peak.



Phred quality score $(30) = -10\ log_{10}\ (0.001)$ (99.9% base call accuracy)
Probability of incorrect base call = 1 in 1000 (0.001)

RNA-seq: a tool for transcriptomics





Paired-End Sequencing

R1.fastq

R2.fastq

Fastq file contains sequence & base quality scores

R1.fastq

@HWI-ST1309F:278:C85EBANXX:5:1101:1497:1996 1:N:0:TGGGAGT
NGGGGAACTCCTGGTGGACCCTAGTGGAAGCCTTCCAGTAATTTCTTGAAGCTGAGCGCTCAGGTGAGTAGGGCGACATCTGGTG
GCCGGTTGTTGAAGG
+
#<<BBFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFB/<F/FFFF
…
…
…
…

R2.fastq

@HWI-ST1309F:278:C85EBANXX:5:1101:1497:1996 2:N:0:TGGGAGT
CTCAGAGGTGAAGTAACTTGCCCAGGGTTGTAGCCCAGGCCCTTCTCAGGACACGGCTCTCCCAGGGCCTCCGCCTCCCTGCAC
TGAGCCCTGCCAGTTC
+
/BBBBFFFBFFFFFFFFBFFFFFFF/FFFFFFFFFF<FFFFF<FFFFFFBFFFFFFBF/F<FFFFF<FFFBBFFFFFFFFFFFFF<FFBFFFFFFFB##
…
…
…
…

Fastq file contains sequence & base quality scores

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

R2.fastq

@HWI-ST1309F:278:C85EBANXX:5:1101:1497:1996 2:N:0:TGGGAGT
CTCAGAGGTGAAGTAACTTGCCCAGGGTTGTAGCCCAGGCCCTTCTCAGGACACGGCTCTCCCAGGGCCTCCGCCTCCCTGCAC
TGAGCCCTGCCAGTTC
+
/BBBBFFFBFFFFFFFFBFFFFFFF/FFFFFFFFFF<FFFFF<FFFFFFBFFFFFFBF/F<FFFFF<FFFBBFFFFFFFFFFFFF<FFBFFFFFFFB##
…
…
…
…

ASCII (/) char = 47 (decimal)
Illumina 1.8+ : 14 (Phred)+33 = 47 (/)

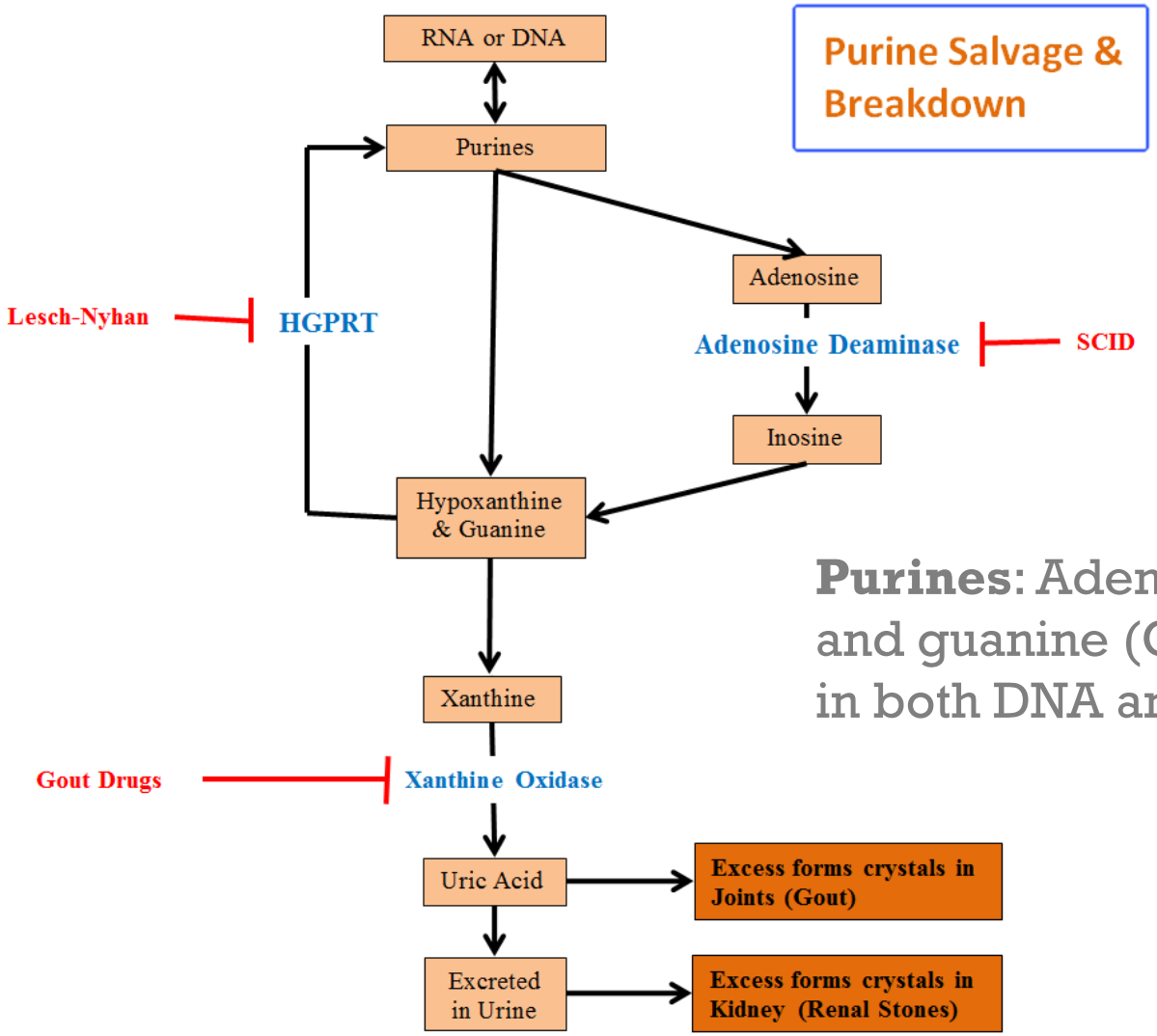
RNA-seq best practices

- **3 or more replicates**
  - Biological replicates are recommended rather than technical replicates
- **Avoid unwanted batch effects**
  - Always process your RNA extractions at the same time.
  - To Avoid lane batch effects, all samples would need to be multiplexed together and run on the same lane

HGPRT deficiency leads to Lesch-Nyhan disease

- HGPRT deficiency (in severely affected) leads to
  - abnormal accumulation of uric acid
  - neurologic disorders and
  - behavioral abnormalities
- Lesch-Nyhan disease affects about 1 in 380,000 live births.
- HGPRT (encoded by HPRT1) is an enzyme in purine metabolism.

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

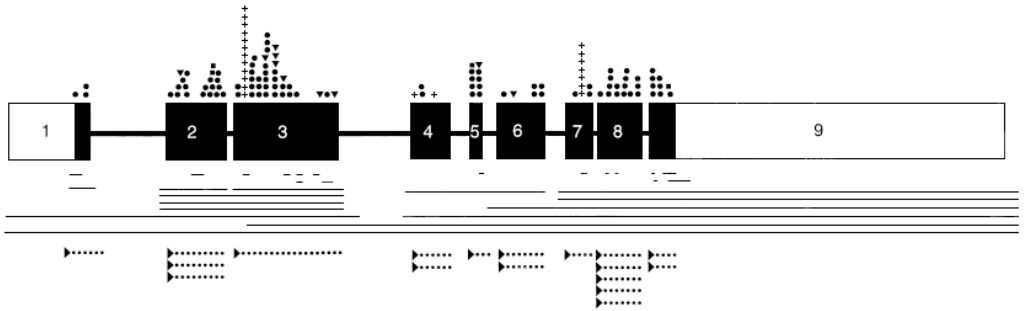


Purines: Adenine (A) and guanine (G) are found in both DNA and RNA.

