Chapter 9:
Analysis of next-generation sequence data

Learning objectives

After studying this chapter you should be able to:

- explain how sequencing technologies generate NGS data;
- describe the FASTQ, SAM/BAM, and VCF data formats;
- compare methods for aligning NGS data to a reference genome;
- describe types of genomic variants and how they are determined;
- explain types of error associated with alignment, assembly, and variant calling; and
- explain methods for predicting the functional consequence of genomic variants in individual genomes.

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

- Sanger sequencing; NGS; Illumina; pyrosequencing;
- ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

- Overview
- Topic 1: Design
- Topic 2: FASTQ
- Topic 3: Assembly
- Topic 4: Alignment
- Topic 5: SAM/BAM
- Topic 6: Variant calling: SNVs
- Topic 7: Variant calling: SVs
- Topic 8: VCF
- Topic 9: Visualizing NGS data
- Topic 10: Significance

Specialized applications of NGS

Perspective

Human genome sequencing

We currently obtain whole genome sequences at 30x to 50x depth of coverage. For a typical individual:

- 2.8 billion base pairs are sequenced
- ~3-4 million single nucleotide variants
- ~600,000 insertions/deletions (SNPs)
- Cost (research basis) is <$2000
- We try to sequence mother/father/child trios

We also can enrich the collection of exons ("whole exome sequencing"). For a typical individual:

- 60 million base pairs are sequenced
- There are ~80,000 variants
- There are ~11,000 nonsynonymous SNPs

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview
Topic 1: Design
Topic 2: FASTQ
Topic 3: Assembly
Topic 4: Alignment
Topic 5: SAM/BAM

Topic 6: Variant calling: SNVs
Topic 7: Variant calling: SVs
Topic 8: VCF
Topic 9: Visualizing NGS data
Topic 10: Significance

Specialized applications of NGS

Perspective

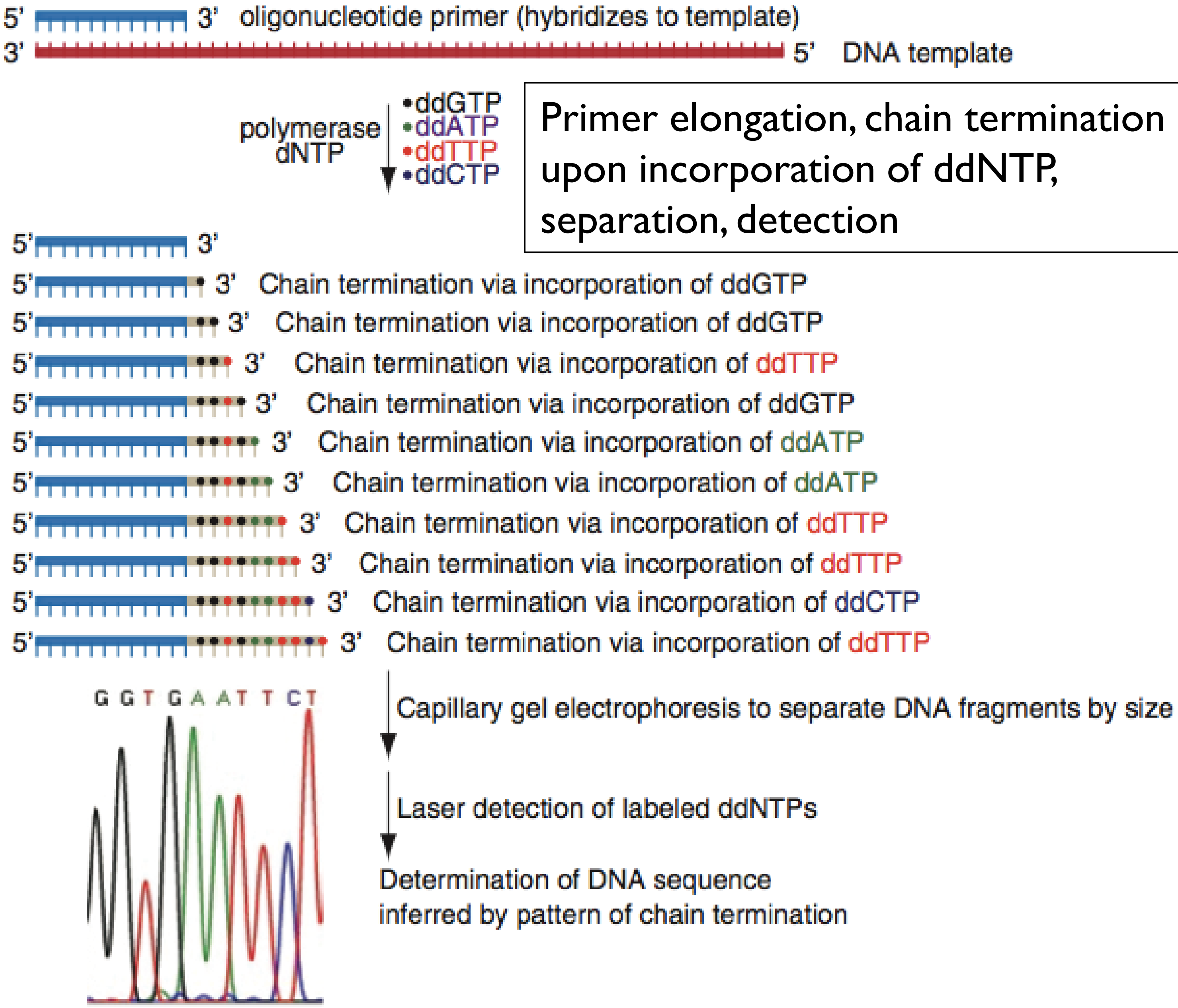
Sanger sequencing: what we had before NGS

Introduced in 1977

A template is denatured to form single strands, and extended with a polymerase in the presence of dideoxynucleotides (ddNTPs) that cause chain termination.

Typical read lengths are up to 800 base pairs. For the sequencing of Craig Venter's genome (2007; first whole genome of an individual), Sanger sequencing was employed because of its relatively long read lengths.

DNA sequencing by the Sanger method



5' 3' oligonucleotide primer (hybridizes to template)

3' 5' DNA template

polymerase
dNTP

•ddGTP
•ddATP
•ddTTP
•ddCTP

Primer elongation, chain termination upon incorporation of ddNTP, separation, detection

5' 3'

5' • 3' Chain termination via incorporation of ddGTP

5' •• 3' Chain termination via incorporation of ddGTP

5' ••• 3' Chain termination via incorporation of ddTTP

5' •••• 3' Chain termination via incorporation of ddGTP

5' ••••• 3' Chain termination via incorporation of ddATP

5' •••••• 3' Chain termination via incorporation of ddATP

5' ••••••• 3' Chain termination via incorporation of ddTTP

5' •••••••• 3' Chain termination via incorporation of ddTTP

5' ••••••••• 3' Chain termination via incorporation of ddCTP

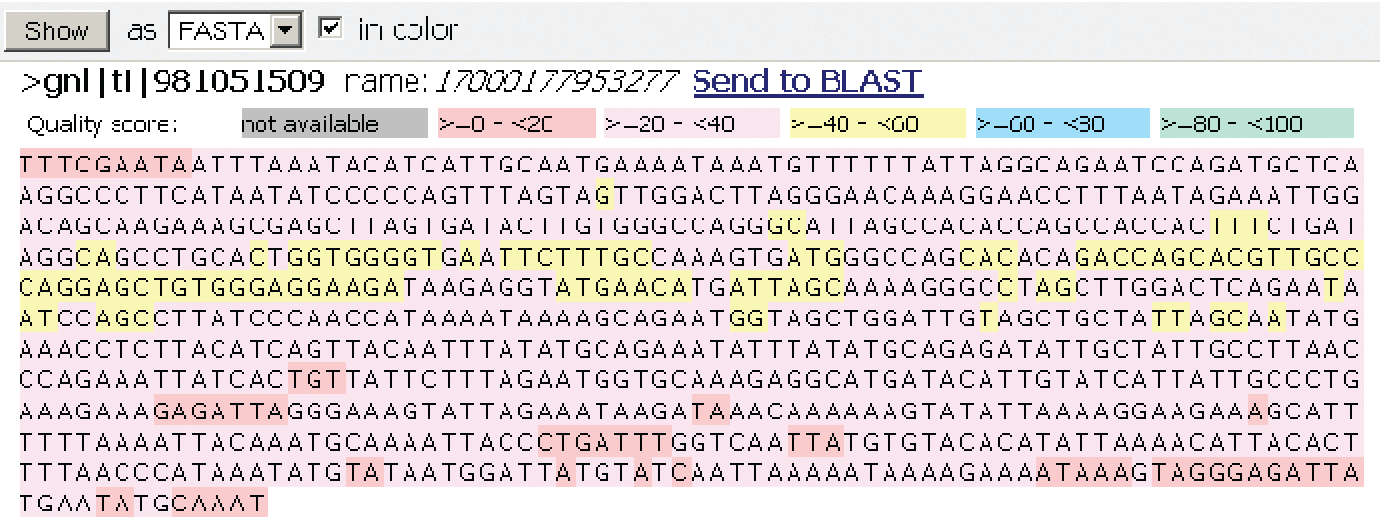
5' •••••••••• 3' Chain termination via incorporation of ddTTP

Capillary gel electrophoresis to separate DNA fragments by size

Laser detection of labeled ddNTPs

Determination of DNA sequence
inferred by pattern of chain termination

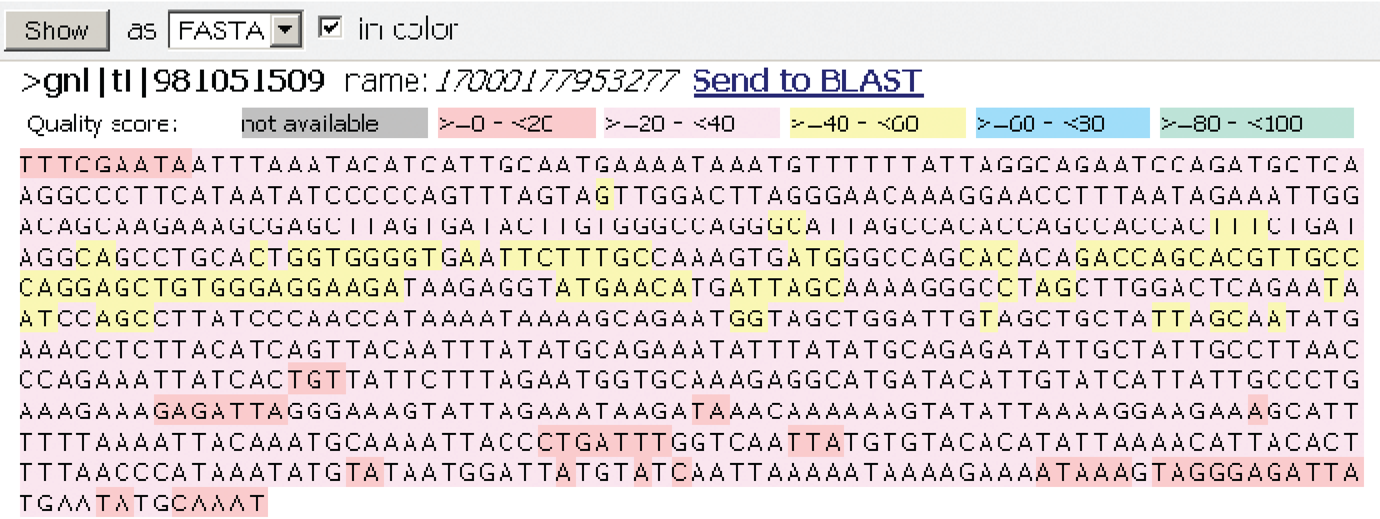
View genomic DNA (here from the beta globin locus) from the Trace Archive at NCBI: FASTA format



Quality score: not available | >=0 - <20 | >=20 - <40 | >=40 - <60 | >=60 - <80 | >=80 - <100

```
>gnl|ti|981051509 name:17000177953277
TTTCGAATAATTTAAATACATCATTGCAATGAAAATAAATGTTTTTTATTAGGCAGAATCCAGATGCTCA
AGGCCCTTCATAATATCCCCCAGTTTAGTAGTTGGACTTAGGGAACAAAGGAACCTTTAATAGAAATTGG
ACAGCAAGAAAGCGAGCTTAGTGATACTTGTGGGCCAGGGCATTAGCCACACCAGCCACCACTTTCTGAT
AGGCAGCCTGCACTGGTGGGGTGAATTCTTTGCCAAAGTGATGGGCCAGCACACAGACCAGCACGTTGCC
CAGGAGCTGTGGGAGGAAGATAAGAGGTATGAACATGATTAGCAAAAGGGCCTAGCTTGGACTCAGAATA
ATCCAGCCTTATCCCAACCATAAAATAAAAGCAGAATGGTAGCTGGATTGTAGCTGCTATTAGCAATATG
AAACCTCTTACATCAGTTACAATTTATATGCAGAAATATTTATATGCAGAGATATTGCTATTGCCTTAAC
CCAGAAATTATCACTGTTATTCTTTAGAATGGTGCAAAGAGGCATGATACATTGTATCATTATTGCCCTG
AAAGAAAGAGATTAGGGAAAGTATTAGAAATAAGATAAACAAAAAAGTATATTAAAAGGAAGAAAGCATT
TTTTAAAATTACAAATGCAAAATTACCCTGATTTGGTCAATTATGTGTACACATATTAAAACATTACACT
TTTAACCCATAAATATGTATAATGGATTATGTATCAATTAAAAATAAAAGAAAATAAAGTAGGGAGATTA
TGAATATGCAAAT
```

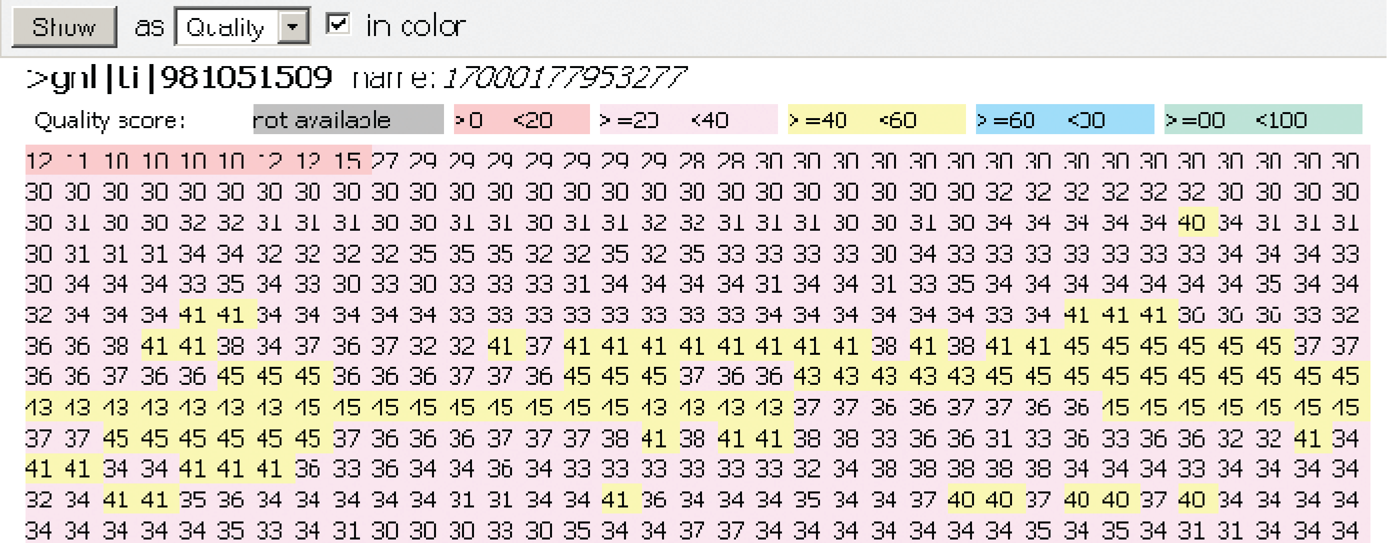
Each DNA base in the Trace Archive has an associated base quality score (best scores highlighted in yellow)



>gnl|ti|981051509 name: *17000177953277*

Quality score:	not available	>=0 - <20	>=20 - <40	>=40 - <60	>=60 - <80	>=80 - <100

```
TTTCGAATAATTTAAATACATCATTGCAATGAAAATAAATGTTTTTTATTAGGCAGAATCCAGATGCTCA
AGGCCCTTCATAATATCCCCCAGTTTAGTAGTTGGACTTAGGGAACAAAGGAACCTTTAATAGAAATTGG
ACAGCAAGAAAGCGAGCTTAGTGATACTTGTGGGCCAGGGCATTAGCCACACCAGCCACCACTTTCTGAT
AGGCAGCCTGCACTGGTGGGGTGAATTCTTTGCCAAAGTGATGGGCCAGCACACAGACCAGCACGTTGCC
CAGGAGCTGTGGGAGGAAGATAAGAGGTATGAACATGATTAGCAAAAGGGCCTAGCTTGGACTCAGAATA
ATCCAGCCTTATCCCAACCATAAAATAAAAGCAGAATGGTAGCTGGATTGTAGCTGCTATTAGCAATATG
AAACCTCTTACATCAGTTACAATTTATATGCAGAAATATTTATATGCAGAGATATTGCTATTGCCTTAAC
CCAGAAATTATCACTGTTATTCTTTAGAATGGTGCAAAGAGGCATGATACATTGTATCATTATTGCCCTG
AAAGAAAGAGATTAGGGAAAGTATTAGAAATAAGATAAACAAAAAAGTATATTAAAAGGAAGAAAGCATT
TTTTAAAATTACAAATGCAAAATTACCCTGATTTGGTCAATTATGTGTACACATATTAAAACATTACACT
TTTAACCCATAAATATGTATAATGGATTATGTATCAATTAAAAATAAAAGAAAATAAAGTAGGGAGATTA
TGAATATGCAAAT
```



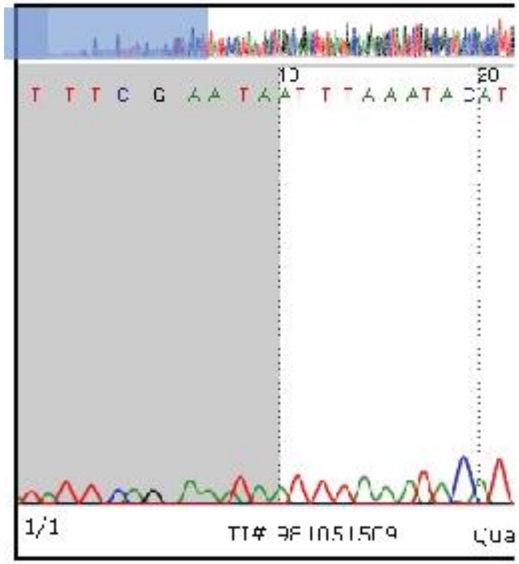
>gnl|ti|981051509 name: *17000177953277*

Quality score:	not available	>0 <20	>=20 <40	>=40 <60	>=60 <80	>=80 <100

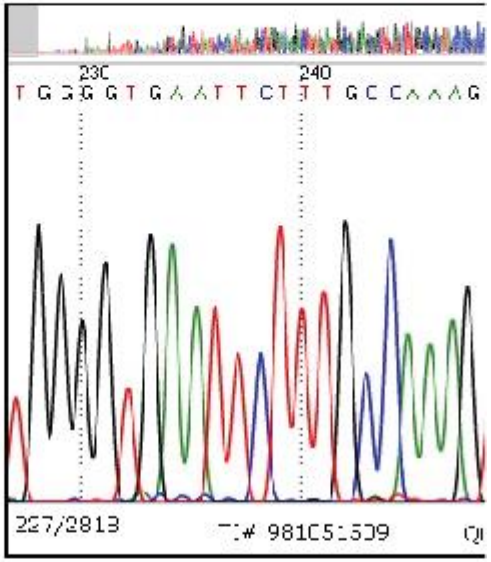
```
12 11 10 10 10 10 12 12 15 27 29 29 29 29 29 29 29 28 28 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 32 32 32 32 32 32 30 30 30 30
30 31 30 30 32 32 31 31 31 30 30 31 31 30 31 31 32 32 31 31 31 30 30 31 30 34 34 34 34 34 40 34 31 31 31
30 31 31 31 34 34 32 32 32 32 35 35 35 32 32 35 32 35 33 33 33 33 30 34 33 33 33 33 33 33 33 34 34 34 33
30 34 34 34 33 35 34 33 30 33 30 33 33 33 31 34 34 34 34 31 34 34 31 33 35 34 34 34 34 34 34 34 35 34 34
32 34 34 34 41 41 34 34 34 34 34 33 33 33 33 33 33 33 33 34 34 34 34 34 34 33 34 41 41 41 36 36 36 33 32
36 36 38 41 41 38 34 37 36 37 32 32 41 37 41 41 41 41 41 41 41 41 38 41 38 41 41 45 45 45 45 45 45 37 37
36 36 37 36 36 45 45 45 36 36 36 37 37 36 45 45 45 37 36 36 43 43 43 43 43 45 45 45 45 45 45 45 45 45 45
43 43 43 43 43 43 43 45 45 45 45 45 45 45 45 45 43 43 43 43 37 37 36 36 37 37 36 36 45 45 45 45 45 45 45
37 37 45 45 45 45 45 45 37 36 36 36 37 37 37 38 41 38 41 41 38 38 33 36 36 31 33 36 33 36 36 32 32 41 34
41 41 34 34 41 41 41 36 33 36 34 34 36 34 33 33 33 33 33 33 32 34 38 38 38 38 38 34 34 34 33 34 34 34 34
32 34 41 41 35 36 34 34 34 34 34 31 31 34 34 41 36 34 34 34 35 34 34 37 40 40 37 40 40 37 40 34 34 34 34
34 34 34 34 34 35 33 34 31 30 30 30 33 30 35 34 34 37 37 34 34 34 34 34 34 34 35 34 35 34 31 31 34 34 34
```

Examples of Sanger sequencing traces

Low quality reads



High quality reads

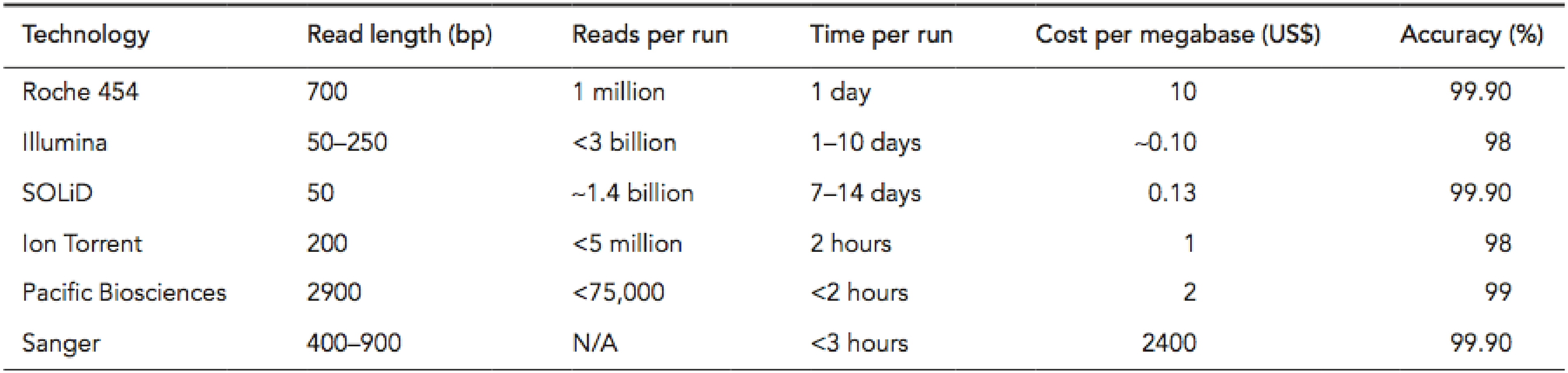


Next-generation sequence technologies

Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase	Accuracy
Roche 454	700	1 million	1 day	$10	99.9%
Illumina	50-250	<3 billion	1-10 days	~$0.10	98%
SOLiD	50	~1.4 billion	7-14 days	$0.13	99.9%
Ion Torrent	200	<5 million	2 hours	$1	98%
Pacific Biosciences	2900	<75,000	<2 hours	$2	99%
Sanger	400-900	N/A	<3 hours	$2400	99.9%

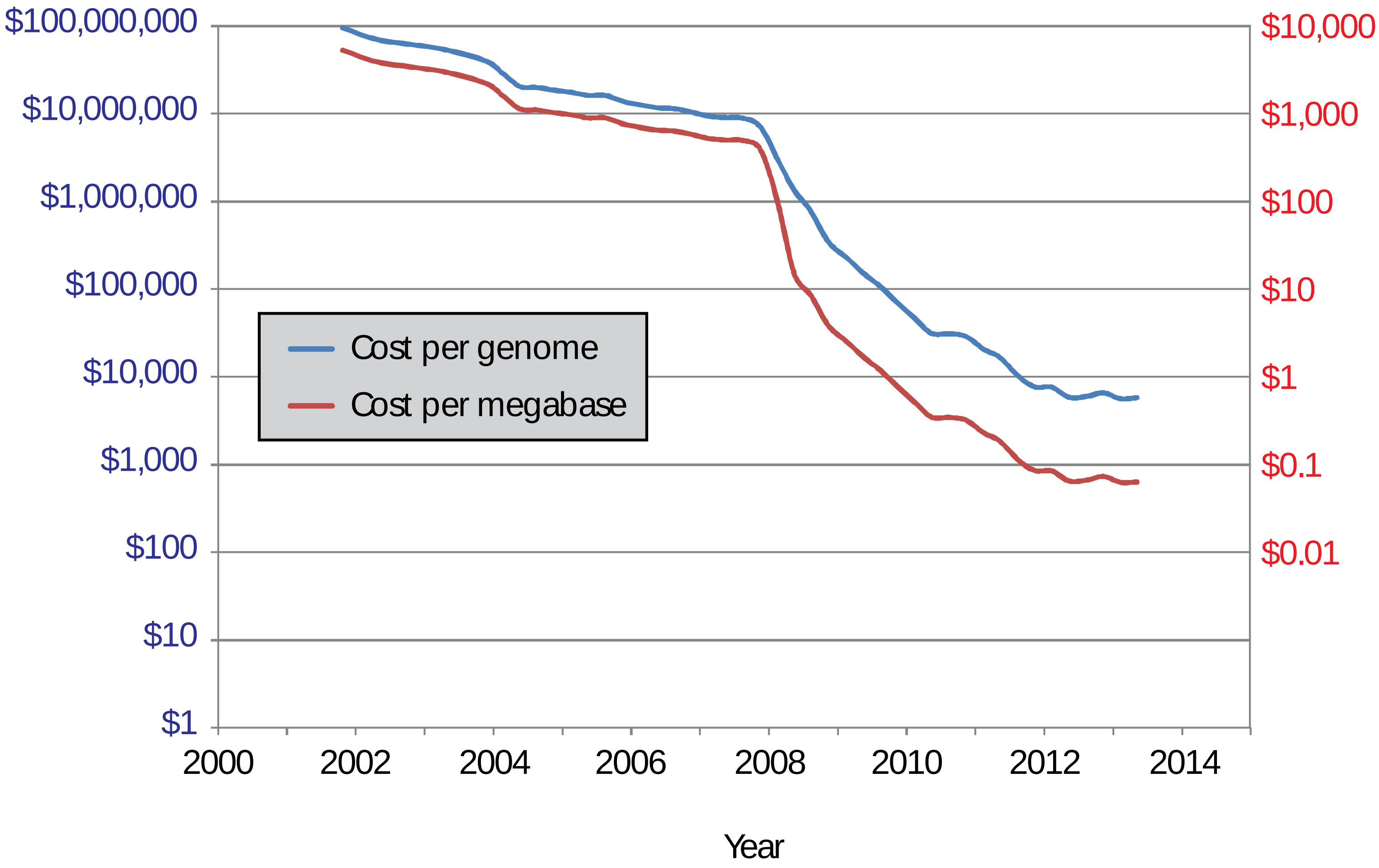
Source: adapted from Wikipedia 1/11/13

NGS technologies compared to Sanger sequencing

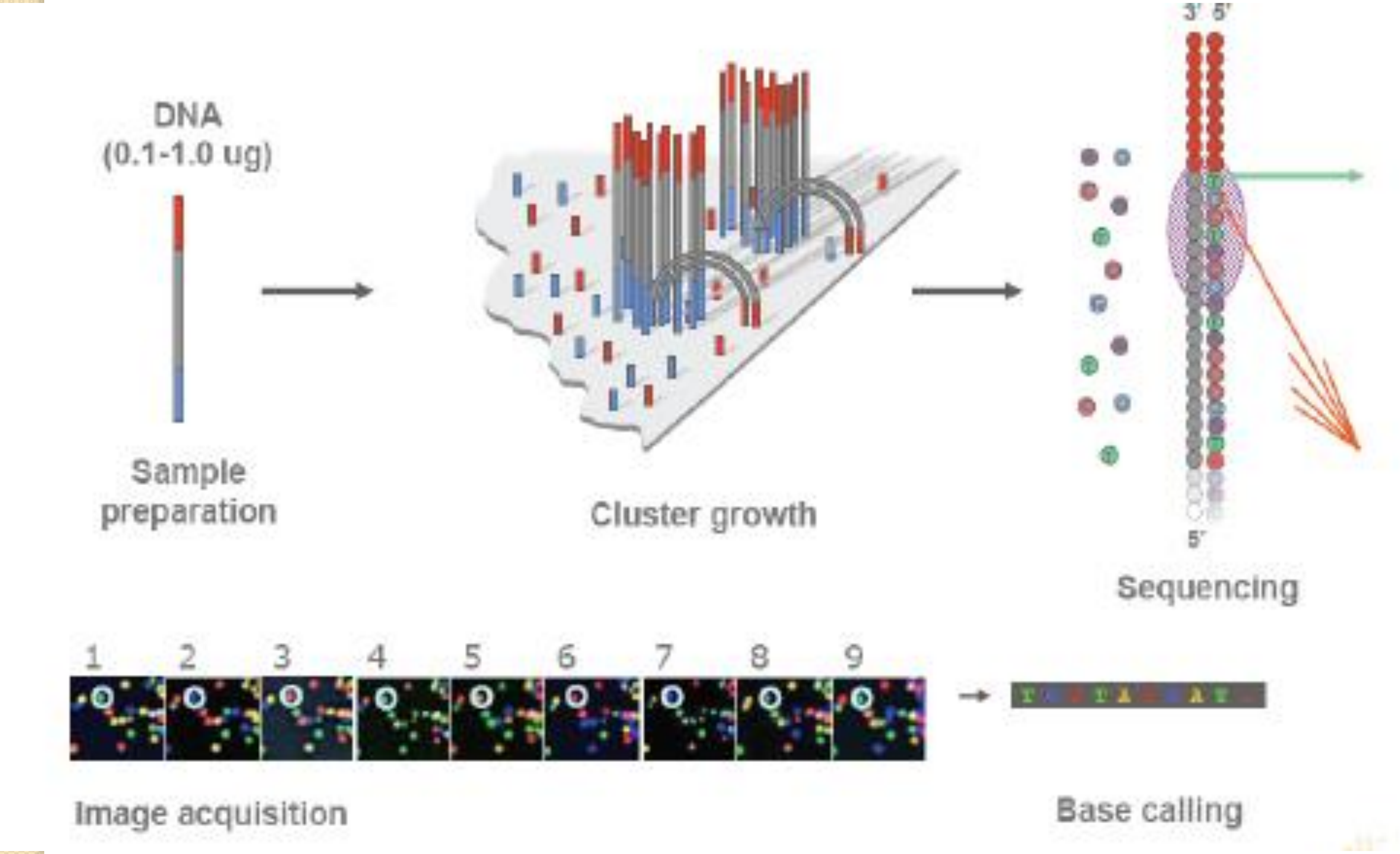


Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US$)	Accuracy (%)
Roche 454	700	1 million	1 day	10	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	98
SOLiD	50	~1.4 billion	7–14 days	0.13	99.90
Ion Torrent	200	<5 million	2 hours	1	98
Pacific Biosciences	2900	<75,000	<2 hours	2	99
Sanger	400–900	N/A	<3 hours	2400	99.90

Whole genome sequencing (WGS) costs have declined dramatically



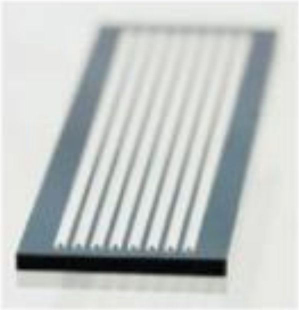
Next-generation sequence technology: Illumina



Sequencing by Illumina technology

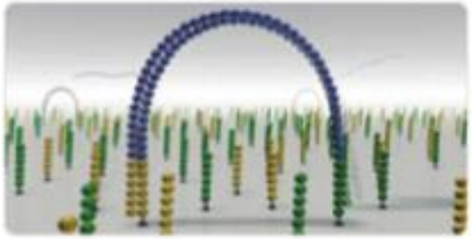
Randomly fragment genomic DNA

Library preparation



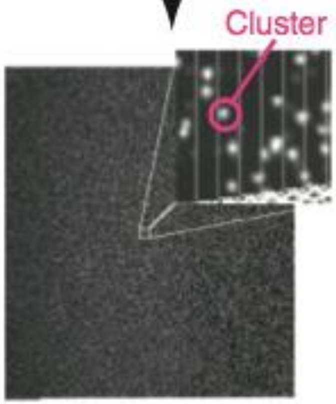
Samples immobilized on surface of a flow cell (8 lanes)

Solid phase amplification



- Bridge amplification (inverted U) generates clusters on surface of flow cell
- ~Ten million single-molecule clusters per square centimeter

Sequencing by synthesis



- Each cycle: add polymerase, one labeled deoxynucleoside triphosphate (dNTP) at a time (four labeled dNTPs per cycle)
- Image fluorescent dyes
- Call nucleotide
- Enzymatic cleavage to remove

Cycle termination sequencing (Illumina)

Disadvantage:

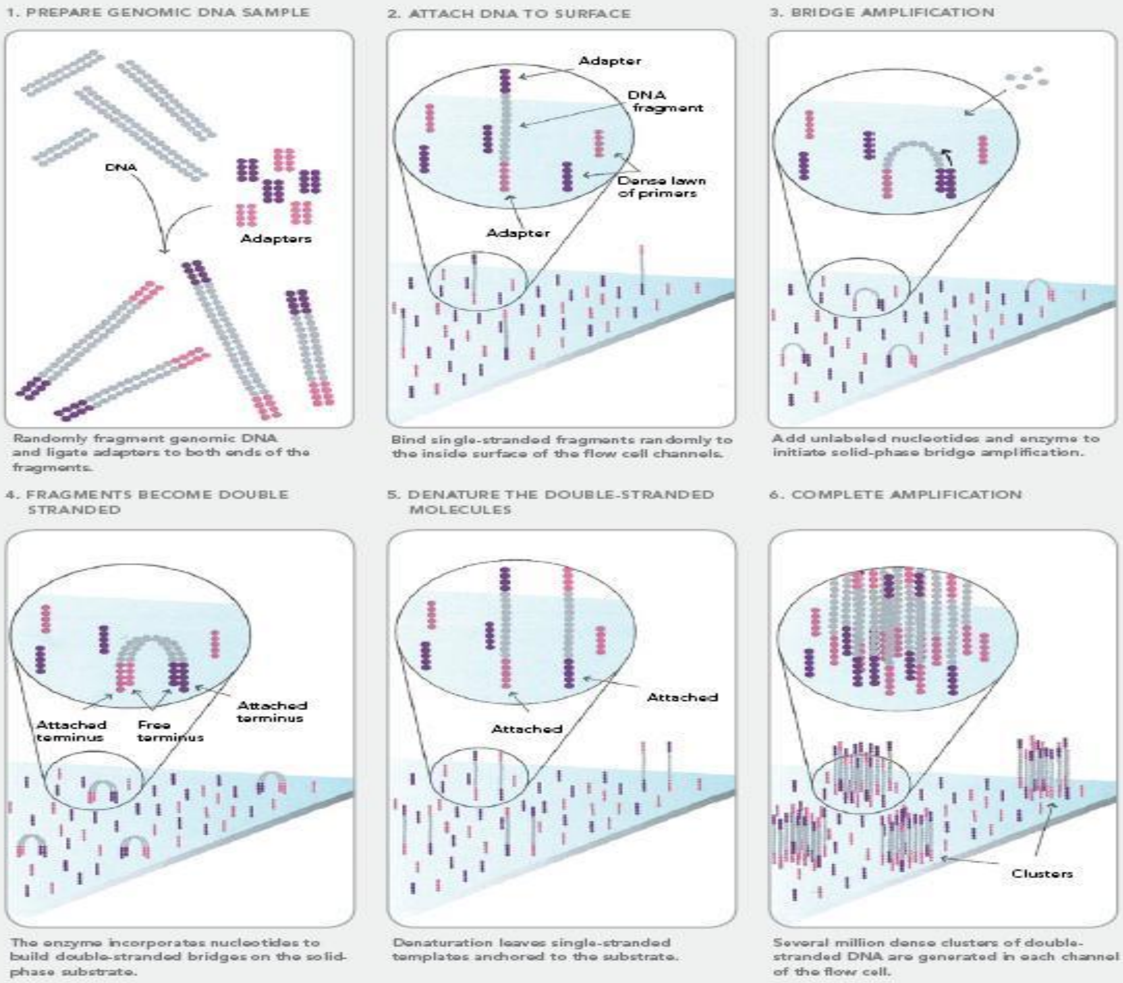
- Short read length (~150 bases)

Advantages:

- Very fast
- Low cost per base
- Large throughput; up to 1 gigabase/epxeriment
- Short read length makes it appropriate for resequencing
- No need for gel electrophoresis
- High accuracy
- All four bases are present at each cycle, with sequential addition of dNTPs. This allows homopolymers to be accurately read.

