



# Low level Processing and Visualization of Genome Sequencing Data

## GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.  
Professor, Division of Infectious Disease  
Director, Center for Data Science  
Rutgers University – New Jersey Medical School

2025-04-30

## Motivating Example: X-linked Disease

# Rare X-linked Disease



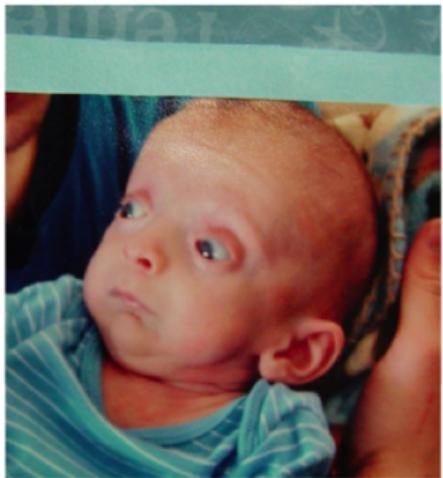
# Rare X-linked Disease



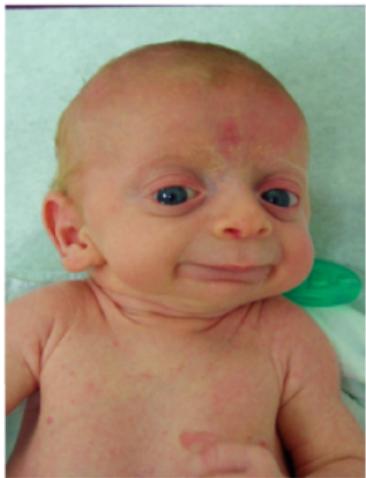
**Uncle #1**



**Uncle #2**

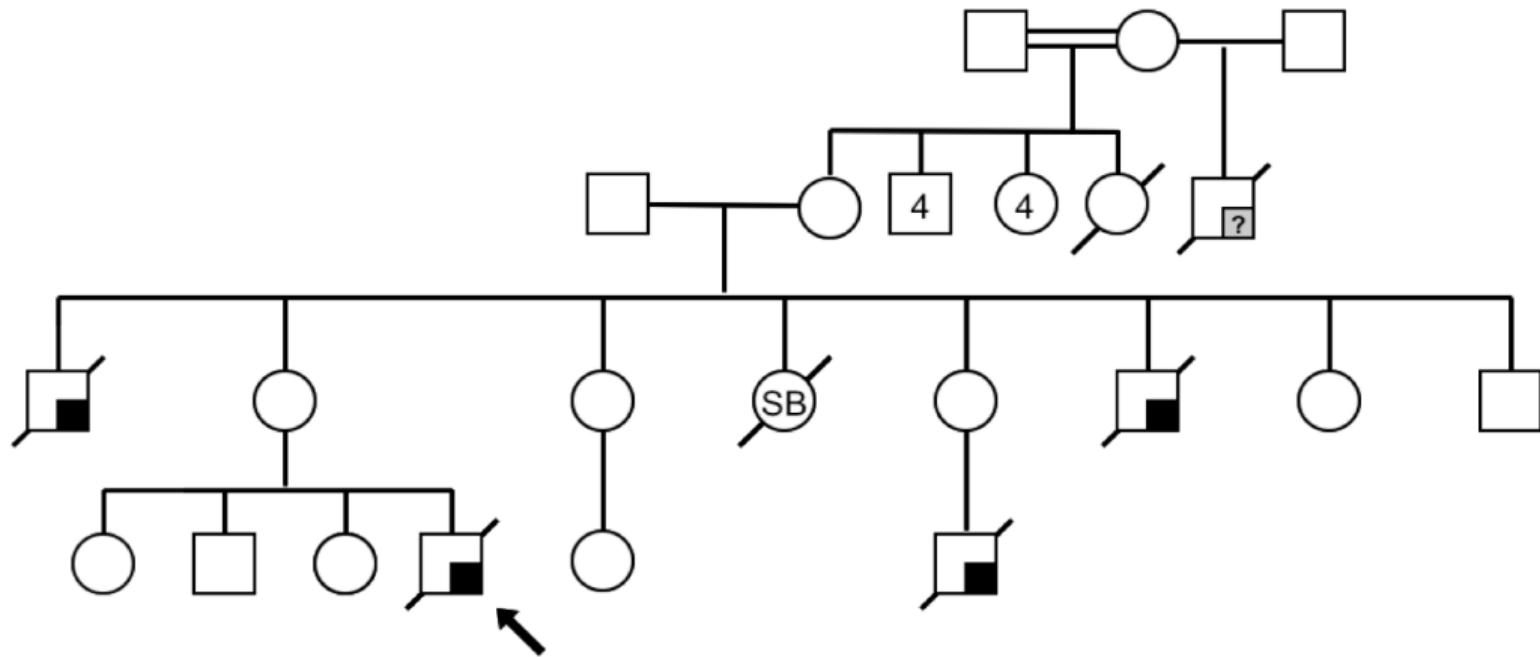


**cousin**

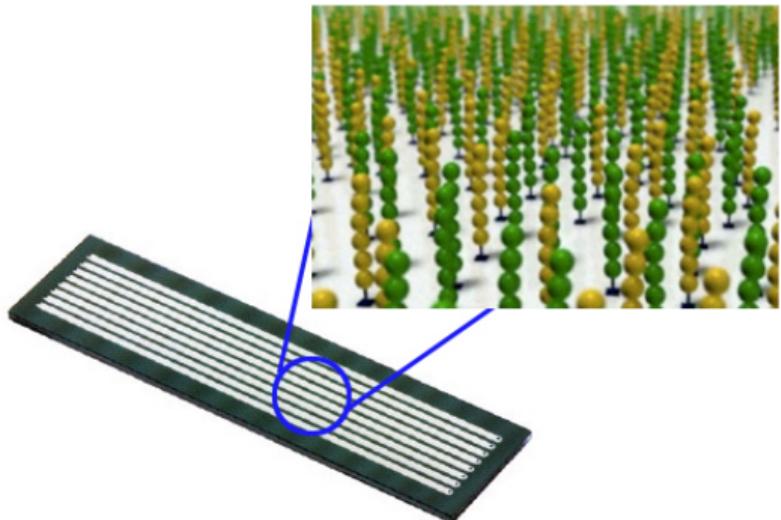


**proband**

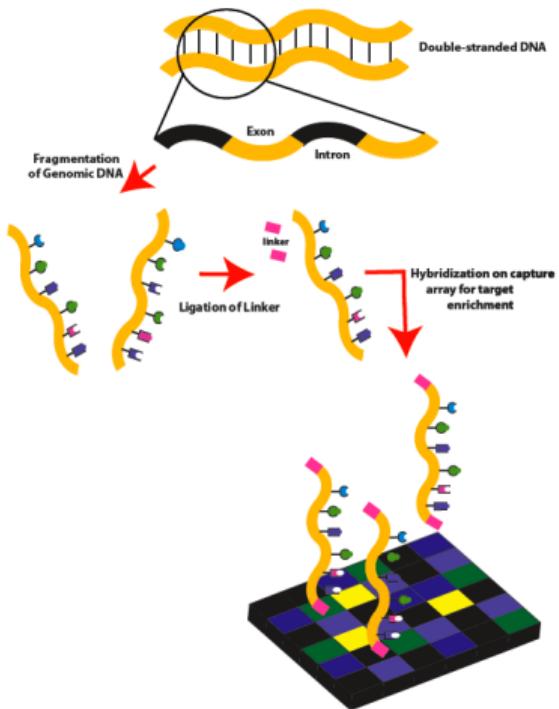
# Rare X-linked Disease



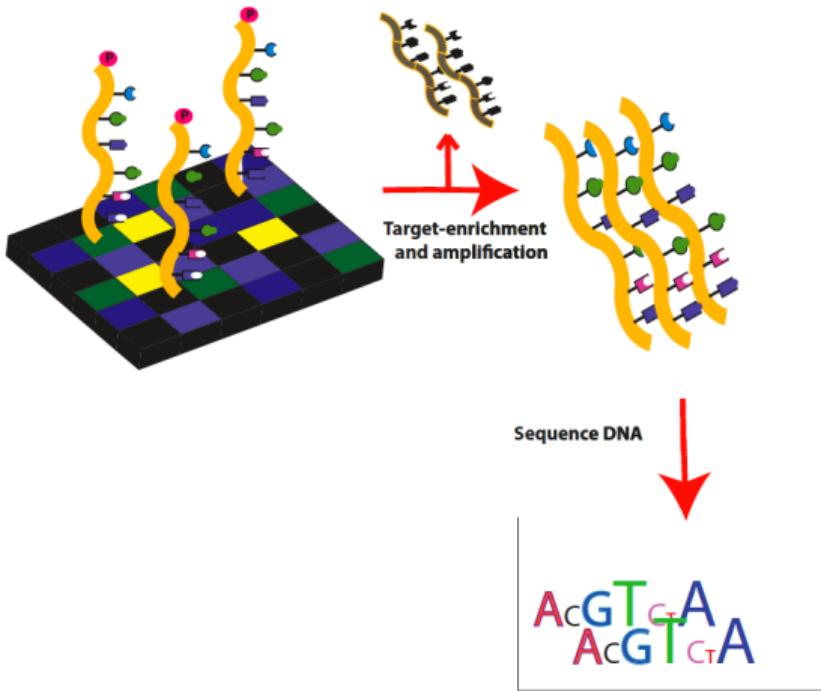