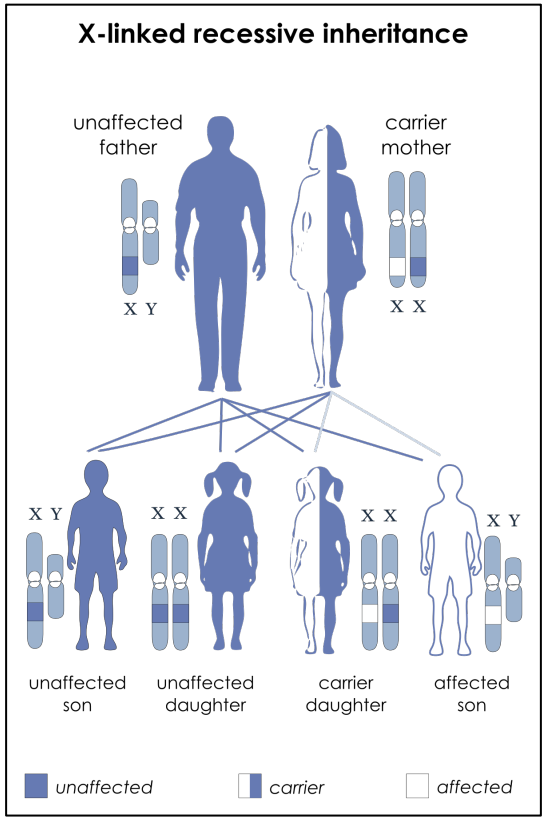
Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

HPRT1 gene mutations in 'Lesch-Nyhan' Disease

- HPRT1 is an X linked gene
- > 600 mutations in HPRT1 are known

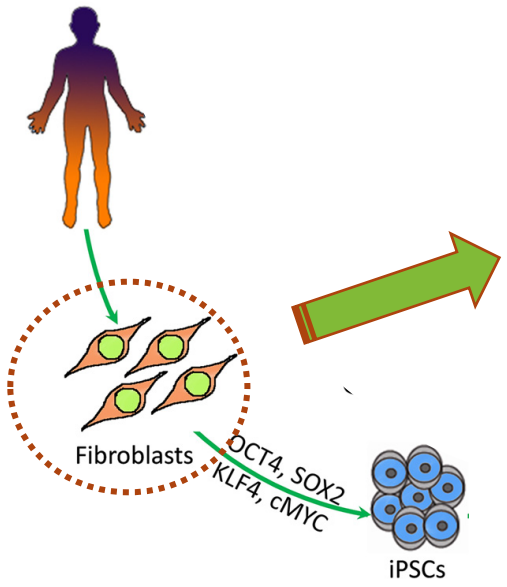


Wikipedia



Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Lesch-nyhan disease specific iPSCs obtained from fibroblasts



RNA-seq

AK
MG
PY
WW

Healthy/ Control

PB — Deletion (Exon 4)

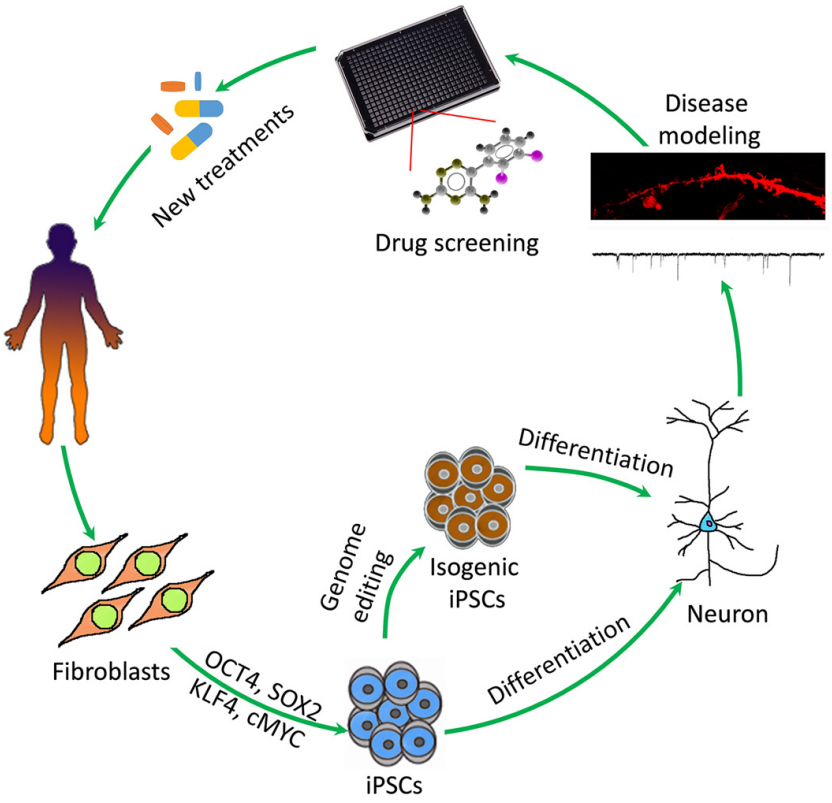
DP2 — 508 C>T (Exon 7)

TH — 151 C>T (Exon 3)

TS — 371 ins TT (Exon 4)