Illumina sequencing technology in 12 steps

FIGURE 2: SEQUENCING TECHNOLOGY OVERVIEW



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

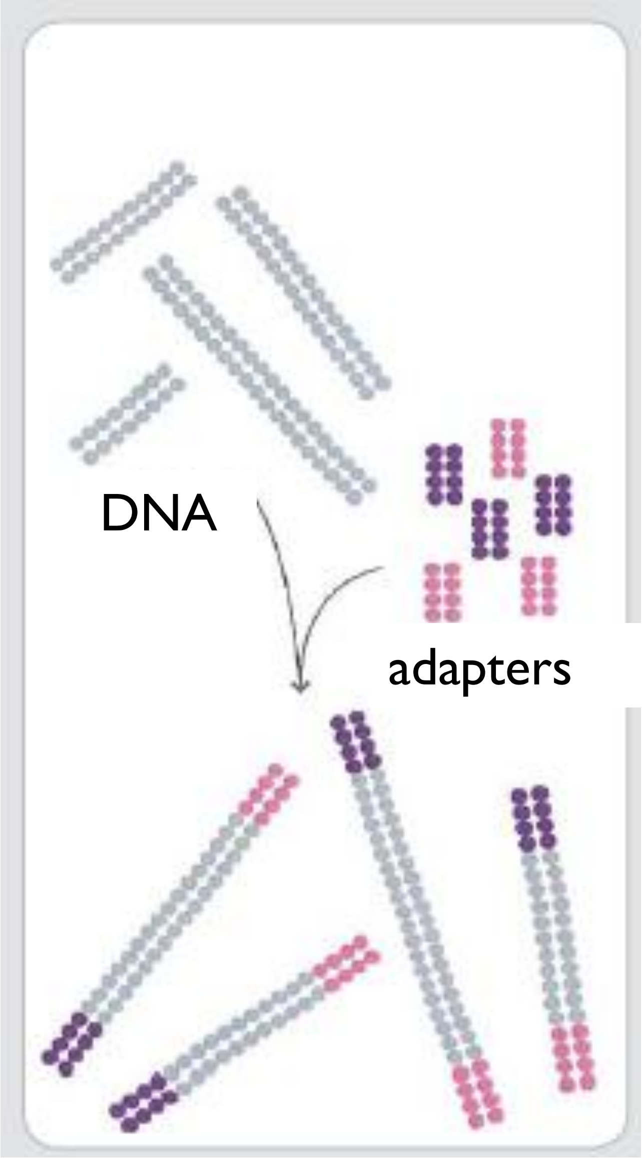
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

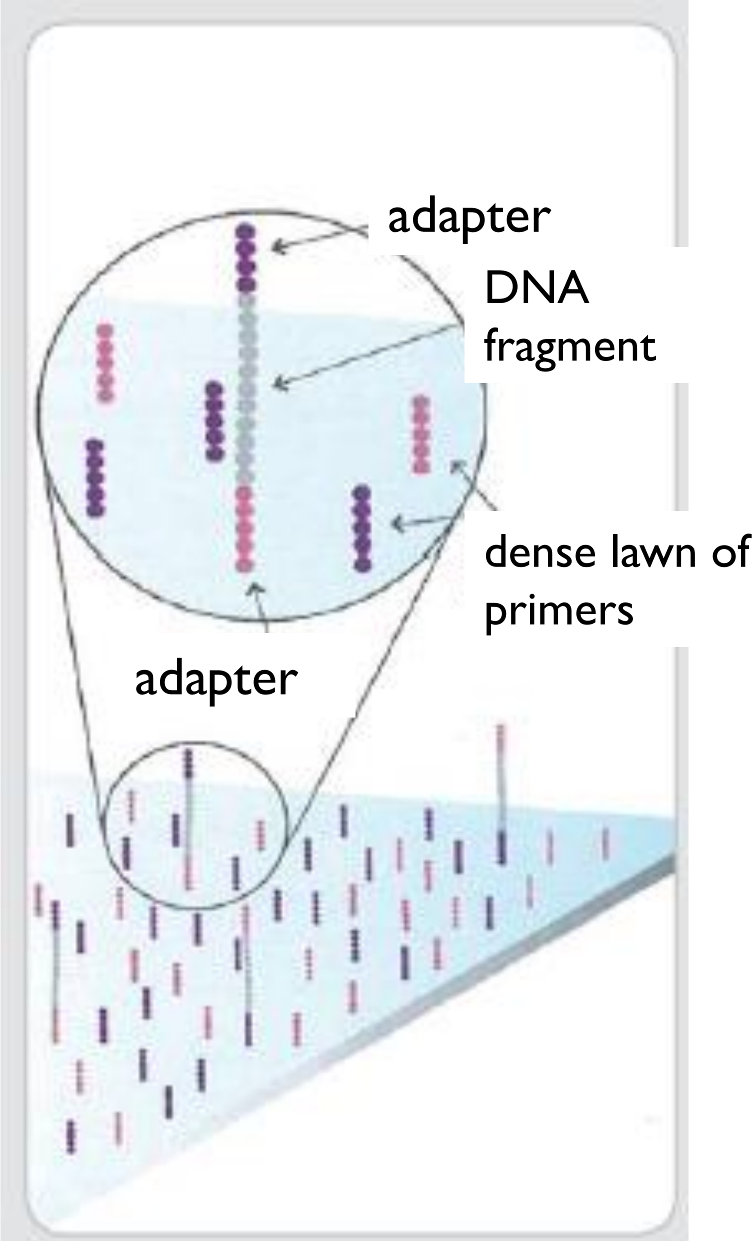
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Source: http://www.illumina.com/downloads/SS_DNAsequencing.pdf



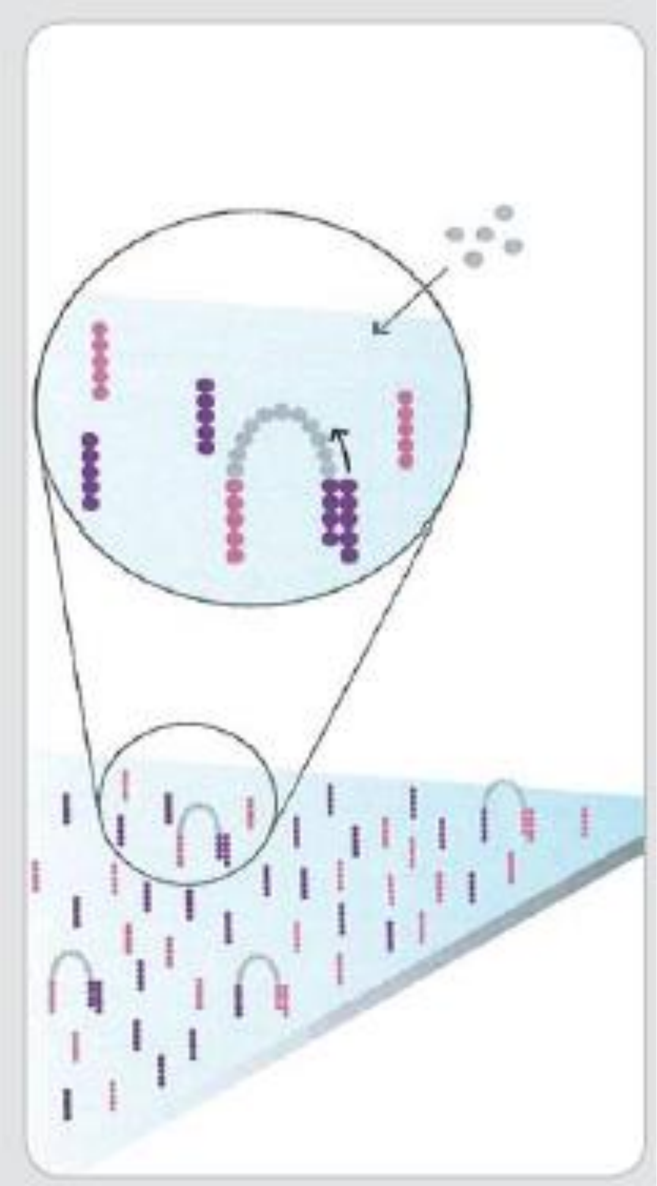
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



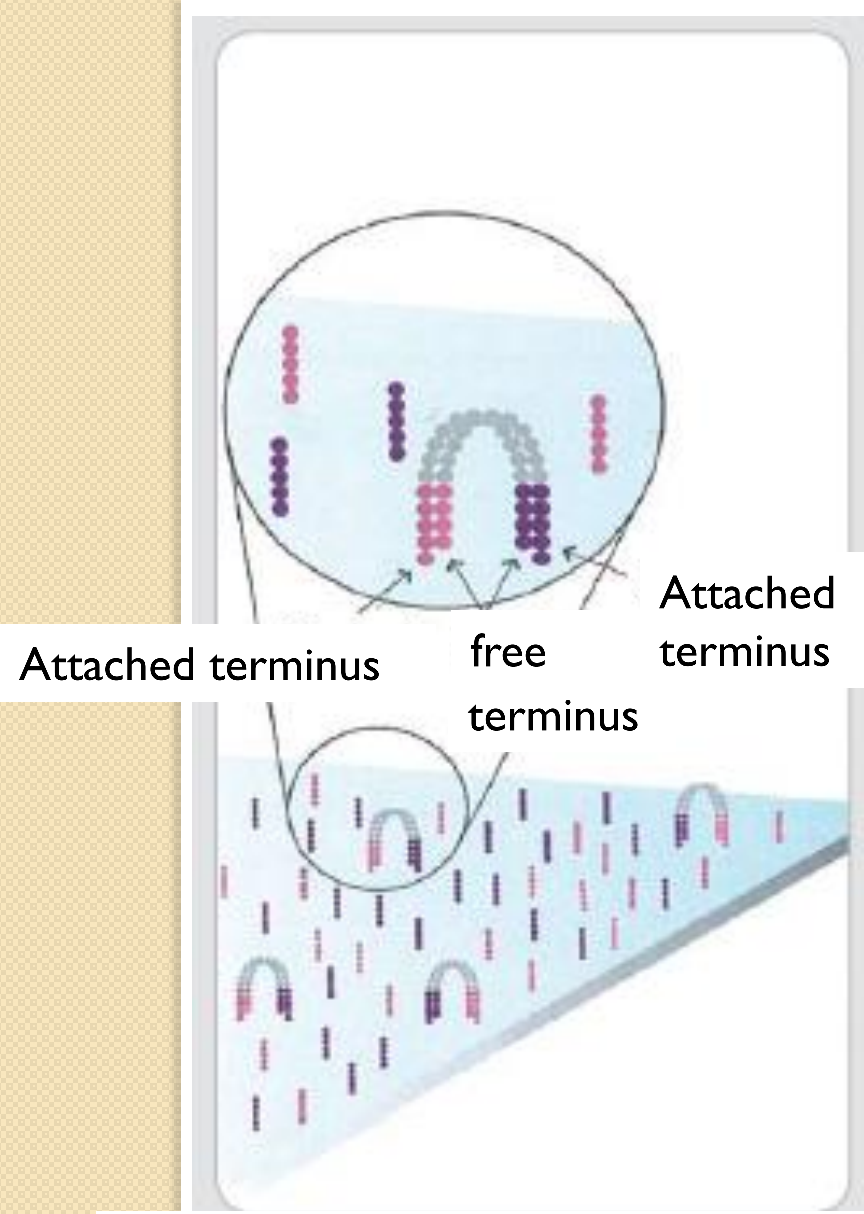
Bind single-stranded fragments randomly to the inside surface of the flow cell channels

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



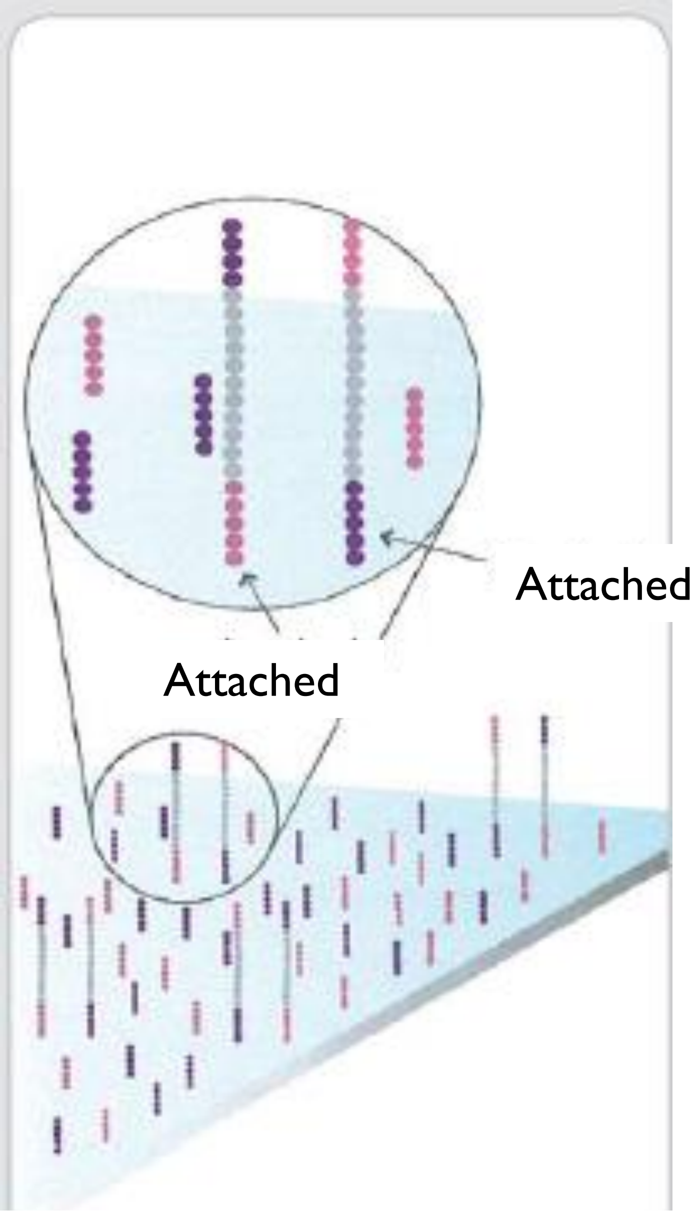
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



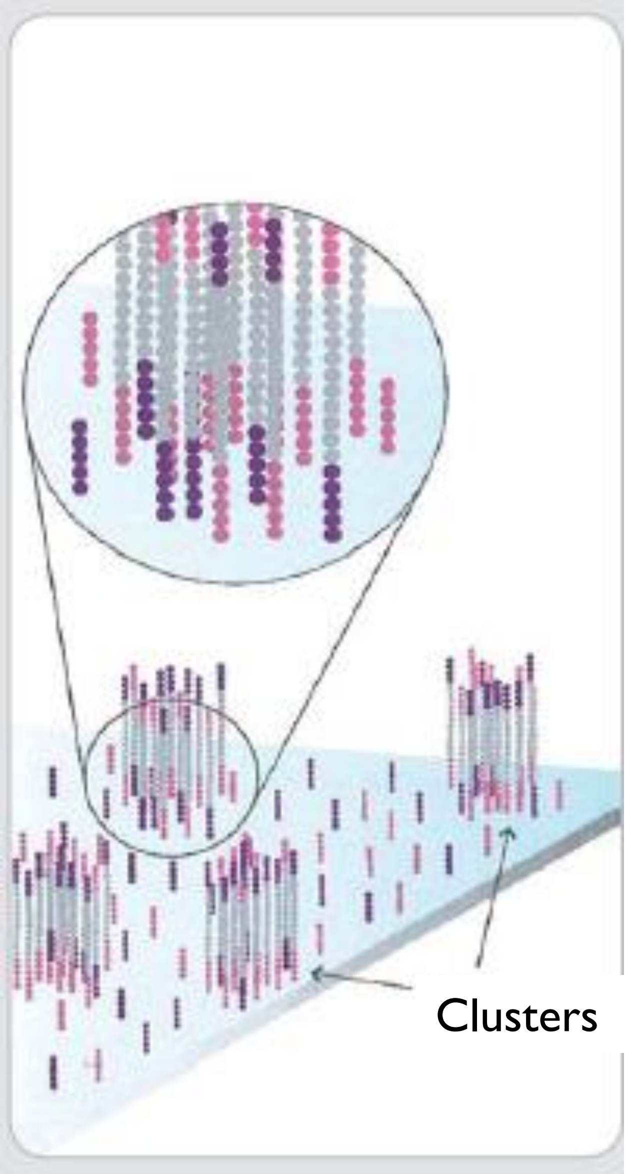
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



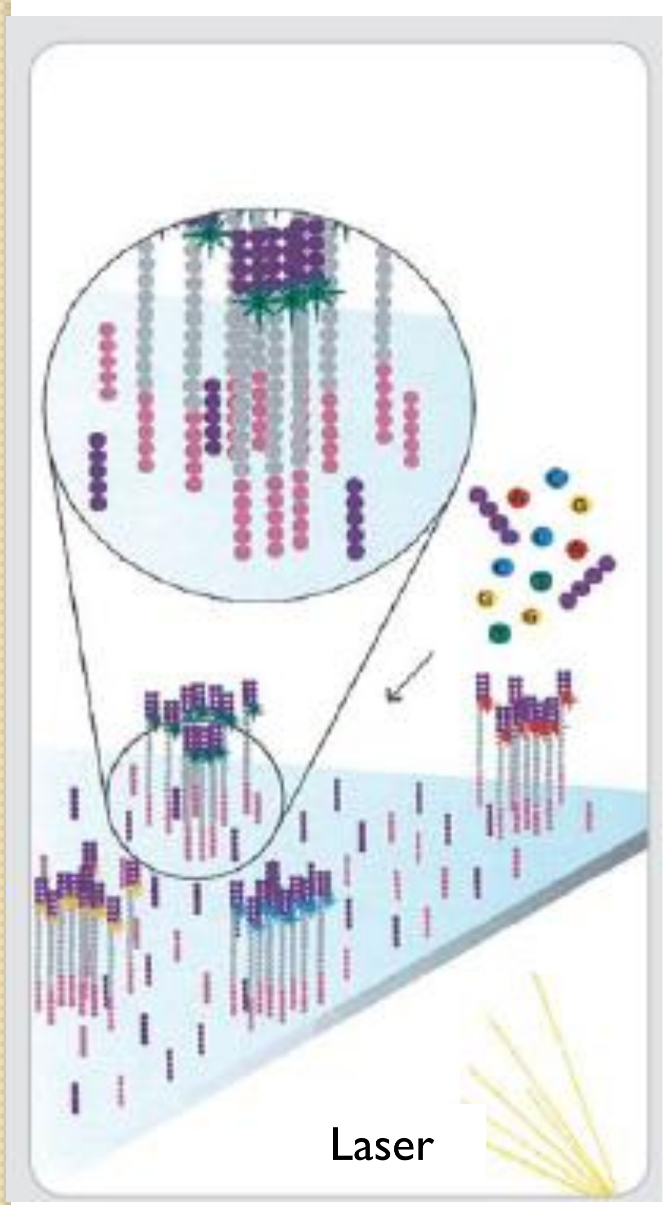
Denaturation leaves single-stranded templates anchored to the substrate

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



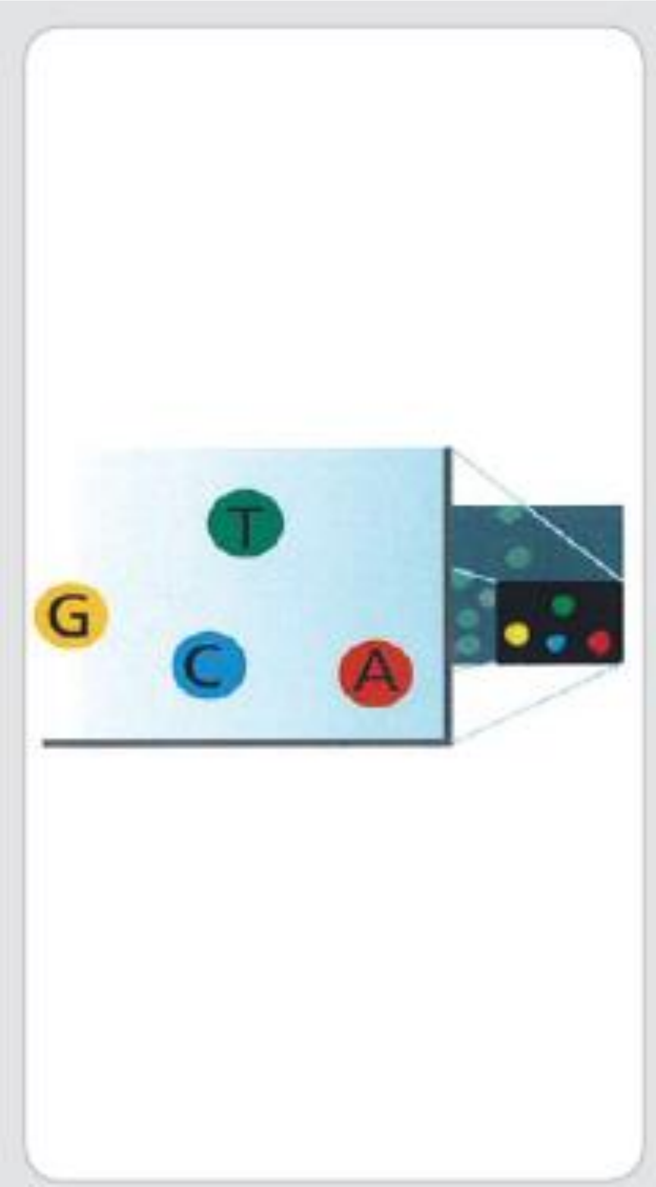
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase

7. Determine first base
8. Image first base
9. Determine second base
10. Image second chemistry cycle
11. Sequencing over multiple chemistry cycles
12. Align data



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified

7. Determine first base

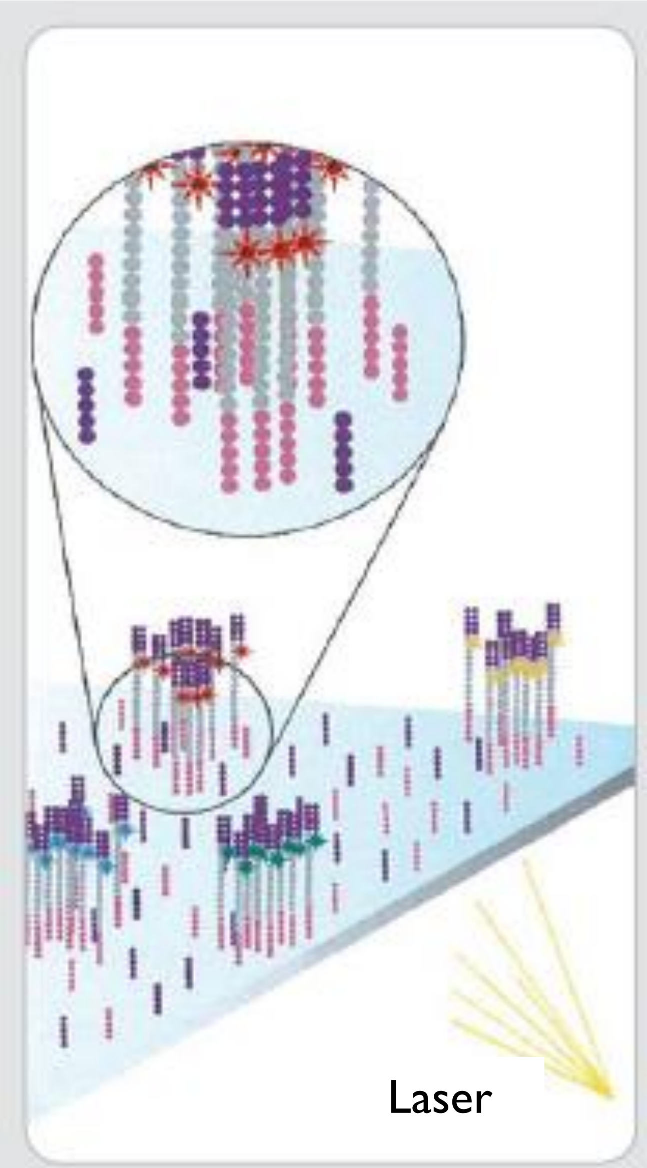
8. Image first base

9. Determine second base

10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

12. Align data



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase

7. Determine first base

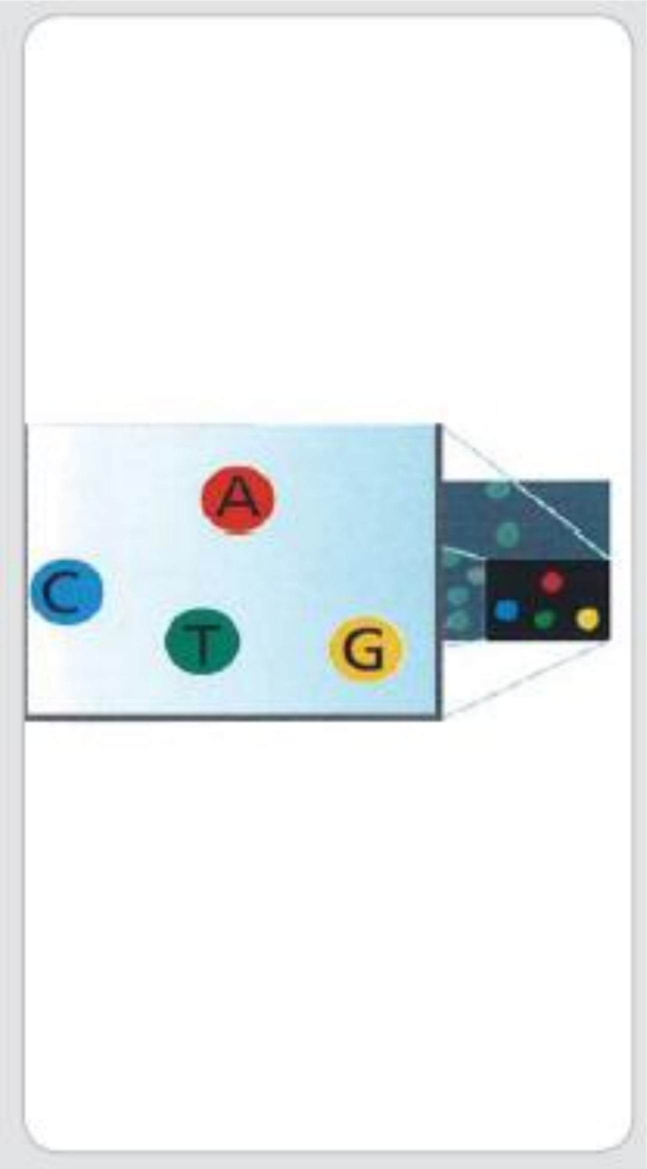
8. Image first base

9. Determine second base

10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

12. Align data



After laser excitation the image is captured as before, and the identity of the second base is recorded.

7. Determine first base

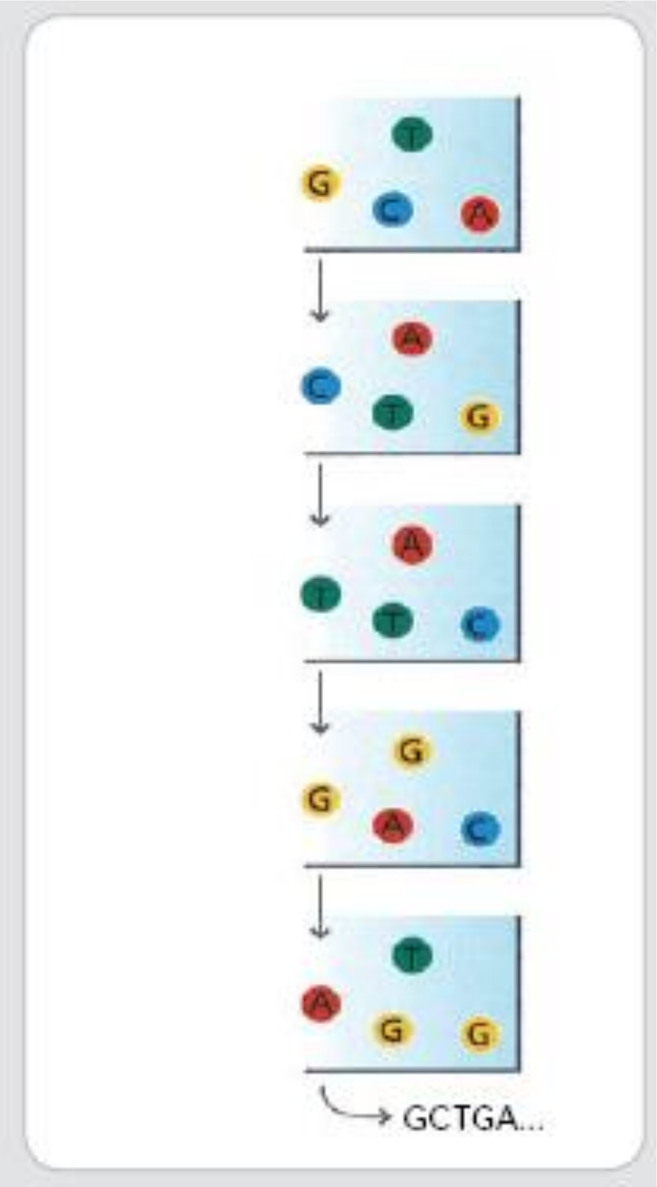
8. Image first base

9. Determine second base

10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

12. Align data



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

7. Determine first base

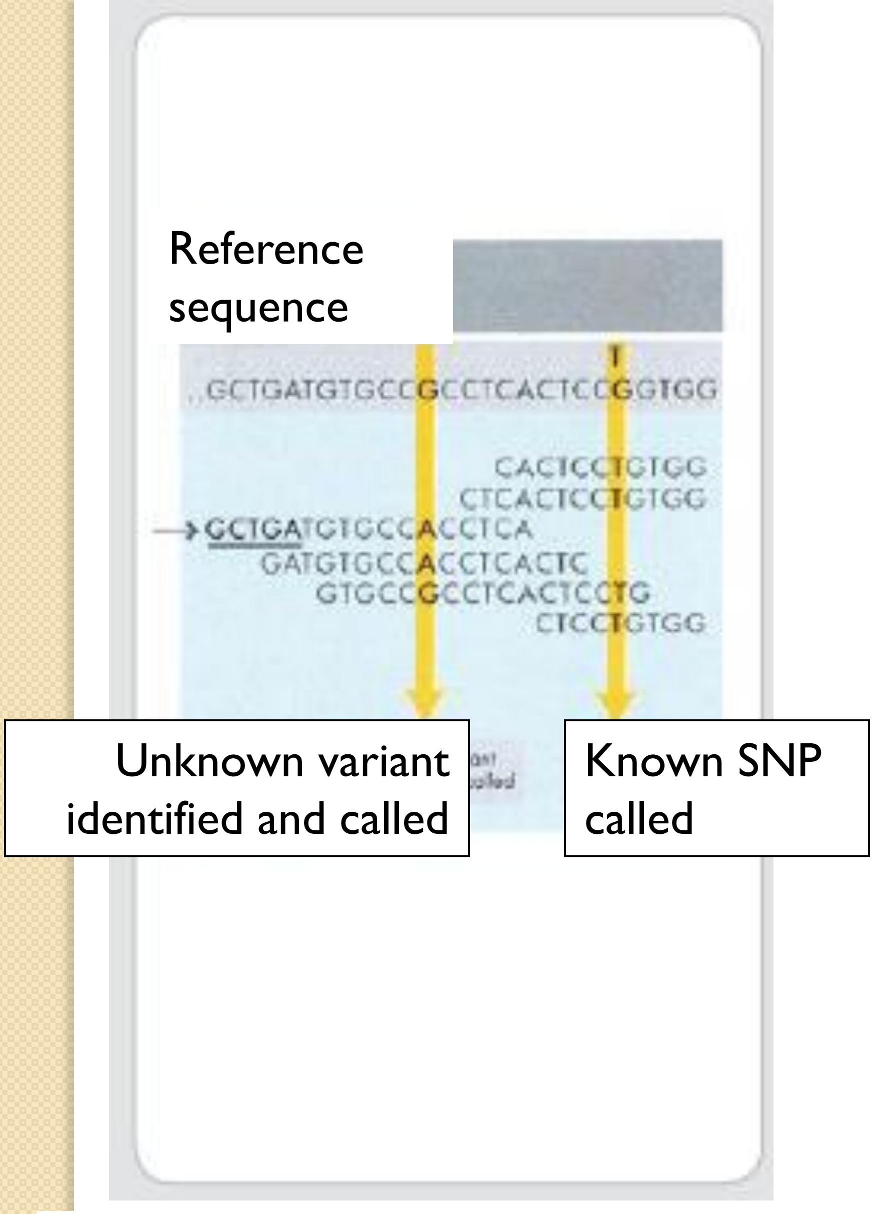
8. Image first base

9. Determine second base

10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

12. Align data



Reference sequence

Unknown variant identified and called

Known SNP called

The data are aligned and compared to a reference, and sequencing differences are identified.

7. Determine first base

8. Image first base

9. Determine second base

10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

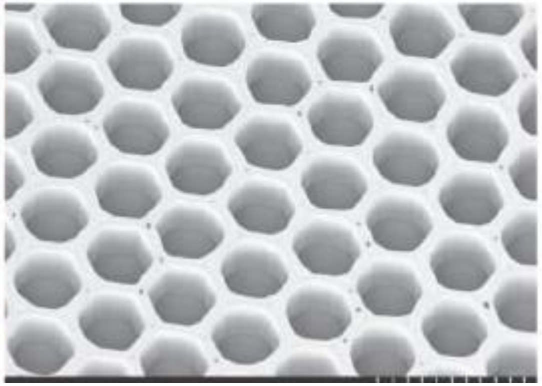
12. Align data

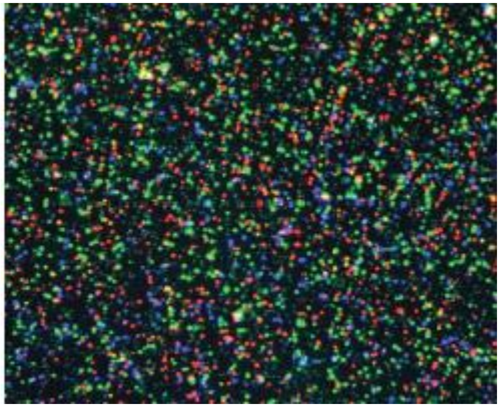
NGS technologies: Roche 454

- Introduced in 2005 (sequenced *Mycoplasma genitalium* genome in one run)

- ~2400 publications (as of Jan. 2013) but now defunct

- Sequencing by synthesis: nucleotide incorporation leads to light emission





Pyrosequencing

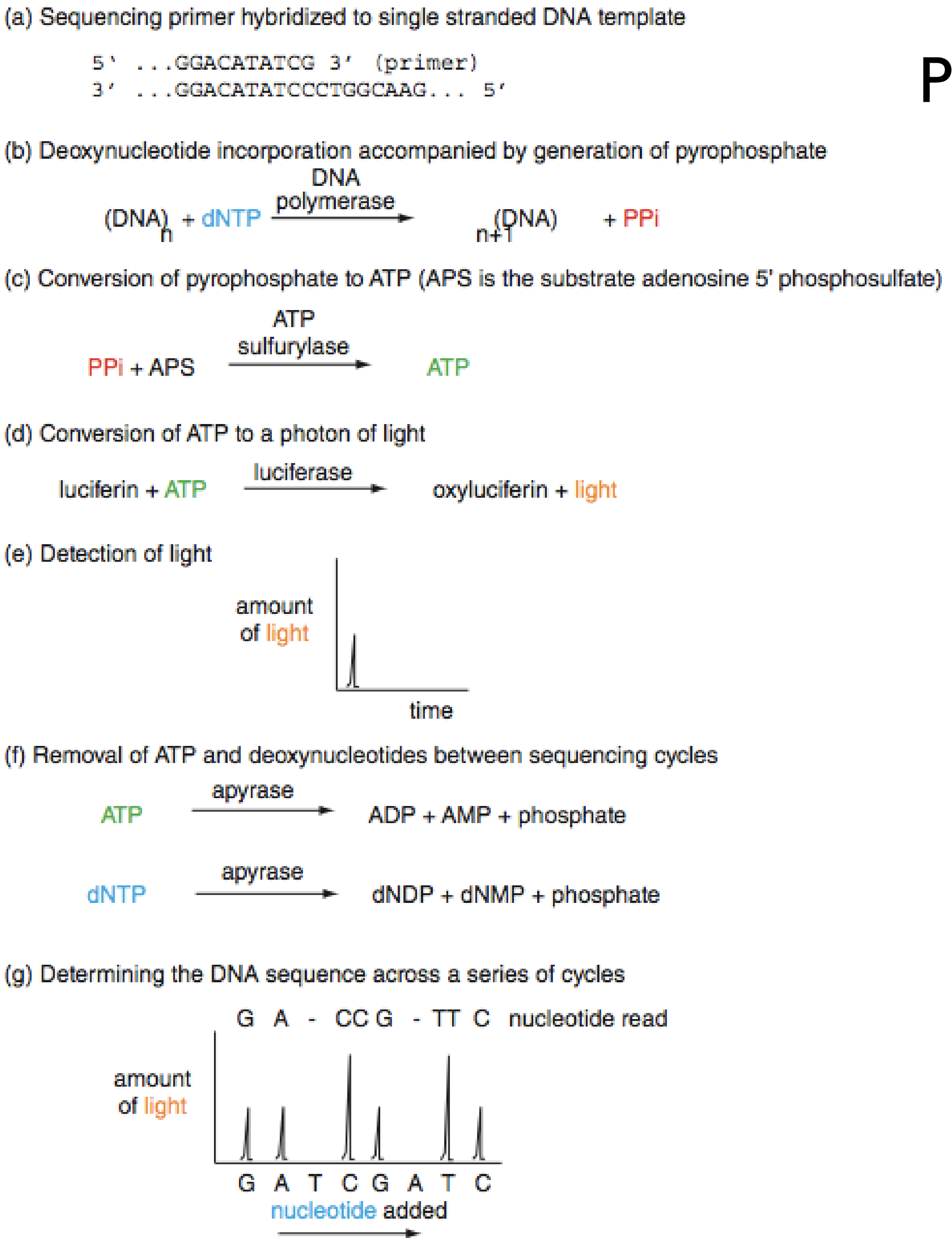
Advantages:

- Very fast
- Low cost per base
- Large throughput; up to 40 megabases/epxeriment
- No need for bacterial cloning (with its associated artifacts); this is especially helpful in metagenomics
- High accuracy

Disadvantages:

- Short read lengths (soon to be extended to ~500 bp)
- Difficulty sequencing homopolymers accurately

Pyrosequencing



(a) Sequencing primer hybridized to single stranded DNA template

```
5' ...GGACATATCG 3' (primer)
3' ...GGACATATCCCTGGCAAG... 5'
```

(b) Deoxynucleotide incorporation accompanied by generation of pyrophosphate

$$(DNA)_n + dNTP \xrightarrow{\text{DNA polymerase}} (DNA)_{n+1} + PPi$$

(c) Conversion of pyrophosphate to ATP (APS is the substrate adenosine 5' phosphosulfate)

$$PPi + APS \xrightarrow{\text{ATP sulfurylase}} ATP$$

(d) Conversion of ATP to a photon of light

$$\text{luciferin} + ATP \xrightarrow{\text{luciferase}} \text{oxyluciferin} + \text{light}$$

(e) Detection of light

(f) Removal of ATP and deoxynucleotides between sequencing cycles

$$ATP \xrightarrow{\text{apyrase}} ADP + AMP + \text{phosphate}$$

$$dNTP \xrightarrow{\text{apyrase}} dNDP + dNMP + \text{phosphate}$$

(g) Determining the DNA sequence across a series of cycles

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview Topic 6: Variant calling: SNVs
Topic 1: Design Topic 7: Variant calling: SVs
Topic 2: FASTQ Topic 8: VCF
Topic 3: Assembly Topic 9: Visualizing NGS data
Topic 4: Alignment Topic 10: Significance
Topic 5: SAM/BAM