# Exon Capture Sequencing



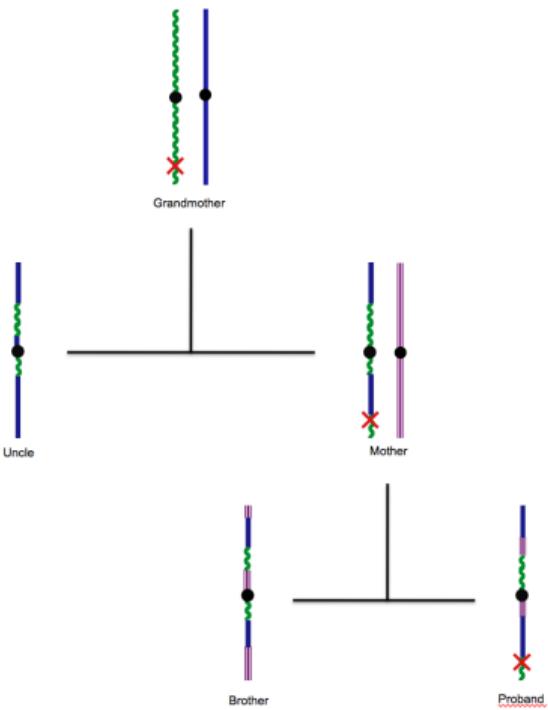
# Exome Capture



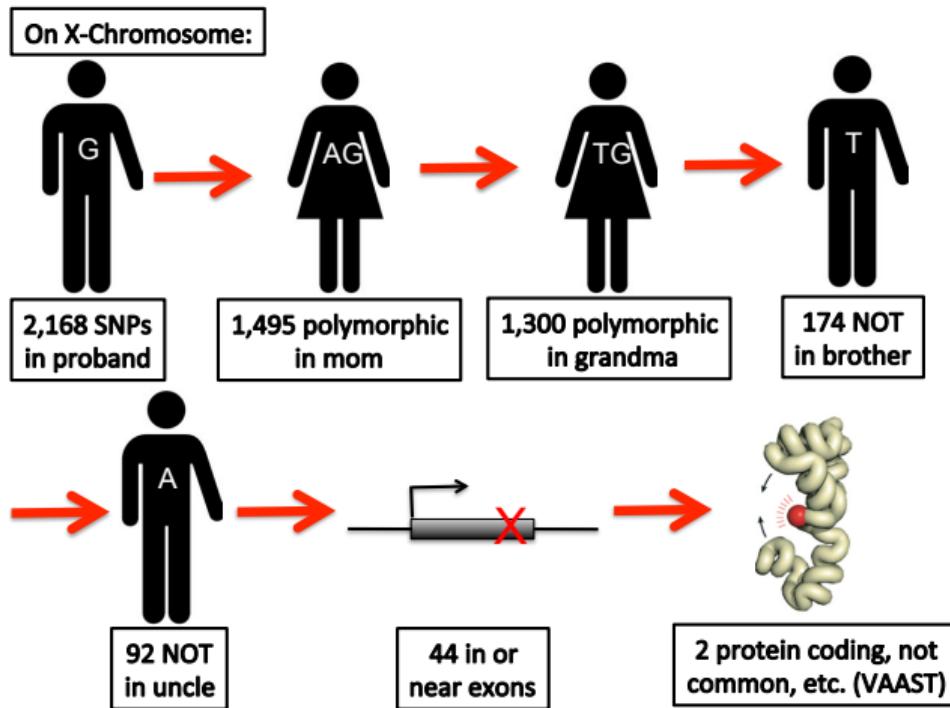
# Exome Capture



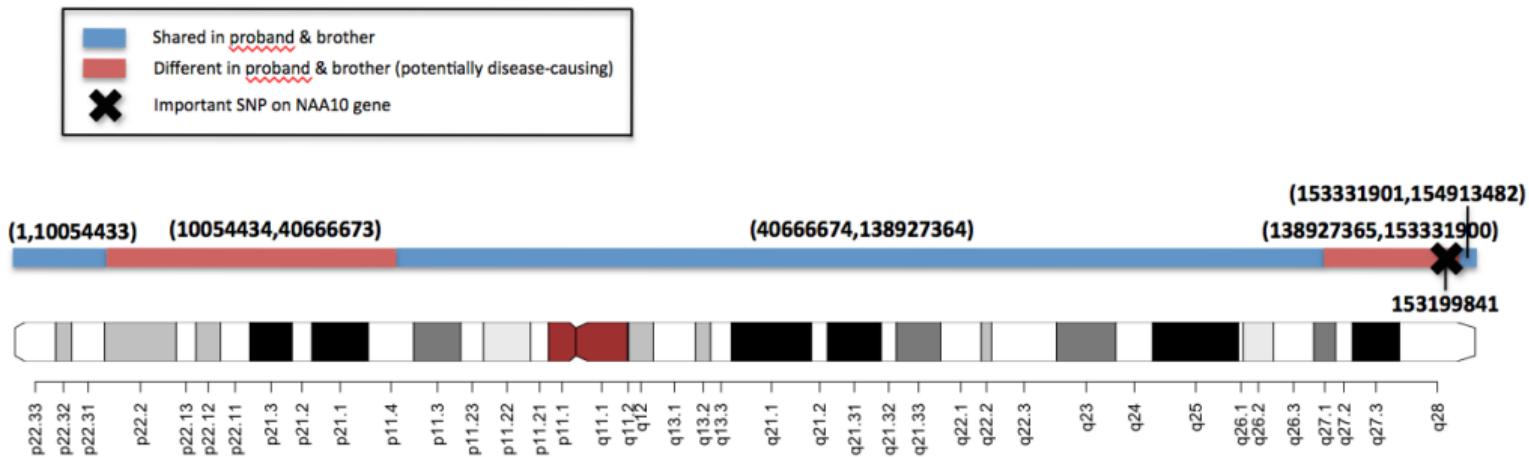
# Rare X-linked Disease



# Rare X-linked Disease



# Rare X-linked Disease



# Rare X-linked Disease

AJHG  Supports open access

Submit Log in Register Subscribe Claim

ARTICLE | VOLUME 89, ISSUE 1, P28-43, JULY 15, 2011 [Download Full Issue](#)

 PDF [603 KB]  Figures  Save  Share  Reprints  Request

## Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope • Kai Wang <sup>19</sup> • Rune Ejventh • ... Mark Yandell • Thomas Arnesen • Gholson J. Lyon                     <img alt="Wiley-Interscience icon" data-bbox="7835

# N-terminal acetyltransferase (NAA10)

N-terminal acetyltransferase (NAT):

- ▶ Common modification (~80-90% of human proteins)
- ▶ Depletion from cancer cells linked to cell cycle arrest and apoptosis (Starheim, *BMC Proc* 2009)
- ▶ NAT genes directly implicated as cause of genetic disease
- ▶ Mutation demonstrated a significantly impaired biochemical activity *in vitro*
- ▶ NAA10 lethal if knocked out of Drosophila

# Generating Sequencing Data

# Next-Generation Sequencing

- ▶ Expensive to purchase (hundreds of thousands \$USD)
- ▶ Expensive to operate (e.g. reagents, flow cells)
- ▶ You can sequence your genome at 30X depth for <\$1000 USD.