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

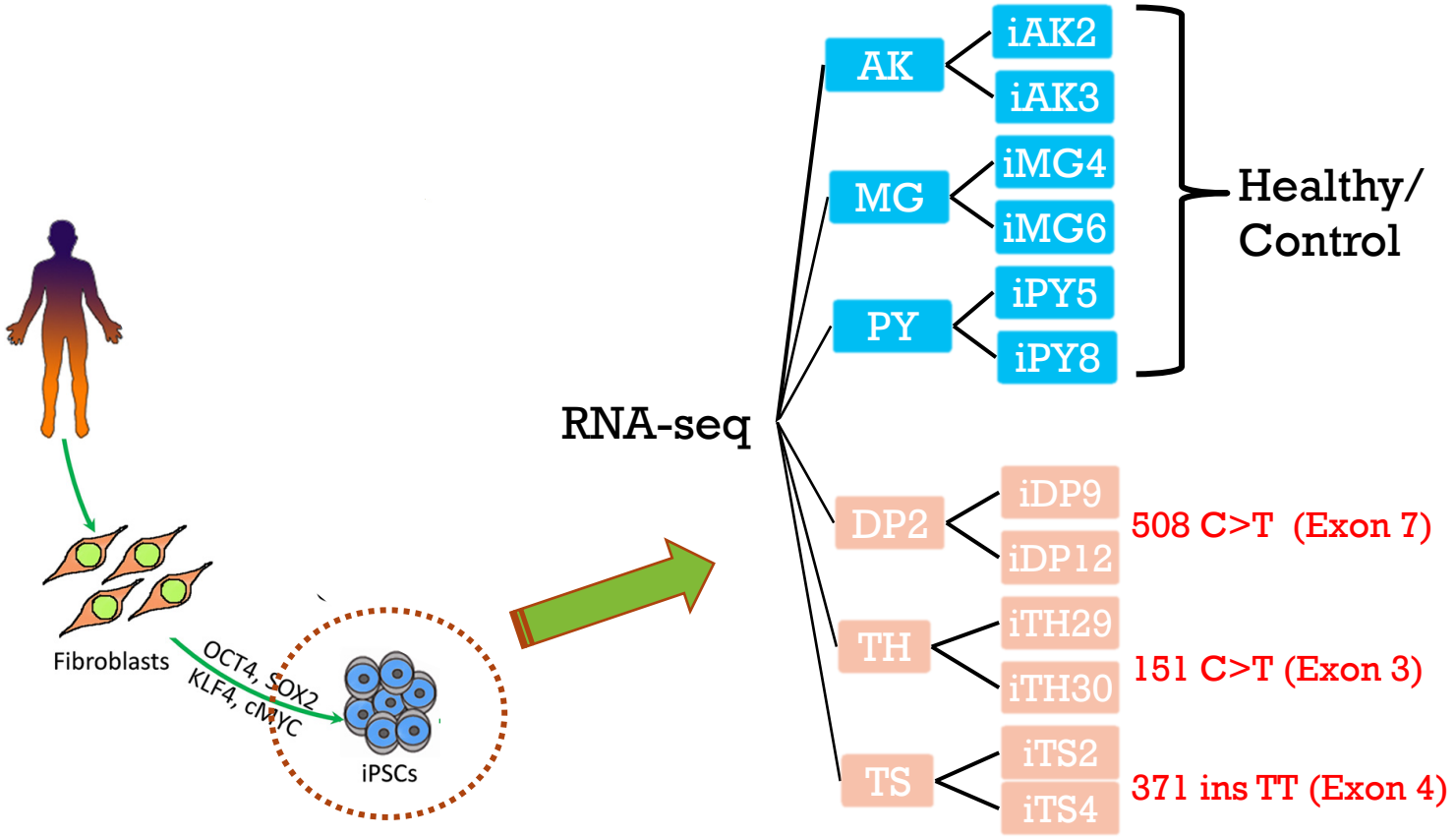
Human iPSCs to model neurologic disorders



- The disorder is so rare that it is **difficult to recruit** sufficient patients for meaningful studies.
- Autopsied **brain samples** are **difficult to obtain**.
- Established **animal models** have been helpful but species **differences in purine metabolism** make it difficult to relate to human condition.

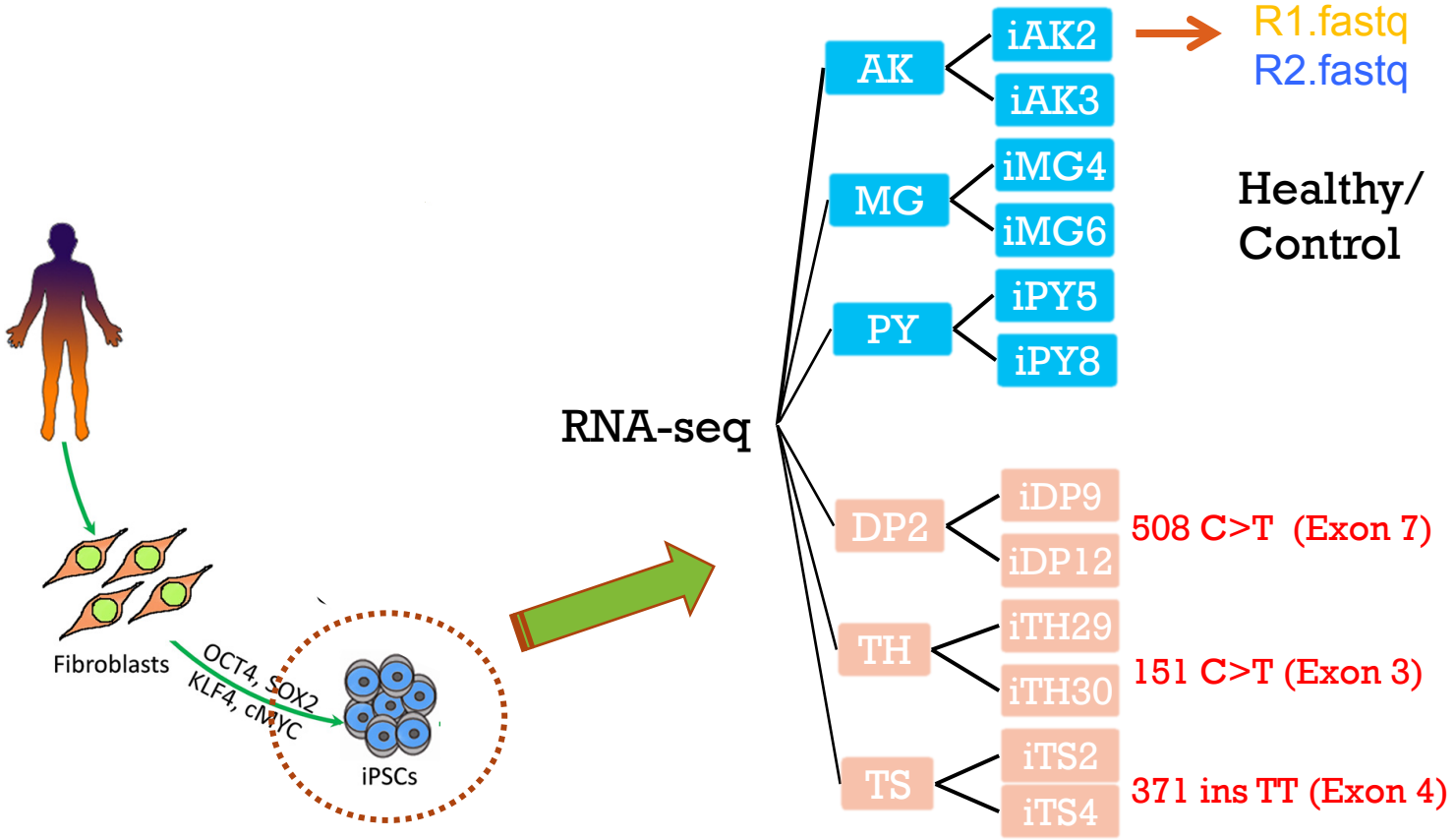
Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Lesch-nyhan disease specific iPSCs obtained from fibroblasts



Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Lesch-nyhan disease specific iPSCs obtained from fibroblasts



Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Data analysis

1. Read quality
2. Alignment / Mapping
3. Gene count & normalization
4. Differential expression

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Data analysis

1. Read quality
2. Alignment / Mapping
3. Gene count & normalization
4. Differential expression

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Read Quality

One read

> 50 **million reads** per sample

Basic Statistics

Measure	Value
Filename	C8DF6ANXX_1_GSLv3-7_SL146145.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	57154438
Sequences flagged as poor quality	0
Sequence length	100
%GC	49

R1.fastq

Measure	Value
Filename	C8DF6ANXX_2_GSLv3-7_SL146145.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	57154438
Sequences flagged as poor quality	0
Sequence length	100
%GC	49