Specialized applications of NGS

Perspective

A workflow for whole genome sequencing (WGS) of individual genomes

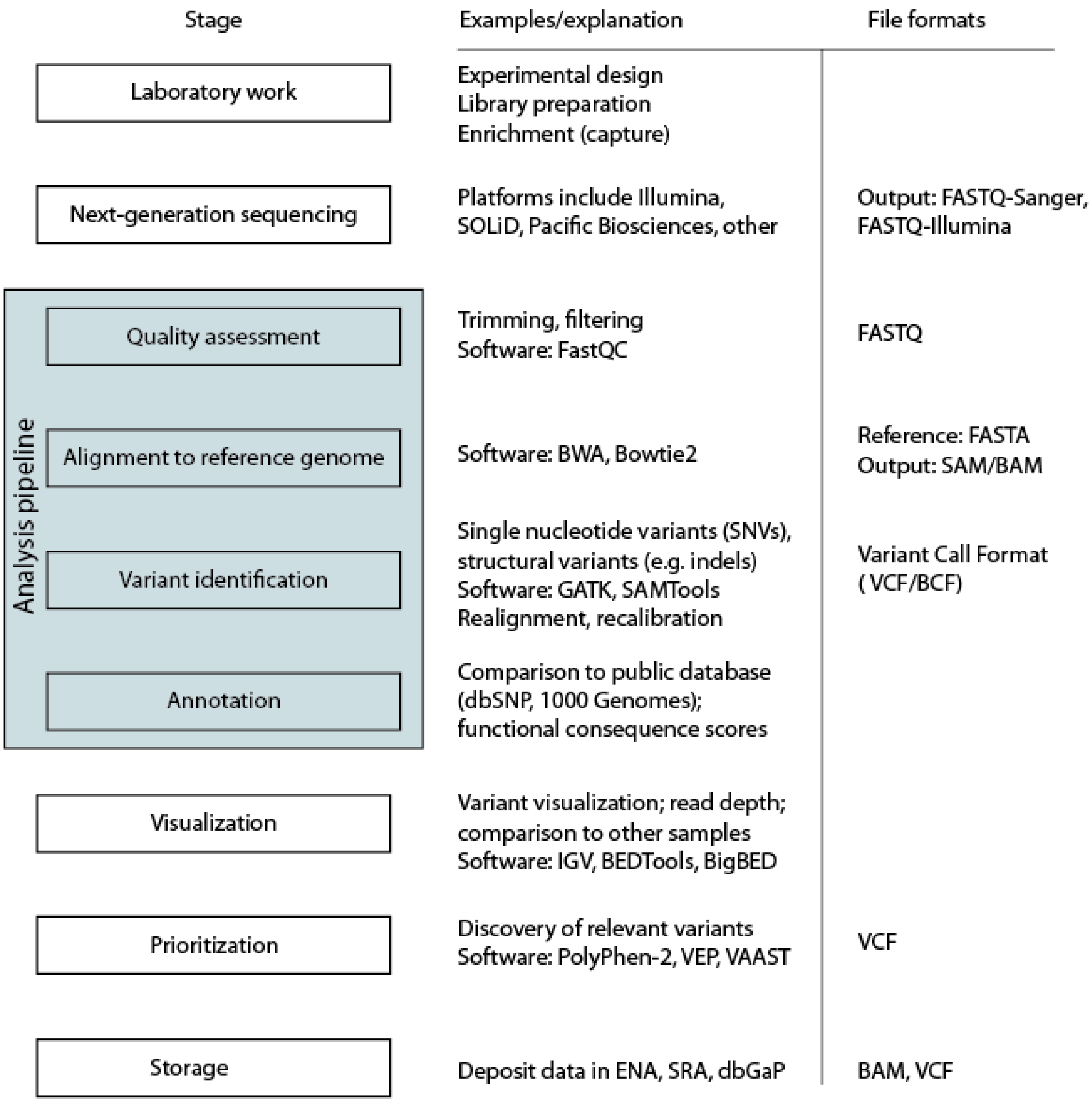
1. Select proband(s)
2. Purify genomic DNA
3. Generate paired-end library
4. Design capture beads (e.g. Agilent SureSelect)
5. Hybridize in solution
6. Elute enriched genomic DNA
7. Amplify
8. Next-generation sequencing
9. Align sequence to a human genome reference
10. Determine coverage (e.g. 30-fold)
11. Identify variants: SNPs, indels (distinguish true variants from sequencing errors)
12. Prioritize variants
13. Validate variants

} optional; used for whole exome sequencing

Broad clinical workflow for WGS of patients

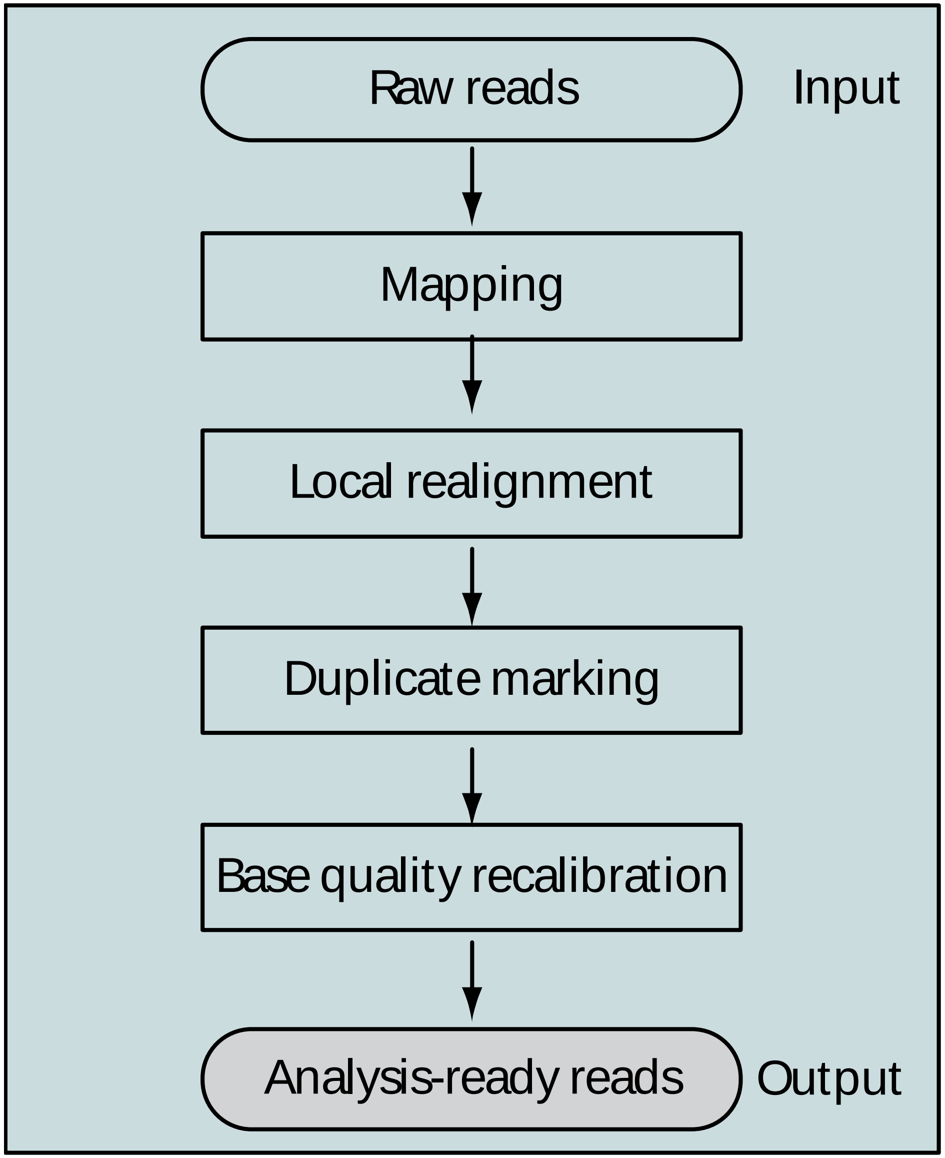
- Overview of the process
  - Motivation to sequence a patient’s genome
  - Oversight, IRB, and informed consent
  - Time frame and costs
  - Inclusion criteria: identifying appropriate patients
  - Exclusion criteria: whose genome not to sequence
- Data acquisition
  - Informed consent, blood, and saliva
  - Obtaining whole genome sequence: the technology
  - The deliverables: catalogs of genetic variants
- Data interpretation
  - Identifying candidate genes
  - Validation

Next-generation sequencing workflow

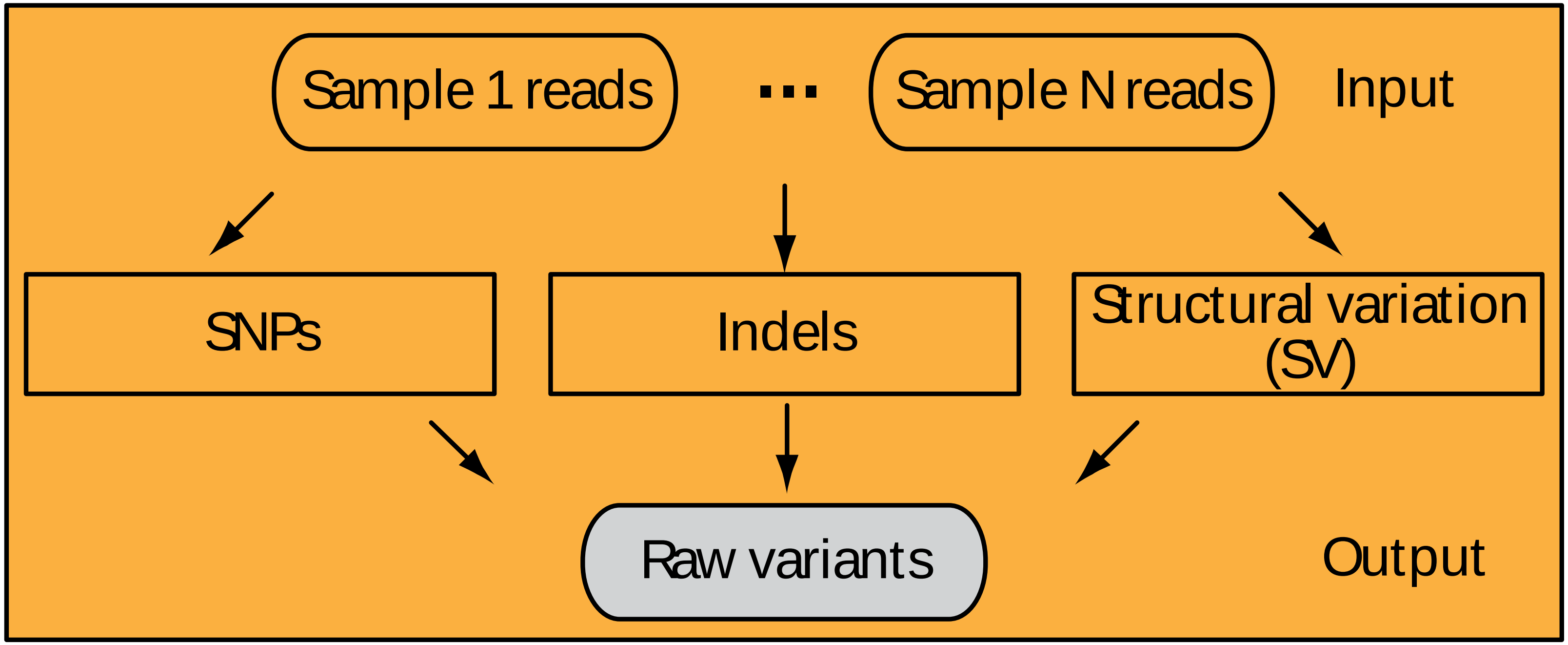


	Stage	Examples/explanation	File formats
	Laboratory work	Experimental design Library preparation Enrichment (capture)	
	Next-generation sequencing	Platforms include Illumina, SOLiD, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Analysis pipeline	Quality assessment	Trimming, filtering Software: FastQC	FASTQ
	Alignment to reference genome	Software: BWA, Bowtie2	Reference: FASTA Output: SAM/BAM
	Variant identification	Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: GATK, SAMTools Realignment, recalibration	Variant Call Format (VCF/BCF)
	Annotation	Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores	
	Visualization	Variant visualization; read depth; comparison to other samples Software: IGV, BEDTools, BigBED	
	Prioritization	Discovery of relevant variants Software: PolyPhen-2, VEP, VAAST	VCF
	Storage	Deposit data in ENA, SRA, dbGaP	BAM, VCF

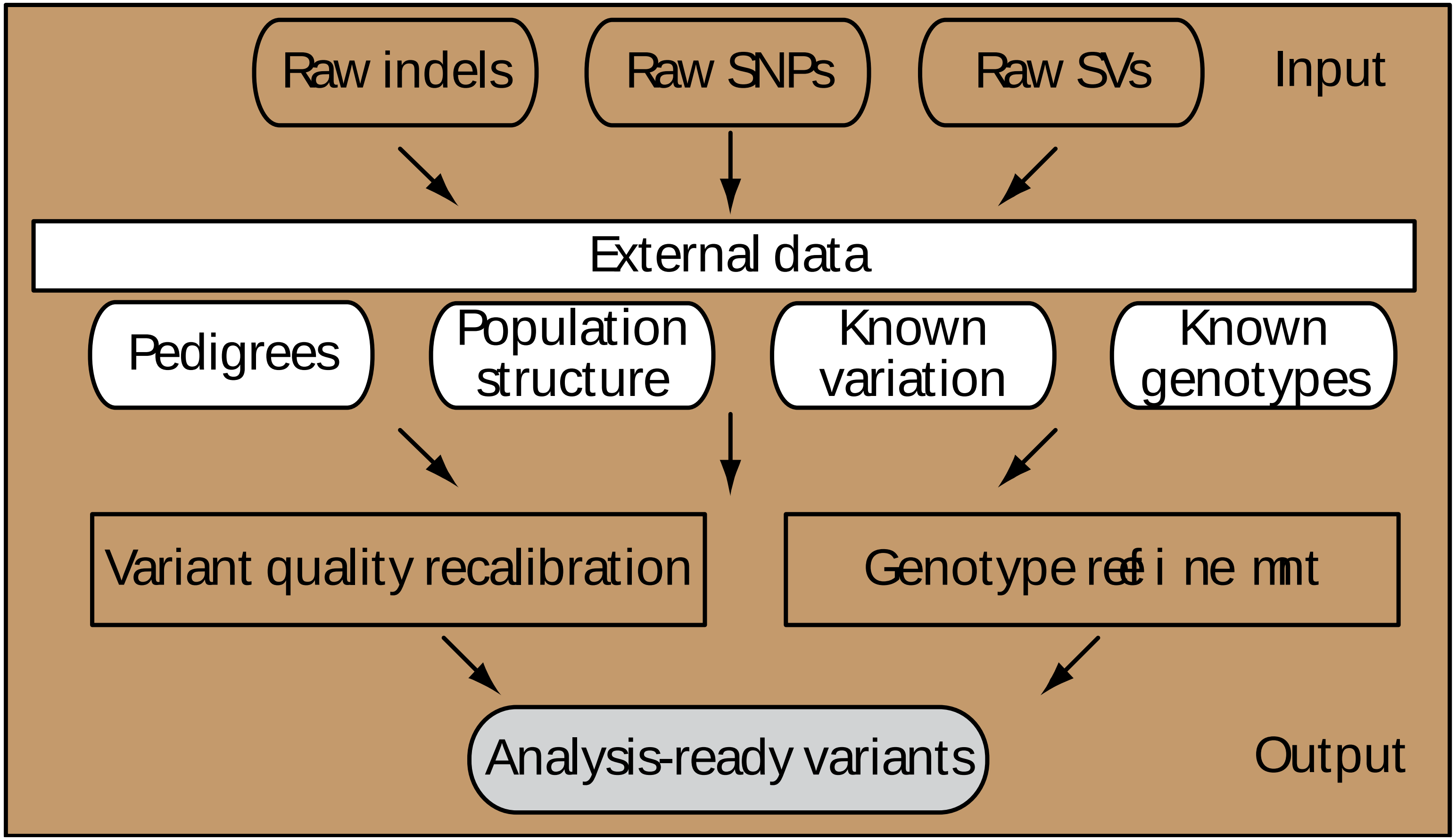
Genome Analysis Toolkit (GATK) workflow
Phase I: data processing



Genome Analysis Toolkit (GATK) workflow
Phase II: variant discovery and genotyping



Genome Analysis Toolkit (GATK) workflow
Phase III: integrative analysis



Raw indels

Raw SNPs

Raw SVs

Input

External data

Pedigrees

Population structure

Known variation

Known genotypes

Variant quality recalibration

Genotype refinement

Analysis-ready variants

Output

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview

Topic 1: Design

Topic 2: FASTQ

Topic 3: Assembly

Topic 4: Alignment

Topic 5: SAM/BAM

Topic 6: Variant calling: SNVs

Topic 7: Variant calling:
SVs

Topic 8: VCF

Topic 9: Visualizing NGS data

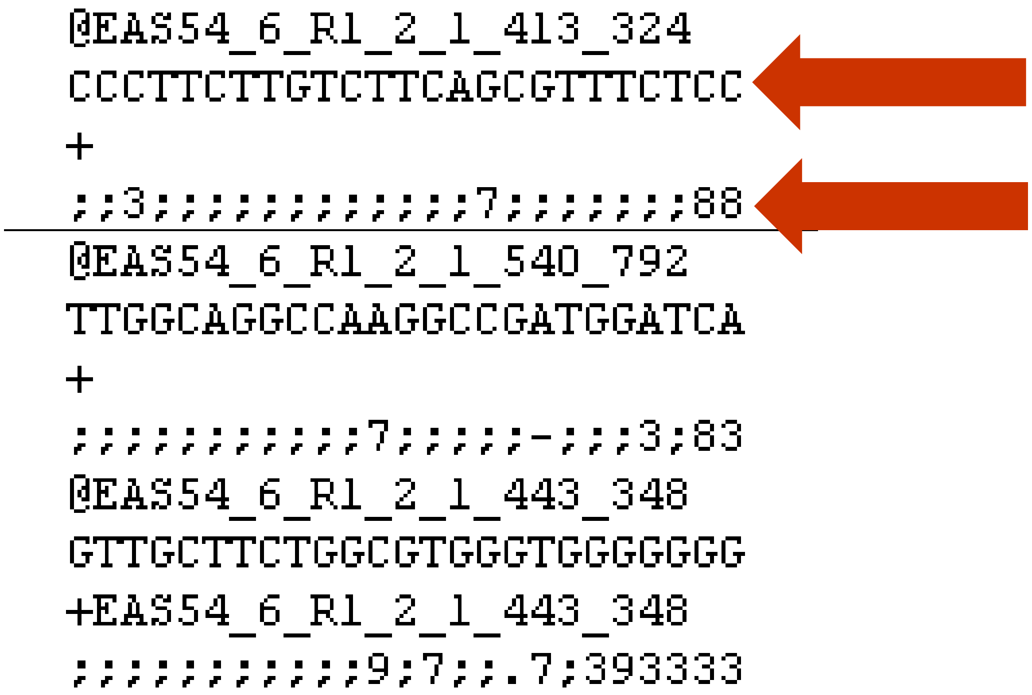
Topic 10: Significance

Specialized applications of NGS

Perspective

FASTQ format

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.

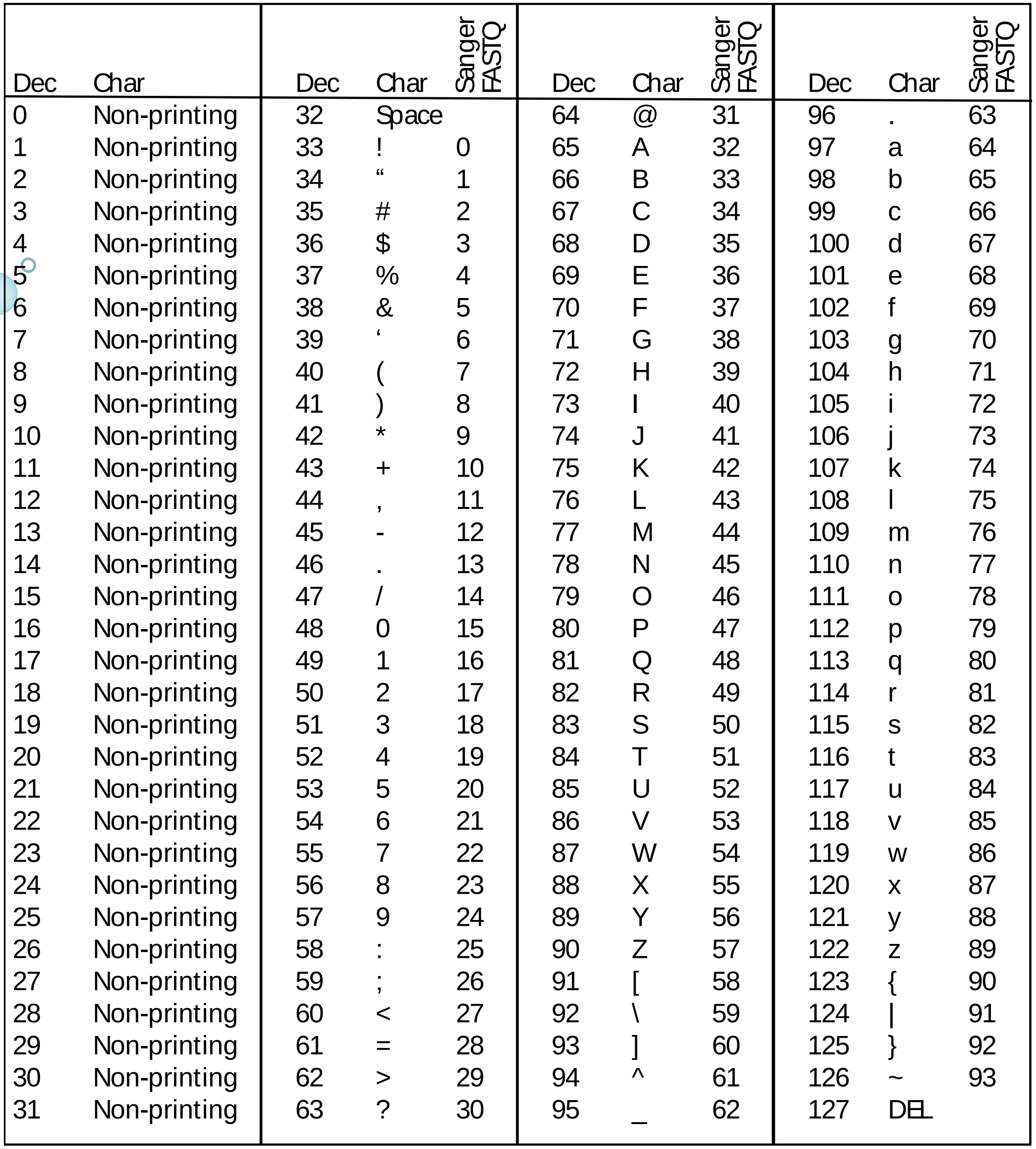


```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

DNA read

Base quality score

FASTQ quality scores use ASCII characters



Dec	Char	Dec	Char	Sanger FASTQ	Dec	Char	Sanger FASTQ	Dec	Char	Sanger FASTQ
0	Non-printing	32	Space		64	@	31	96	.	63
1	Non-printing	33	!	0	65	A	32	97	a	64
2	Non-printing	34	“	1	66	B	33	98	b	65
3	Non-printing	35	#	2	67	C	34	99	c	66
4	Non-printing	36	$	3	68	D	35	100	d	67
5	Non-printing	37	%	4	69	E	36	101	e	68
6	Non-printing	38	&	5	70	F	37	102	f	69
7	Non-printing	39	‘	6	71	G	38	103	g	70
8	Non-printing	40	(	7	72	H	39	104	h	71
9	Non-printing	41	)	8	73	I	40	105	i	72
10	Non-printing	42	*	9	74	J	41	106	j	73
11	Non-printing	43	+	10	75	K	42	107	k	74
12	Non-printing	44	,	11	76	L	43	108	l	75
13	Non-printing	45	-	12	77	M	44	109	m	76
14	Non-printing	46	.	13	78	N	45	110	n	77
15	Non-printing	47	/	14	79	O	46	111	o	78
16	Non-printing	48	0	15	80	P	47	112	p	79
17	Non-printing	49	1	16	81	Q	48	113	q	80
18	Non-printing	50	2	17	82	R	49	114	r	81
19	Non-printing	51	3	18	83	S	50	115	s	82
20	Non-printing	52	4	19	84	T	51	116	t	83
21	Non-printing	53	5	20	85	U	52	117	u	84
22	Non-printing	54	6	21	86	V	53	118	v	85
23	Non-printing	55	7	22	87	W	54	119	w	86
24	Non-printing	56	8	23	88	X	55	120	x	87
25	Non-printing	57	9	24	89	Y	56	121	y	88
26	Non-printing	58	:	25	90	Z	57	122	z	89
27	Non-printing	59	;	26	91	[	58	123	{	90
28	Non-printing	60	<	27	92	\	59	124	|	91
29	Non-printing	61	=	28	93	]	60	125	}	92
30	Non-printing	62	>	29	94	^	61	126	~	93
31	Non-printing	63	?	30	95	_	62	127	DEL	

…relating quality scores (e.g. Q30 for 1 in 10^{-3} error rate) to a compact, one character symbol

B&FG 3e
Fig.9-8
Page 392

You do not need to learn the one character symbols, but you should know the importance of base quality scores in sequence analysis.

FASTQ format: Phred scores define quality

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_{\mathrm{e}})$$

Phred quality score	**Probability of incorrect base call**	**Base call accuracy**
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

FASTQ format: Phred scores define quality

Phred quality scores of each base are usually defined:

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_{\mathrm{e}})$$

There have been alternative base quality definitions:

$$Q_{\mathrm{Solexa}} = -10 \times \log_{10}\left(\frac{P_{\mathrm{e}}}{1 - P_{\mathrm{e}}}\right).$$

$$Q_{\mathrm{PHRED}} = 10 \times \log_{10}(10^{Q_{\mathrm{Solexa}}/10} + 1).$$

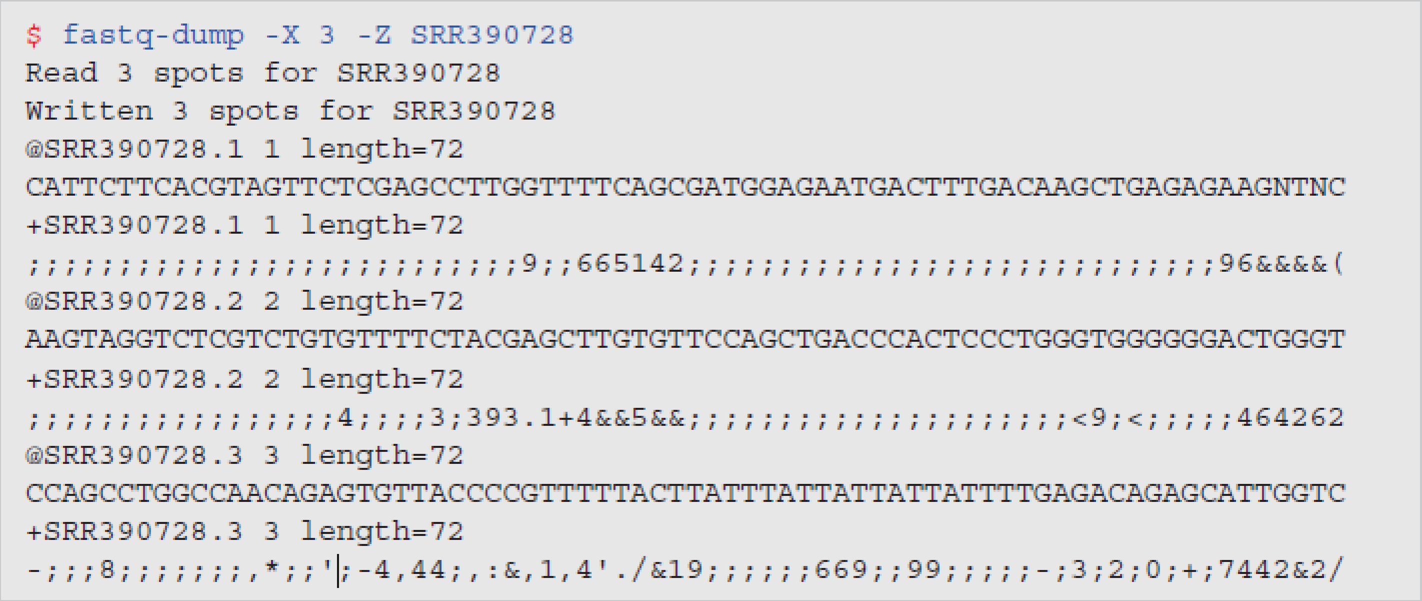
99% of sequence analysis is on the command line (Linux or Mac)

Most next-generation sequence (NGS) analysis is done on the command line. Command line software (using Linux or the Unix-like platform on a Mac terminal) is capable of handling the data analysis tasks, and most NGS software is written for the Unix operating system.

Many people access a Linux (or related Unix) environment while working on a PC or Mac. For example, you can do “cloud computing” in which you pay someone (Amazon, Google, Microsoft) to access their servers. Johns Hopkins has Linux servers you can access (https://www.marcc.jhu.edu).

The next three slides provide examples of command-line tools to look at FASTQ-formatted files.

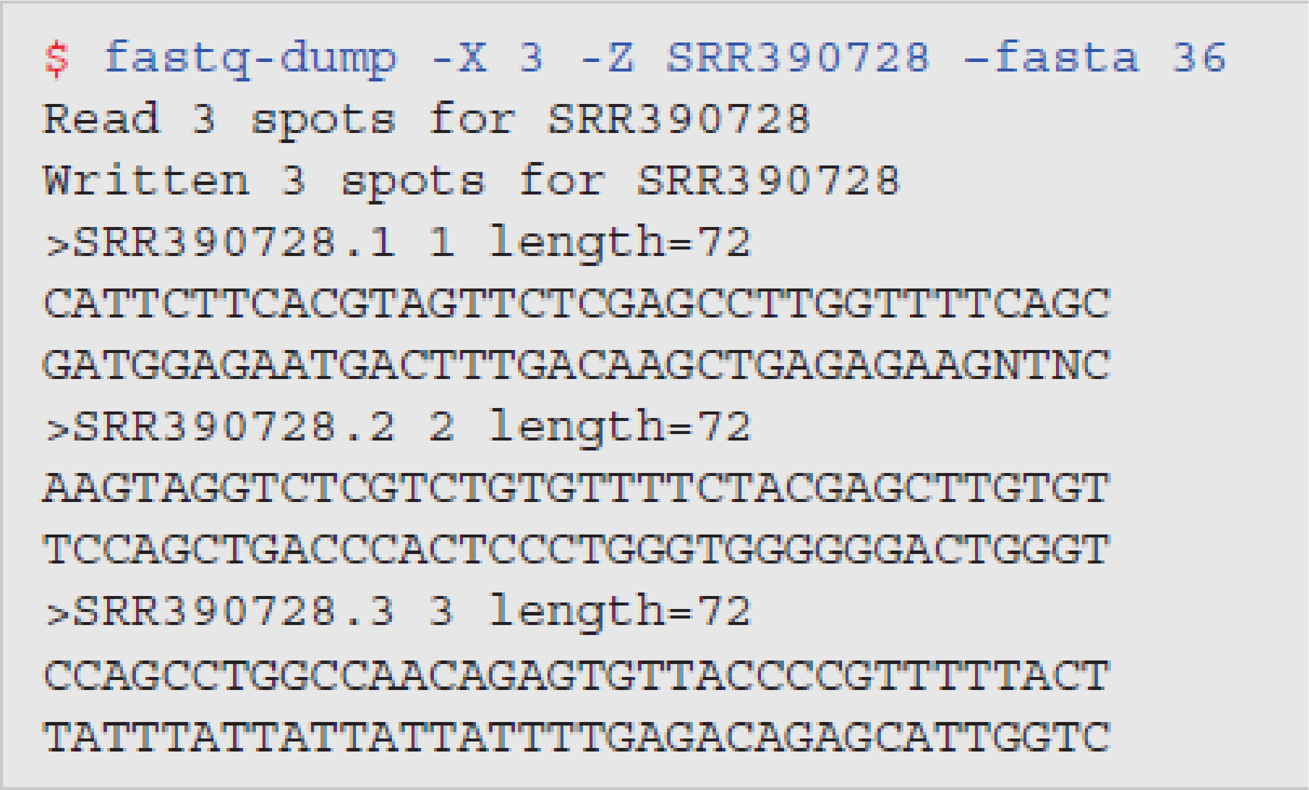
SRA toolkit: `fastq-dump` to obtain FASTQ formatted data



```
$ fastq-dump -X 3 -Z SRR390728
Read 3 spots for SRR390728
Written 3 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
@SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTTACCCCGTTTTTACTTATTTATTATTATTATTTTGAGACAGAGCATTGGTC
+SRR390728.3 3 length=72
-;;;8;;;;;;;,*;;';-4,44;,:&,1,4'./&19;;;;;;669;;99;;;;;-;3;2;0;+;7442&2/
```

NCBI offers the SRA Toolkit to manipulate sequence data. The `fastq-dump` command can pull FASTQ-formatted data from an accession number (such as SRR390728).

SRA toolkit:
`fastq-dump` to obtain FASTA formatted data



```
$ fastq-dump -X 3 -Z SRR390728 -fasta 36
Read 3 spots for SRR390728
Written 3 spots for SRR390728
>SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGC
GATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
>SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGT
TCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
>SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTTACCCCGTTTTTACT
TATTTATTATTATTATTTTGAGACAGAGCATTGGTC
```

Finding FASTQ files

There are two main places you can find FASTQ files.

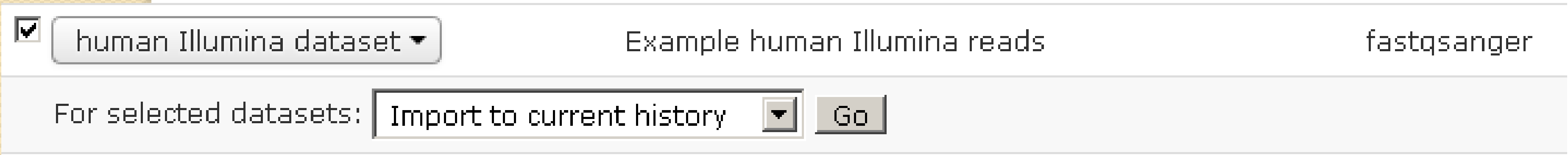
(1) The central repositories at NCBI and EBI

(2) A sequencing core: data are often returned to investigators in the FASTQ format. (In some cases the data are returned in the BAM format, discussed next, from which FASTQ-formatted data can be retrieved.)

FASTQ format: where to learn more

- FASTQ project page
http://maq.sourceforge.net/fastq.shtml

- You can look at FASTQ files in Galaxy > Shared data > Data libraries > Sample NGS Datasets > Human Illumina dataset. Check the box, click Go, and the data are entered in Galaxy (see the Analyze Data tab where you usually begin a Galaxy session).

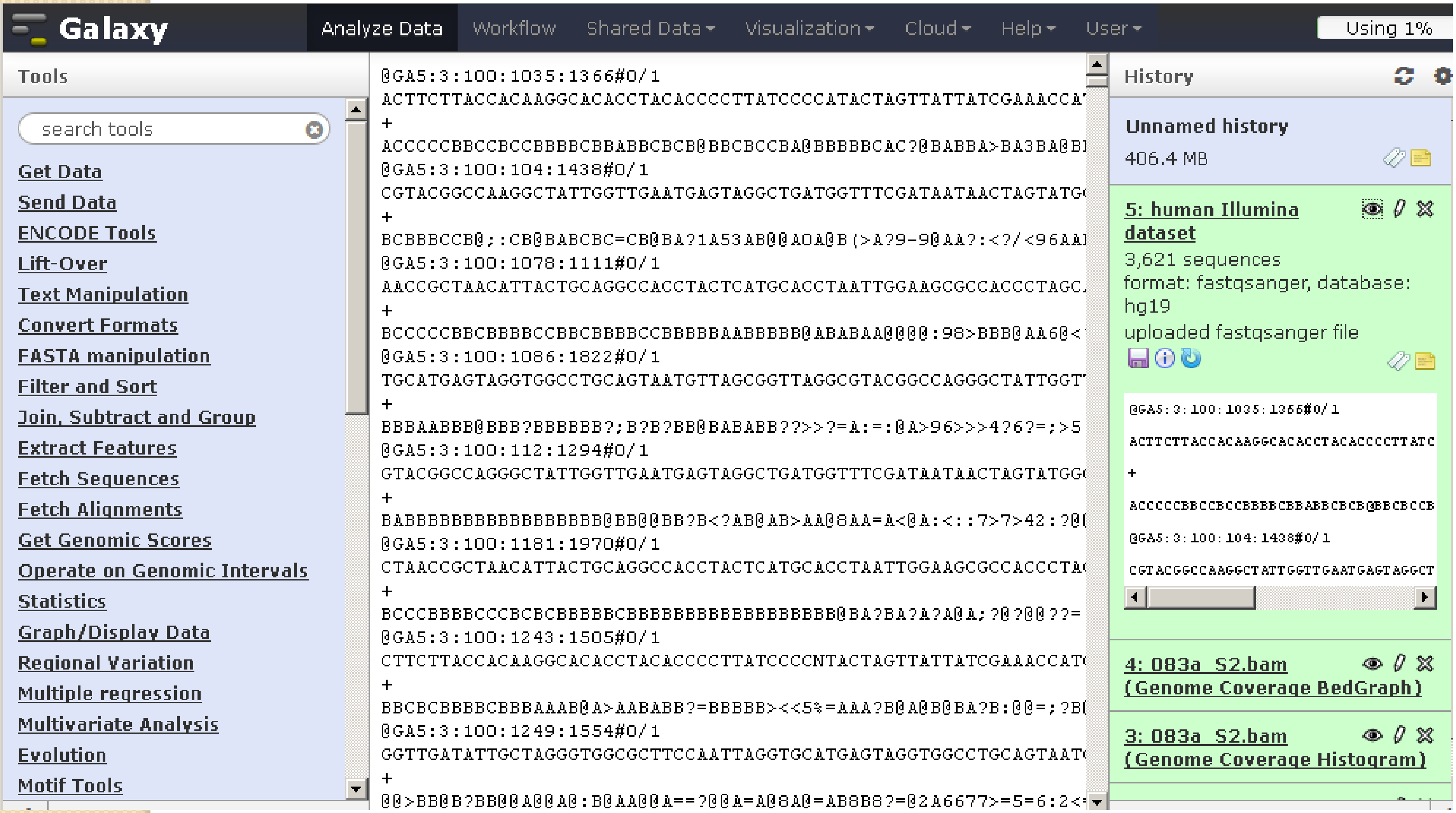


☑ human Illumina dataset ▾ Example human Illumina reads fastqsanger

For selected datasets: Import to current history ▾ Go

- Galaxy also offers helpful videocasts about manipulating FASTQ files.

Example of FASTQ data in Galaxy