Roche 454



Ion Torrent



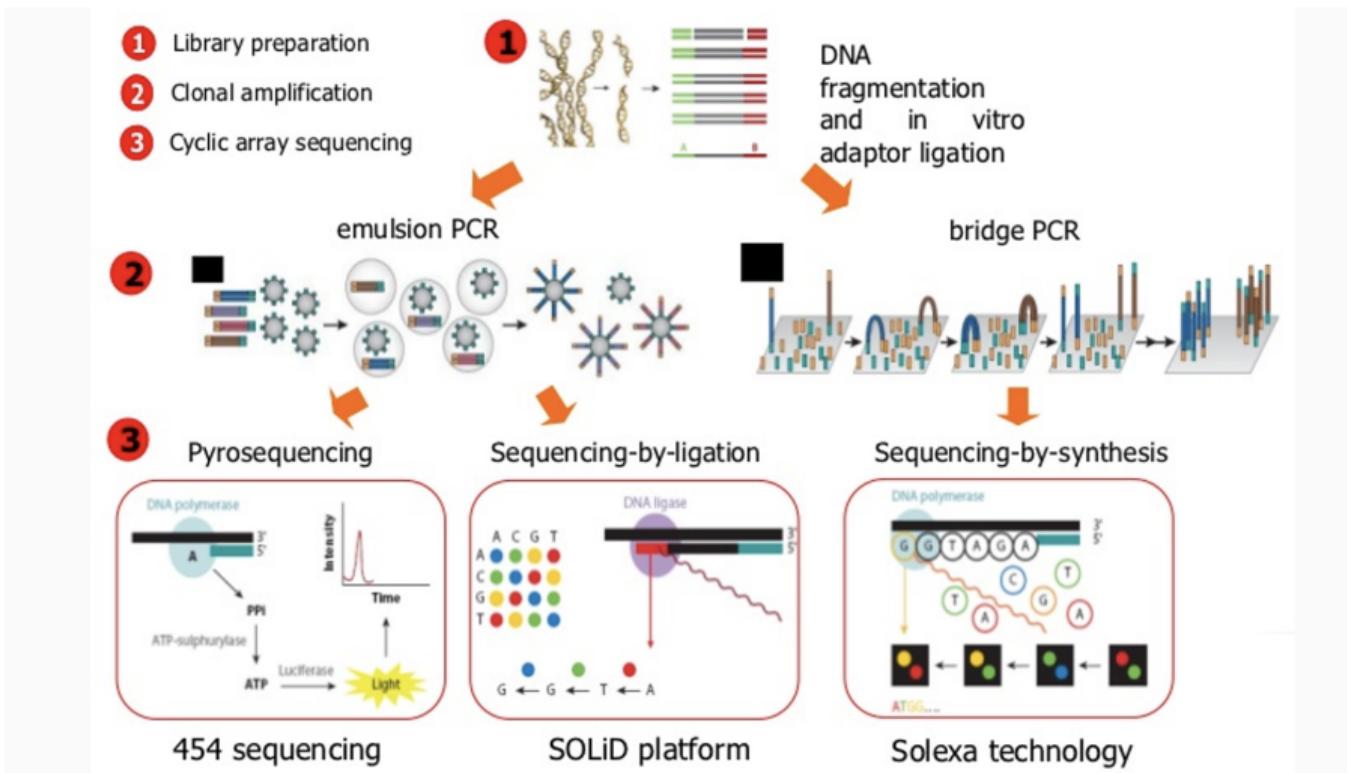
Illumina HiSeq 2500



Illumina NovaSeq 6000



# Next-Generation Sequencing



# Illumina Sequencing

- ▶ Most common sequencing technology today
- ▶ Sequences any DNA
- ▶ Sequencing by synthesis method
- ▶ Sequences (reads) are short (<300bp)
- ▶ 2 gigabases - 6 terabases per run
- ▶ Hours to days to complete one run

For more information, you can watch the following:

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

# Illumina Sequencing

- Sequencing occurs on a **flow cell**
- Each flow cell has 1 to 8 **lanes**
- Number of reads for overall flow cell varies
- Length of reads is fixed (e.g. 250 bp)
- Read format:
  - **Single end** - one read per molecule
  - **Paired end** - two reads per molecule
- **Multiplexing:** sequence many samples at once using molecular barcodes



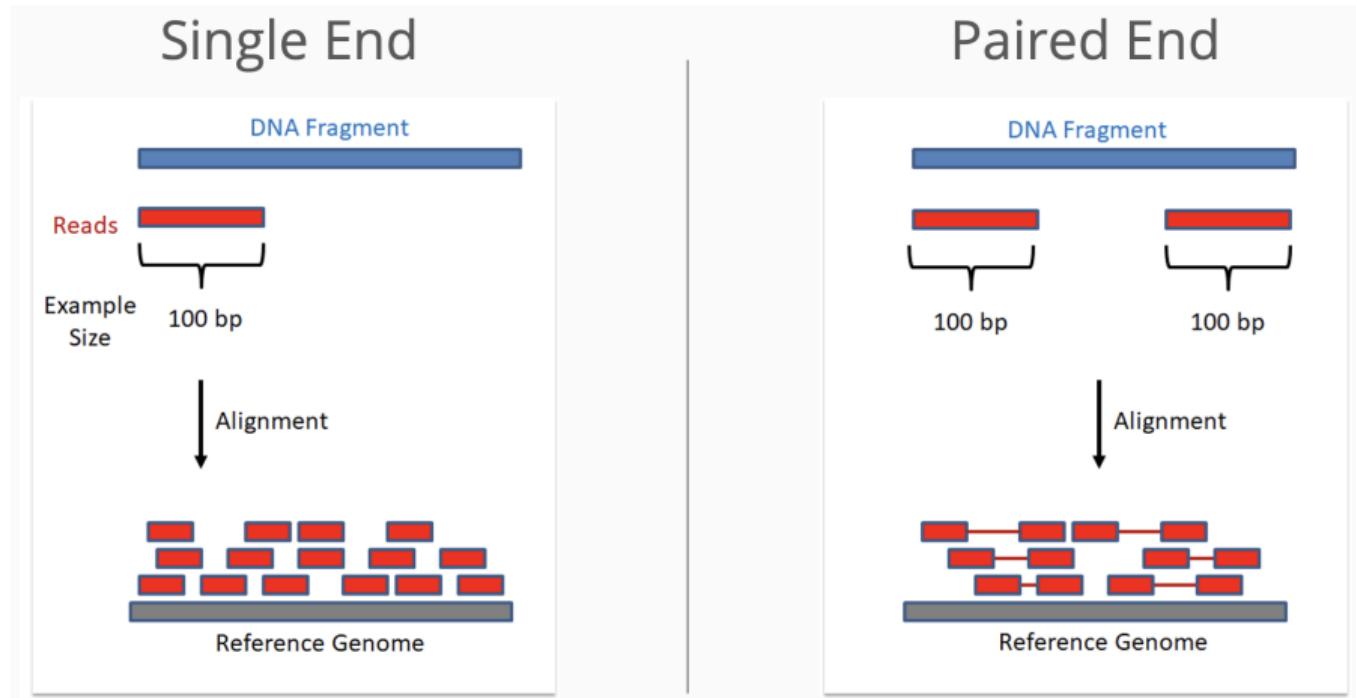
NovaSeq Flowcell  
Courtesy of Illumina, Inc.