R2.fastq

Read Quality

Per base & Per sequence quality scores

1 @HWI-ST1309F:276:C8DF6ANXX:1:1101:3694:1997 1:N:0:TAGCACC
NAAGAAAAAGAAGAAACAGAAGAGAAAAAAGGAGAAACCAAATTCCGGAGGCACCAAGTCAGACTCGGCATCTGATTCCCAGGAGATTAAAATTCAGCAG
+
#<<BBBFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFBFFFFFFBFFFFFFFFFFFFF<FF<F<FFFFFFFF<BB/7F/7BFF7FFFFF</FF
2 @HWI-ST1309F:276:C8DF6ANXX:1:1101:4615:1997 1:N:0:TAGCACC
NTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCCGTGATGCCCAGGGAAGACAGGGCGACCTGGAAGTCCAACTACTTCCTTAAGATC
+
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFBFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFBFFBFF#
3 @HWI-ST1309F:276:C8DF6ANXX:1:1101:7169:2000 1:N:0:TAGCACC
NTCCTCCGGGTAAAACACATTCCTTTGGTCCAATCTGGATGAGGAGAGCCTCAAGATTGGAGAACTGATCATTATCAGGGAATTCACACAGTCCTAGTTT
+
#<BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
…

…

…

50 million

100 bp read
Position = 15
Base = A
Q score = ‘F’ (as ASCII, 70 in decimal)

Read Quality

Per base & Per sequence quality scores

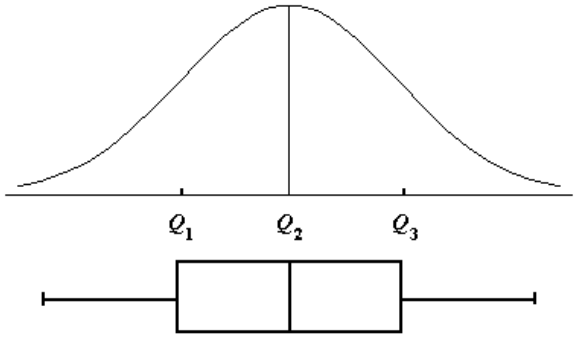
```
1 @HWI-ST1309F:276:C8DF6ANXX:1:1101:3694:1997 1:N:0:TAGCACC
  NAAGAAAAAGAAGAAACAGAAGAGAAAAAAGGAGAAACCAAATTCCGGAGGCACCAAGTCAGACTCGGCATCTGATTC
  +
  #<<BBBFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFBFFFFFFFFFFFFF<FF<F<FFFFFFF
2 @HWI-ST1309F:276:C8DF6ANXX:1:1101:4615:1997 1:N:0:TAGCACC
  NTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCCGTGATGCCCAGGGAAGACAGGGCGACCTGGAA
  +
  #<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFBFFFFFFFFFFF<FFFFFFFFFFFFFFFFFF
3 @HWI-ST1309F:276:C8DF6ANXX:1:1101:7169:2000 1:N:0:TAGCACC
  NTCCTCCGGGTAAAACACATTCCTTTGGTCCAATCTGGATGAGGAGAGCCTCAAGATTGGAGAACTGATCATTATCAG
  +
  #<BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
...

...

...
```

50 million

100 bp read
Position = 15
Base = A
Q score = 'F' (as ASCII, 70 in de
Q score = 70-33 = 37

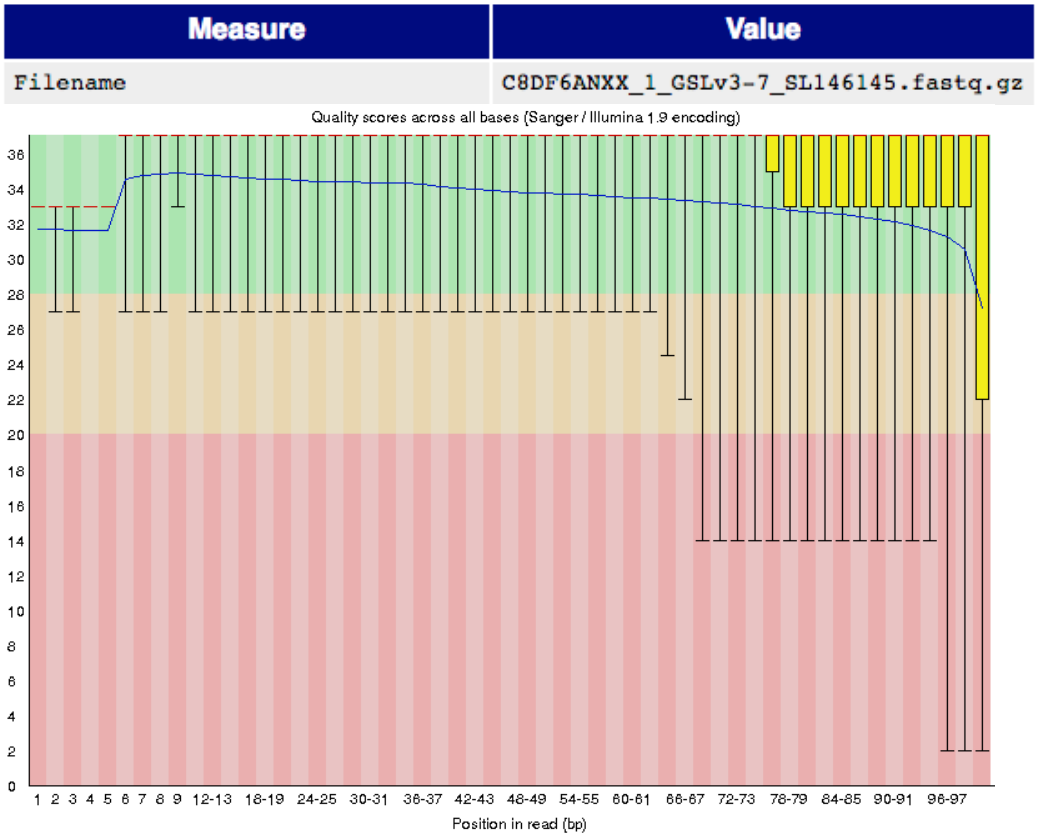
- Phred quality score
  - Error probability
  - ASCII encoded
  - Phred +33
    - Sanger [0,40]
    - Illumina 1.8 [0,41]
    - Illumina 1.9 [0,41]
  - Phred +64
    - Illumina 1.3 [0,40]
    - Illumina 1.5 [3,40]