```
@GA5:3:100:1035:1366#0/1
ACTTCTTACCACAAGGCACACCTACACCCCTTATCCCCATACTAGTTATTATCGAAACCA
+
ACCCCCBBCCBCCBBBBCBBABBCBCB@BBCBCCBA@BBBBBCAC?@BABBA>BA3BA@B
@GA5:3:100:104:1438#0/1
CGTACGGCCAAGGCTATTGGTTGAATGAGTAGGCTGATGGTTTCGATAATAACTAGTATG
+
BCBBBCCB@;:CB@BABCBC=CB@BA?1A53AB@@AOA@B(>A?9-9@AA?:<?/<96AA
@GA5:3:100:1078:1111#0/1
AACCGCTAACATTACTGCAGGCCACCTACTCATGCACCTAATTGGAAGCGCCACCCTAGC
+
BCCCCCBBCBBBBCCBBCBBBBCCBBBBBAABBBBB@ABABAA@@@@:98>BBB@AA6@<
@GA5:3:100:1086:1822#0/1
TGCATGAGTAGGTGGCCTGCAGTAATGTTAGCGGTTAGGCGTACGGCCAGGGCTATTGGT
+
BBBAABBB@BBB?BBBBBB?;B?B?BB@BABABB??>>?=A:=:@A>96>>>4?6?=;>5
@GA5:3:100:112:1294#0/1
GTACGGCCAGGGCTATTGGTTGAATGAGTAGGCTGATGGTTTCGATAATAACTAGTATGG
+
BABBBBBBBBBBBBBBBBB@BB@@BB?B<?AB@AB>AA@8AA=A<@A:<::7>7>42:?@
@GA5:3:100:1181:1970#0/1
CTAACCGCTAACATTACTGCAGGCCACCTACTCATGCACCTAATTGGAAGCGCCACCCTA
+
BCCCBBBBCCCBCBCBBBBBCBBBBBBBBBBBBBBBBB@BA?BA?A?A@A;?@?@@??=
@GA5:3:100:1243:1505#0/1
CTTCTTACCACAAGGCACACCTACACCCCTTATCCCCNTACTAGTTATTATCGAAACCAT
+
BBCBCBBBBCBBBAAAB@A>AABABB?=BBBBB><<5%=AAA?B@A@B@BA?B:@@=;?B
@GA5:3:100:1249:1554#0/1
GGTTGATATTGCTAGGGTGGCGCTTCCAATTAGGTGCATGAGTAGGTGGCCTGCAGTAAT
+
@@>BB@B?BB@@A@@A@:B@AA@@A==?@@A=A@8A@=AB8B8?=@2A6677>=5=6:2<
```

Unnamed history
406.4 MB

5: human Illumina dataset
3,621 sequences
format: fastqsanger, database: hg19
uploaded fastqsanger file

```
@GA5:3:100:1035:1366#0/1
ACTTCTTACCACAAGGCACACCTACACCCCTTATC
+
ACCCCCBBCCBCCBBBBCBBABBCBCB@BBCBCCB
@GA5:3:100:104:1438#0/1
CGTACGGCCAAGGCTATTGGTTGAATGAGTAGGCT
```

4: 083a_S2.bam (Genome Coverage BedGraph)

3: 083a_S2.bam (Genome Coverage Histogram)

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview Topic 6: Variant calling: SNVs
Topic 1: Design Topic 7: Variant calling:
SVs
Topic 2: FASTQ Topic 8: VCF
Topic 3: Assembly Topic 9: Visualizing NGS data
Topic 4: Alignment Topic 10: Significance
Topic 5: SAM/BAM

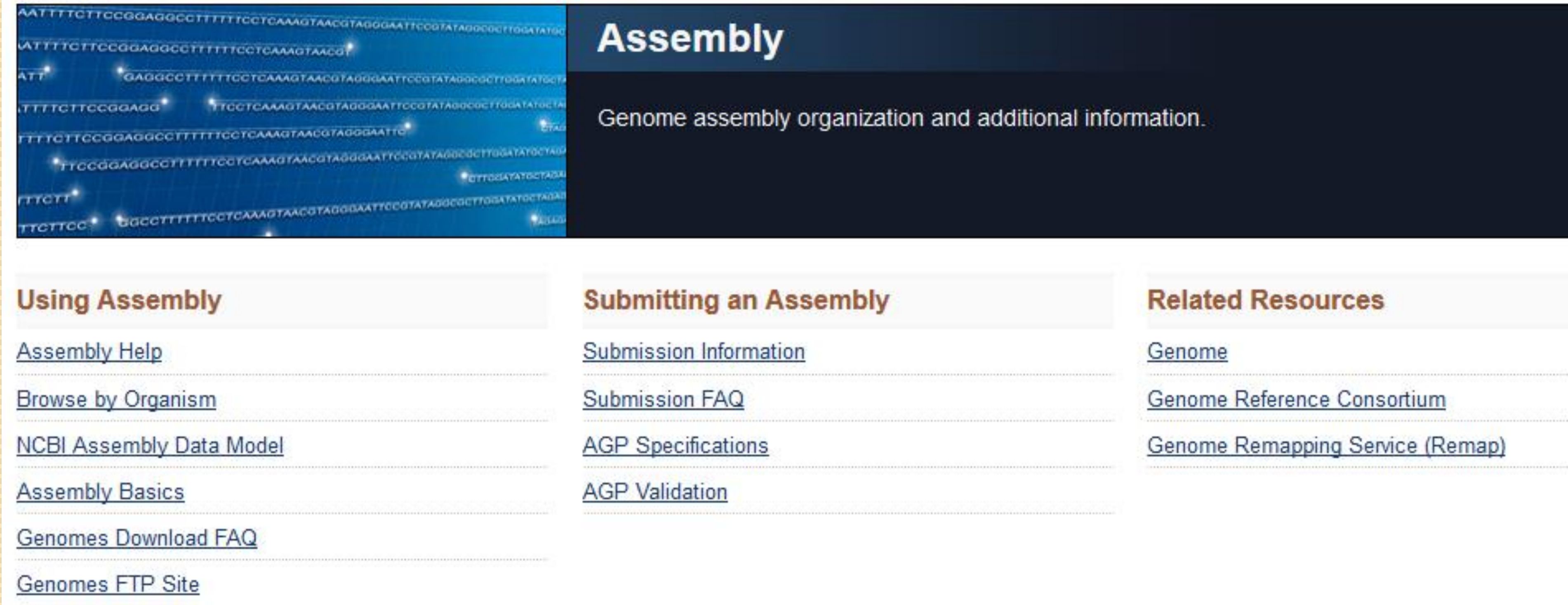
Specialized applications of NGS

Perspective

Genome assembly

Genome assembly is the process of converting short reads into a detailed set of sequences corresponding to the chromosome(s) of an organism.

To learn more about assembly visit
http://www.ncbi.nlm.nih.gov/assembly/
http://www.ncbi.nlm.nih.gov/assembly/basics/



Assembly

Genome assembly organization and additional information.

Using Assembly

- Assembly Help
- Browse by Organism
- NCBI Assembly Data Model
- Assembly Basics
- Genomes Download FAQ
- Genomes FTP Site

Submitting an Assembly

- Submission Information
- Submission FAQ
- AGP Specifications
- AGP Validation

Related Resources

- Genome
- Genome Reference Consortium
- Genome Remapping Service (Remap)

Genome assembly: relevance

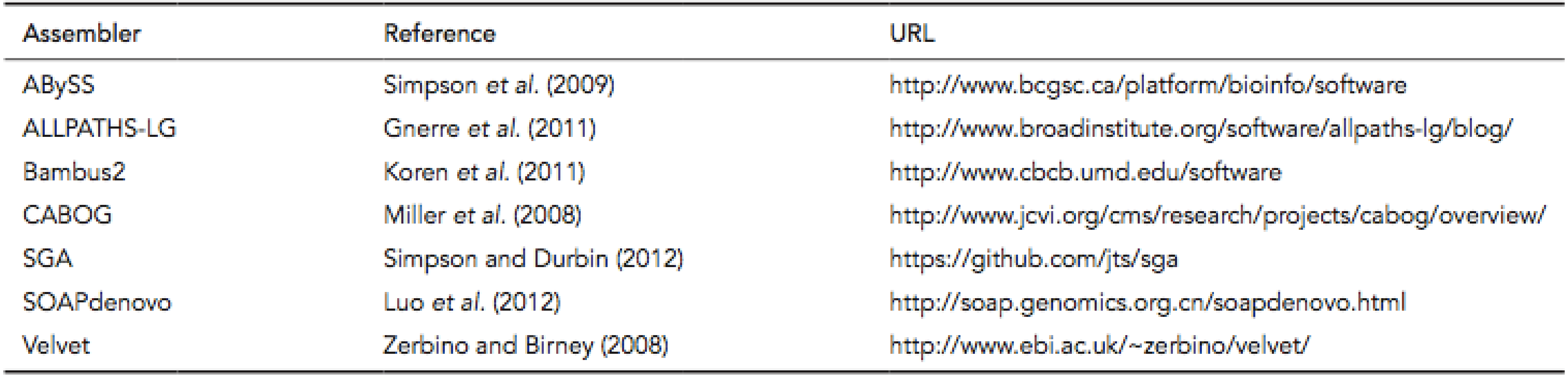
- Genome assembly is needed when a genome is first sequenced. We can relate reads to chromosomes.

- For the human genome, the assembly is "frozen" as a snapshot every few years. The current assembly is GRCh38. (GRC refers to Genome Reference Consortium at http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)

- For most human genome work we do not need to do "de novo" (from anew) assembly. Instead we map reads to a reference genome—one that is already assembled.

- Genome assembly is a crucial behind-the-scenes part of calling human genome (or other) variants.

Software for genome assembly



Assembler	Reference	URL
ABySS	Simpson *et al.* (2009)	http://www.bcgsc.ca/platform/bioinfo/software
ALLPATHS-LG	Gnerre *et al.* (2011)	http://www.broadinstitute.org/software/allpaths-lg/blog/
Bambus2	Koren *et al.* (2011)	http://www.cbcb.umd.edu/software
CABOG	Miller *et al.* (2008)	http://www.jcvi.org/cms/research/projects/cabog/overview/
SGA	Simpson and Durbin (2012)	https://github.com/jts/sga
SOAPdenovo	Luo *et al.* (2012)	http://soap.genomics.org.cn/soapdenovo.html
Velvet	Zerbino and Birney (2008)	http://www.ebi.ac.uk/~zerbino/velvet/

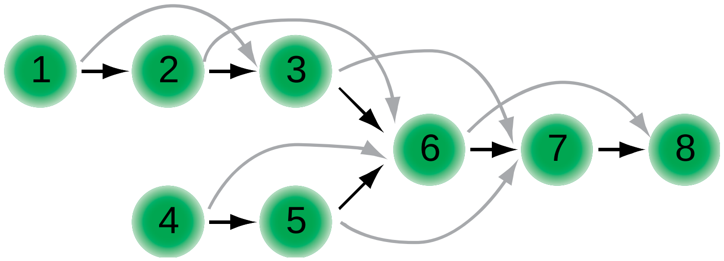
Velvet for assembly.

Genome assembly methods: overlap graph, de Bruijn graph, string graph

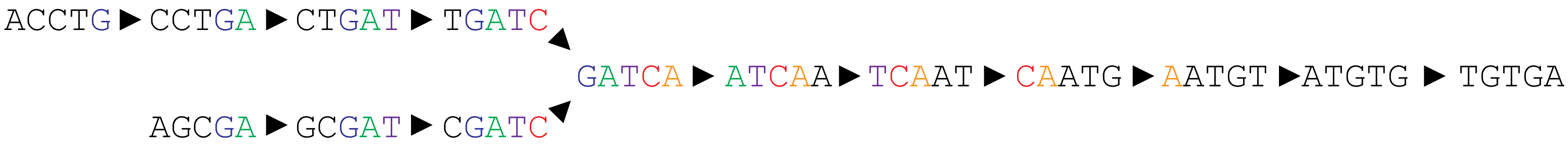
```
1  ACCTGATC
2    CTGATCAA
3     TGATCAAT
4   AGCGATCA
5     CGATCAAT
6      GATCAATG
7        TCAATGTG
8         CAATGTGA
```

reads

overlap graph

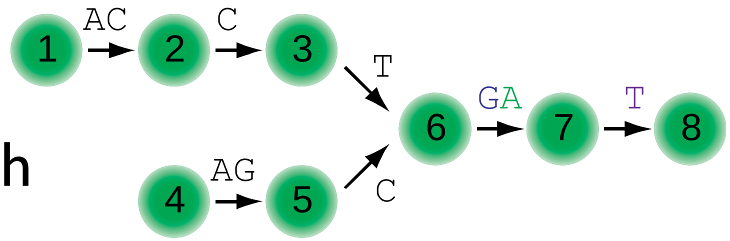


de Bruijn graph


ACCTG ▶ CCTGA ▶ CTGAT ▶ TGATC

GATCA ▶ ATCAA ▶ TCAAT ▶ CAATG ▶ AATGT ▶ATGTG ▶ TGTGA

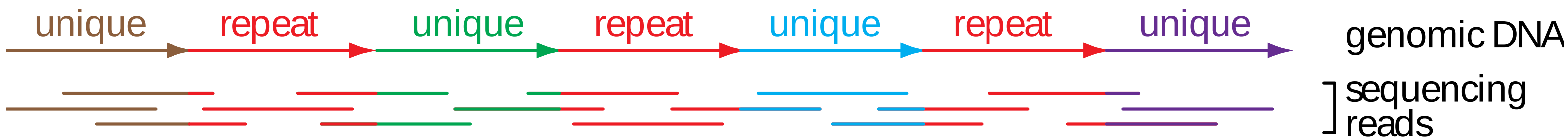
AGCGA ▶ GCGAT ▶ CGATC



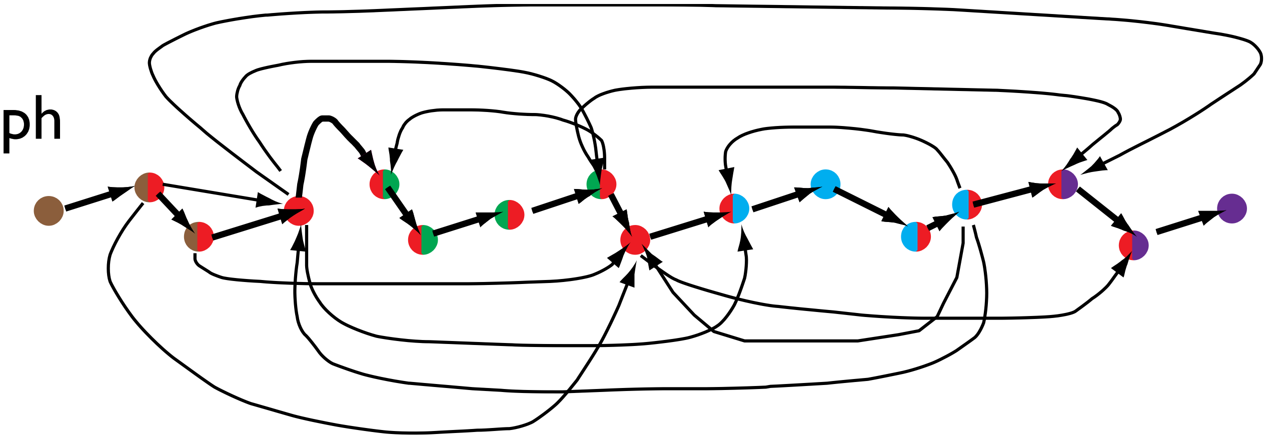
string graph

Genome assembly with overlap graph and de Bruijn graph

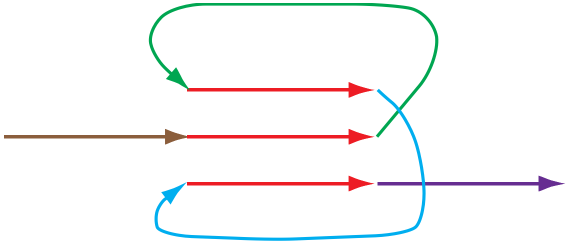
DNA sequence with a triple repeat



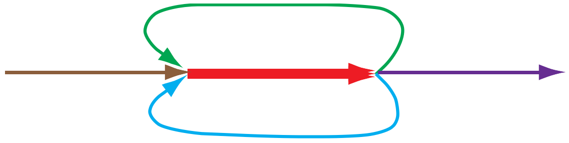
Layout graph



Construction of de Bruijn graph by gluing repeats

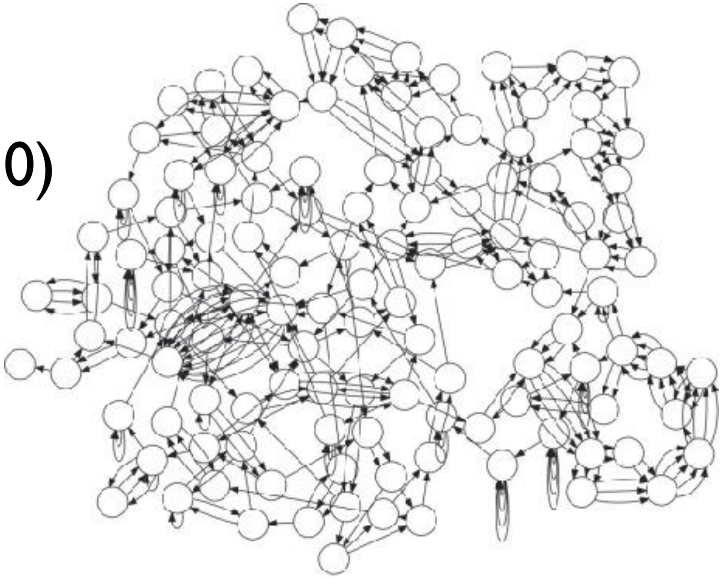


de Bruijn graph

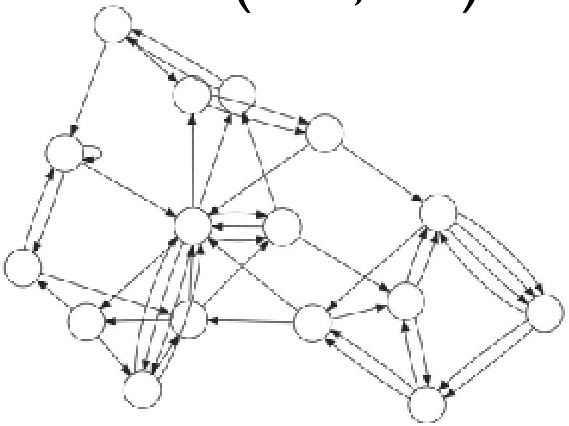


de Bruijn graphs resolve assembly with higher *k* values

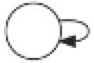
E. coli K12 (k=50)



E. coli K12 (k=1,000)



E. coli K12 (k=5,000)



Source: PMID 24034426

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

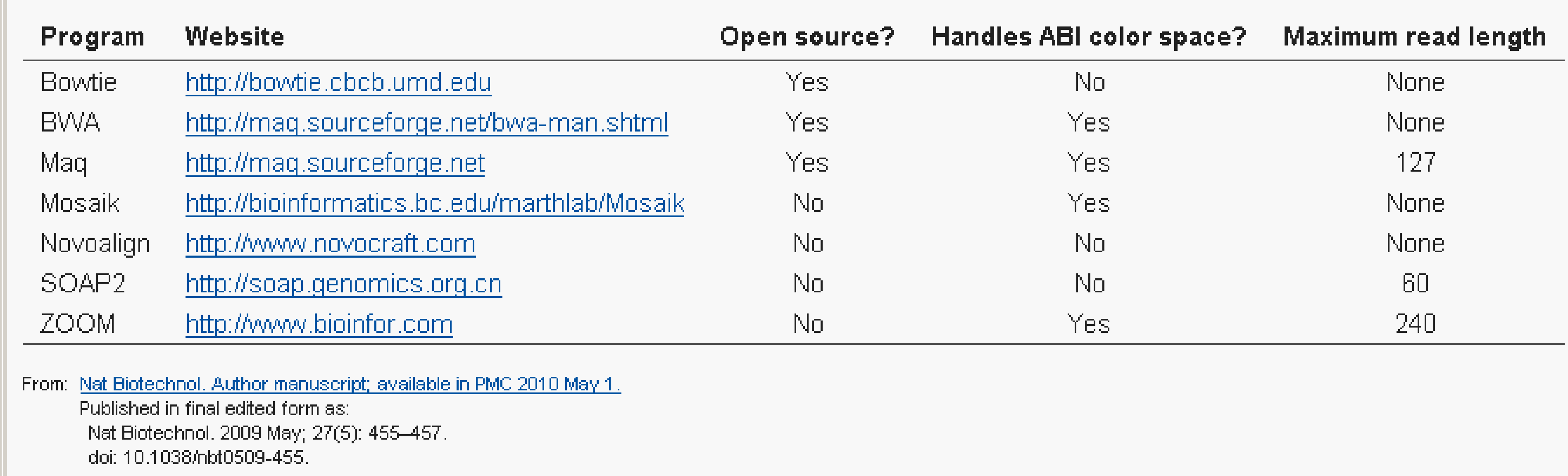
Analysis of NGS sequencing of genomic DNA

Overview
Topic 6: Variant calling: SNVs
Topic 1: Design
Topic 7: Variant calling:
SVs
Topic 2: FASTQ
Topic 8: VCF
Topic 3: Assembly
Topic 9: Visualizing NGS data
Topic 4: Alignment
Topic 10: Significance
Topic 5: SAM/BAM

Specialized applications of NGS

Perspective

Next-generation sequence: the problem of alignment



Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

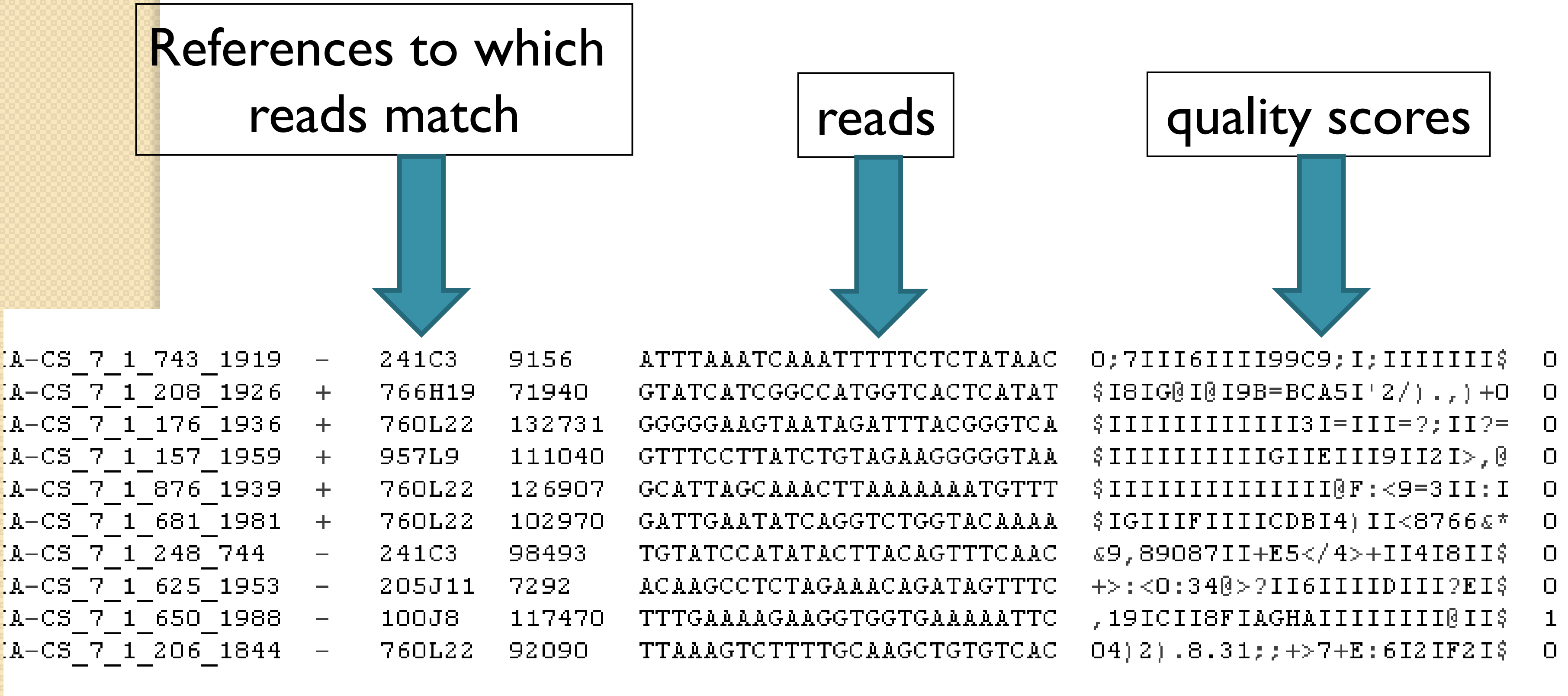
From: Nat Biotechnol. Author manuscript; available in PMC 2010 May 1.
Published in final edited form as:
Nat Biotechnol. 2009 May; 27(5): 455–457.
doi: 10.1038/nbt0509-455.

Recent software tools allow the mapping (alignment) of millions or billions of short reads to a reference genome.

--For the human genome, this would take thousands of hours using BLAST.

--Reads may come from regions of repetitive DNA (exacerbated by sequencing errors)

Alignment to a reference genome: example of short-read alignment (Bowtie) results

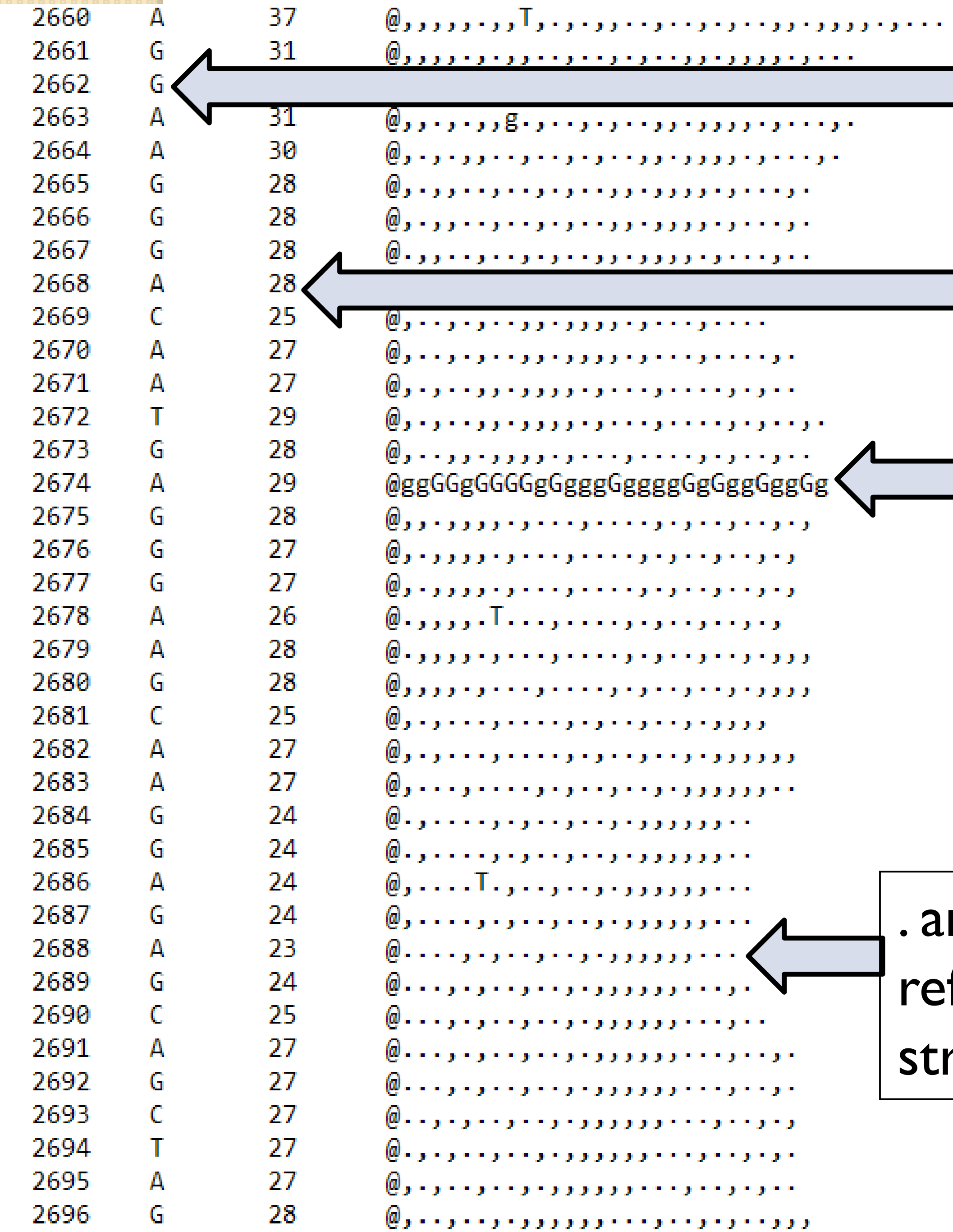


References to which reads match

reads

quality scores

```
A-CS_7_1_743_1919  -   241C3   9156    ATTTAAATCAAATTTTTCTCTATAAC  0;7III6IIII99C9;I;IIIIIII$  0
A-CS_7_1_208_1926  +   766H19  71940   GTATCATCGGCCATGGTCACTCATAT  $I8IG@I@I9B=BCA5I'2/).,)+0  0
A-CS_7_1_176_1936  +   760L22  132731  GGGGGAAGTAATAGATTTACGGGTCA  $IIIIIIIIIII3I=III=?;II?=  0
A-CS_7_1_157_1959  +   957L9   111040  GTTTCCTTATCTGTAGAAGGGGGTAA  $IIIIIIIIIGIIEIII9II2I>,@  0
A-CS_7_1_876_1939  +   760L22  126907  GCATTAGCAAACTTAAAAAAATGTTT  $IIIIIIIIIIIII@F:<9=3II:I  0
A-CS_7_1_681_1981  +   760L22  102970  GATTGAATATCAGGTCTGGTACAAAA  $IGIIIFIIIICDBI4)II<8766&*  0
A-CS_7_1_248_744   -   241C3   98493   TGTATCCATATACTTACAGTTTCAAC  &9,89087II+E5</4>+II4I8II$  0
A-CS_7_1_625_1953  -   205J11  7292    ACAAGCCTCTAGAAACAGATAGTTTC  +>:<0:34@>?II6IIIIDIII?EI$  0
A-CS_7_1_650_1988  -   100J8   117470  TTTGAAAAGAAGGTGGTGAAAAATTC  ,19ICII8FIAGHAIIIIIIII@II$  1
A-CS_7_1_206_1844  -   760L22  92090   TTAAAGTCTTTTGCAAGCTGTGTCAC  04)2).8.31;;+>7+E:6I2IF2I$  0
```



```
2660    A    37    @,,,,,.,,T,.,.,,..,..,.,..,,.,,,,.,...
2661    G    31    @,,,,.,.,,..,..,.,..,,.,,,,.,...
2662    G
2663    A    31    @,,.,.,,g.,..,.,..,,.,,,,.,...,.
2664    A    30    @,.,.,,..,..,.,..,,.,,,,.,...,.
2665    G    28    @,.,,..,..,.,..,,.,,,,.,...,.
2666    G    28    @,.,,..,..,.,..,,.,,,,.,...,.
2667    G    28    @.,,..,..,.,..,,.,,,,.,...,..
2668    A
2669    C    25    @,..,.,..,,.,,,,.,...,....
2670    A    27    @,..,.,..,,.,,,,.,...,....,.
2671    A    27    @,.,..,,.,,,,.,...,....,.,..
2672    T    29    @,.,..,,.,,,,.,...,....,.,..,.
2673    G    28    @,..,,.,,,,.,...,....,.,..,..
2674    A    29    @ggGGgGGGGgGgggGggggGgGggGggGg
2675    G    28    @,,.,,,,.,...,....,.,..,..,.,
2676    G    27    @,.,,,,.,...,....,.,..,..,.,
2677    G    27    @,.,,,,.,...,....,.,..,..,.,
2678    A    26    @.,,,,.T...,....,.,..,..,.,
2679    A    28    @.,,,,.,...,....,.,..,..,.,,,
2680    G    28    @,,,,.,...,....,.,..,..,.,,,,
2681    C    25    @,.,...,....,.,..,..,.,,,,
2682    A    27    @,.,...,....,.,..,..,.,,,,,,
2683    A    27    @,...,....,.,..,..,.,,,,,,..
2684    G    24    @.,....,.,..,..,.,,,,,,..
2685    G    24    @.,....,.,..,..,.,,,,,,..
2686    A    24    @,....T.,..,..,.,,,,,,...
2687    G    24    @,....,.,..,..,.,,,,,,...
2688    A    23    @....,.,..,..,.,,,,,,...
2689    G    24    @...,.,..,..,.,,,,,,...,.
2690    C    25    @...,.,..,..,.,,,,,,...,..
2691    A    27    @...,.,..,..,.,,,,,,...,..,.
2692    G    27    @...,.,..,..,.,,,,,,...,..,.
2693    C    27    @..,.,..,..,.,,,,,,...,..,.,
2694    T    27    @.,.,..,..,.,,,,,,...,..,.,.
2695    A    27    @,.,..,..,.,,,,,,...,..,.,..
2696    G    28    @,..,..,.,,,,,,...,..,.,..,,,
```

Reference sequence (5 Mb, fasta format)

Read depth

Reference sequence A; Sample has G 29 times

. and , denote agreement with reference on top, bottom strands

MAQ analysis

BWA: a popular short-read aligner

- Aligns short reads (<200 base pairs) to a reference genome

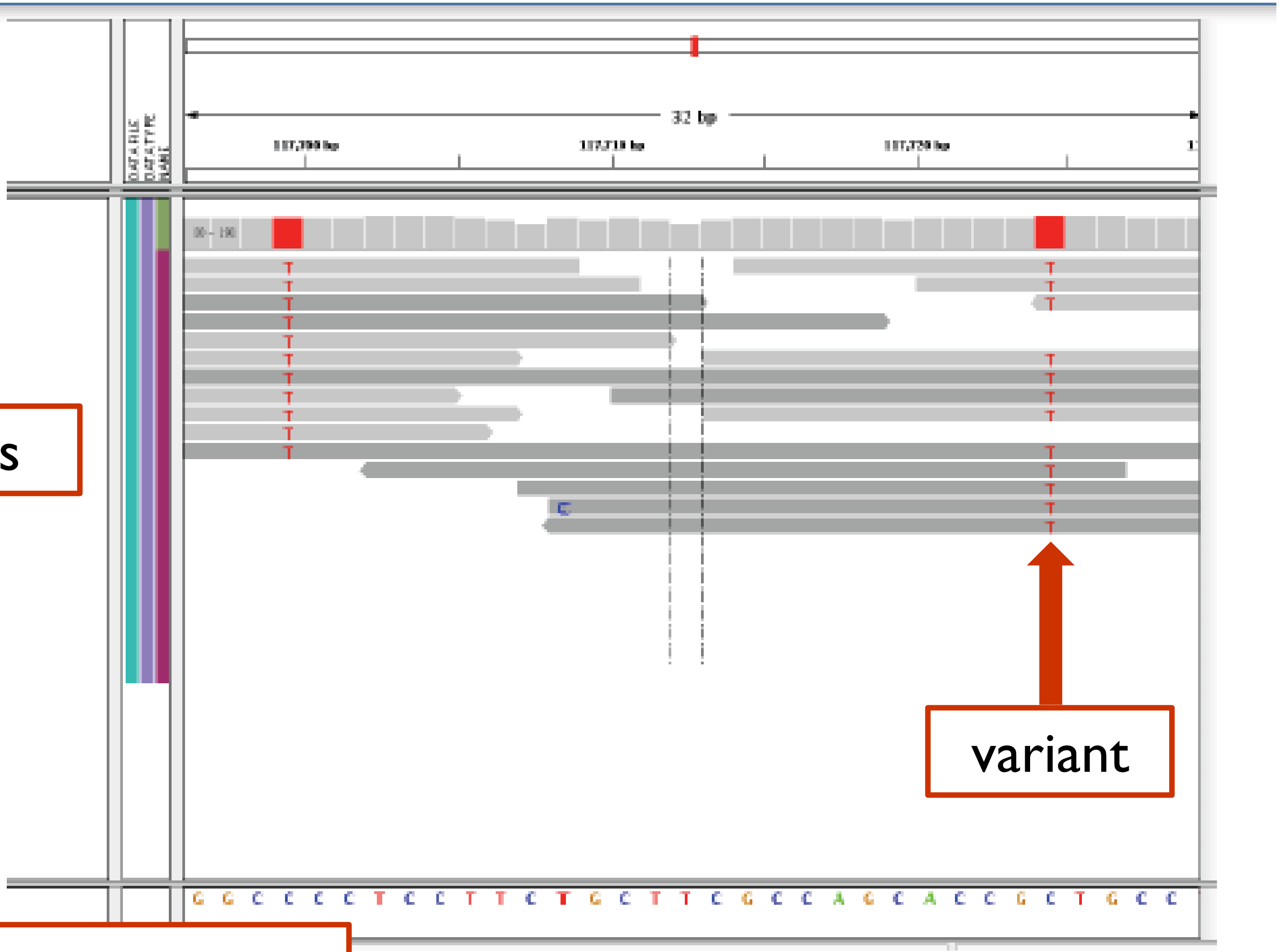
- Fast, accurate

- Learn more at http://bio-bwa.sourceforge.net/

- Command-line software for the Linux environment (like essentially all NGS tools)

- Try it in a web-accessible version! Go to Galaxy > see list of tools on left sidebar > NGS Toolbox beta > NGS: Mapping > Map with BWA for Illumina

Next-generation sequence data: visualizing of short reads aligned to a reference genome



reads

variant

reference genome

Reads (FASTQ format) can be mapped to a reference genome using software tools such as BWA

- There are dozens of aligners to choose from.

- Each aligner has many parameters you can choose.

- BWA is a popular aligner. It stands for "Burroughs-Wheeler Aligner" referring to the algorithmic approach. See http://bio-bwa.sourceforge.net

Reads (FASTQ format) can be mapped to a reference genome using software tools such as BWA (cont.)

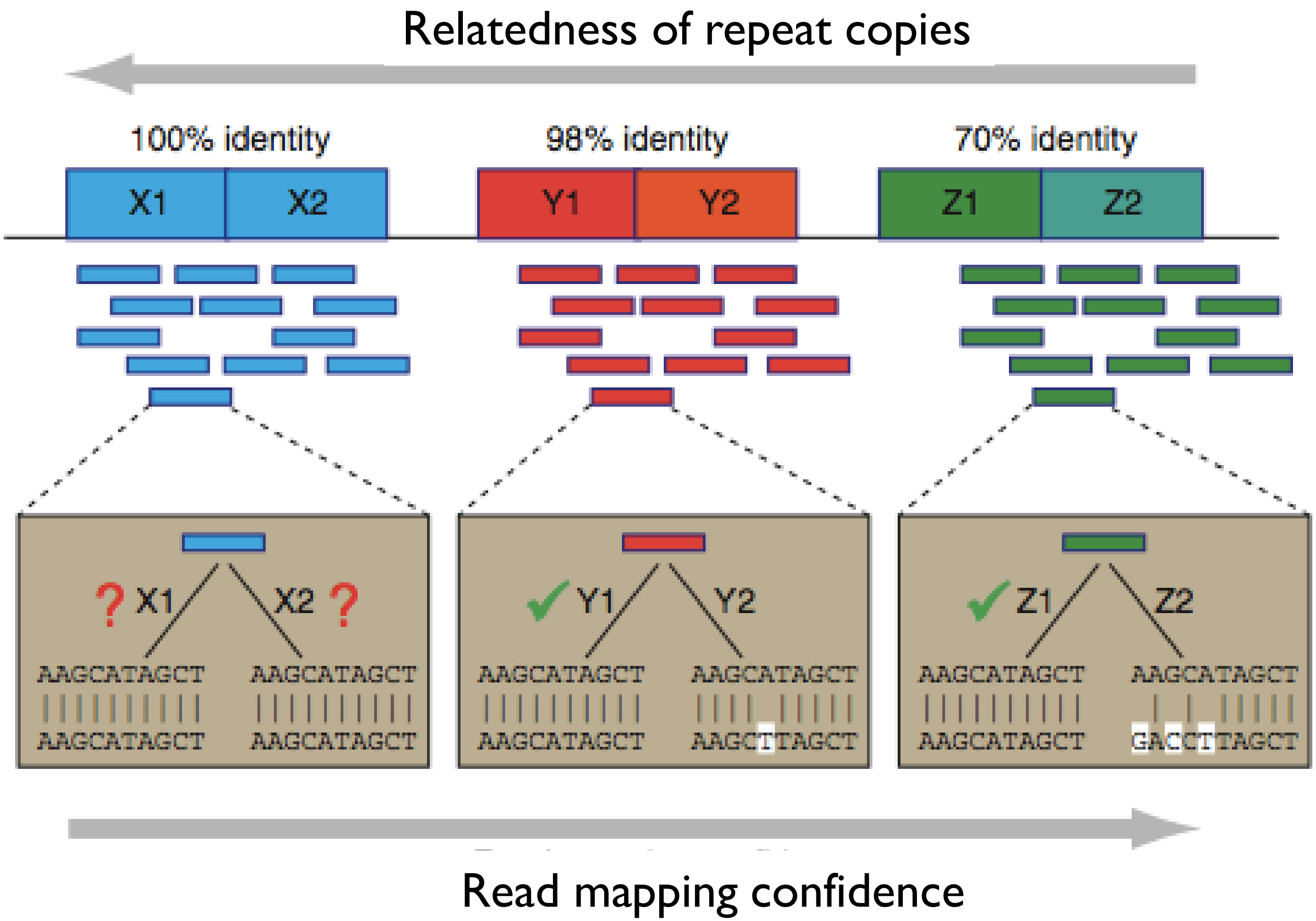
- Considerations are speed and sensitivity.

- For all software we measure error rates: using some gold standard we define true positive (TP) and true negative (TN) results, and we then define sensitivity and specificity.

- A standard format has been introduced called Sequence Alignment/Map (SAM). Its binary version (which is compressed) is called BAM.

- Google SAM/BAM for specifications & more information.

As repeat regions share *lower* identity, read mapping gains *higher* confidence



Relatedness of repeat copies

Read mapping confidence

Source: PMID 22124482

There is ambiguity mapping a read with a mismatch versus a deletion