# Sequencing Library Generation Workflow

## Sequencing Workflow:

- ▶ Extract RNA/DNA from sample
  - ▶ If RNA, reverse transcribe to cDNA
- ▶ Size select using gel cut or random shearing
- ▶ PCR amplify DNA if concentration is low
- ▶ Add sequencing adapters
  - ▶ If multiplexing, use barcoded adapters
- ▶ Pool samples, load across flow lanes for sequencing
- ▶ Typically only perform 1, sequencing cores

# Design Choice: Single End vs Paired End



# Design Choice: Number of Reads

**Table 1: Coverage and Read Recommendations by Application**

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELS	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev et al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELS	not recommended	Feng et al., 2014
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
DNA Target-Based Sequencing	ChIP-Seq	10-14M (sharp peaks); 20-40M (broad marks)	Rozowsky et al., 2009; ENCODE 2011 Genome; Landt et al., 2012

<https://genohub.com/recommended-sequencing-coverage-by-application/>

# Design Choice: Sequencing Depth

## Whole Exome

- ▶ Less expensive
- ▶ Nearly complete ascertainment of variation in the coding ~1% of the genome (i.e. exons)
- ▶ Will miss functional variants outside of the coding region

# Design Choice: Sequencing Depth

Whole Exome Low Coverage Whole Genome

- ▶ Less expensive
- ▶ Reasonably good ascertainment of shared variation, but not unique variation

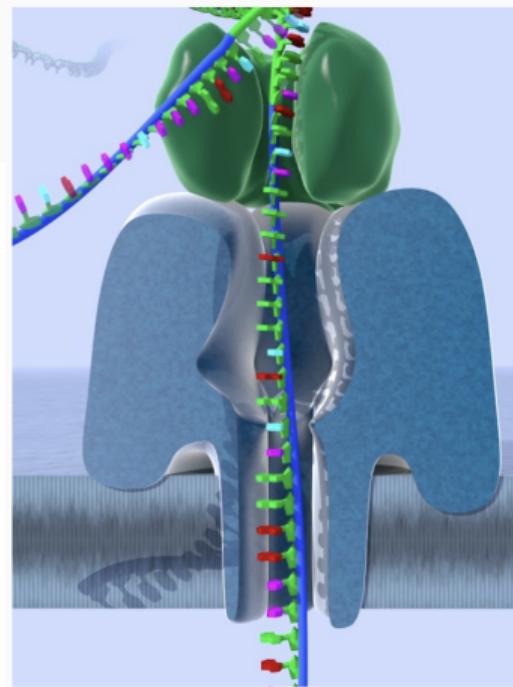
# Design Choice: Sequencing Depth

Whole Exome Deep Whole Genome

- ▶ More expensive
- ▶ Capture most of the genetic information
- ▶ Sequence the entire genome of each subject

# 3rd Gen Sequencing: Oxford Nanopore

- “Real time single molecule”



<https://www.youtube.com/watch?v=GUb1TZvMWsw>

# Sequencing Data Formats

# Biological Data Formats are standardized

- ▶ Many different types of data
- ▶ Standard formats exist for many of them
- ▶ Always use/extend standard formats
- ▶ Don't save data in non standard formats when standards are available - No .xlsx!
- ▶ For complete list of formats see:  
<https://genome.ucsc.edu/FAQ/FAQformat.html>

# Sequencing Data Formats

format	data	tool(s)
FASTA	sequence of nucleotides	Samtools, Biopython
FASTQ	sequenced reads	FASTQC, Biopython
SAM/BAM/CRAM	aligned reads	Samtools, Deeptools, PySam
VCF	variant calls	vcftools bedtools
BED / BED-PE	genomic regions	bedtools
GFF	general features	Biopython, Manual Parsing
GTF	gene features	Biopython, Manual Parsing

# Sequencing Data Formats

format	data	tool(s)
FASTA	sequence of nucleotides	samtools faidx
FASTQ	sequenced reads	-
SAM/BAM/CRAM	aligned reads	samtools
VCF	variant calls	vcftools bedtools
BED / BED-PE		bedtools
GFF		-
GTF	gene features	-

Contains sequences  
of nucleotides / AA /  
bases

# Sequencing Data Formats

format	data	
FASTA	sequences	fastq
FASTQ	sequences	fastq
SAM/BAM/CRAM	aligned reads	samtools
VCF	variant calls	vcftools bedtools
BED / BED-PE	genomic regions	bedtools
GFF	general features	-
GTF	gene features	-

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome Y overlapping with all the Alu elements.

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs on chromosome Y overlapping with all the Alu elements.

FASTQ format

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs and output the SNPs on chromosome Y overlapping with all the Alu elements.

FASTA format

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome 1 save in BAM format with all the Alu elements.

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome VCF format with all the Alu elements.

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome Y overlapping with all the Alu elements.

Read from GTF/GFF/BED

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome Y overlapping with all the Alu elements.

BWA/BOWTIE2

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome 1 using samtools, and then merge with all the Alu elements.

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome Y over the Alu elements.

samtools / GATK

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome Y overlapping with all the Alu elements.

samtools view

# Data Formats Example

## EXAMPLE

Align reads to the reference, sort and index, call SNPs, extract the SNPs on chromosome Y **overlapping** with all the Alu elements.

bedtools intersect

## Raw Sequencing Data and QC

# History and Evolution of Illumina Data Output

Illumina sequencers have given output in many different formats:

- ▶ Illumina .PRB and .INT files
  - ▶ Better access to raw data.
  - ▶ Base calling algorithms (Bravo and Irizarry, *Biometrics*, 2010)
  - ▶ Mapping algorithms (GNUMAP, NOVO)
  - ▶ Confusing formats; Large data files
- ▶ Illumina .FASTQ files
- ▶ Sanger .FASTQ files

# Illumina .INT and .PRB

1	1	125	771	1651.8	2189.6	228.1	549.9	219.0	202.5	48.4	3016.8	127.8	6.1	2046.0	1709.3	155.4	215.1	1936.2	1472.2		
1	1	478	16	1050.0	969.9	149.5	311.7	0.0	0.0	39.3	134.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	470.5	0.0	
1	1	780	553	639.6	980.8	555.2	6412.8	1040.1	4408.7	750.3	638.1	946.2	4351.1	712.6	1244.1	2808.9	4518.5	329.0	946.2		
1	1	123	685	-116.5	341.5	-14.0	-985.9	231.5	1090.0	88.3	-102.2	240.2	513.6	3.1	-282.3	19.1	-86.6	41.2	-6.5		
1	1	61	934	40.7	87.3	38.5	21.3	16.4	31.7	100.9	68.8	41.4	29.4	40.4	30.1	79.1	3.5	52.3	75.9	53.0	43.6
1	1	866	972	2820.5	4698.9	435.8	8502.2	4740.3	4890.2	1491.5	1241.7	2137.5	2505.6	6653.8	4579.8	2243.2	3102.6	5157.2	2819.4		