Read Quality



Positional box-and-whisker plot

Per base sequence quality



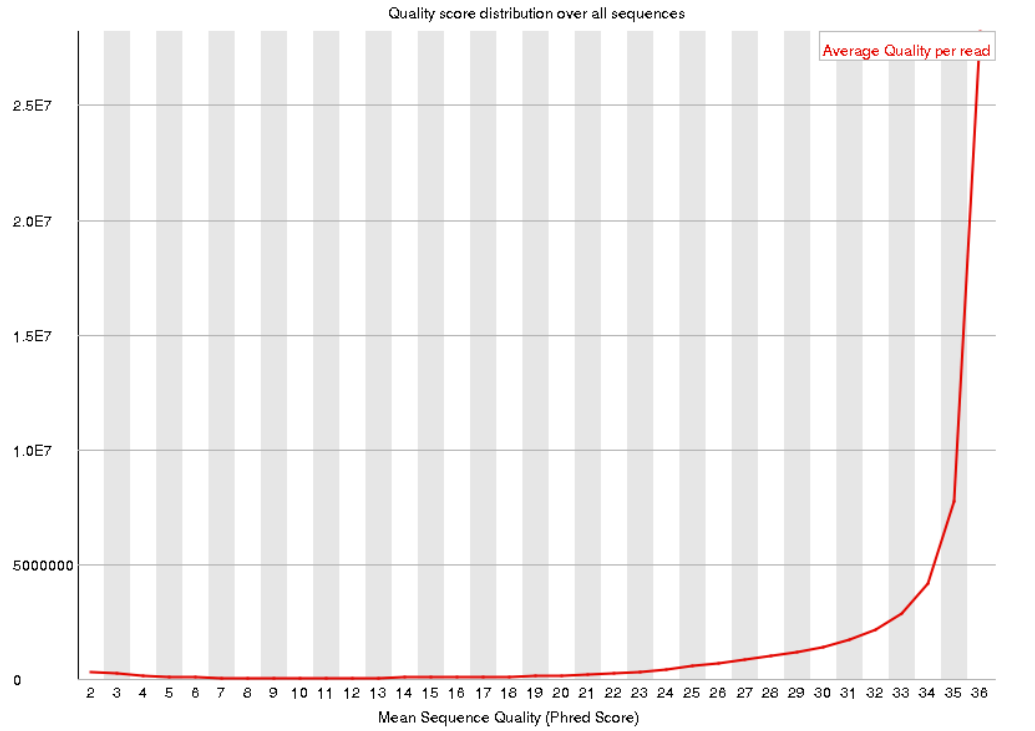
Software: FastQC (Java 1.8)

Read Quality

Software: FastQC (Java 1.8)

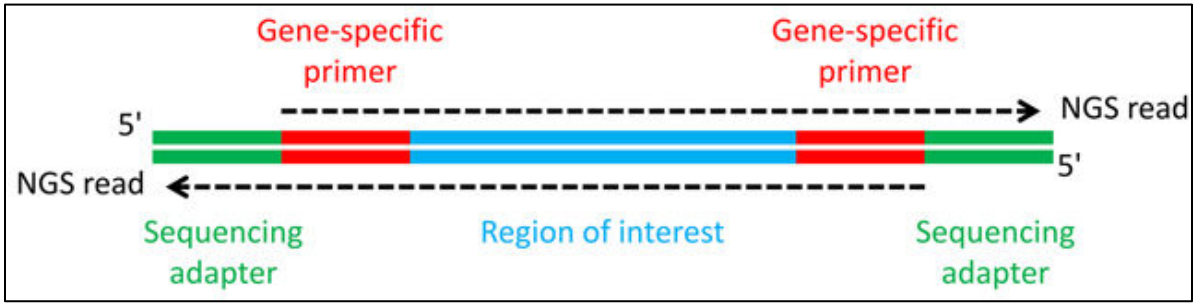
Per sequence quality scores

Measure	Value
Filename	C8DF6ANXX_1_GSLv3-7_SL146145.fastq.gz



Read Trimming & Filtering

Softwares: Cutadapt, Trimmomatic, etc.



- Adapter trimming
  - May increase mapping rates
  - Probably improves *de novo* assembly
- Quality trimming & Read filtering (erroneous base calls)
  - May increase mapping rates
  - May also lead to loss of information

Data analysis

1. Read quality
2. **Alignment / Mapping**
3. Gene count & normalization
4. Differential expression

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Read Mapping to Reference Genome

Reference-based RNA-Seq mapping

$$\text{Human genome length} = 3.3 \times 10^9\ bp$$

Release name	Date of release	Equivalent UCSC version
GRCh38 New	Dec 2013	hg38
GRCh37	Feb 2009	hg19 Old
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

Read Mapping to Reference Genome

Reference-based RNA-Seq mapping

Splice aware-aligner

Release name	Date of release	Equivalent UCSC version
GRCh38 New	Dec 2013	hg38
GRCh37	Feb 2009	hg19 Old
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

Read Mapping to Reference Genome using STAR

- STAR (Spliced Transcripts Alignment to Reference) is a splice aware-aligner.
- STAR aligns RNA-seq reads to a reference genome using uncompressed suffix arrays.

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 15–21
doi:10.1093/bioinformatics/bts635

Sequence analysis

Advance Access publication October 25, 2012

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA

Associate Editor: Inanc Birol

Alignment – output file

SAM (Sequence Alignment/Map) File: A tab-delimited text file that contains aligned data information (human readable). BAM format is binary version of SAM format.

Each alignment line has 11 fields contain information such as

- Mapping position
- Mapping quality
- Segment sequence etc.,

1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33 = Phred base quality)

```
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
    AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
    NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
    ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<7<<<< \
    MF:i:18 RG:Z:L2
```

Alignment stats summary by STAR aligner

RNASeq File ID: SL146145

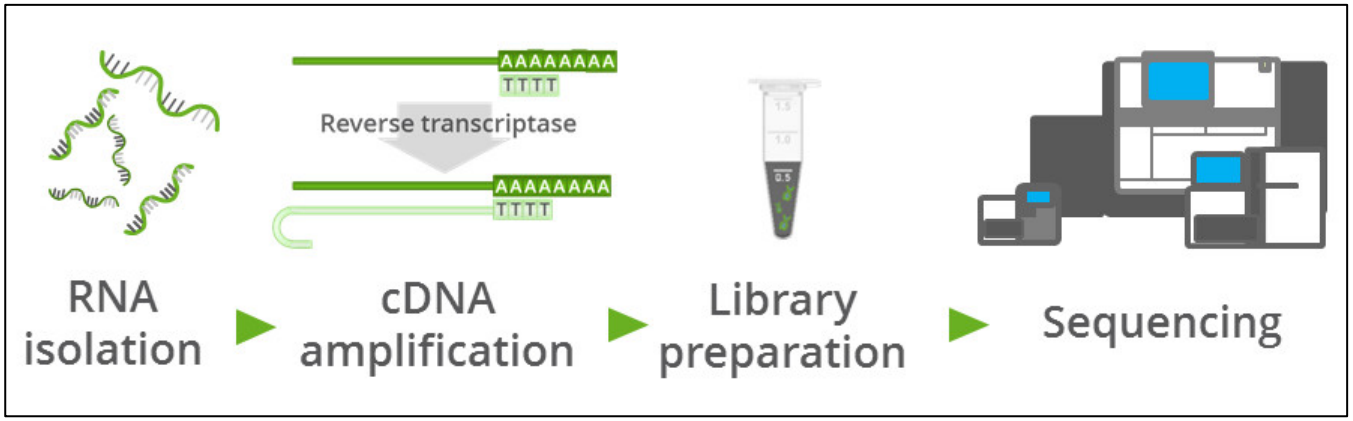
Sample ID: iDP9

Lesch-Nyhan (iPSC)

A homozygous mutation in **HPRT1 gene** (**c.508C>T**; Exon 7)

Number of input reads =
Uniquely mapped reads number +
Number of reads mapped to multiple loci +
Number of reads unmapped from BAM