```
      location 1 (mismatch)               location 2 (deletion)
...TTTAGAATGAGCCGAGTTCGCGCGGGTAGAAT-AGCCGAGTT...     genomic DNA
      ||||| |||||||           ||||| |||||||
      AGAATTAGCCGAG           AGAATTAGCCGAG
        13 bp read              13 bp read
```

Source: PMID xxxxxxxx

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

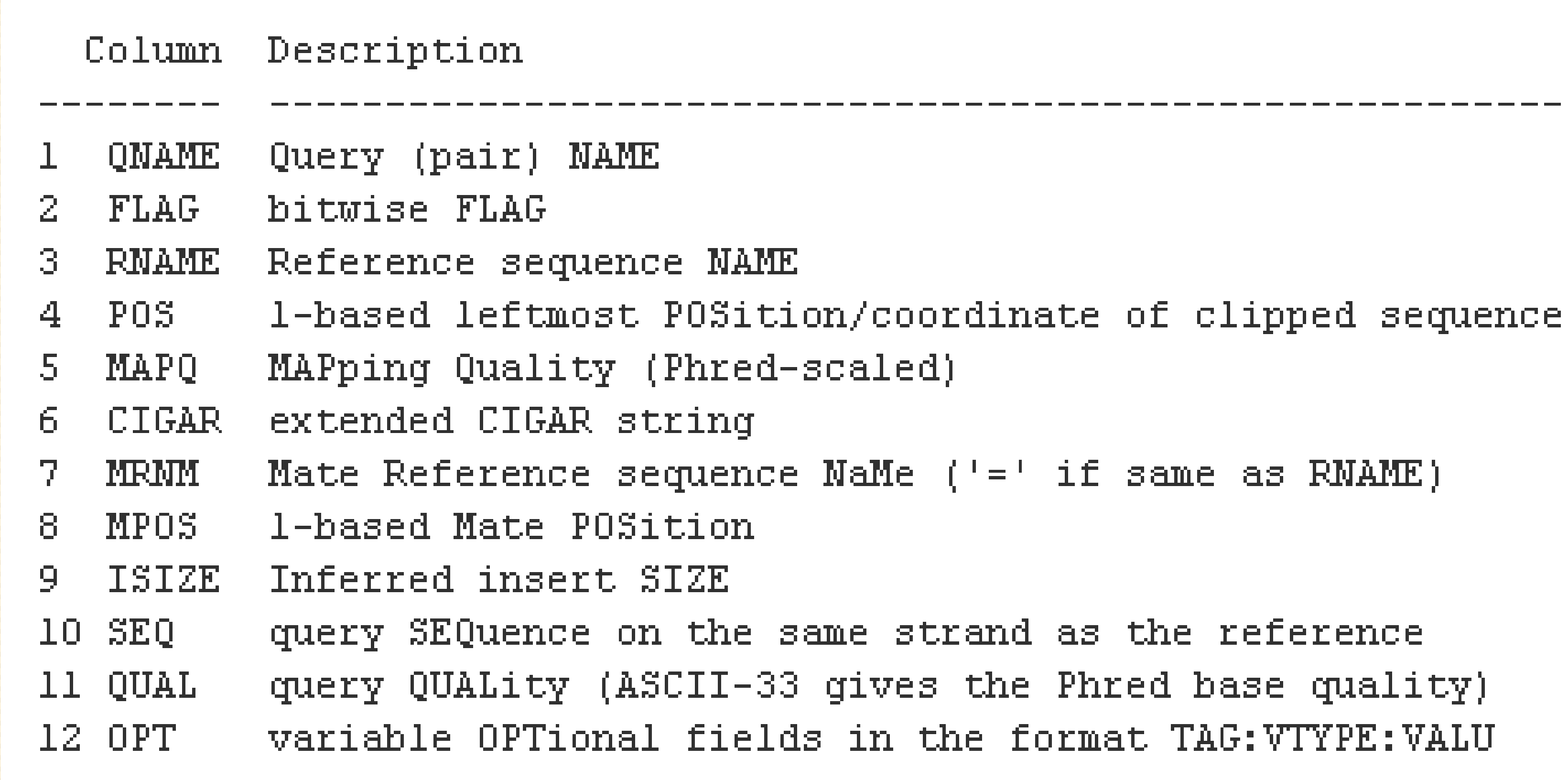
Overview
Topic 1: Design
Topic 2: FASTQ
Topic 3: Assembly
Topic 4: Alignment
Topic 5: SAM/BAM

Topic 6: Variant calling: SNVs
Topic 7: Variant calling: SVs
Topic 8: VCF
Topic 9: Visualizing NGS data
Topic 10: Significance

Specialized applications of NGS

Perspective

BWA and other aligners produce output in the SAM format



```
  Column  Description
--------  ---------------------------------------------------------
1  QNAME  Query (pair) NAME
2  FLAG   bitwise FLAG
3  RNAME  Reference sequence NAME
4  POS    1-based leftmost POSition/coordinate of clipped sequence
5  MAPQ   MAPping Quality (Phred-scaled)
6  CIGAR  extended CIGAR string
7  MRNM   Mate Reference sequence NaMe ('=' if same as RNAME)
8  MPOS   1-based Mate POSition
9  ISIZE  Inferred insert SIZE
10 SEQ    query SEQuence on the same strand as the reference
11 QUAL   query QUALity (ASCII-33 gives the Phred base quality)
12 OPT    variable OPTional fields in the format TAG:VTYPE:VALU
```

Sequence alignment/map format (SAM) and BAM

- SAM is a common format having sequence reads and their alignment to a reference genome.

- BAM is the binary form of a SAM file.

- Aligned BAM files are available at repositories (Sequence Read Archive at NCBI, ENA at Ensembl)

- SAMTools is a software package commonly used to analyze SAM/BAM files.

- Visit http://samtools.sourceforge.net/

Anatomy of a Sequence Alignment/Map (SAM) file

(1) The query name of the read is given (M01121...)

(2) The flag value is 163 (this equals 1+2+32+128)

(3) The reference sequence name, chrM, refers to the mitochondrial genome

(4) Position 480 is the left-most coordinate position of this read

(5) The Phred-scaled mapping quality is 60 (an error rate of 1 in 10^6)

(6) The CIGAR string (148M2S) shows 148 matches and 2 soft-clipped (unaligned) bases

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571     163     chrM
480      60       148M2S   =        524      195      AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG  BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFH00E@EGFGGEEE1FFEEEHBGEFFFGGGG@</0
1BG212222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-
RG:Z:Sample7    XC:i:148        XT:A:U  NM:i:3  SM:i:37
AM:i:37 X0:i:1  X1:i:0  XM:i:3  XO:i:0  XG:i:0  MD:Z:19C109C0A17
```

(7) An = sign shows that the mate reference matches the reference name

(8) The 1-based left position is 524

(9) The insert size is 195 bases

(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

(11) Each base is assigned a quality score (from BBBBB ending FHC.-)

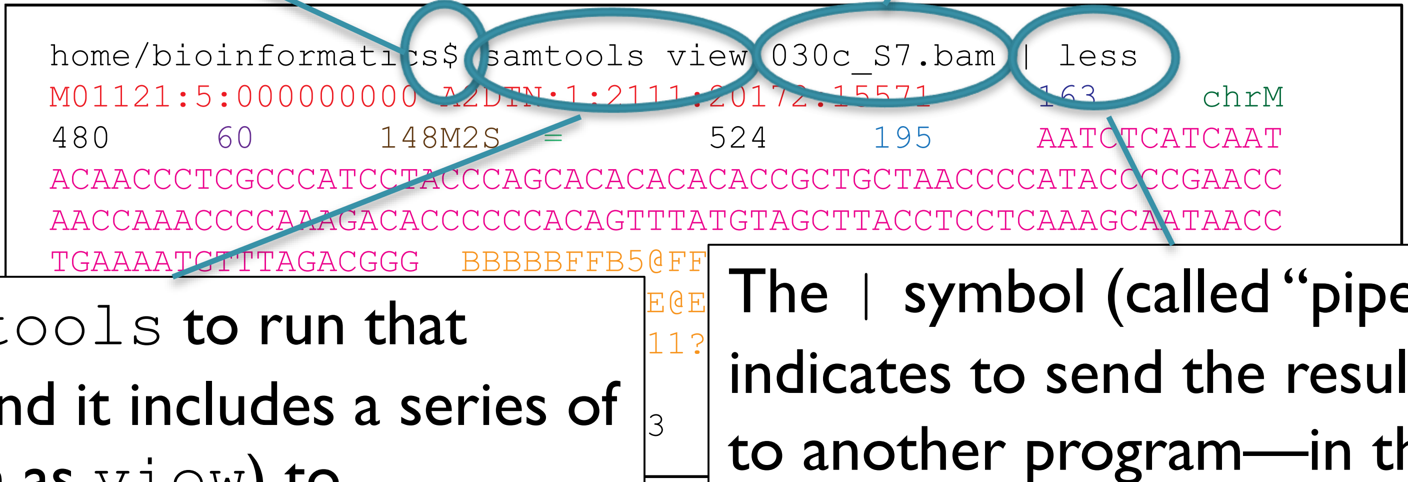
(12) This read has additional, optional fields that accompany the MiSeq analysis

Anatomy of a Sequence Alignment/Map (SAM) file

(1) The query name of the read is given (`M01121…`)

The `$` symbol indicates a command prompt in Unix

In this example we'll look at a file called `030c_s7.bam`. It is a BAM file (the binary of a SAM). Most software manipulates BAM files rather than SAM.



```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571          163        chrM
480       60        148M2S    =            524        195        AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG   BBBBBFFB5@FF
                                                           E@E
                                                           11?

                                                           3
```

Type `samtools` to run that program, and it includes a series of tools (such as `view`) to accomplish particular tasks—here, to view the contents of a file

based

s524

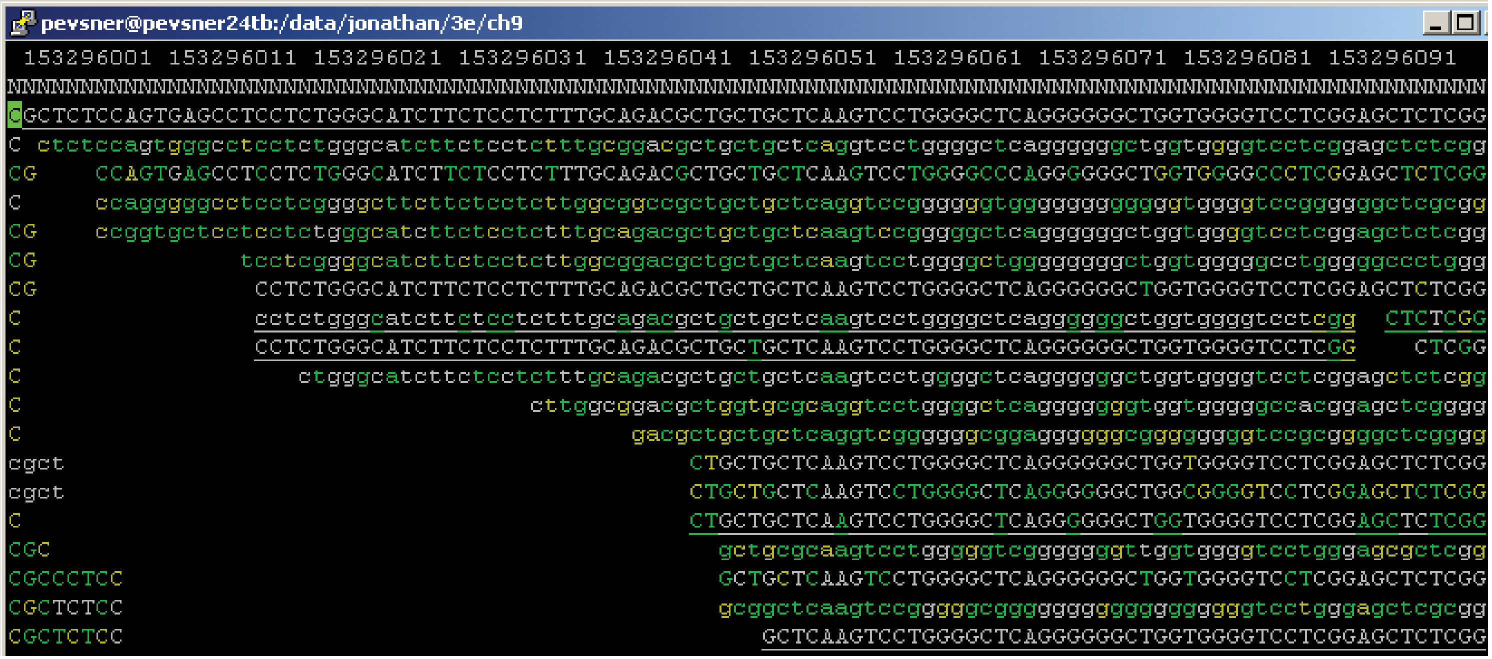
The `|` symbol (called "pipe") indicates to send the results to another program—in this case to the utility called `less` that displays one page at a time on your terminal.

(10) The sequence begins `AATCT` and ends `ACGGG` (its length is 150 bases)

(11) Each base is assigned a quality score (from BBBBB ending `FHC.-`)

(12) This read has additional, optional fields that accompany the MiSeq analysis

SAMTools tview visualization of reads from a BAM file

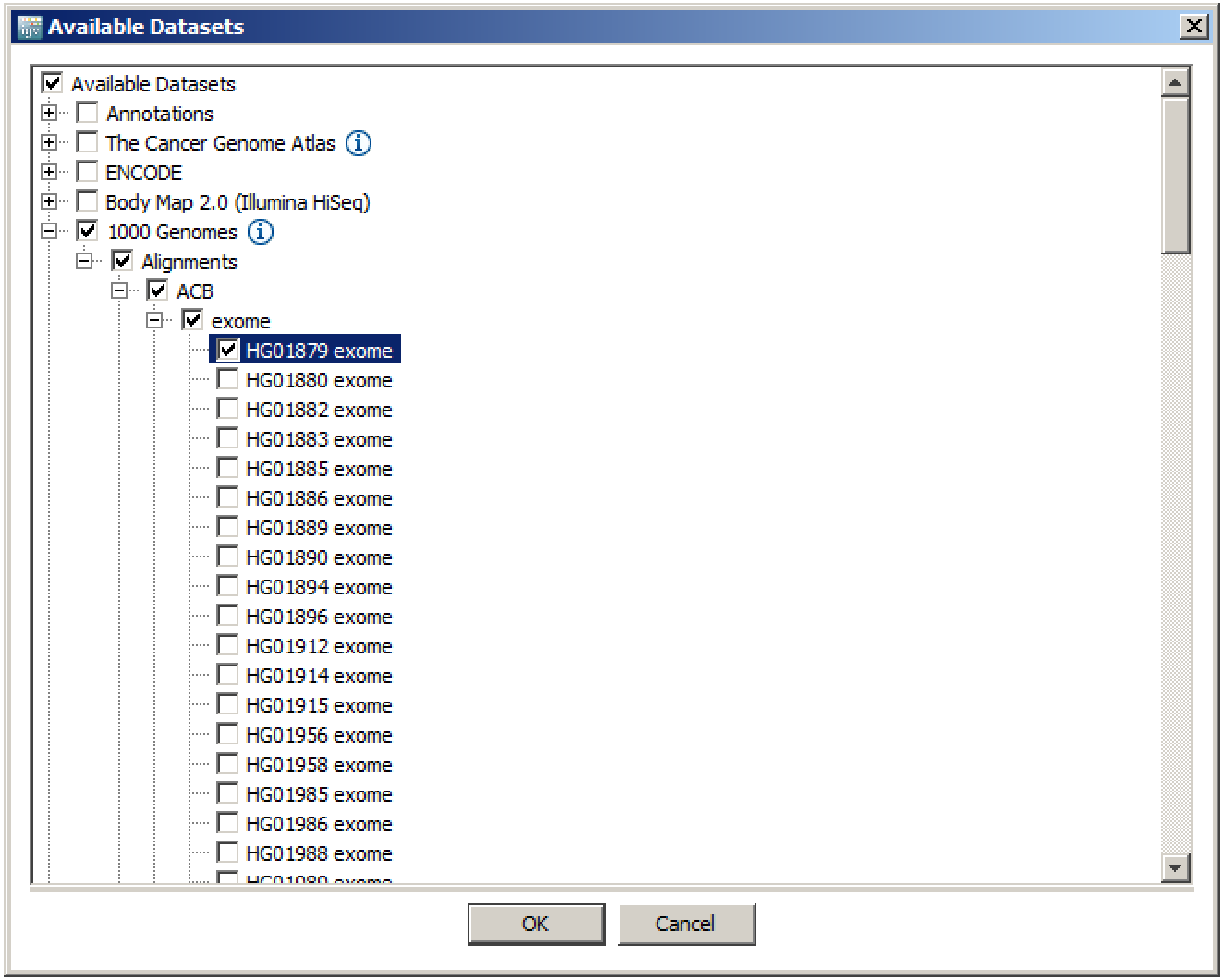


pevsner@pevsner24tb:/data/jonathan/3e/ch9

```
 153296001 153296011 153296021 153296031 153296041 153296051 153296061 153296071 153296081 153296091
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CGCTCTCCAGTGAGCCTCCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
C ctctccagtgggcctcctctgggcatcttctcctctttgcggacgctgctgctcaggtcctggggctcaggggggctggtggggtcctcggagctctcgg
CG    CCAGTGAGCCTCCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCCCAGGGGGGCTGGTGGGGCCCTCGGAGCTCTCGG
C     ccagggggcctcctcggggcttcttctcctcttggcggccgctgctgctcaggtccgggggggtggggggggggggtggggtccgggggctcgcgg
CG    ccggtgctcctcctctgggcatcttctcctctttgcagacgctgctgctcaagtccgggggctcaggggggctggtggggtcctcggagctctcgg
CG              tcctcggggcatcttctcctcttggcggacgctgctgctcaagtcctggggctggggggggctggtgggggcctgggggccctggg
CG               CCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
C                cctctgggcatcttctcctctttgcagacgctgctgctcaagtcctggggctcaggggggctggtggggtcctcgg  CTCTCGG
C                CCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGG    CTCGG
C                   ctgggcatcttctcctctttgcagacgctgctgctcaagtcctggggctcaggggggctggtggggtcctcggagctctcgg
C                                   cttggcggacgctggtgcgcaggtcctggggctcaggggggtggtgggggccacggagctcgggg
C                                          gacgctgctgctcaggtcggggggcggaggggggcggggggggtccgcggggctcgggg
cgct                                           CTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
cgct                                           CTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGCGGGGTCCTCGGAGCTCTCGG
C                                              CTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
CGC                                              gctgcgcaagtcctgggggtcgggggggttggtggggtcctgggagcgctcgg
CGCCCTCC                                         GCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
CGCTCTCC                                         gcggctcaagtccgggggcgggggggggggggggggtcctgggagctcgcgg
CGCTCTCC                                            GCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
```

There are many tools to view SAM/BAM files. A popular software package (SAMTools, used in Linux) includes `tview` visualization of reads from a BAM file

IGV visualization of reads from a BAM file



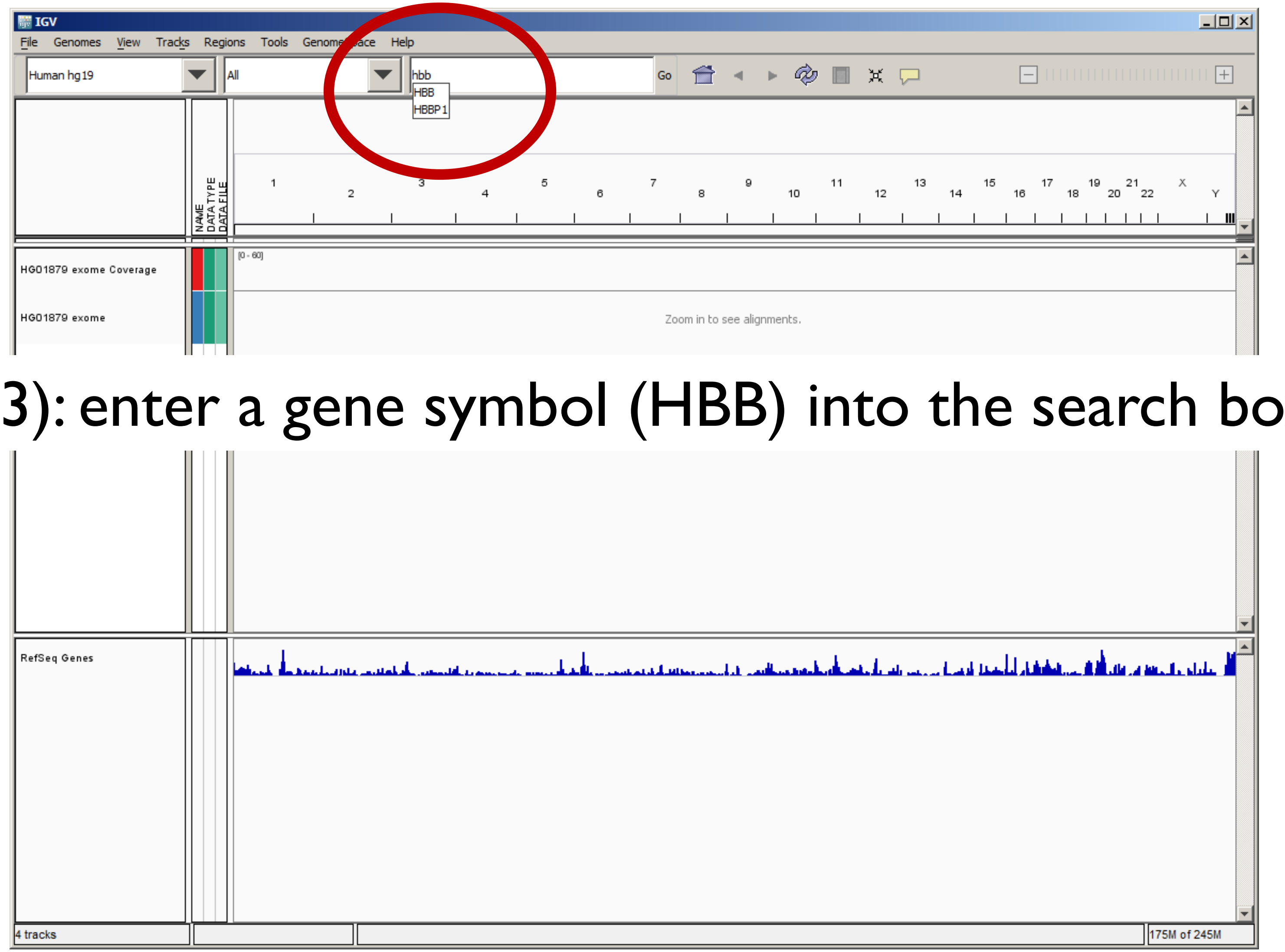
Available Datasets

- ☑ Available Datasets
  - ☐ Annotations
  - ☐ The Cancer Genome Atlas
  - ☐ ENCODE
  - ☐ Body Map 2.0 (Illumina HiSeq)
  - ☑ 1000 Genomes
    - ☑ Alignments
      - ☑ ACB
        - ☑ exome
          - ☑ HG01879 exome
          - ☐ HG01880 exome
          - ☐ HG01882 exome
          - ☐ HG01883 exome
          - ☐ HG01885 exome
          - ☐ HG01886 exome
          - ☐ HG01889 exome
          - ☐ HG01890 exome
          - ☐ HG01894 exome
          - ☐ HG01896 exome
          - ☐ HG01912 exome
          - ☐ HG01914 exome
          - ☐ HG01915 exome
          - ☐ HG01956 exome
          - ☐ HG01958 exome
          - ☐ HG01985 exome
          - ☐ HG01986 exome
          - ☐ HG01988 exome

Step (1): open IGV (Mac or PC) from its website

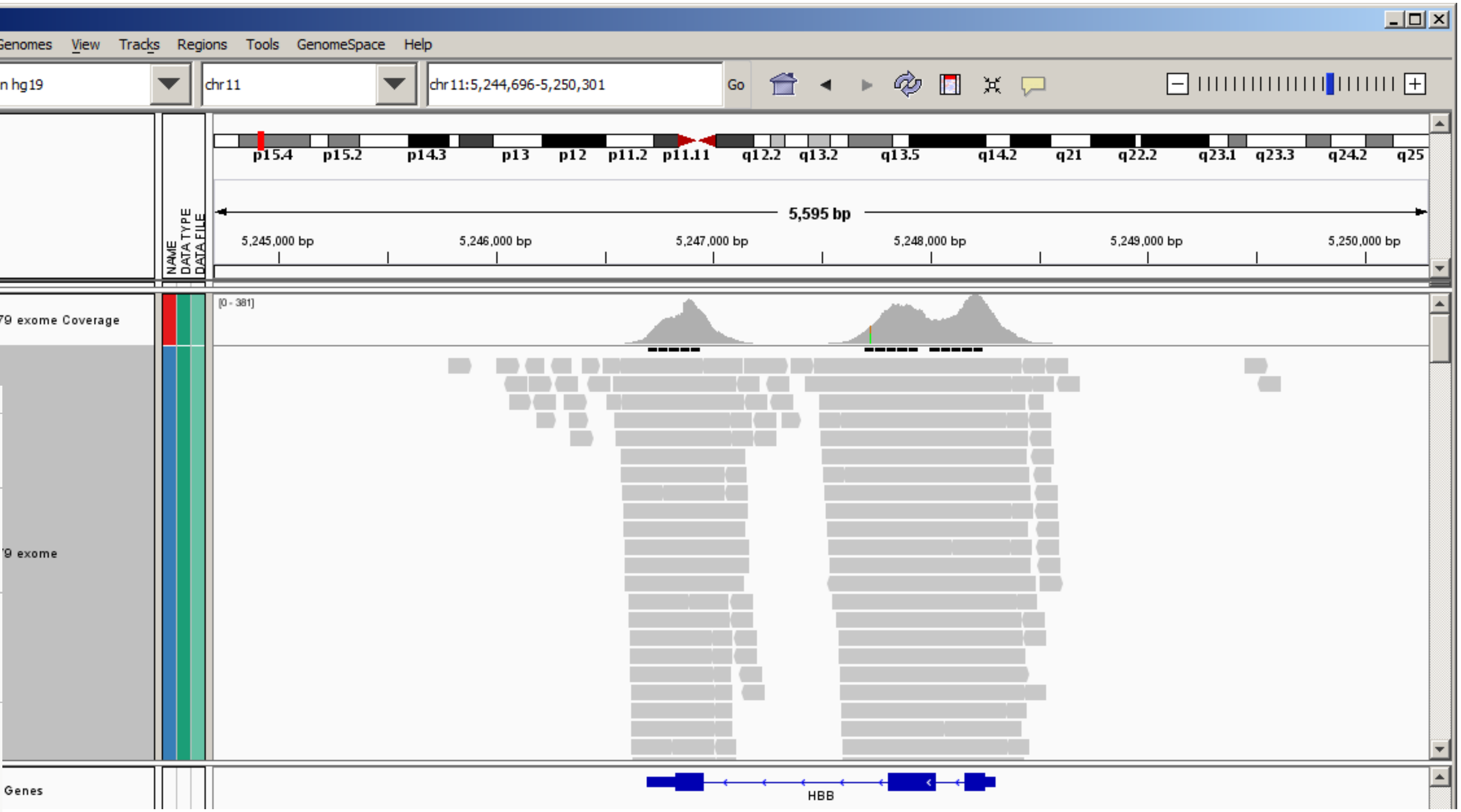
Step (2): File > Load from server > load one exome

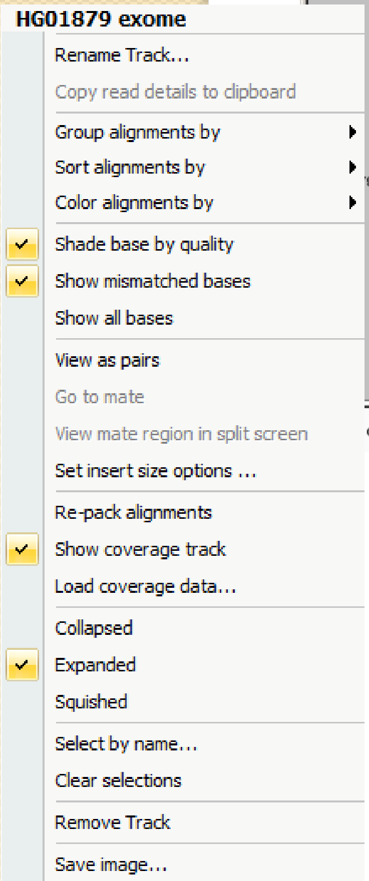
IGV visualization of reads from a BAM file


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
All
hbb
HBB
HBBP1
Go
NAME
DATA TYPE
DATA FILE
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
HG01879 exome Coverage
HG01879 exome
[0 - 60]
Zoom in to see alignments.
RefSeq Genes
4 tracks
175M of 245M

Step (3): enter a gene symbol (HBB) into the search box.

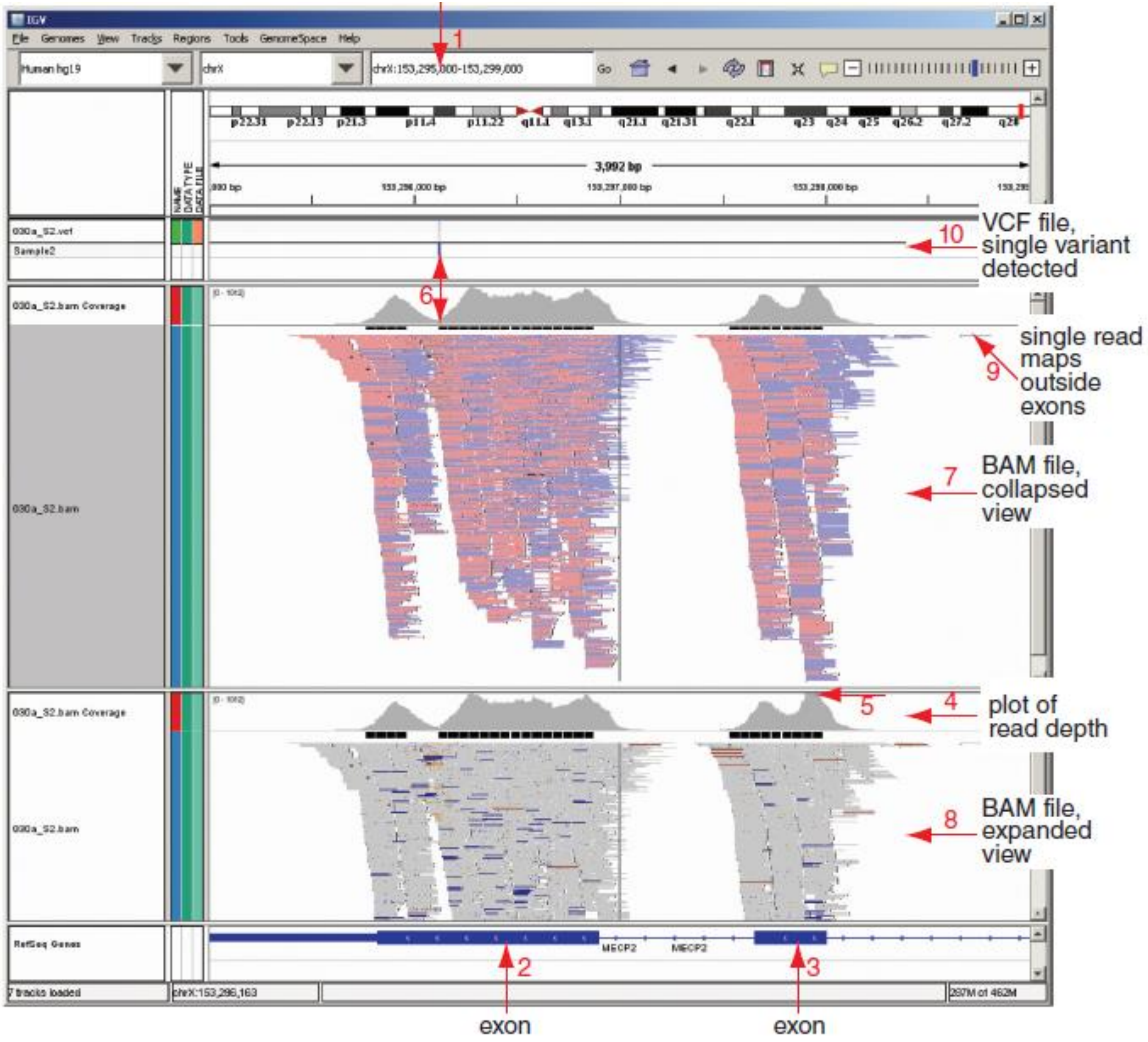
IGV visualization of reads from a BAM file


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr11
chr11:5,244,696-5,250,301
Go
p15.4 p15.2 p14.3 p13 p12 p11.2 p11.11 q12.2 q13.2 q13.5 q14.2 q21 q22.2 q23.1 q23.3 q24.2 q25
5,595 bp
5,245,000 bp
5,246,000 bp
5,247,000 bp
5,248,000 bp
5,249,000 bp
5,250,000 bp
NAME
DATA TYPE
DATA FILE
HG01879 exome Coverage
[0 - 381]
9 exome
Genes
HBB



Step (4): explore this gene. Zoom in. Click the left sidebar to change the display to squished. Color the alignments. Find variants.