40	-40	-40	-40	-40	-40	-40	40	-40	-40	40	-40	-40	-40	-40	-40	-40	-40	-40	40
28	-28	-40	-40	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
-17	-16	-30	13	-40	26	-26	-40	-21	17	-20	-27	21	-21	-40	-40	-40	-40	-40	-26
-40	40	-40	-40	-40	40	-40	-40	-1	1	-40	-40	-0	-40	-0	-14	26	-26	-40	-40
-1	-40	-0	-10	-20	-18	16	-35	-1	-8	-2	-13	12	-34	-13	-26	5	-10	-12	-11
-27	-40	-40	27	40	-40	-40	-40	-40	-40	40	-40	-40	-40	-40	-40	40	-40	40	-40

.fasta or .fa

```
>chrX
ttgaactcctgacctcaggtgatccgcggccttgcacctccaaagcgct
cctgcctcagcctcccagtagctggactacaggtgcctgccaccatgc
....
caggctaatttttgtatttttagtagagacggggtttcaccatgttagc
cagcatggtctcaatctcctgacctcatgatccgcctgcctcggccccc
>chrY
tgtacacttaaatgggtgaatttatggaatgtgaattataCGTGTG
CTTGTAAAAAAATGATGGAGATGGAGACGTGACTCTAGCGTGAAGGGG
...
GTGGGGAGAGTAGATCTAGAGTGGAGACACCACCTTTAGGAGGTATGATC
cctgcaccatgcctaaattttgtatttttagtagagacacggat
```

# .fastq or .fq (or .fq.gz)

```
@HWI-EAS240_0001:2:1:1142:17571#0/1
CTCTCTTTCTCCCCANGTCTCCTCATGACCATATCCNTGTTGTCCATTGTGTANGNNNCTTCGCGCTGCAACATC
+HWI-EAS240_0001:2:1:1142:17571#0/1
bababb`abbbbabbYB^[^`_`_bbbbbbbbb`_B`_``^`^bbbba^`^B00BBBN0000bY^`^bbbbbbb
@HWI-EAS240_0001:2:1:1142:6453#0/1
CAGGACGTGCACTATGCATCCACGATGCAAGTCTTANCATTATTAGGATACANATANNGCAAACGTGTAATT
+HWI-EAS240_0001:2:1:1142:6453#0/1
_U] [aa_aa^`^`^`[B] [] []] S``Y`a`aaZ]B] ]^`^`[`^`^`[OY]XBNNNBB]]]]]L^`^`^`^`^`[V`^
@HWI-EAS240_0001:2:1:1142:19443#0/1
TCAGAAAACAGAAAGGTCTTTCTTACTTCTTGCANGATGCCACCCTCCAGANCAGNNATTGCTACTTGCAC
+HWI-EAS240_0001:2:1:1142:19443#0/1
bbbabb^`bbb^`bbabb]bb]^`]]] bbbbbbYYBYX00000bb`bbVY^HB000BB0J0L0_bbbb]b_b`a
```

# .FASTQ Paired End

fastq\_SRR1997469\_1.fastq

```
@SRR1997469.1 1 length=36
CAGTCTTCTTAGAAATATCCACTTCGGAATAAAAGA
+SRR1997469.1 1 length=36
BBBBBFFFFF<FFFFFFFFFFFFFFBFFFFFFFFFFF
@SRR1997469.2 2 length=36
ACAGTTAACGATCCTTACAGANAGNAGNCTNGTA
+SRR1997469.2 2 length=36
<BBBBFFFBBF<BFF/<FFFFF##########
...
```

fastq\_SRR1997469\_2.fastq

```
@SRR1997469.1 1 length=36
AGATAAGATGGTAATCTTGATGGAGAACATTAAGA
+SRR1997469.1 1 length=36
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR1997469.2 2 length=36
ACTGGGAANCCTTCTGTCTAGCCTTATATGAAAAAA
+SRR1997469.2 2 length=36
BBB/BFFB#BB/FFFF/FFFFFBBFBBBBBBBBBF
...
```



## .FASTQ Comparison

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)

[https://en.wikipedia.org/wiki/Phred\\_quality\\_score](https://en.wikipedia.org/wiki/Phred_quality_score)

# Sequence Quality Score (PHRED)

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

# Sequence Quality Score (PHRED)

## Quality scores (Phred)

- ▶ Sanger Phred: Range=(0,40),  $P = 1 - 10^{-(ASCII-33)/10}$
- ▶ Solexa: Range= (-5,40),  $P = \frac{10^{(ASCII-64)/10}}{1+10^{(ASCII-64)/10}}$
- ▶ Illumina 1.3: (0,40),  $P = \frac{10^{(ASCII-64)/10}}{1+10^{(ASCII-64)/10}}$
- ▶ Illumina 1.5: Range=(2,40),  $P = 1 - 10^{-(ASCII-64)/10}$
- ▶ Illumina 1.8: Same as Sanger except Range=(0,41)

# .FASTQ Comparison

> Sanger			> Solexa			> Illumina1.3			> Illumina1.5		
	ASCII	Quality		ASCII	Quality		ASCII	Quality		ASCII	Quality
!	33	0.0000	;	59	0.2403	@	64	0.5000	B	66	0.3690
"	34	0.2057	<	60	0.2847	A	65	0.5573	C	67	0.4988
#	35	0.3690	=	61	0.3339	B	66	0.6131	D	68	0.6019
\$	36	0.4988	>	62	0.3869	C	67	0.6661	E	69	0.6838
%	37	0.6019	?	63	0.4427	D	68	0.7153	F	70	0.7488
&	38	0.6838	@	64	0.5000	E	69	0.7597	G	71	0.8005
'	39	0.7488	A	65	0.5573	F	70	0.7992	H	72	0.8415
(	40	0.8005	B	66	0.6131	G	71	0.8337	I	73	0.8741
)	41	0.8415	C	67	0.6661	H	72	0.8632	J	74	0.9000
*	42	0.8741	D	68	0.7153	I	73	0.8882	K	75	0.9206
+	43	0.9000	E	69	0.7597	J	74	0.9091	L	76	0.9369
,	44	0.9206	F	70	0.7992	K	75	0.9264	M	77	0.9499
-	45	0.9369	G	71	0.8337	L	76	0.9406	N	78	0.9602
			H	72	0.8632	M	77	0.9523	O	79	0.9684
			I	73	0.8882	N	78	0.9617	P	80	0.9749