Started job on	|	Feb 10 16:43:39
Started mapping on	|	Feb 10 16:49:45
Finished on	|	Feb 10 18:16:40
Mapping speed, Million of reads per hour	|	39.45
Number of input reads	|	**57154438**
Average input read length	|	**200**
UNIQUE READS:		
Uniquely mapped reads number	|	52395014
Uniquely mapped reads %	|	91.67%
Average mapped length	|	197.65
Number of splices: Total	|	37757343
Number of splices: Annotated (sjdb)	|	37017481
Number of splices: GT/AG	|	37386868
Number of splices: GC/AG	|	263606
Number of splices: AT/AC	|	29999
Number of splices: Non-canonical	|	76870
Mismatch rate per base, %	|	0.46%
Deletion rate per base	|	0.01%
Deletion average length	|	1.55
Insertion rate per base	|	0.01%
Insertion average length	|	1.36
MULTI-MAPPING READS:		
Number of reads mapped to multiple loci	|	1915136
% of reads mapped to multiple loci	|	3.35%
Number of reads mapped to too many loci	|	25254
% of reads mapped to too many loci	|	0.04%
UNMAPPED READS:		
% of reads unmapped: too many mismatches	|	0.00%
% of reads unmapped: too short	|	4.91%
% of reads unmapped: other	|	0.03%
CHIMERIC READS:		
Number of chimeric reads	|	0
% of chimeric reads	|	0.00%



FastQ

```
@HWI-ST1309F:276:C8DF6ANXX:1:1101:3694:1997  1:N:0:TAGCACC
NAAGAAAAAGAAGAACAGAAGAGAAAAAGGAGAAACCAAATTCCGGAGGCACCAAGTCAGACTCGGCATCTGATTCCCAGGAGATTAAAATTCAGCAG
+
#<<BBBFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFFBFFFFFFBFFFFFFFFFFFFF<FF<F<FFFFFFFF<BB/7F/7BFF7FFFFF</FF
@HWI-ST1309F:276:C8DF6ANXX:1:1101:4615:1997  1:N:0:TAGCACC
NTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCCGTGATGCCCAGGGAAGACAGGGCGACCTGGAAGTCCAACTACTTCCTTAAGATC
+
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFBFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFBFFBFF#
```

mapping of reads to reference genome

SAM/BAM

```
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
    AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
    NM:i:1 RG:Z:L1
```

Data analysis

1. Read quality
2. Alignment / Mapping
3. **Gene count & normalization**
4. Differential expression

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Count the number of reads mapped to each gene

- In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it.
- Count the number of reads mapped to each gene.
- Software tools: HTSeq, Cufflinks, MULTICOM etc.,

HTSeq

- Python code that converts aligned reads to counts
- Give alignment file and associated transcript file, and it will output a list of counts by feature.

Counts data by HTSeq filtering of genes

Remove lowly expressed genes

a gene is retained if it has 10 or greater counts for at least 25% of the samples.

26,490 -> 17,721 genes remained after filter

Gene	iDP9	iDP12	iTS2	iTS4	iMG3	iMG4
A1BG	195	228	207	233	142	131
A1BG-AS1	22	34	37	36	23	16
A1CF	8	8	14	16	17	16
A2M	81	72	55	36	49	107
A2M-AS1	34	28	118	69	60	24
A2ML1	446	460	83	97	335	347
A2MP1	1	1	6	8	5	4
A3GALT2	1	0	1	4	0	0
A4GALT	313	362	815	734	640	317
A4GNT	1	1	0	1	2	0
AA06	0	0	0	0	0	0
AAAS	3209	3382	3391	3443	3888	2772
AACS	3134	3235	1853	1874	3006	2725
AACSP1	12	11	82	56	31	2
AADAC	1	0	0	0	0	0
AADACL2	1	0	0	0	0	0
AADACL2-AS1	0	0	0	0	0	0
AADACL3	25	19	0	1	3	12
AADACL4	0	1	0	0	1	3
AADACP1	0	0	0	0	1	0
AADAT	828	703	1086	1102	982	882
AAED1	201	185	180	208	192	183
AAGAB	2505	2343	2279	2401	3221	2819
AAK1	1888	1747	1186	1131	1794	1864
AAMDC	429	506	565	638	526	328
AAMP	6000	5742	4680	4920	6981	5441
AANAT	13	11	11	13	19	9
AAR2	1669	1636	1670	1624	2051	1550
AARD	115	90	87	123	129	85
...	...	...	...	...	...	...
...	...	...	...	...	...	...
26490 genes						

Counts per million data

- **Trimmed Method of M-values (TMM, Bioconductor package EdgeR)**

"TMM" is the weighted trimmed mean of M-values (to the reference) proposed by Robinson and Oshlack (2010), where the weights are from the delta method on Binomial data. If refColumn is unspecified, the library whose upper quartile is closest to the mean upper quartile is used.

sample	group	lib.size	norm.factors
iDP9	LN	57925500	1.1010002
iDP12	LN	59649711	1.0504658
iTS2	LN	59491288	1.0735295
iTS4	LN	62872205	1.0482534
iMG3	Control	71008326	1.072294
iMG4	Control	55958843	1.117327
iPY5	Control	69161577	0.9692181
iPY8	Control	60218776	0.9487689
iAK2	Control	64587555	0.9914103
iAK3	Control	88089677	0.7580807
iTH29	LN	68281883	1.0151973
iTH30	LN	74956893	0.9140185

Counts per million data

Generate CPM data

$$= \frac{Counts\ per\ feature}{total\ reads} * 1000000$$

Gene	iDP9	iDP12	iTS2	iTS4	iMG3	iMG4
A1BG	3.058	3.639	3.241	3.535	1.865	2.095
A1BG-AS1	0.345	0.543	0.579	0.546	0.302	0.256
A1CF	0.125	0.128	0.219	0.243	0.223	0.256
A2M	1.270	1.149	0.861	0.546	0.644	1.711
A2M-AS1	0.533	0.447	1.848	1.047	0.788	0.384
A2ML1	6.993	7.341	1.300	1.472	4.400	5.550
A4GALT	4.908	5.777	12.761	11.137	8.405	5.070
AAAS	50.317	53.974	53.096	52.241	51.063	44.335
AACS	49.141	51.628	29.014	28.434	39.479	43.583
AACSP1	0.188	0.176	1.284	0.850	0.407	0.032
AADACL3	0.392	0.303	0.000	0.015	0.039	0.192
AADAT	12.983	11.219	17.004	16.721	12.897	14.107
AAED1	3.152	2.952	2.818	3.156	2.522	2.927
AAGAB	39.278	37.392	35.684	36.431	42.303	45.086
AAK1	29.604	27.881	18.570	17.161	23.561	29.812
AAMDC	6.727	8.075	8.847	9.680	6.908	5.246
AAMP	94.079	91.637	73.279	74.652	91.684	87.022
AANAT	0.204	0.176	0.172	0.197	0.250	0.144
AAR2	26.170	26.109	26.149	24.641	26.937	24.790
AARD	1.803	1.436	1.362	1.866	1.694	1.359

Counts per million data

Generate log2(CPM) data

Gene	iDP9	iDP12	iTS2	iTS4	iMG3	iMG4
A1BG	1.614	1.865	1.698	1.823	0.902	1.070
A1BG-AS1	-1.520	-0.872	-0.778	-0.862	-1.709	-1.945
A1CF	-2.952	-2.927	-2.165	-2.020	-2.139	-1.945
A2M	0.349	0.205	-0.209	-0.862	-0.627	0.778
A2M-AS1	-0.897	-1.150	0.889	0.071	-0.337	-1.367
A2ML1	2.807	2.877	0.382	0.561	2.139	2.473
A4GALT	2.296	2.531	3.674	3.478	3.072	2.343
AAAS	5.653	5.754	5.731	5.707	5.674	5.470
AACS	5.619	5.690	4.859	4.830	5.303	5.446
AACSP1	-2.381	-2.479	0.365	-0.229	-1.283	-4.804
AADACL3	-1.337	-1.704	-8.037	-5.719	-4.533	-2.353
AADAT	3.699	3.488	4.088	4.064	3.689	3.819
AAED1	1.658	1.564	1.497	1.660	1.337	1.551
AAGAB	5.296	5.225	5.157	5.187	5.403	5.495
AAK1	4.888	4.801	4.215	4.101	4.559	4.898
AAMDC	2.751	3.014	3.146	3.276	2.789	2.392
AAMP	6.556	6.518	6.195	6.222	6.519	6.443
AANAT	-2.268	-2.479	-2.506	-2.314	-1.981	-2.759
AAR2	4.710	4.707	4.709	4.623	4.752	4.632
AARD	0.854	0.526	0.450	0.903	0.764	0.447