Integrative Genomics Viewer (IGV): display of a BAM file (at two resolutions) and a VCF



Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview Topic 6: Variant calling: SNVs

Topic 1: Design Topic 7: Variant calling:
SVs

Topic 2: FASTQ Topic 8: VCF

Topic 3: Assembly Topic 9: Visualizing NGS data

Topic 4: Alignment Topic 10: Significance

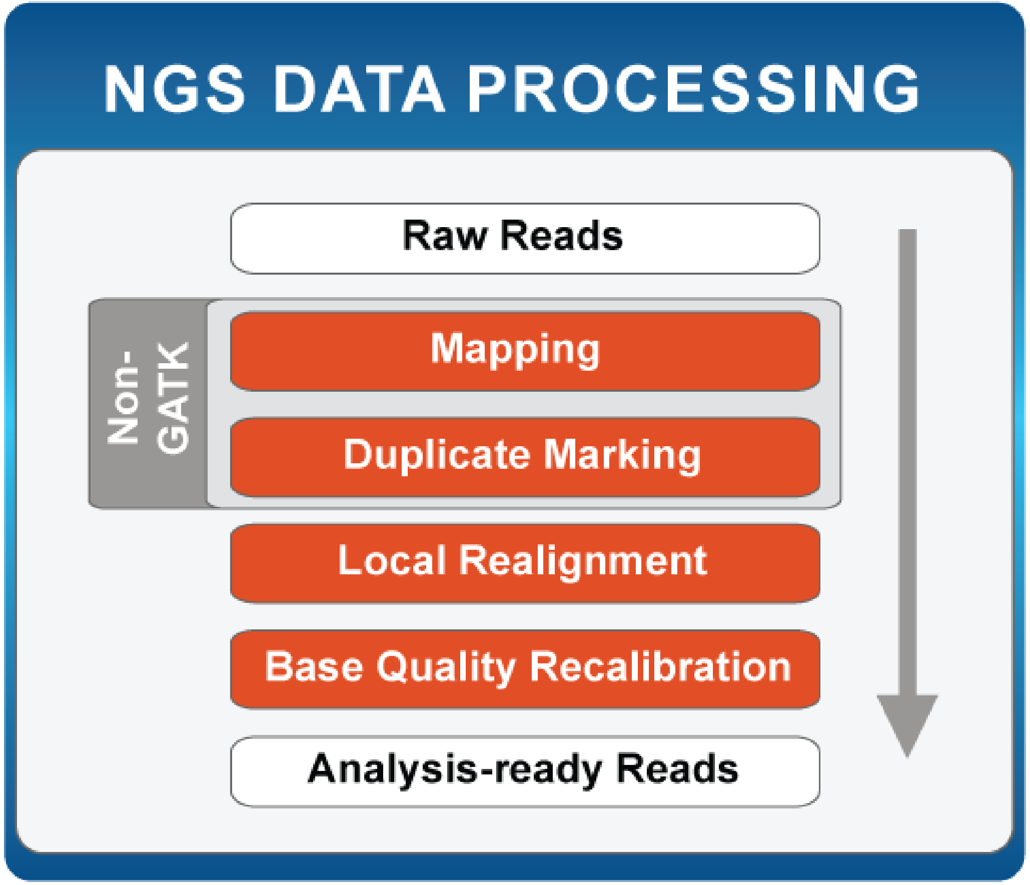
Topic 5: SAM/BAM

Specialized applications of NGS

Perspective

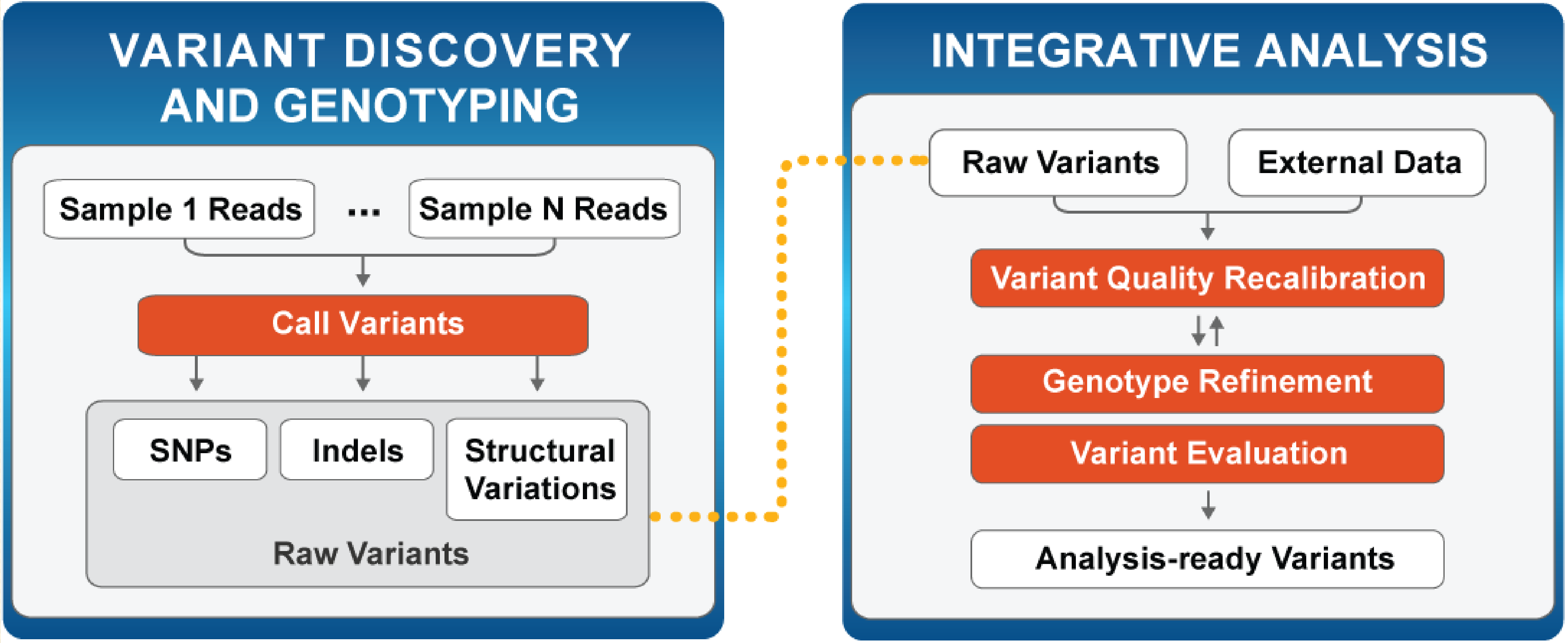
Genotyping with Genome Analysis Toolkit (GATK)

Popular suite of tools used for genotyping and variant discovery



http://www.broadinstitute.org/gatk/

Genotyping with Genome Analysis Toolkit (GATK)



http://www.broadinstitute.org/gatk/

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

- Sanger sequencing; NGS; Illumina; pyrosequencing; ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

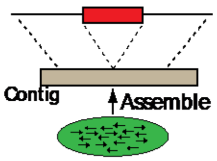
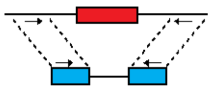
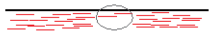
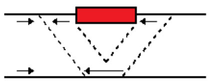
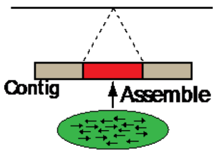
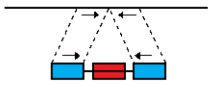
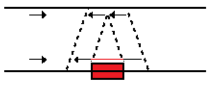
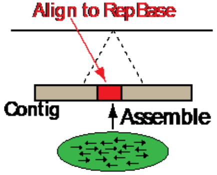
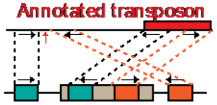
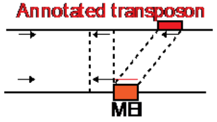
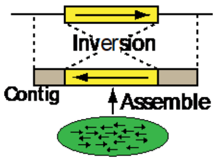
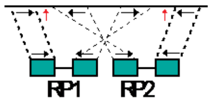
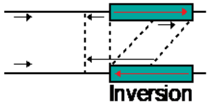
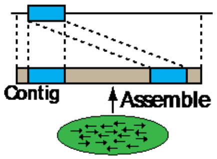
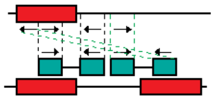
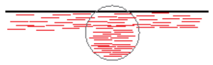
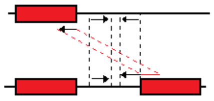
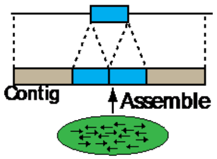
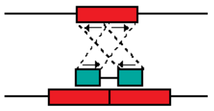
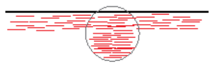
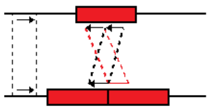
Analysis of NGS sequencing of genomic DNA

- Overview
- Topic 1: Design
- Topic 2: FASTQ
- Topic 3: Assembly
- Topic 4: Alignment
- Topic 5: SAM/BAM
- Topic 6: Variant calling: SNVs
- Topic 7: Variant calling: SVs
- Topic 8: VCF
- Topic 9: Visualizing NGS data
- Topic 10: Significance

Specialized applications of NGS

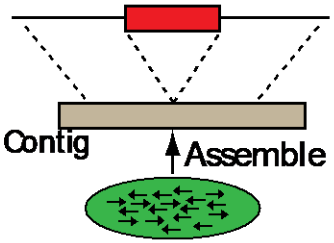
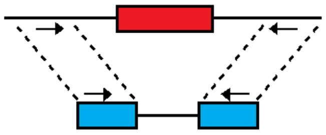
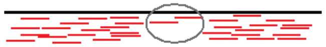
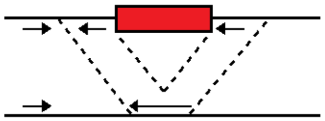
Perspective

Categories of structural variation (SV)

SV class	Assembly	Read pair	Read depth	Split end
Deletion				
Novel sequence insertion			Not applicable	
Mobile-element insertion			Not applicable	
Inversion			Not applicable	
Interspersed duplication				
Tandem duplication				

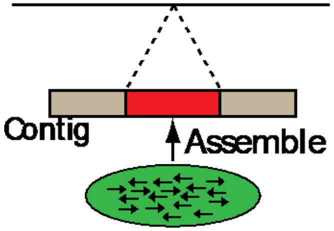
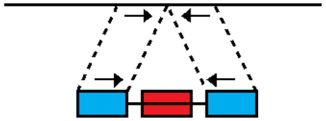
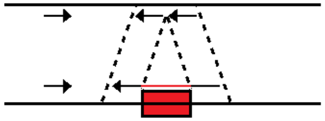
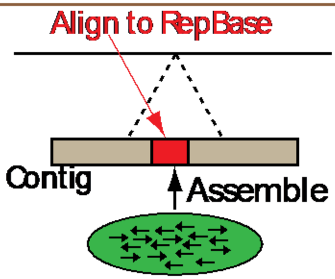
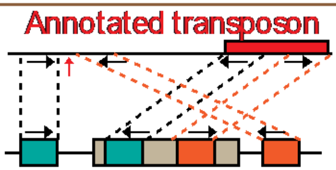
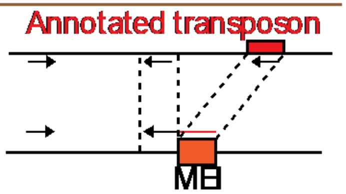
Source: PMID 21358748

Categories of structural variation (SV): deletions

SV class	Assembly	Read pair	Read depth	Split end
Deletion				

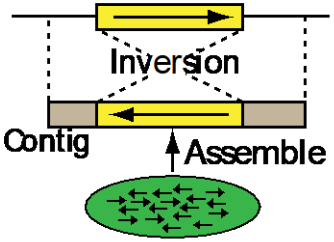
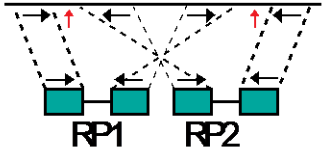
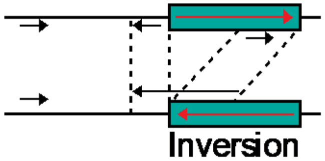
Source: PMID 21358748

Categories of structural variation (SV): insertions

SV class	Assembly	Read pair	Read depth	Split end
Novel sequence insertion			not applicable	
Mobile-element insertion			not applicable	

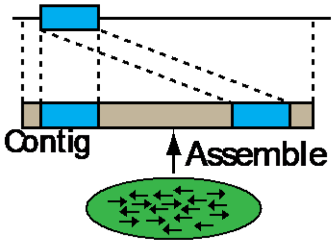
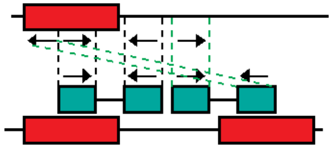
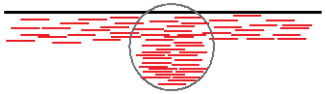
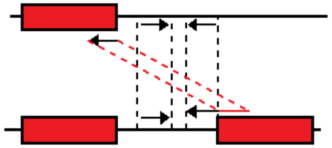
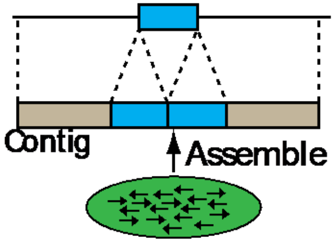
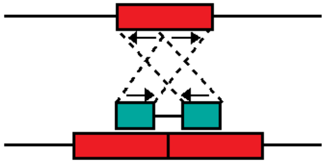
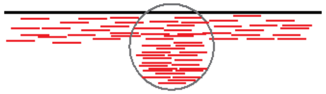
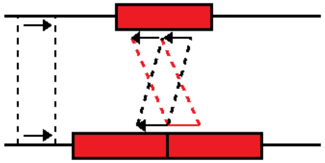
Source: PMID 21358748

Categories of structural variation (SV): inversions

SV class	Assembly	Read pair	Read depth	Split end
Inversion			not applicable	

Source: PMID 21358748

Categories of structural variation (SV): duplications

SV class	Assembly	Read pair	Read depth	Split end
Interspersed duplication				
Tandem duplication				

Source: PMID 21358748

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

- Sanger sequencing; NGS; Illumina; pyrosequencing; ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

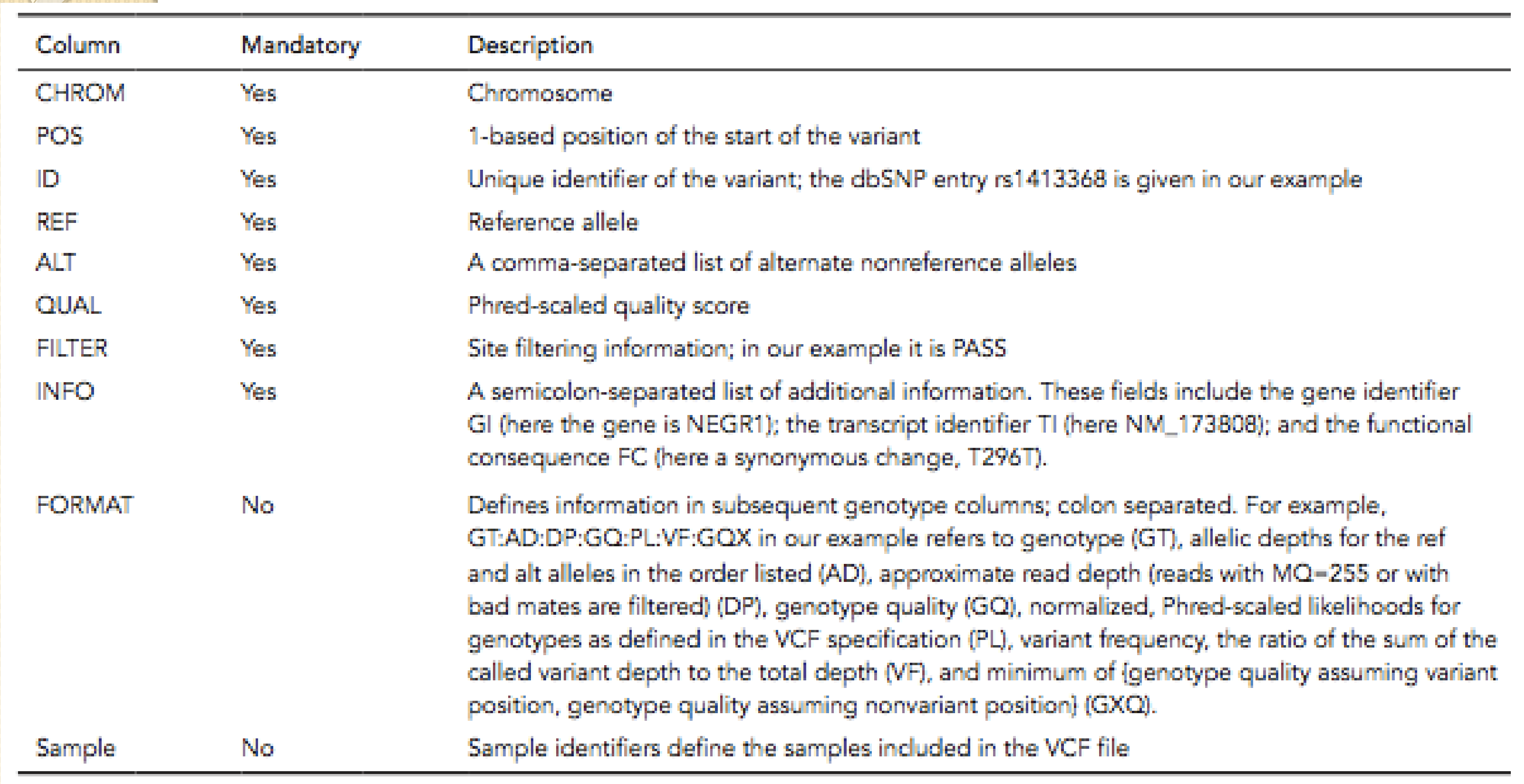
- Overview
- Topic 1: Design
- Topic 2: FASTQ
- Topic 3: Assembly
- Topic 4: Alignment
- Topic 5: SAM/BAM
- Topic 6: Variant calling: SNVs
- Topic 7: Variant calling: SVs
- Topic 8: VCF
- Topic 9: Visualizing NGS data
- Topic 10: Significance

Specialized applications of NGS

Perspective

Variant Call Format (VCF) file summarizes variation

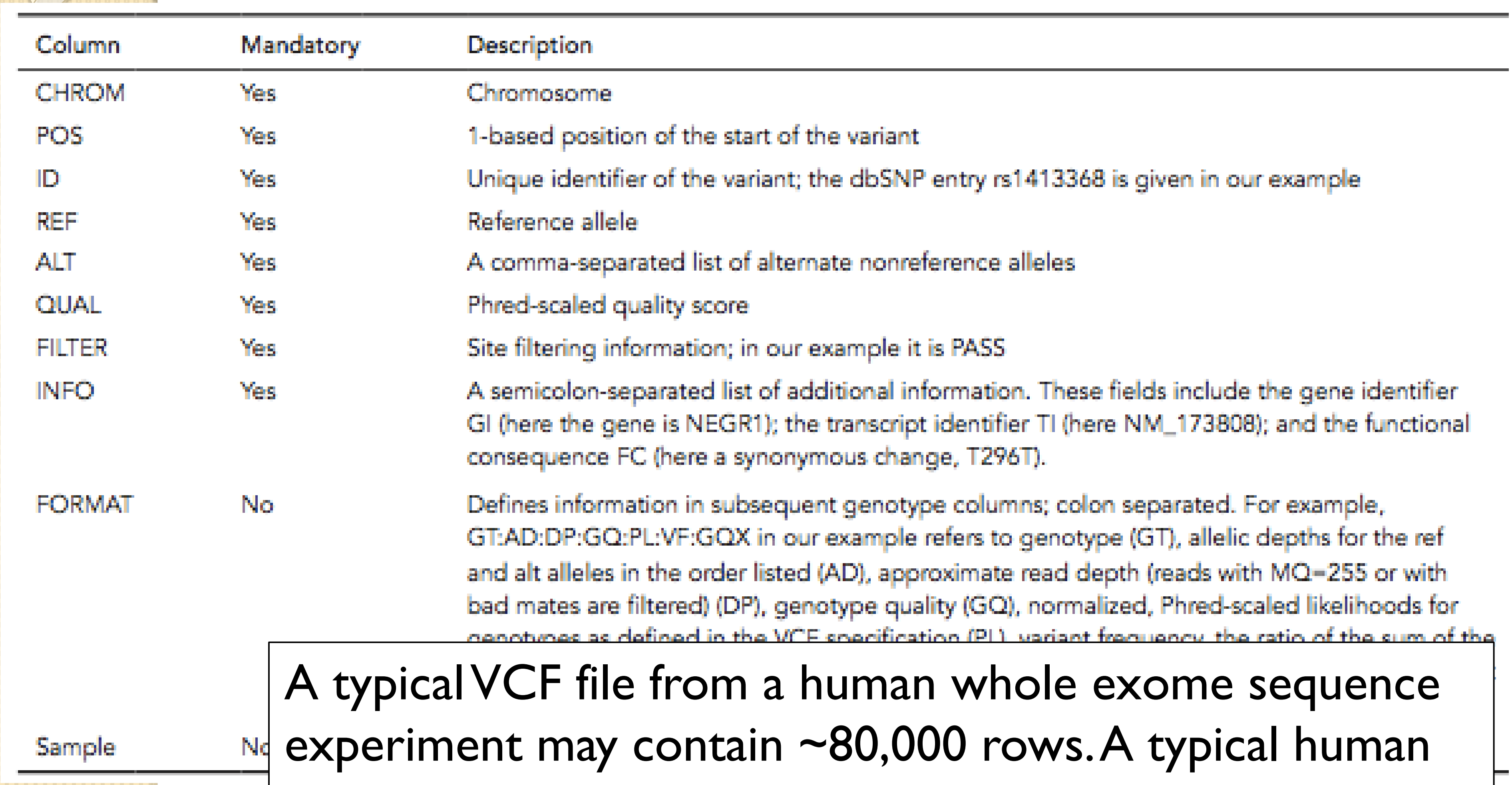
A VCF file includes the following information:



Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GQX in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification (PL), variant frequency, the ratio of the sum of the called variant depth to the total depth (VF), and minimum of {genotype quality assuming variant position, genotype quality assuming nonvariant position} (GXQ).
Sample	No	Sample identifiers define the samples included in the VCF file

Variant Call Format (VCF) file summarizes variation

A VCF file includes the following information:



Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GQX in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification (PL), variant frequency, the ratio of the sum of the
Sample	No	

A typical VCF file from a human whole exome sequence experiment may contain ~80,000 rows. A typical human whole genome sequence experiment produces a VCF with ~4 million rows.

Variant Call Format (VCF) file summarizes variation

- VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths...
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth...
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=VF,Number=1,Type=Float,Description="Variant Frequency...
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
##INFO=<ID=GI,Number=.,Type=String,Description="Gene ID">
##INFO=<ID=FC,Number=.,Type=String,Description="Functional Consequence">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth...
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##FILTER=<ID=R8,Description="IndelRepeatLength is greater than 8">
##FILTER=<ID=SB,Description="Strand bias (SB) is greater than than -10">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=...
##contig=<ID=chr1,length=249250621>
##contig=<ID=chr10,length=135534747>
```

Variant Call Format (VCF) file summarizes variation

VCF field definition line and first row of body

```
#CHROM  POS      ID       REF      ALT      QUAL     FILTER   INFO     FORMAT   Sample7
chr1    72058552          rs1413368         G        A        7398.69 PASS
AC=2;AF=1.00;AN=2;DP=250;DS;Dels=0.00;FS=0.000;HRun=1;HaplotypeScore=3.8533;
MQ=50.89;MQ0=0;QD=29.59;SB=-4337.33;TI=NM_173808;GI=NEGR1;FC=Synonymous_
T296T     GT:AD:DP:GQ:PL:VF:GQX   1/1:0,250:250:99:7399,536,0:1.000:99
```

Fields include chromosome (CHROM), position, identifier (e.g. rsID), reference allele, alternate allele, quality score, and extensive data (e.g. haplotypes, read depth, quality scores, functional consequences, accession numbers)

Variant Call Format (VCF) file summarizes variation

SNP

```
Alignment     VCF representation
1234          POS   REF   ALT
ACGT          2     C     T
ATGT
 ↑
```

Insertion

```
Alignment     VCF representation
12345         POS   REF   ALT
AC-GT         2     C     CT
ACTGT
  ↑
```

Deletion

```
Alignment     VCF representation
1234          POS   REF   ALT
ACGT          1     ACG   A
A--T
 ↑↑
```

Replacement

```
Alignment     VCF representation
1234          POS   REF   ALT
ACGT          1     ACG   AT
A-TT
 ↑↑
```

Large structural variant

```
Alignment                                                  VCF representation
   100    110          120    290          300             POS   REF   ALT     INFO
    .      .            .      .            .
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC 100   T     <DEL>  SVTYPE=DEL;END=299
ATGT----------------[...]----------GTAC
```

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

- Sanger sequencing; NGS; Illumina; pyrosequencing; ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

- Overview
- Topic 1: Design
- Topic 2: FASTQ
- Topic 3: Assembly
- Topic 4: Alignment
- Topic 5: SAM/BAM
- Topic 6: Variant calling: SNVs
- Topic 7: Variant calling: SVs
- Topic 8: VCF
- Topic 9: Visualizing NGS data
- Topic 10: Significance

Specialized applications of NGS

Perspective

Visualizing and tabulating next-generation sequence data

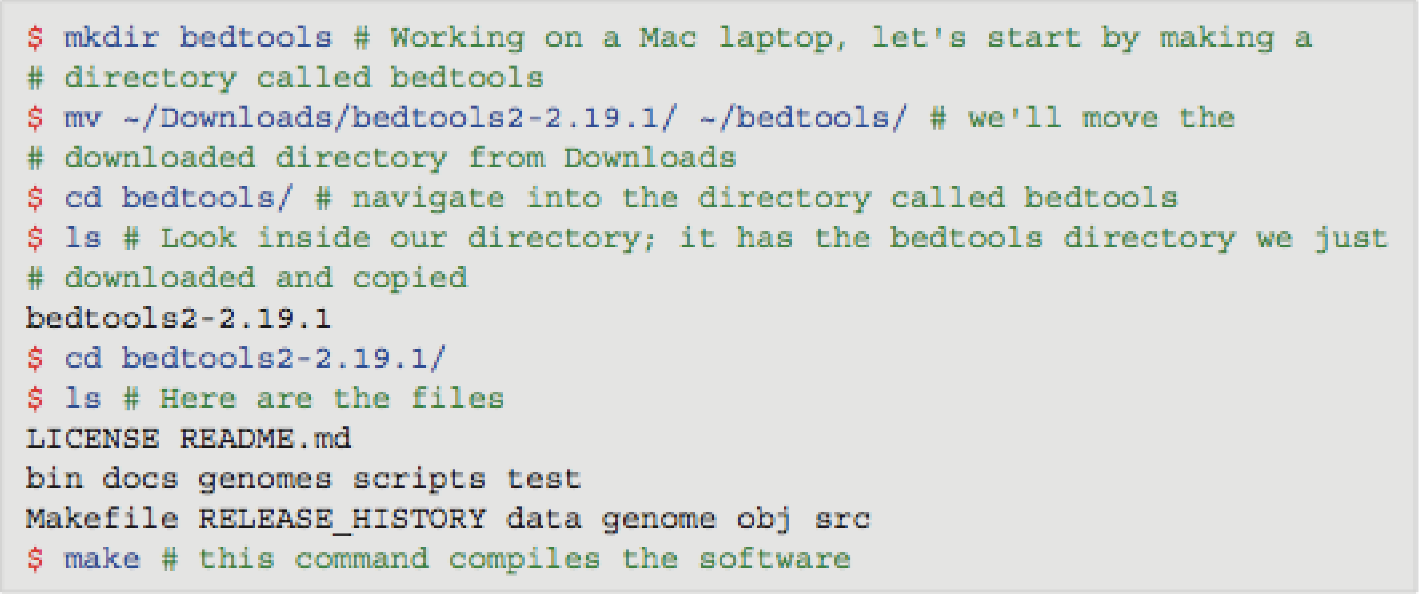
There are many ways to visualize BAM files.

- Try Genome Workbench from NCBI
- Upload your BAM file to a server and point to it using the UCSC Genome Browser
- Use Integrative Genomics Viewer (IGV)
- Use `samtools tview`

We will next explore BEDtools, a set of programs used to analyze BAM, GTF, BED, VCF, and other file types.

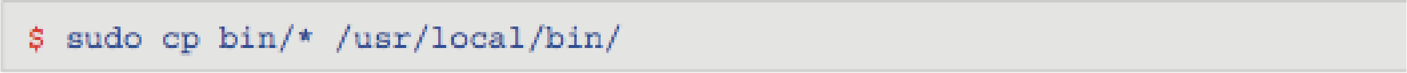
BEDtools to explore genomics data

Download and install bedtools:



```
$ mkdir bedtools # Working on a Mac laptop, let's start by making a
# directory called bedtools
$ mv ~/Downloads/bedtools2-2.19.1/ ~/bedtools/ # we'll move the
# downloaded directory from Downloads
$ cd bedtools/ # navigate into the directory called bedtools
$ ls # Look inside our directory; it has the bedtools directory we just
# downloaded and copied
bedtools2-2.19.1
$ cd bedtools2-2.19.1/
$ ls # Here are the files
LICENSE README.md
bin docs genomes scripts test
Makefile RELEASE_HISTORY data genome obj src
$ make # this command compiles the software
```

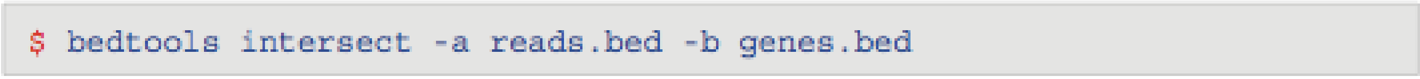
Place the executables in your path:



```
$ sudo cp bin/* /usr/local/bin/
```

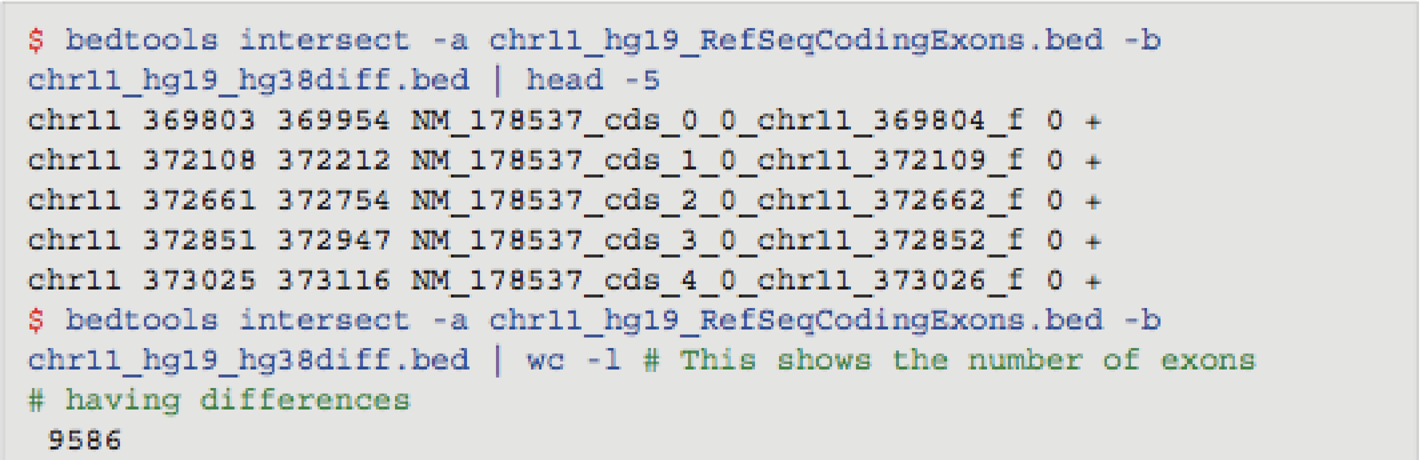
BEDtools example 1: What RefSeq coding exons differ between GRCh37 and GRCh38?

Use BEDtools intersect. General format of a query:



```
$ bedtools intersect -a reads.bed -b genes.bed
```

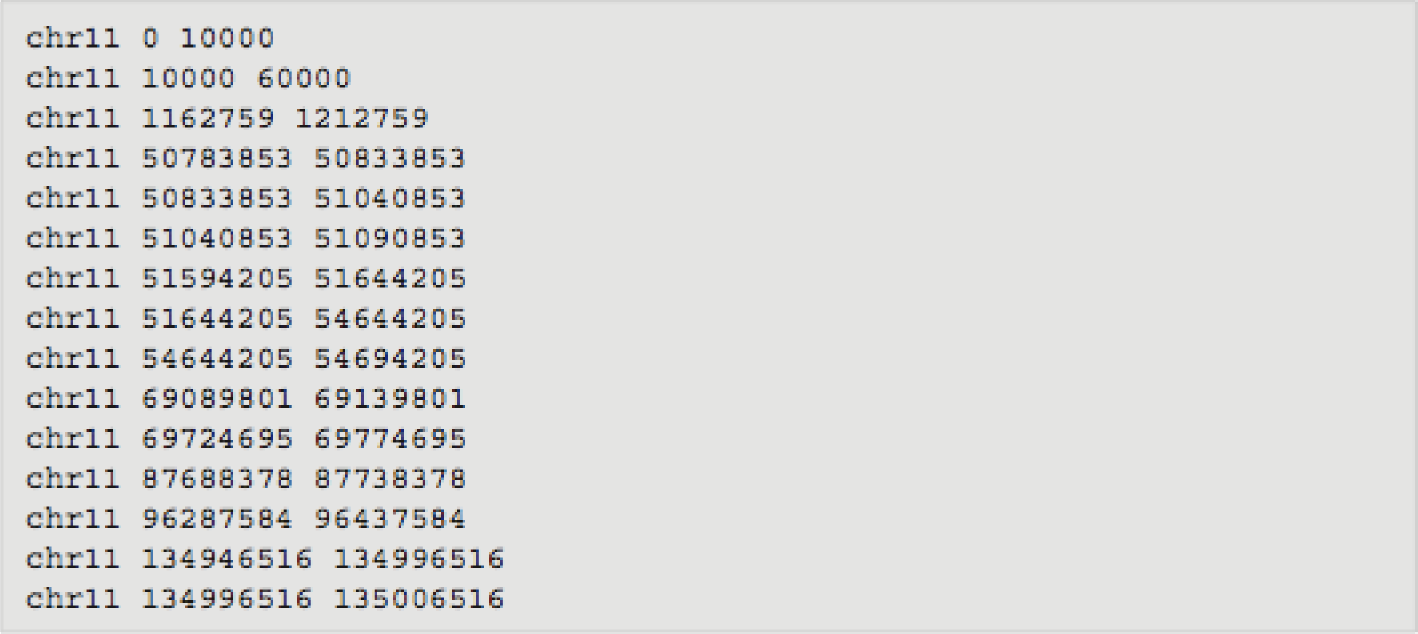
Our query:



```
$ bedtools intersect -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_hg38diff.bed | head -5
chr11 369803 369954 NM_178537_cds_0_0_chr11_369804_f 0 +
chr11 372108 372212 NM_178537_cds_1_0_chr11_372109_f 0 +
chr11 372661 372754 NM_178537_cds_2_0_chr11_372662_f 0 +
chr11 372851 372947 NM_178537_cds_3_0_chr11_372852_f 0 +
chr11 373025 373116 NM_178537_cds_4_0_chr11_373026_f 0 +
$ bedtools intersect -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_hg38diff.bed | wc -l # This shows the number of exons
# having differences
 9586
```

BEDtools example 2: What is the closest chromosomal gap to every RefSeq exon?

Here is a BED file of all gaps on chromosome 11:

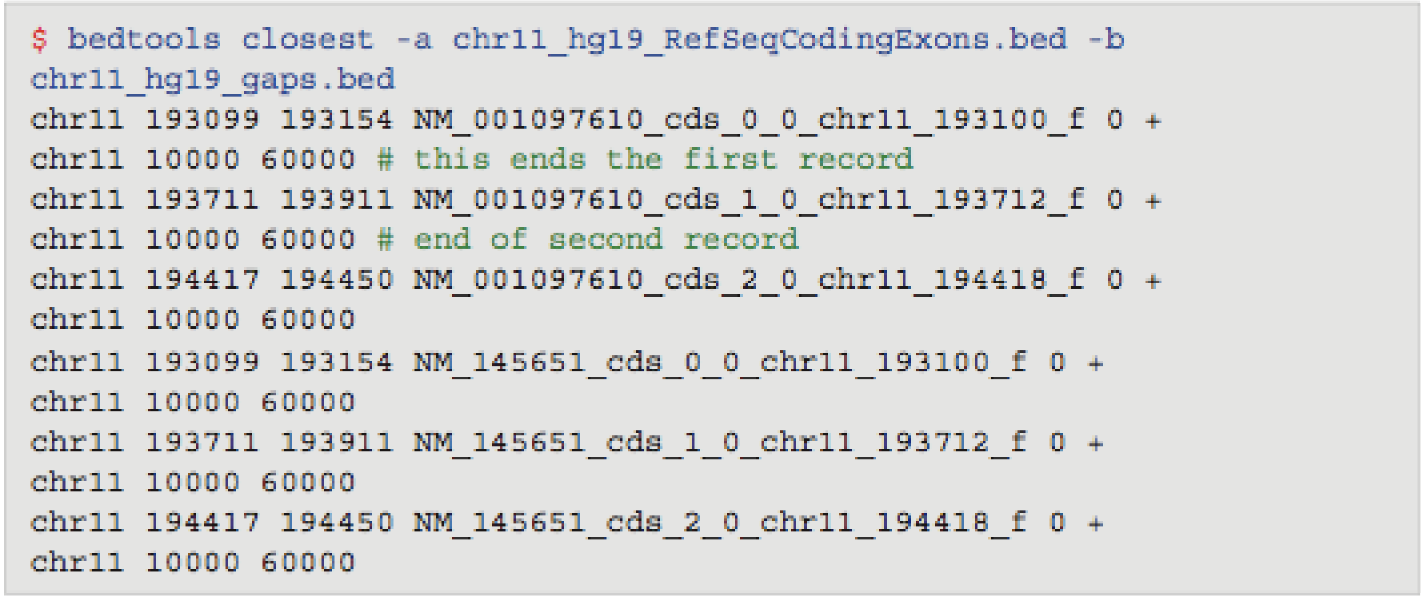


```
chr11 0 10000
chr11 10000 60000
chr11 1162759 1212759
chr11 50783853 50833853
chr11 50833853 51040853
chr11 51040853 51090853
chr11 51594205 51644205
chr11 51644205 54644205
chr11 54644205 54694205
chr11 69089801 69139801
chr11 69724695 69774695
chr11 87688378 87738378
chr11 96287584 96437584
chr11 134946516 134996516
chr11 134996516 135006516
```

Each chromosome has gaps at the telomeres, at the centromere, and at other locations that have been too challenging to sequence.

BEDtools example 2: What is the closest chromosomal gap to every RefSeq exon?

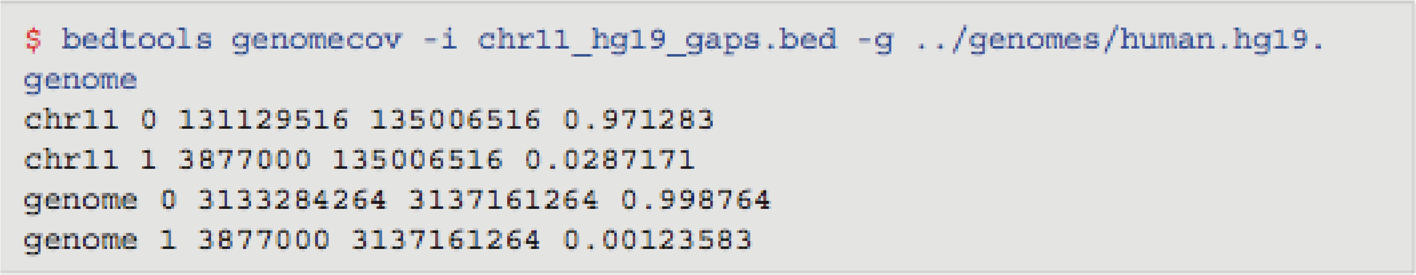
We use the `bedtools closest` utility. Here are the results:



```
$ bedtools closest -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_gaps.bed
chr11 193099 193154 NM_001097610_cds_0_0_chr11_193100_f 0 +
chr11 10000 60000 # this ends the first record
chr11 193711 193911 NM_001097610_cds_1_0_chr11_193712_f 0 +
chr11 10000 60000 # end of second record
chr11 194417 194450 NM_001097610_cds_2_0_chr11_194418_f 0 +
chr11 10000 60000
chr11 193099 193154 NM_145651_cds_0_0_chr11_193100_f 0 +
chr11 10000 60000
chr11 193711 193911 NM_145651_cds_1_0_chr11_193712_f 0 +
chr11 10000 60000
chr11 194417 194450 NM_145651_cds_2_0_chr11_194418_f 0 +
chr11 10000 60000
```

BEDtools example 3: How much of a chromosome (or a genome) is spanned by gaps?

We use the `genomecov` (genome coverage) utility, and use the `-g` argument to specify a genome. Here are the results:



```
$ bedtools genomecov -i chr11_hg19_gaps.bed -g ../genomes/human.hg19.
genome
chr11 0 131129516 135006516 0.971283
chr11 1 3877000 135006516 0.0287171
genome 0 3133284264 3137161264 0.998764
genome 1 3877000 3137161264 0.00123583
```

2.87% of the chromosome (0.0287), and 0.1% of the genome is spanned by gaps.

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

- Sanger sequencing; NGS; Illumina; pyrosequencing; ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

- Overview
- Topic 1: Design
- Topic 2: FASTQ
- Topic 3: Assembly
- Topic 4: Alignment
- Topic 5: SAM/BAM
- Topic 6: Variant calling: SNVs
- Topic 7: Variant calling: SVs
- Topic 8: VCF
- Topic 9: Visualizing NGS data
- Topic 10: Significance

Specialized applications of NGS

Perspective

Prioritizing variants and assessing functional significance

This section is organized in two parts.

(1) We will look at software that is used to assess which variants are functionally significant. Over 50 programs have been introduced. We will mention three: SIFT, PolyPhen, and VAAST.

(2) NCBI offers databases, browsers and software tools to understand functionally important variants. We will introduce four NCBI resources.

Neutral versus deleterious variation

For each genome, we can expect to identify ~4 million variants that are exonic, intronic, or intergenic. We first focus on exonic variants. Of these, there are ~11,000 synonymous SNPs (not changing the amino acid specified by the codon; likely to be benign) and ~11,000 nonsynonymous SNPs.

We also consider indels (some of which introduce stop codons), homozygous deletions, splice site mutations, or other changes that may disrupt gene function.

Approaches to distinguish neutral from deleterious nonsynonymous variants

Most DNA is under neutral selection (not under positive or negative selection). Some variants are deleterious. How can we classify 11,000 nonsynonymous SNPs in a genome?

--Conservation: determine conservation of an amino acid across species
--Structure: determine (or predict) effect of a variant on protein structure
--True positives: train algorithms on a database of known disease-associated mutations (OMIM)
--True negatives: train algorithms of a set of variants in 'apparently normal' individuals (1000 Genomes)

Software to distinguish neutral from deleterious nonsynonymous variants

PolyPhen2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

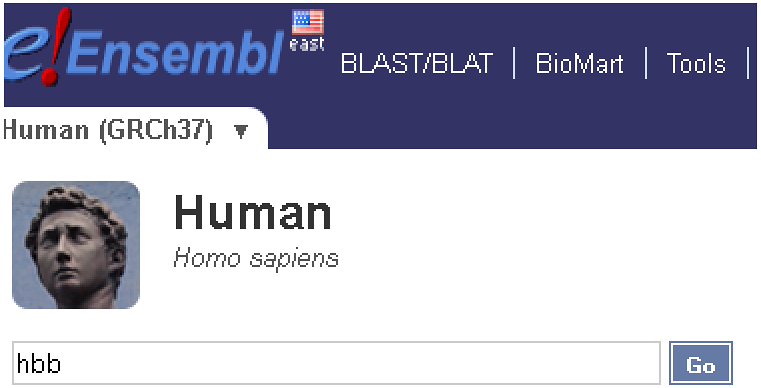
http://genetics.bwh.harvard.edu/pph2/

SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.

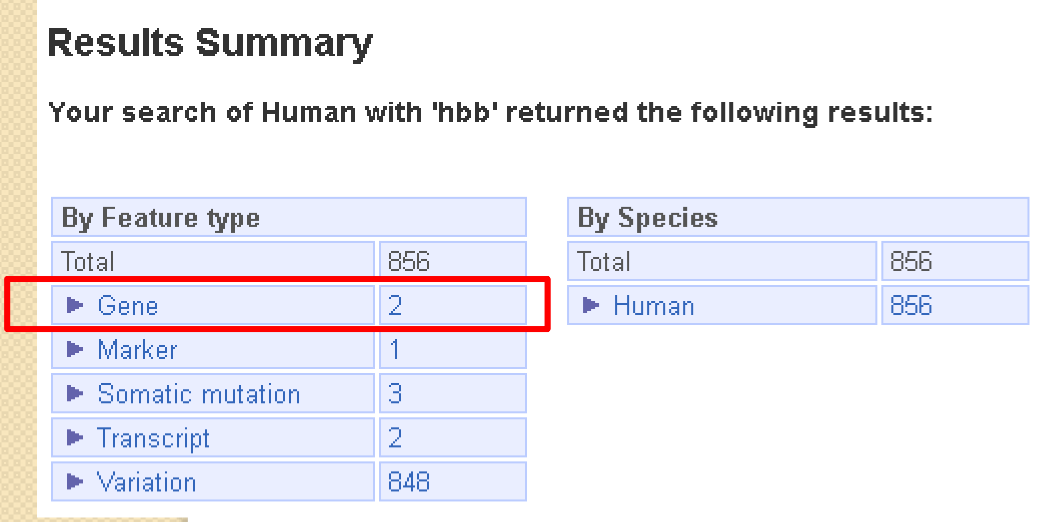
http://sift.jcvi.org/

Example: SIFT and Polyphen scores for HBB

- [1] Visit http://www.ensembl.org/human
- [2] Enter hbb in the search box
- [3] Follow the link to the gene



hbb

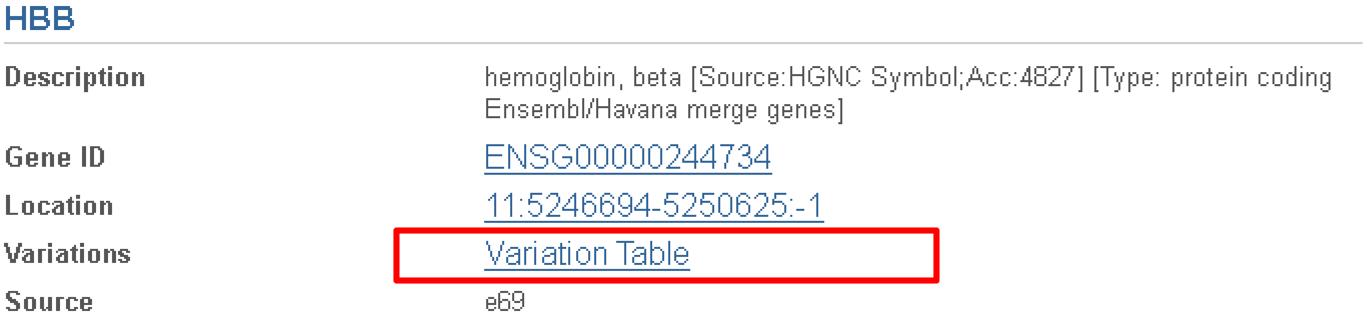


Results Summary

Your search of Human with 'hbb' returned the following results:

By Feature type	
Total	856
▶ Gene	2
▶ Marker	1
▶ Somatic mutation	3
▶ Transcript	2
▶ Variation	848

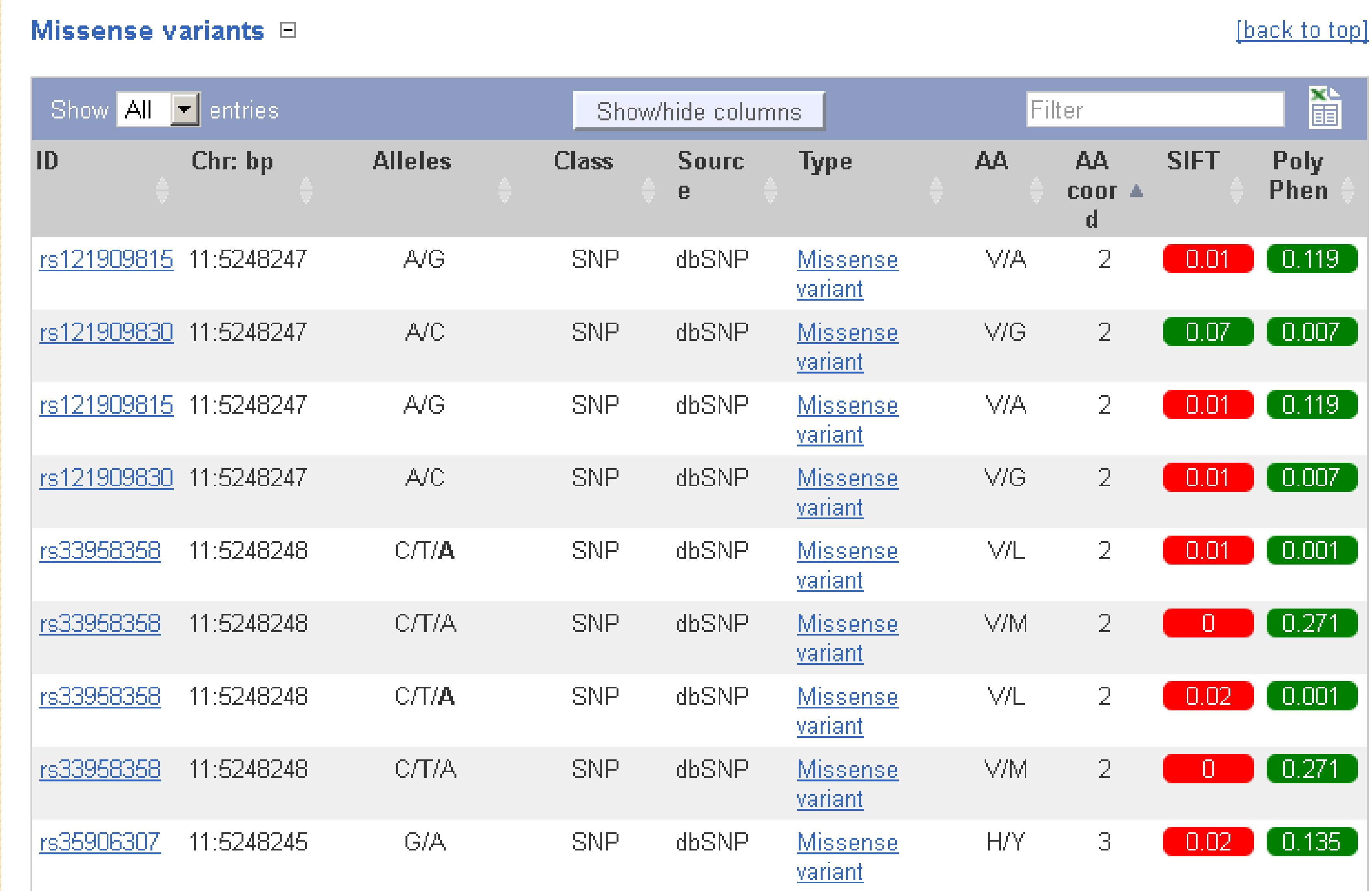
By Species	
Total	856
▶ Human	856



HBB

Description	hemoglobin, beta [Source:HGNC Symbol;Acc:4827] [Type: protein coding Ensembl/Havana merge genes]
Gene ID	ENSG00000244734
Location	11:5246694-5250625:-1
Variations	Variation Table
Source	e69

Ensembl "Variation Table" for *HBB* shows SIFT and PolyPhen scores for nonsynonymous variants (note they disagree)



Missense variants ⊟

[back to top]

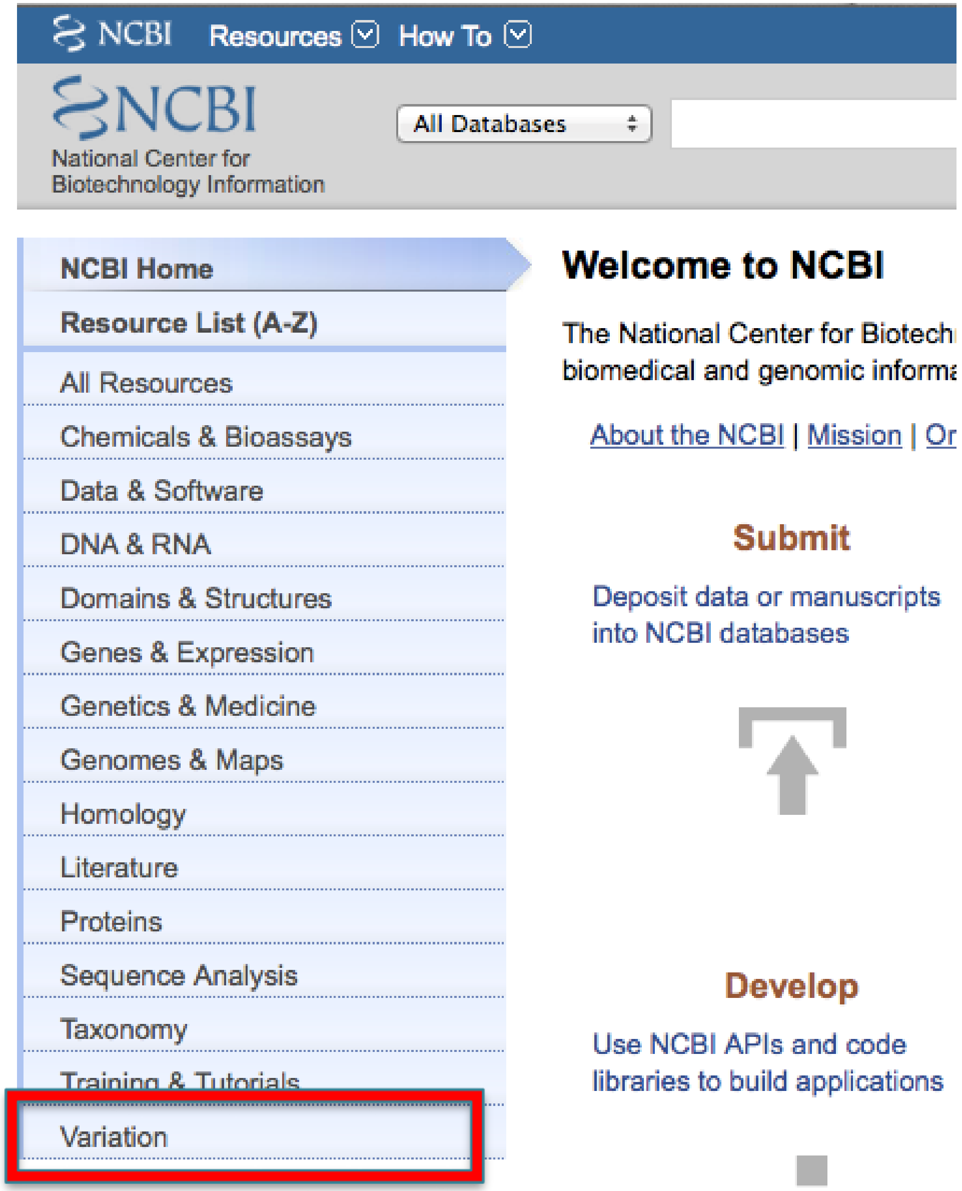
ID	Chr: bp	Alleles	Class	Source	Type	AA	AA coord ▲	SIFT	PolyPhen
rs121909815	11:5248247	A/G	SNP	dbSNP	Missense variant	V/A	2	0.01	0.119
rs121909830	11:5248247	A/C	SNP	dbSNP	Missense variant	V/G	2	0.07	0.007
rs121909815	11:5248247	A/G	SNP	dbSNP	Missense variant	V/A	2	0.01	0.119
rs121909830	11:5248247	A/C	SNP	dbSNP	Missense variant	V/G	2	0.01	0.007
rs33958358	11:5248248	C/T/**A**	SNP	dbSNP	Missense variant	V/L	2	0.01	0.001
rs33958358	11:5248248	C/**T**/A	SNP	dbSNP	Missense variant	V/M	2	0	0.271
rs33958358	11:5248248	C/T/**A**	SNP	dbSNP	Missense variant	V/L	2	0.02	0.001
rs33958358	11:5248248	C/**T**/A	SNP	dbSNP	Missense variant	V/M	2	0	0.271
rs35906307	11:5248245	G/A	SNP	dbSNP	Missense variant	H/Y	3	0.02	0.135

VAAST: probabilistic tool for disease variants

- VAAST (Variant Annotation, Analysis & Search Tool) is a probabilistic search tool used to identify disease-causing variants.

- VAAST calculates amino acid substitution frequencies for healthy genomes and disease genomes (both of these differ from standard BLOSUM62).

NCBI tools to understand variation



NCBI Resources How To

NCBI
National Center for
Biotechnology Information

All Databases

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotech
biomedical and genomic informa

About the NCBI | Mission | Or

Submit

Deposit data or manuscripts into NCBI databases

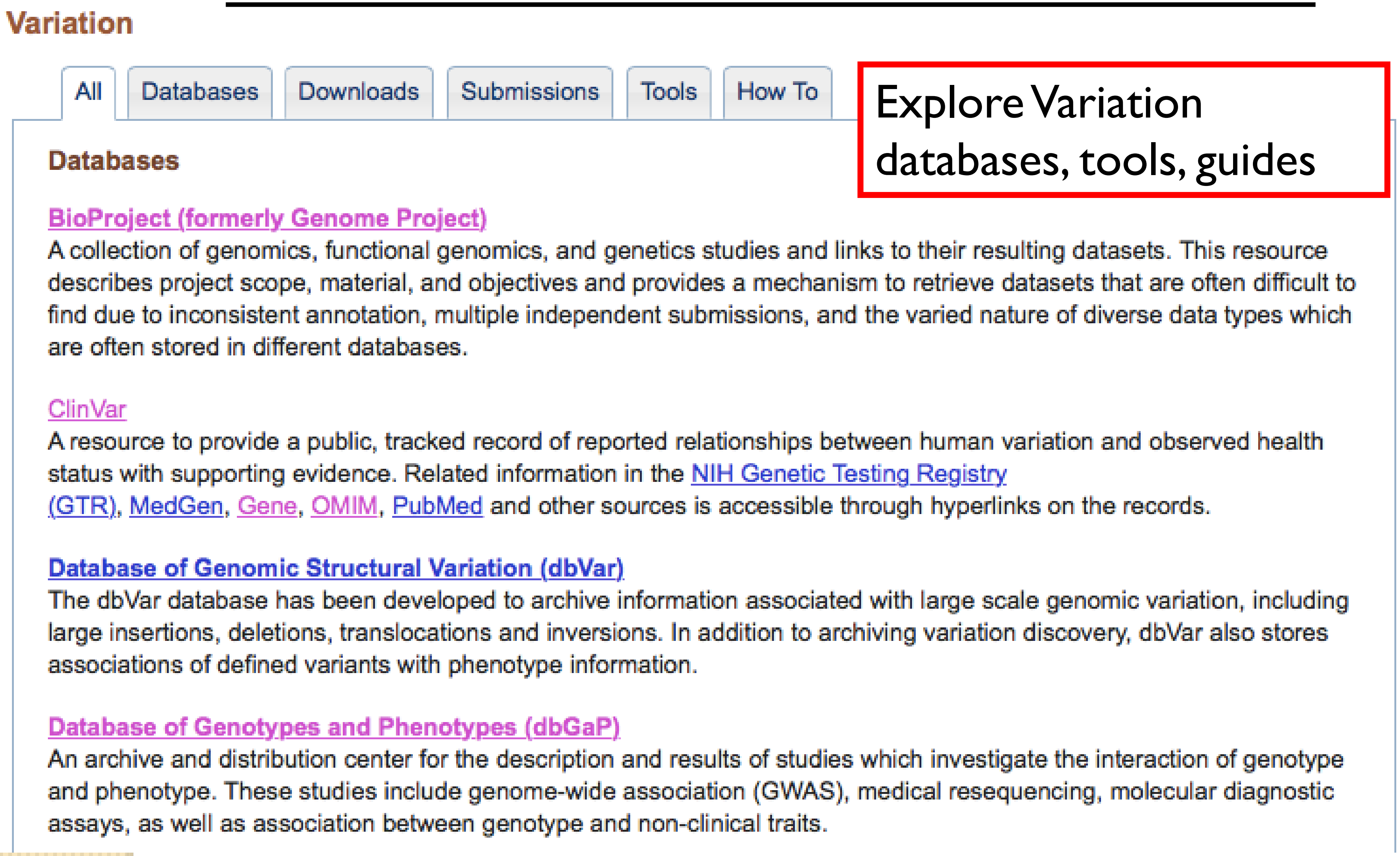
Develop

Use NCBI APIs and code libraries to build applications

From the home page of NCBI choose Variation

NCBI tools to understand variation

Explore Variation databases, tools, guides



Variation

Databases

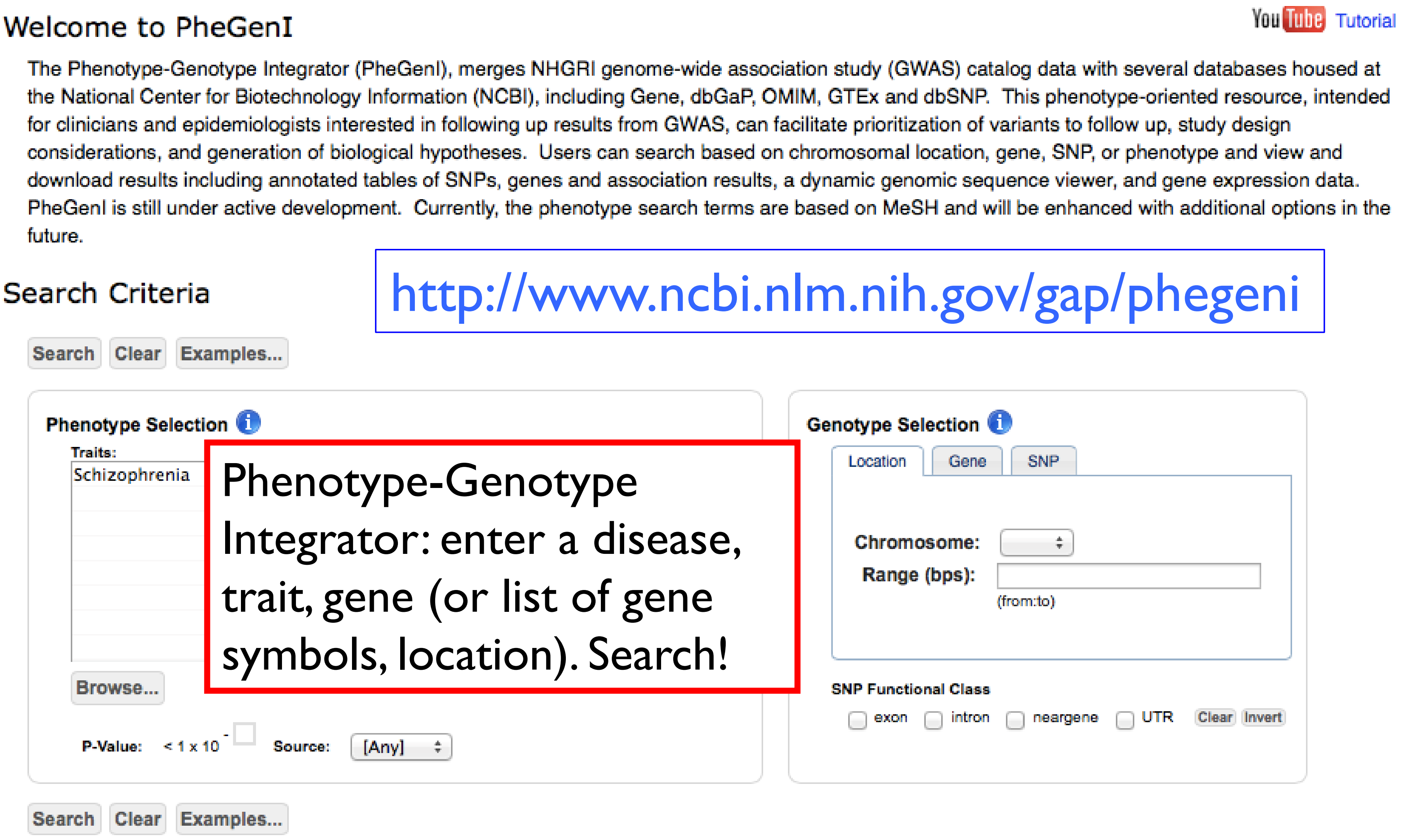
BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

ClinVar
A resource to provide a public, tracked record of reported relationships between human variation and observed health status with supporting evidence. Related information in the NIH Genetic Testing Registry (GTR), MedGen, Gene, OMIM, PubMed and other sources is accessible through hyperlinks on the records.

Database of Genomic Structural Variation (dbVar)
The dbVar database has been developed to archive information associated with large scale genomic variation, including large insertions, deletions, translocations and inversions. In addition to archiving variation discovery, dbVar also stores associations of defined variants with phenotype information.

Database of Genotypes and Phenotypes (dbGaP)
An archive and distribution center for the description and results of studies which investigate the interaction of genotype and phenotype. These studies include genome-wide association (GWAS), medical resequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits.

NCBI tools to understand variation: (1) PheGenI



Welcome to PheGenI

The Phenotype-Genotype Integrator (PheGenI), merges NHGRI genome-wide association study (GWAS) catalog data with several databases housed at the National Center for Biotechnology Information (NCBI), including Gene, dbGaP, OMIM, GTEx and dbSNP. This phenotype-oriented resource, intended for clinicians and epidemiologists interested in following up results from GWAS, can facilitate prioritization of variants to follow up, study design considerations, and generation of biological hypotheses. Users can search based on chromosomal location, gene, SNP, or phenotype and view and download results including annotated tables of SNPs, genes and association results, a dynamic genomic sequence viewer, and gene expression data. PheGenI is still under active development. Currently, the phenotype search terms are based on MeSH and will be enhanced with additional options in the future.

Search Criteria

Phenotype Selection

Traits: Schizophrenia

P-Value: < 1 x 10- □ Source: [Any]

Genotype Selection

Location | Gene | SNP

Chromosome:

Range (bps): (from:to)

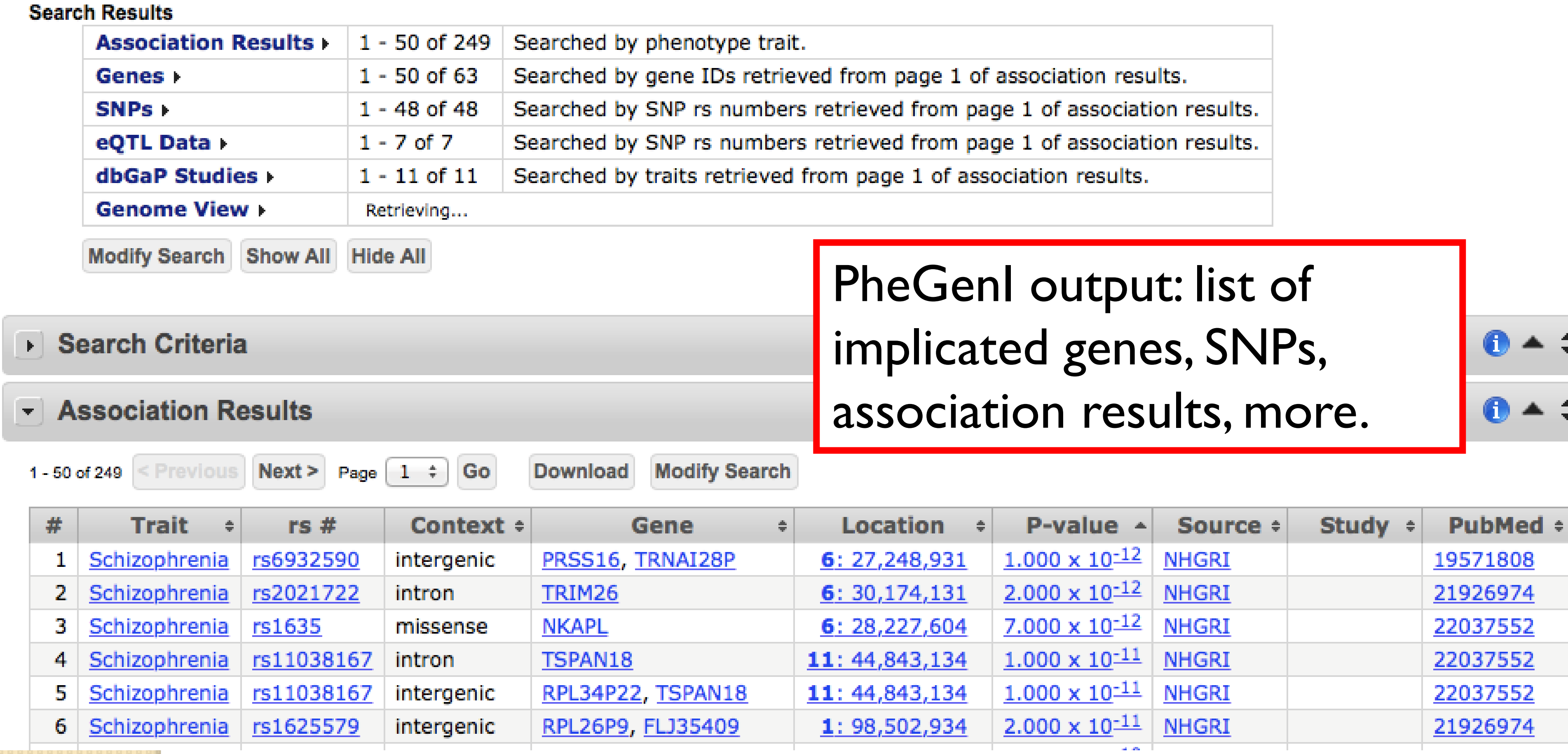
SNP Functional Class

□ exon □ intron □ neargene □ UTR

http://www.ncbi.nlm.nih.gov/gap/phegeni

Phenotype-Genotype Integrator: enter a disease, trait, gene (or list of gene symbols, location). Search!

NCBI tools to understand variation: (1) PheGenI



Search Results

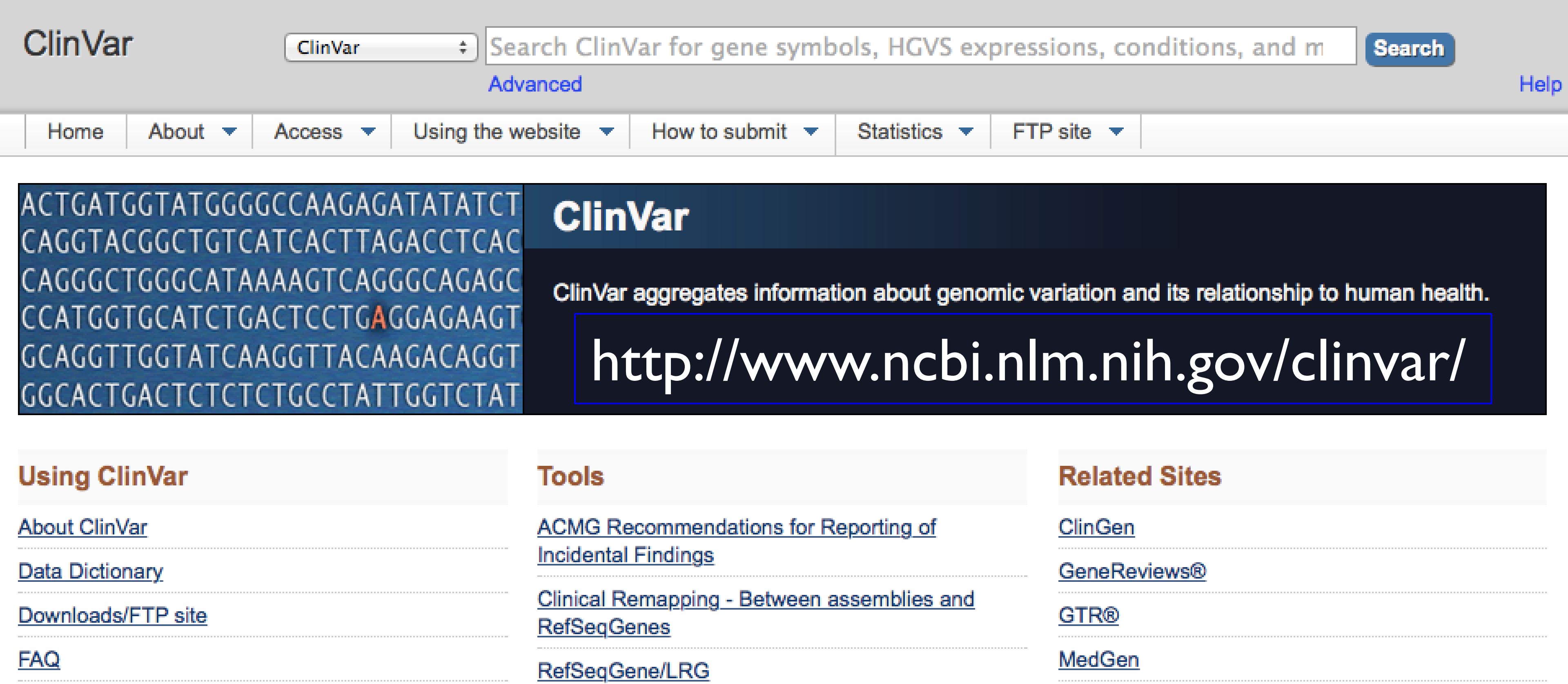
Association Results	1 - 50 of 249	Searched by phenotype trait.
Genes	1 - 50 of 63	Searched by gene IDs retrieved from page 1 of association results.
SNPs	1 - 48 of 48	Searched by SNP rs numbers retrieved from page 1 of association results.
eQTL Data	1 - 7 of 7	Searched by SNP rs numbers retrieved from page 1 of association results.
dbGaP Studies	1 - 11 of 11	Searched by traits retrieved from page 1 of association results.
Genome View	Retrieving...	

1 - 50 of 249

#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	PubMed
1	Schizophrenia	rs6932590	intergenic	PRSS16, TRNAI28P	6: 27,248,931	1.000×10^{-12}	NHGRI		19571808
2	Schizophrenia	rs2021722	intron	TRIM26	6: 30,174,131	2.000×10^{-12}	NHGRI		21926974
3	Schizophrenia	rs1635	missense	NKAPL	6: 28,227,604	7.000×10^{-12}	NHGRI		22037552
4	Schizophrenia	rs11038167	intron	TSPAN18	11: 44,843,134	1.000×10^{-11}	NHGRI		22037552
5	Schizophrenia	rs11038167	intergenic	RPL34P22, TSPAN18	11: 44,843,134	1.000×10^{-11}	NHGRI		22037552
6	Schizophrenia	rs1625579	intergenic	RPL26P9, FLJ35409	1: 98,502,934	2.000×10^{-11}	NHGRI		21926974

PheGenI output: list of implicated genes, SNPs, association results, more.

NCBI tools to understand variation: (2) ClinVar


ClinVar
ClinVar
Search ClinVar for gene symbols, HGVS expressions, conditions, and m
Search
Advanced
Help
Home
About
Access
Using the website
How to submit
Statistics
FTP site
ACTGATGGTATGGGGCCAAGAGATATATCT
CAGGTACGGCTGTCATCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCCTGAGGAGAAGT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTCTGCCTATTGGTCTAT

ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health.

http://www.ncbi.nlm.nih.gov/clinvar/

Using ClinVar

- About ClinVar
- Data Dictionary
- Downloads/FTP site
- FAQ

Tools

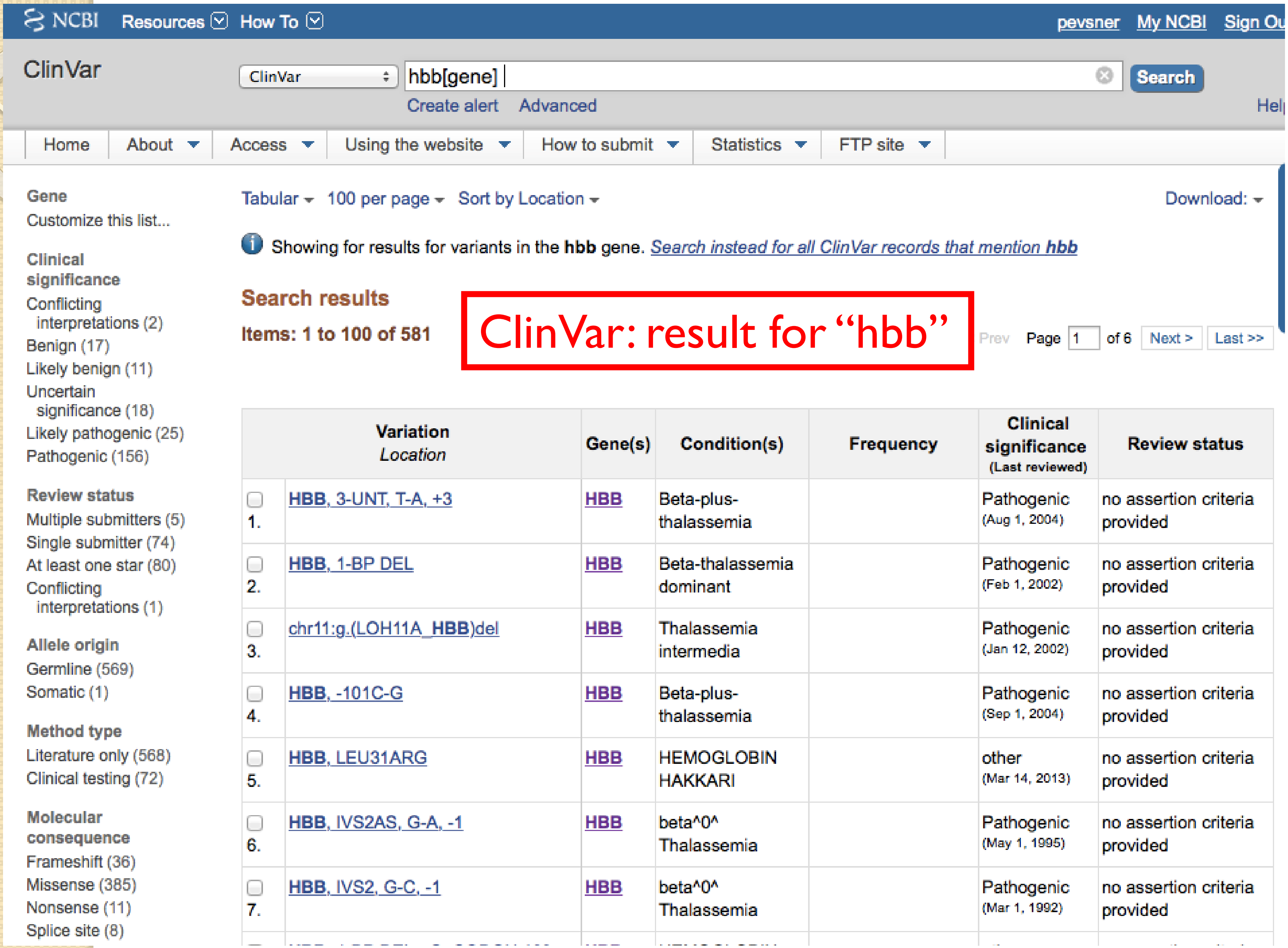
- ACMG Recommendations for Reporting of Incidental Findings
- Clinical Remapping - Between assemblies and RefSeqGenes
- RefSeqGene/LRG

Related Sites

- ClinGen
- GeneReviews®
- GTR®
- MedGen

ClinVar: "A resource to provide a public, tracked record of reported relationships between human variation and observed health status with supporting evidence. Related information in the NIH Genetic Testing Registry (GTR), MedGen, Gene, OMIM, PubMed and other sources is accessible through hyperlinks on the records."

NCBI tools to understand variation: (2) ClinVar



hbb[gene]

Showing for results for variants in the **hbb** gene. *Search instead for all ClinVar records that mention **hbb***