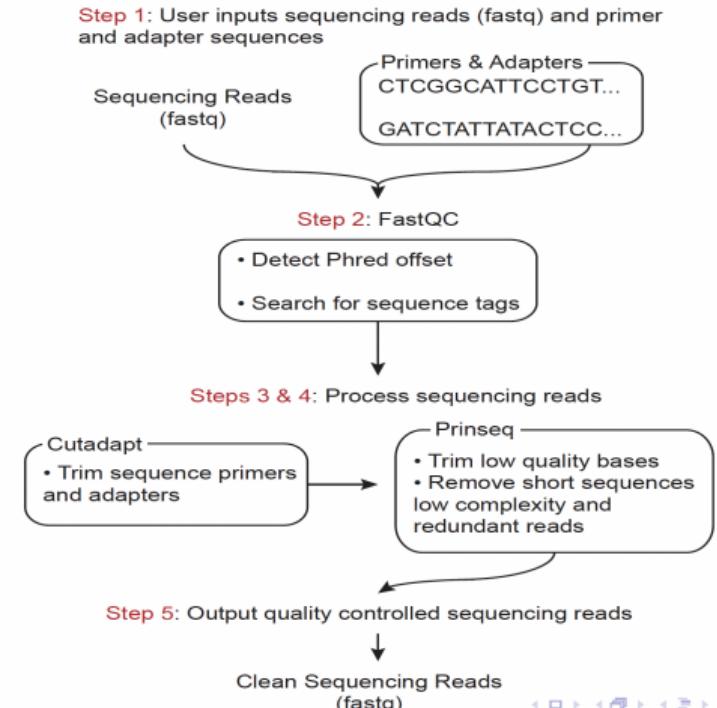
# Quality Control

Need to preprocess the reads:

- ▶ Check for quality (FASTQC)
- ▶ Trim adapter and (Cutadapt, others)
- ▶ Remove duplicate reads, trim low complexity/quality bases/reads (Prinseq)
- ▶ Complete pipelines: NCBI Toolkit, QC-Chain, PathoQC ([pathoscope.sourceforge.net](http://pathoscope.sourceforge.net)), others

**Note:** Not comprehensive or updated!

# Read Quality Checks

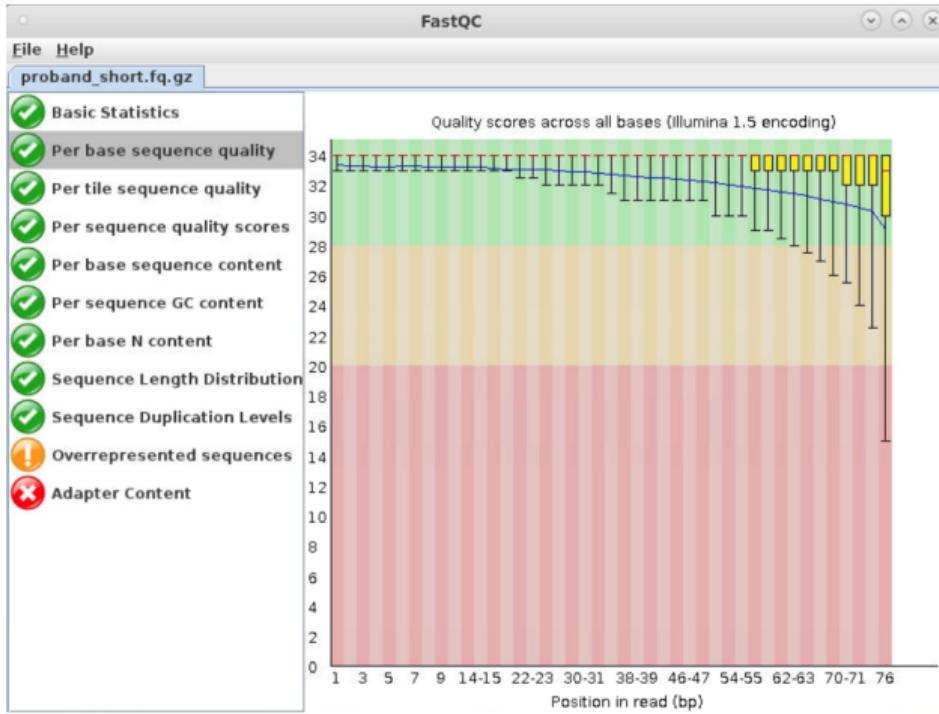


# Read Quality Checks (Outdated!)

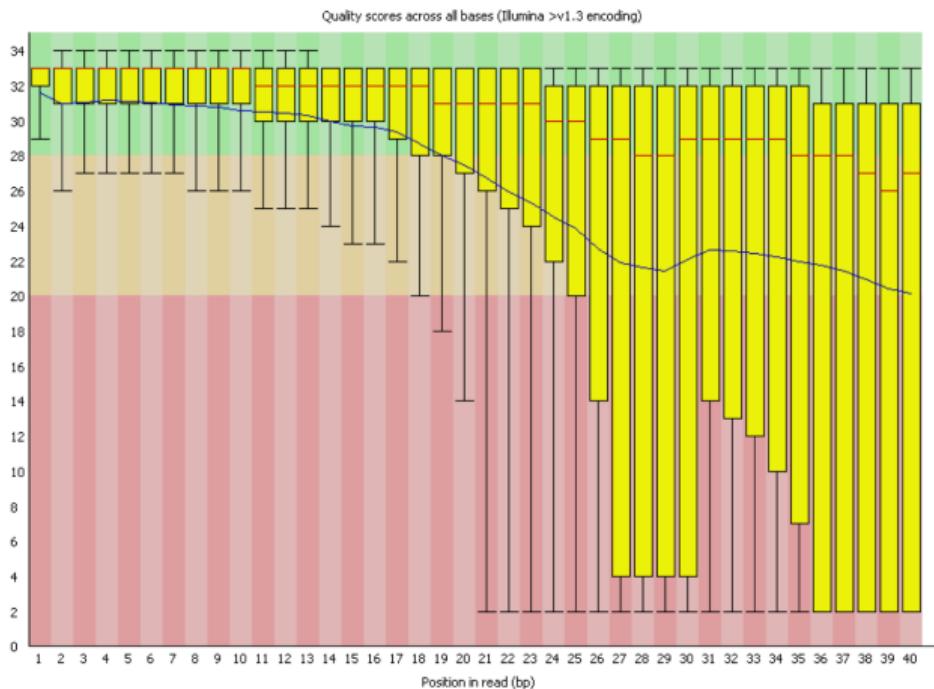
Comparison of some of the more popular QC pipelines:

QC Features	QCToolkit	QC-Chain	Cutadapt	Prinseq
Parallel computation	X	X		
Phred offset detection	X	X		
Tag sequence removal	X	X	X	
Poly-A/T tail trimming			X	X
PCR duplication filtering		X		X
Low complexity filtering				X
Homopolymer removal	X			X
GC content filtering		X		X
N/X content filtering				X

# FASTQC



# FASTQC



# FastQC Example

Interactive GUI from Amarel Desktop:

```
module spider fastqc  
#module avail fastqc  
module load FastQC  
fastqc
```

# FastQC Example

Running FastQC from command-line (single file):

```
fastqc myfastqfile.fq.gz --outdir=output/
```

Running FastQC from command-line (multiple files):

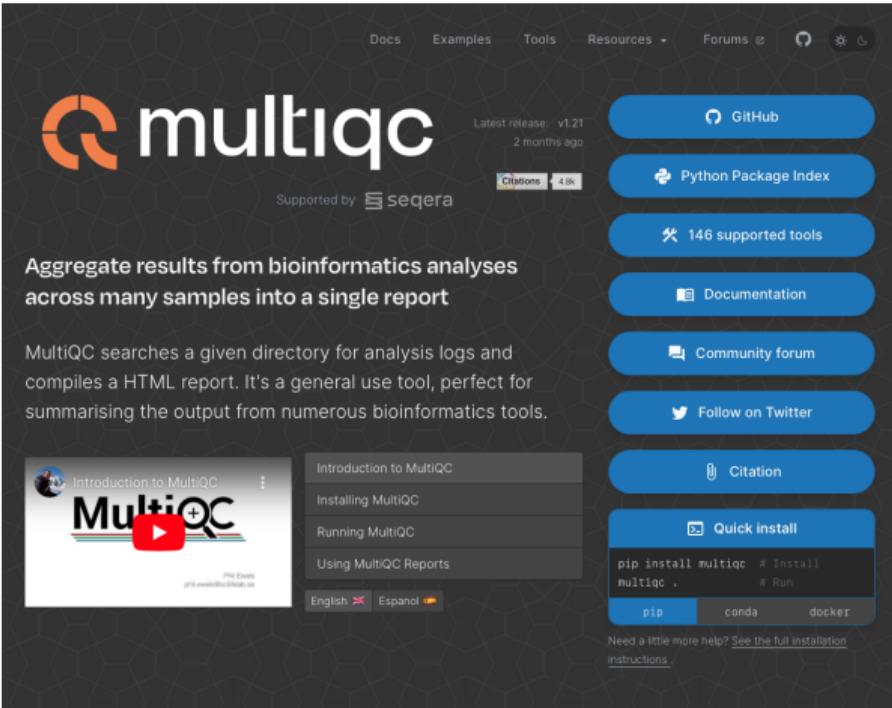
```
fastqc *.fq.gz --outdir=output/
```

# MultiQC Example

You can use multiqc to combine FastQC results. On Amarel Desktop:

```
module load miniconda
pip install multiqc
cd my_fastqc_dir
multiqc .
```

# MultiQC Example



The screenshot shows the official MultiQC website. At the top, there's a navigation bar with links for Docs, Examples, Tools, Resources, Forums, and a search bar. Below the header, the MultiQC logo is prominently displayed, followed by the text "Supported by seqera". To the right of the logo, it says "Latest release: v1.21 2 months ago" and "Citations: 4.8k". A vertical sidebar on the right contains links to GitHub, Python Package Index, 146 supported tools, Documentation, Community forum, Follow on Twitter, Citation, and Quick install. The Quick install section shows the command: `pip install multiqc # Install  
multiqc . # Run`, with options for pip, conda, and docker. On the left side, there's a video player showing a "Introduction to MultiQC" video and a sidebar with links to Introduction to MultiQC, Installing MultiQC, Running MultiQC, and Using MultiQC Reports. Language options for English and Spanish are also present.

# MultiQC Example

**multicq**

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-05-08, 00:23 MDT based on data in:

- fastQC
- Sequence Counts
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distributions
- Sequence Duplication Levels
- Oversampled sequences by sample
- Top oversampled sequences
- Adapter Content
- Status Checks
- Software Versions

Welcome! Not sure where to start? [Get a summary view](#) (0.00)

**General Statistics**

General Statistics

Sample Name    % Depth    % GC    M Seqs

Sample Name	% Depth	% GC	M Seqs
brother_short	71.1%	31%	1.2M
grandmother_short	72.6%	31%	1.2M
mother_short	72.0%	32%	1.2M
proband_29	98.6%	31%	0.8M
proband_short	97.9%	47%	1.2M
wife_short	71.9%	31%	1.7M

[Export as CSV](#)

**FastQC** Version: 0.13.0

FastQC is a quality control tool for high-throughput sequence data, written by Simon Andrews at the Wellcome Sanger Institute in Cambridge.

**Sequence Counts**

Sequence counts for each sample. Duplicate read counts are an estimate only.

Percentages

FastQC: Sequence Counts

Number of reads

Uniques Reads    Duplicate Reads

mother\_short    grandmother\_short    mother\_short    proband\_29    proband\_short    wife\_short

[Export Plot](#)

**Sequence Quality Histograms** 8

The mean quality value across each base position in the read.

FastQC: Mean Quality Scores

R Center for Data Science

# Session info

```
sessionInfo()

## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.4.1
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics   grDevices  utils       datasets   methods    base
##
## loaded via a namespace (and not attached):
## [1] compiler_4.4.2    fastmap_1.2.0    cli_3.6.5      tools_4.4.2
## [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10    rmarkdown_2.29
## [9] knitr_1.50        xfun_0.52       digest_0.6.37   rlang_1.1.6
## [13] evaluate_1.0.3
```