Other methods used to normalize counts data by length of genes & total number of reads

- RPKM (Reads per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset
- FPKM (Fragments per kilobase of transcript per million reads of library)
  - Only relevant for paired end libraries
  - Pairs are not independent observations
  - RPKM/2
- TPM (transcripts per million)
  - Normalizes to transcript copies instead of reads
  - Corrects for cases where the average transcript length differs between samples

Normalize counts by length of genes & total number of reads

- RPKM (Reads Per Kilobase per Million of reads)

$$RPKM = \frac{number\ of\ reads\ of\ the\ region}{\frac{total\ reads}{1000000} \times \frac{region\ length}{1000}}$$

- 2000 kb transcript with 500 alignments in a sample of 55 million reads (out of which 50 million reads can be mapped)

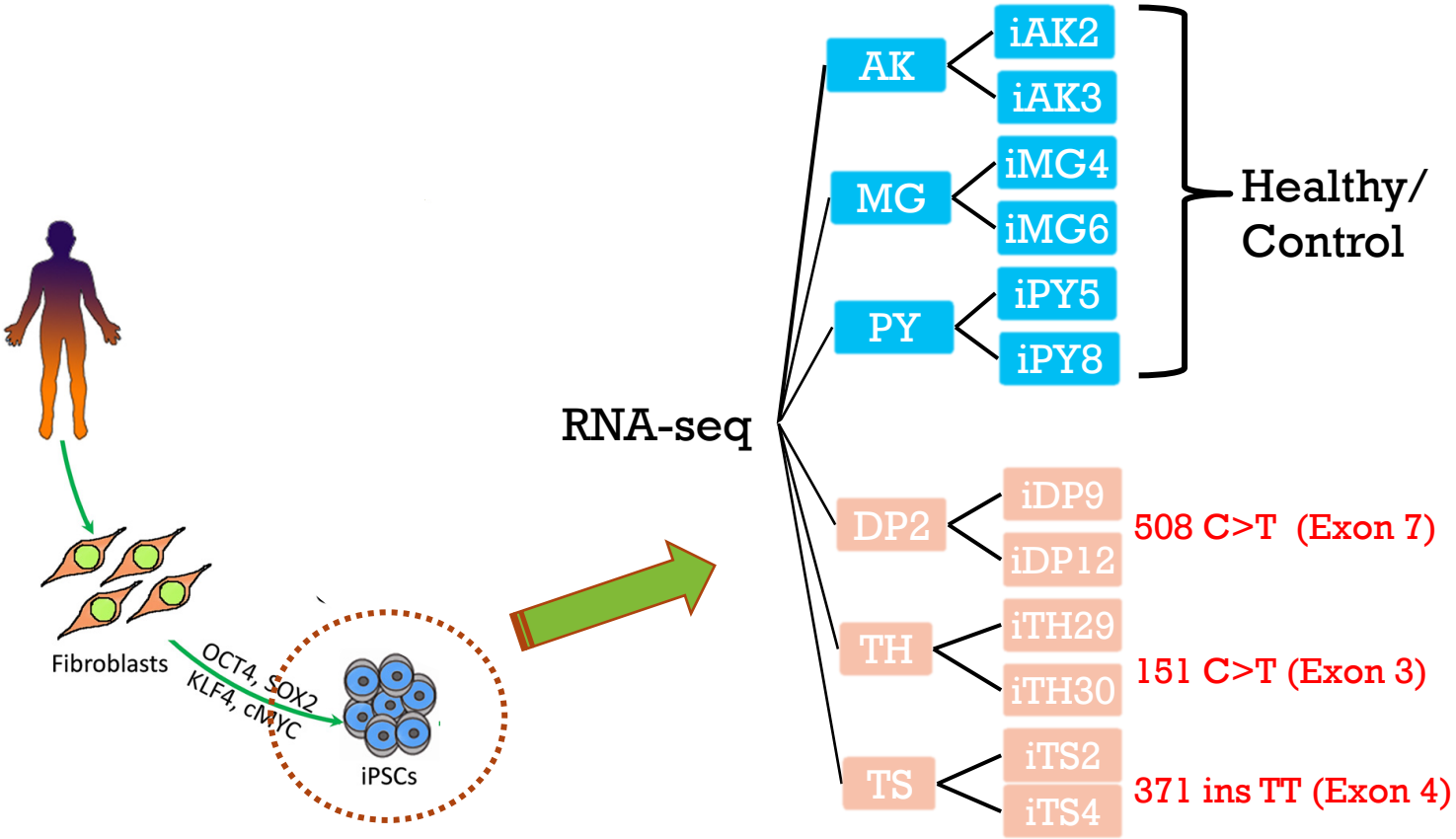
$$RPKM = \frac{500}{\frac{50000000}{1000000} \times \frac{2000}{1000}}$$

Data analysis

1. Read quality
2. Alignment / Mapping
3. Gene count & normalization
4. **Differential expression**

Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Lesch-nyhan disease specific iPSCs obtained from fibroblasts



Hyder A. (Buz) Jinnah, MD Ph.D (Professor, Neurology) & Mike Zwick, Ph.D (Asso. Professor, Human Genetics)

Differential expression analysis for sequencing count data

The methods for differential gene expression analysis from RNA-Seq can be grouped into **parametric** & **non-parametric**.

- When parametric methods are applied to differential gene expression **each expression value for a given gene is mapped into a particular distribution**, such as **Normal**, **Poisson** or **negative binomial**.

- Non-parametric methods can capture more details about the data distribution, i.e., not imposing a rigid model to be fitted.

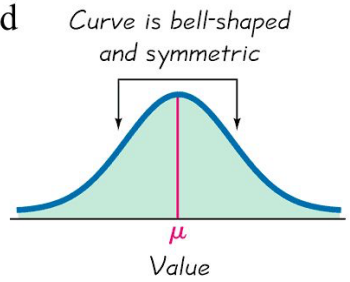
STATISTICAL DISTRIBUTIONS - GAUSSIAN

- Gaussian (normal) distributions
  - nice and easy to work with
  - describe smooth distributions
  - underlie the **t-test** (among others)

If a continuous random variable has a distribution with a graph that is symmetric and bell-shaped and can be described by the equation

$$y = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

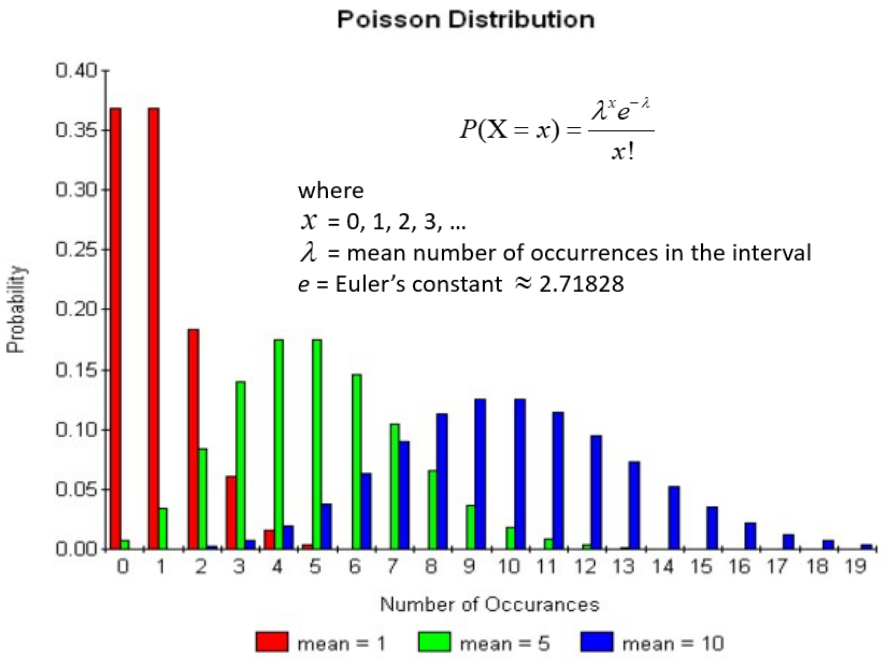
we say that it has a **normal distribution**.



- The dispersion in this case is equal to the standard deviation
- You completely specify this distribution by the mean (μ) and the standard deviation (σ).

POISSON DISTRIBUTIONS

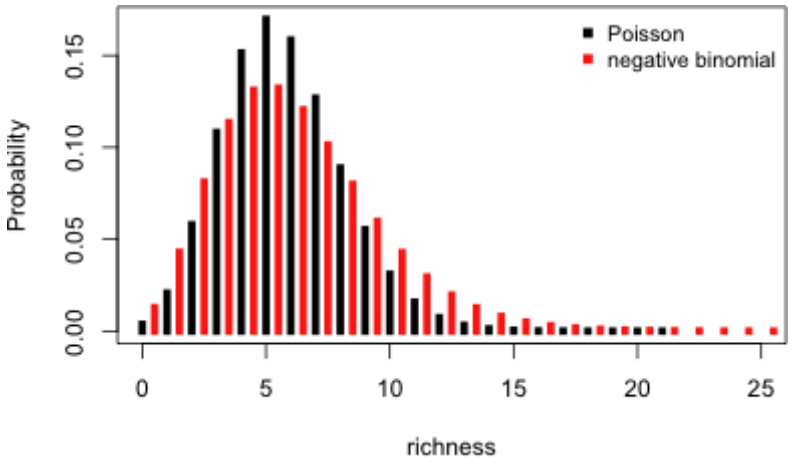
- Poisson distributions
  - like a gaussian for non-smooth distributions
  - describes things like stars in a small area on the sky
  - for very large numbers, this looks like a gaussian distribution



- The dispersion in this case is equal to the mean (λ)
- You completely specify this distribution by the mean

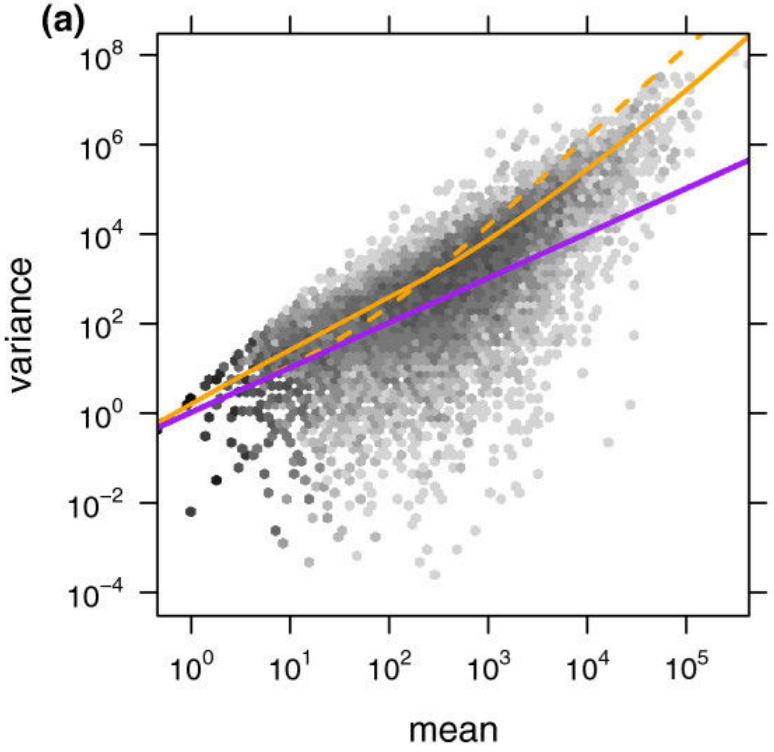
NEGATIVE BINOMIAL

- Negative Binomial distributions
  - like a poisson but allows the variance to be different from the mean
  - often called "over-dispersed" poisson distribution
  - for very large numbers, this looks like a gaussian distribution



- The dispersion in this case is measured empirically from the data

RNA-Seq data fits a Negative Binomial distribution



poisson

negative binomial fit

Anders and **Huber**, **G enome Biology**, **2010**)

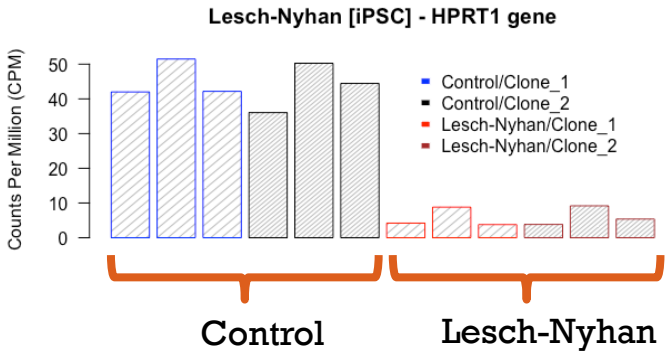
Molly Hammell

- How do we know? Because, when you measure variance (per gene, between replicates), it's not equal to the mean, and it's not even a good linear fit.

HPRT1 expression pattern in iPSCs

iPSC			
Gene	Fold Change	p-Value	FDR
FAR2P1	-4.156	3.34E-08	0.0006
PPIP5K1	0.504	2.27E-06	0.0193
AAGAB	-0.243	3.21E-06	0.0193
RNF39	2.027	8.57E-06	0.0381
VDAC3	-0.114	1.06E-05	0.0381
FAXC	-0.633	1.57E-05	0.0472
HPRT1	-3.011	1.92E-05	0.0494
LY75	1.712	2.63E-05	0.0517
MED6	-0.269	2.92E-05	0.0517
WASH3P	0.645	3.08E-05	0.0517
ITIH4	0.873	3.52E-05	0.0517
ALOX12B	0.825	3.67E-05	0.0517
POMZP3	1.044	3.74E-05	0.0517
MTRNR2L2	-1.215	4.67E-05	0.0576
DLEU2	-0.617	4.80E-05	0.0576
TUBB3	-0.668	6.92E-05	0.0778

FDR < 0.1, 16 Genes



- HPRT1 gene expression pattern reveal how iPSCs retained a "memory" of their tissue of origin even after undergoing reprogramming.

THANK YOU