ClinVar: result for "hbb"

Search results

Items: 1 to 100 of 581

Gene
Customize this list...

Clinical significance
- Conflicting interpretations (2)
- Benign (17)
- Likely benign (11)
- Uncertain significance (18)
- Likely pathogenic (25)
- Pathogenic (156)

Review status
- Multiple submitters (5)
- Single submitter (74)
- At least one star (80)
- Conflicting interpretations (1)

Allele origin
- Germline (569)
- Somatic (1)

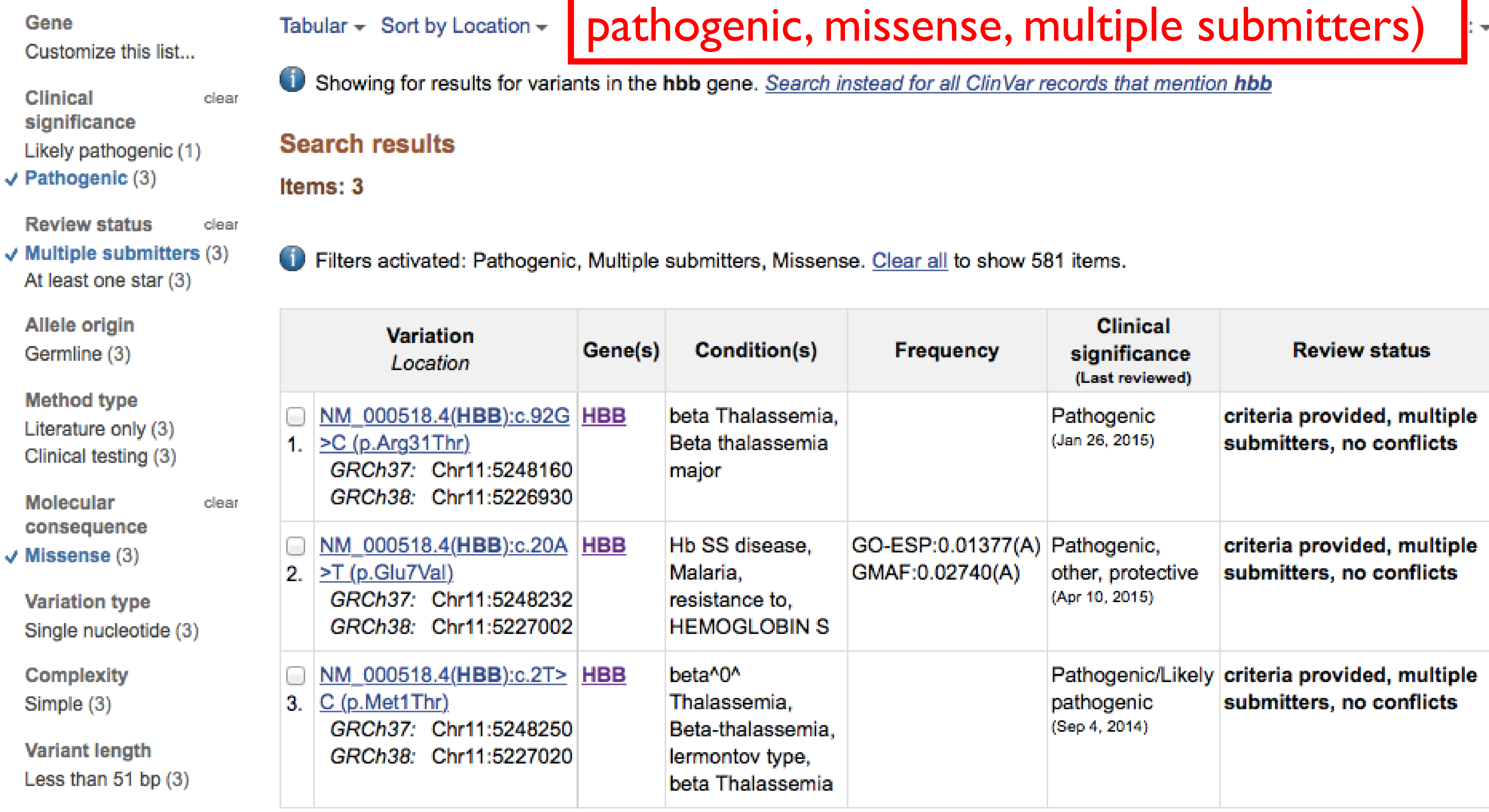
Method type
- Literature only (568)
- Clinical testing (72)

Molecular consequence
- Frameshift (36)
- Missense (385)
- Nonsense (11)
- Splice site (8)

	Variation *Location*	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
□ 1.	HBB, 3-UNT, T-A, +3	HBB	Beta-plus-thalassemia		Pathogenic (Aug 1, 2004)	no assertion criteria provided
□ 2.	HBB, 1-BP DEL	HBB	Beta-thalassemia dominant		Pathogenic (Feb 1, 2002)	no assertion criteria provided
□ 3.	chr11:g.(LOH11A_HBB)del	HBB	Thalassemia intermedia		Pathogenic (Jan 12, 2002)	no assertion criteria provided
□ 4.	HBB, -101C-G	HBB	Beta-plus-thalassemia		Pathogenic (Sep 1, 2004)	no assertion criteria provided
□ 5.	HBB, LEU31ARG	HBB	HEMOGLOBIN HAKKARI		other (Mar 14, 2013)	no assertion criteria provided
□ 6.	HBB, IVS2AS, G-A, -1	HBB	beta^0^ Thalassemia		Pathogenic (May 1, 1995)	no assertion criteria provided
□ 7.	HBB, IVS2, G-C, -1	HBB	beta^0^ Thalassemia		Pathogenic (Mar 1, 1992)	no assertion criteria provided

NCBI tools to understand variation: (2) ClinVar

ClinVar: use facets to limit results (here pathogenic, missense, multiple submitters)



Gene
Customize this list...

Clinical significance clear
Likely pathogenic (1)
✓ Pathogenic (3)

Review status clear
✓ Multiple submitters (3)
At least one star (3)

Allele origin
Germline (3)

Method type
Literature only (3)
Clinical testing (3)

Molecular consequence clear
✓ Missense (3)

Variation type
Single nucleotide (3)

Complexity
Simple (3)

Variant length
Less than 51 bp (3)

Showing for results for variants in the **hbb** gene. *Search instead for all ClinVar records that mention **hbb***

Search results

Items: 3

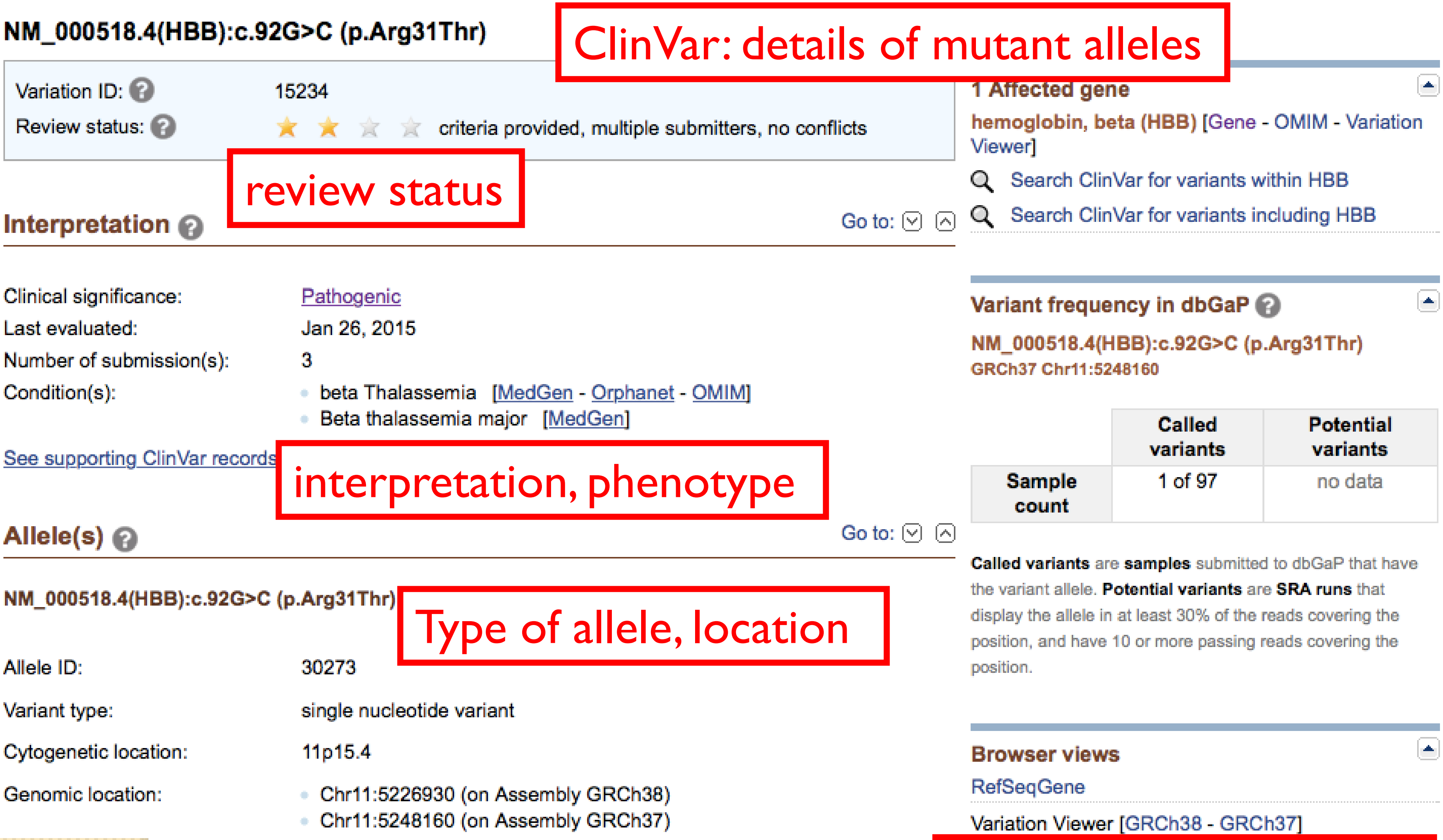
Filters activated: Pathogenic, Multiple submitters, Missense. Clear all to show 581 items.

	Variation *Location*	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
□ 1.	NM_000518.4(HBB):c.92G>C (p.Arg31Thr) *GRCh37:* Chr11:5248160 *GRCh38:* Chr11:5226930	HBB	beta Thalassemia, Beta thalassemia major		Pathogenic (Jan 26, 2015)	**criteria provided, multiple submitters, no conflicts**

□ 2.	NM_000518.4(HBB):c.20A>T (p.Glu7Val)
	GRCh37: Chr11:5248232
	GRCh38: Chr11:5227002
		HBB	Hb SS disease, Malaria, resistance to, HEMOGLOBIN S	GO-ESP:0.01377(A)
				GMAF:0.02740(A)
					Pathogenic, other, protective
					(Apr 10, 2015)
						criteria provided, multiple submitters, no conflicts

□ 3.	NM_000518.4(HBB):c.2T>C (p.Met1Thr) *GRCh37:* Chr11:5248250 *GRCh38:* Chr11:5227020	HBB	beta^0^ Thalassemia, Beta-thalassemia, lermontov type, beta Thalassemia		Pathogenic/Likely pathogenic (Sep 4, 2014)	**criteria provided, multiple submitters, no conflicts**

NCBI tools to understand variation: (2) ClinVar



NM_000518.4(HBB):c.92G>C (p.Arg31Thr)

Variation ID: ? 15234

Review status: ? ★★☆☆ criteria provided, multiple submitters, no conflicts

Interpretation ? Go to:

Clinical significance: Pathogenic

Last evaluated: Jan 26, 2015

Number of submission(s): 3

Condition(s):
- beta Thalassemia [MedGen - Orphanet - OMIM]
- Beta thalassemia major [MedGen]

See supporting ClinVar records

Allele(s) ? Go to:

NM_000518.4(HBB):c.92G>C (p.Arg31Thr)

Allele ID: 30273

Variant type: single nucleotide variant

Cytogenetic location: 11p15.4

Genomic location:
- Chr11:5226930 (on Assembly GRCh38)
- Chr11:5248160 (on Assembly GRCh37)

1 Affected gene

hemoglobin, beta (HBB) [Gene - OMIM - Variation Viewer]

Search ClinVar for variants within HBB

Search ClinVar for variants including HBB

Variant frequency in dbGaP ?

NM_000518.4(HBB):c.92G>C (p.Arg31Thr)
GRCh37 Chr11:5248160

	Called variants	Potential variants
Sample count	1 of 97	no data

Called variants are **samples** submitted to dbGaP that have the variant allele. **Potential variants** are **SRA runs** that display the allele in at least 30% of the reads covering the position, and have 10 or more passing reads covering the position.

Browser views

RefSeqGene

Variation Viewer [GRCh38 - GRCh37]

ClinVar: details of mutant alleles

review status

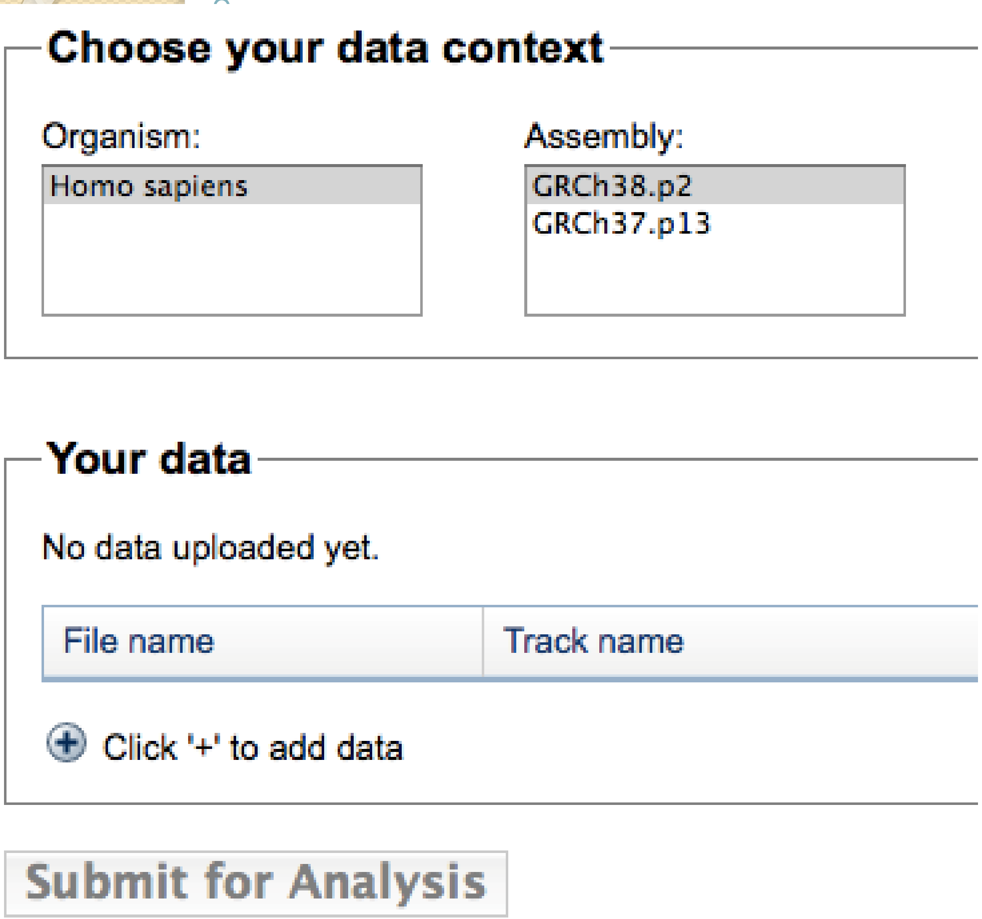
interpretation, phenotype

Type of allele, location

Link to Variation Viewer

http://www.ncbi.nlm.nih.gov/clinvar/variation/15234/

NCBI tools to understand variation: (3) Variation Reporter



Choose your data context

Organism:

Homo sapiens

Assembly:

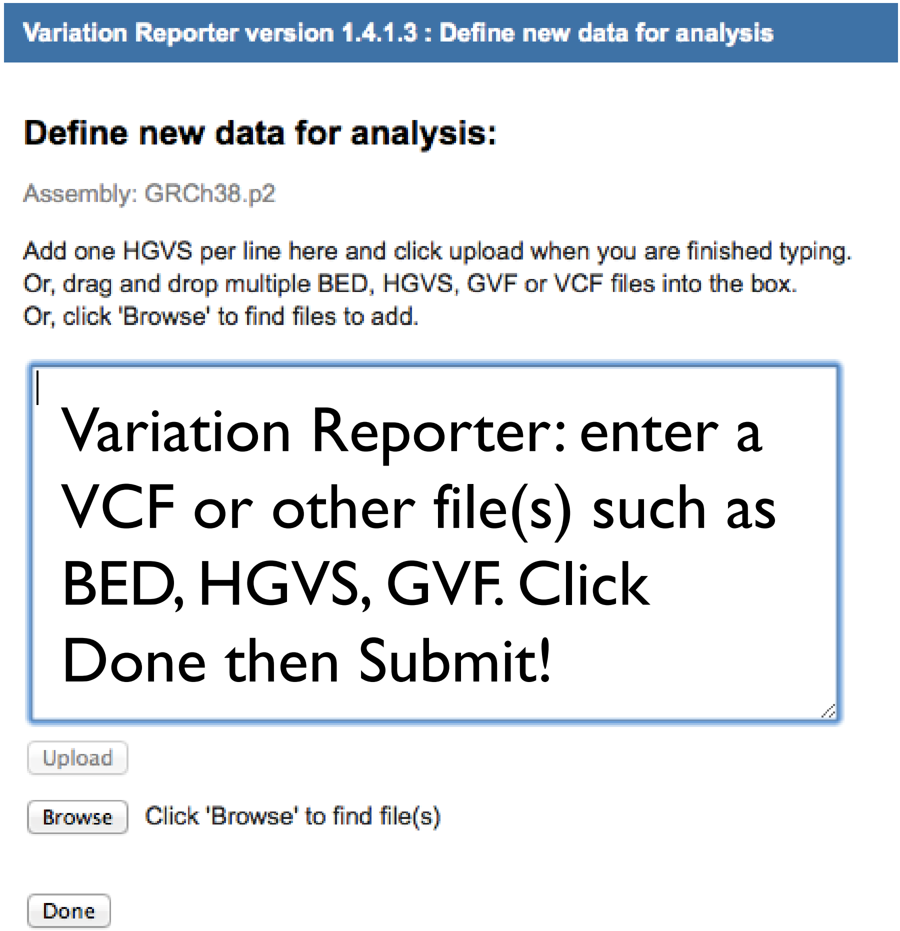
GRCh38.p2
GRCh37.p13

Your data

No data uploaded yet.

File name	Track name

⊕ Click '+' to add data



Variation Reporter version 1.4.1.3 : Define new data for analysis

Define new data for analysis:

Assembly: GRCh38.p2

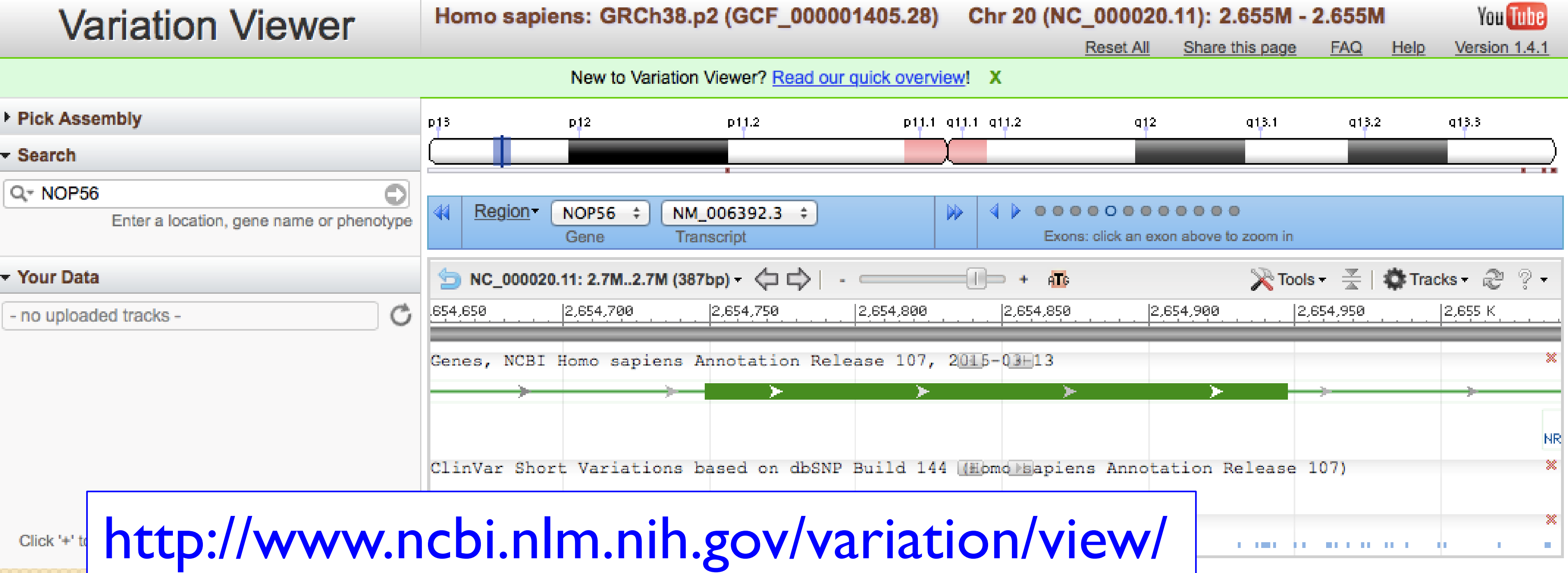
Add one HGVS per line here and click upload when you are finished typing.
Or, drag and drop multiple BED, HGVS, GVF or VCF files into the box.
Or, click 'Browse' to find files to add.

Variation Reporter: enter a VCF or other file(s) such as BED, HGVS, GVF. Click Done then Submit!

Click 'Browse' to find file(s)

http://www.ncbi.nlm.nih.gov/variation/tools/reporter

NCBI tools to understand variation: (4) Variation Viewer

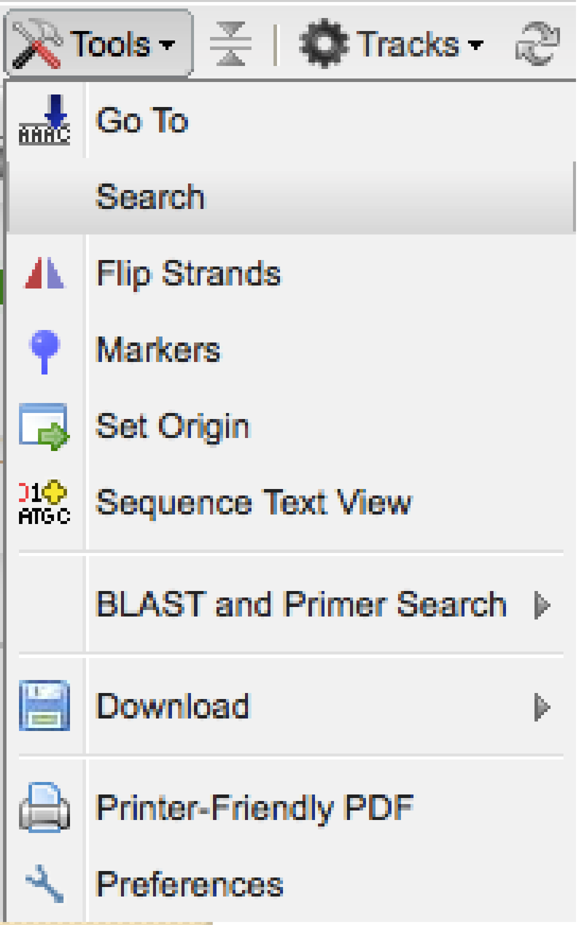

Variation Viewer
Homo sapiens: GRCh38.p2 (GCF_000001405.28)
Chr 20 (NC_000020.11): 2.655M - 2.655M
Reset All
Share this page
FAQ
Help
Version 1.4.1
New to Variation Viewer? Read our quick overview!
Pick Assembly
Search
NOP56
Enter a location, gene name or phenotype
Your Data
- no uploaded tracks -
Region
NOP56
Gene
NM_006392.3
Transcript
Exons: click an exon above to zoom in
NC_000020.11: 2.7M..2.7M (387bp)
Tools
Tracks
Genes, NCBI Homo sapiens Annotation Release 107, 2015-03-13
ClinVar Short Variations based on dbSNP Build 144 (Homo sapiens Annotation Release 107)

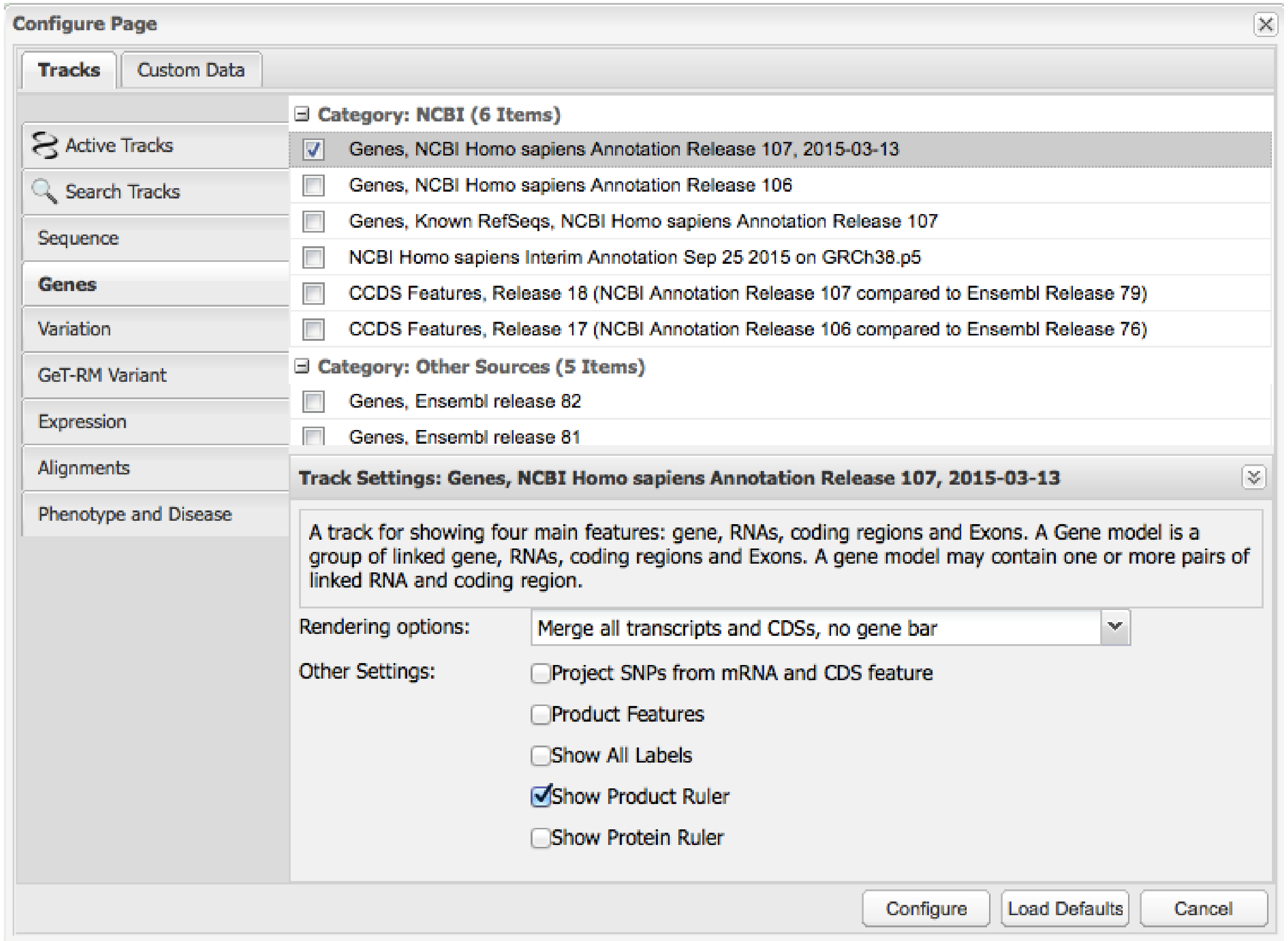
http://www.ncbi.nlm.nih.gov/variation/view/

Variation Viewer: "A genomic browser to search and view genomic variations listed in dbSNP, dbVar, and ClinVar databases. Searches can be performed using chromosomal location, gene symbol, phenotype, or variant IDs from dbSNP and dbVar. The browser enables exploration of results in a dynamic graphical sequence viewer with annotated tables of variations."

NCBI tools to understand variation: (4) Variation Viewer

Variation Viewer: vast options in tools and tracks (the gear icon)





⊟ Category: NCBI (6 Items)

☑ Genes, NCBI Homo sapiens Annotation Release 107, 2015-03-13
☐ Genes, NCBI Homo sapiens Annotation Release 106
☐ Genes, Known RefSeqs, NCBI Homo sapiens Annotation Release 107
☐ NCBI Homo sapiens Interim Annotation Sep 25 2015 on GRCh38.p5
☐ CCDS Features, Release 18 (NCBI Annotation Release 107 compared to Ensembl Release 79)
☐ CCDS Features, Release 17 (NCBI Annotation Release 106 compared to Ensembl Release 76)

⊟ Category: Other Sources (5 Items)

☐ Genes, Ensembl release 82
☐ Genes, Ensembl release 81

Track Settings: Genes, NCBI Homo sapiens Annotation Release 107, 2015-03-13

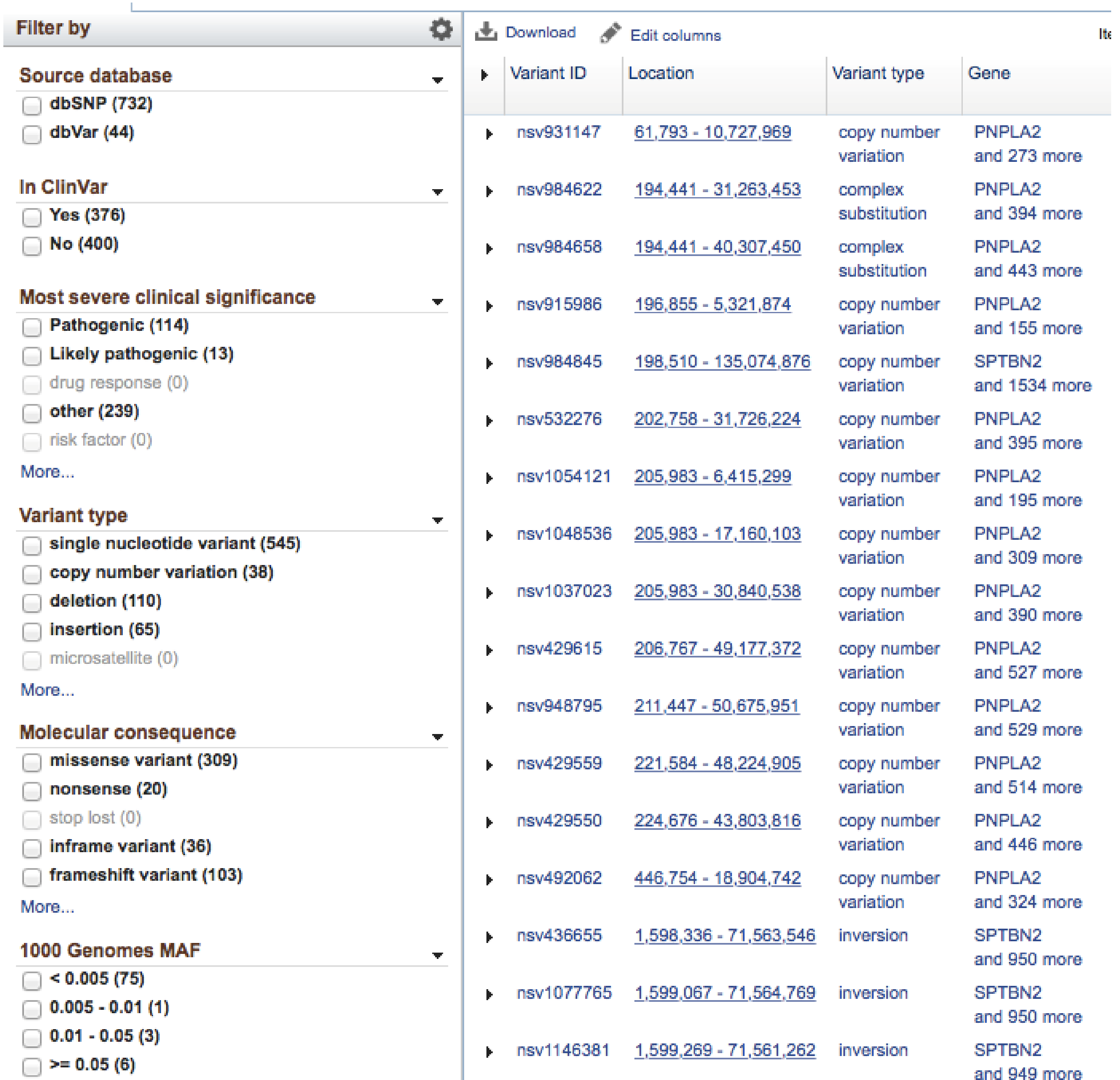
A track for showing four main features: gene, RNAs, coding regions and Exons. A Gene model is a group of linked gene, RNAs, coding regions and Exons. A gene model may contain one or more pairs of linked RNA and coding region.

Rendering options: Merge all transcripts and CDSs, no gene bar

Other Settings:
☐ Project SNPs from mRNA and CDS feature
☐ Product Features
☐ Show All Labels
☑ Show Product Ruler
☐ Show Protein Ruler

NCBI tools to understand variation: (4) Variation Viewer

Variation Viewer: note extensive faceted searches



Source database

- □ dbSNP (732)
- □ dbVar (44)

In ClinVar

- □ Yes (376)
- □ No (400)

Most severe clinical significance

- □ Pathogenic (114)
- □ Likely pathogenic (13)
- □ drug response (0)
- □ other (239)
- □ risk factor (0)

More...

Variant type

- □ single nucleotide variant (545)
- □ copy number variation (38)
- □ deletion (110)
- □ insertion (65)
- □ microsatellite (0)

More...

Molecular consequence

- □ missense variant (309)
- □ nonsense (20)
- □ stop lost (0)
- □ inframe variant (36)
- □ frameshift variant (103)

More...

1000 Genomes MAF

- □ < 0.005 (75)
- □ 0.005 - 0.01 (1)
- □ 0.01 - 0.05 (3)
- □ >= 0.05 (6)

Variant ID	Location	Variant type	Gene
nsv931147	61,793 - 10,727,969	copy number variation	PNPLA2 and 273 more
nsv984622	194,441 - 31,263,453	complex substitution	PNPLA2 and 394 more
nsv984658	194,441 - 40,307,450	complex substitution	PNPLA2 and 443 more
nsv915986	196,855 - 5,321,874	copy number variation	PNPLA2 and 155 more
nsv984845	198,510 - 135,074,876	copy number variation	SPTBN2 and 1534 more
nsv532276	202,758 - 31,726,224	copy number variation	PNPLA2 and 395 more
nsv1054121	205,983 - 6,415,299	copy number variation	PNPLA2 and 195 more
nsv1048536	205,983 - 17,160,103	copy number variation	PNPLA2 and 309 more
nsv1037023	205,983 - 30,840,538	copy number variation	PNPLA2 and 390 more
nsv429615	206,767 - 49,177,372	copy number variation	PNPLA2 and 527 more
nsv948795	211,447 - 50,675,951	copy number variation	PNPLA2 and 529 more
nsv429559	221,584 - 48,224,905	copy number variation	PNPLA2 and 514 more
nsv429550	224,676 - 43,803,816	copy number variation	PNPLA2 and 446 more
nsv492062	446,754 - 18,904,742	copy number variation	PNPLA2 and 324 more
nsv436655	1,598,336 - 71,563,546	inversion	SPTBN2 and 950 more
nsv1077765	1,599,067 - 71,564,769	inversion	SPTBN2 and 950 more
nsv1146381	1,599,269 - 71,561,262	inversion	SPTBN2 and 949 more

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview Topic 6: Variant calling: SNVs
Topic 1: Design Topic 7: Variant calling:
SVs
Topic 2: FASTQ Topic 8: VCF
Topic 3: Assembly Topic 9: Visualizing NGS data
Topic 4: Alignment Topic 10: Significance
Topic 5: SAM/BAM

Specialized applications of NGS

Perspective

Specialized next-generation sequence (NGS) applications

There are many useful applications of NGS technology. These include:

- RNA-seq to measure RNA levels ("gene expression" of genes and isoforms)
- Chromatin immunoprecipitation sequencing (ChIP-Seq) to measure protein– DNA interactions
- Methyl-seq
- FAIRE-seq
- many others

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction

DNA sequencing technologies

Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA

Overview
Topic 1: Design
Topic 2: FASTQ
Topic 3: Assembly
Topic 4: Alignment
Topic 5: SAM/BAM

Topic 6: Variant calling: SNVs
Topic 7: Variant calling: SVs
Topic 8: VCF
Topic 9: Visualizing NGS data
Topic 10: Significance

Specialized applications of NGS

Perspective

Perspective

Next-generation sequencing (NGS) technology is revolutionizing biology. We are now able to catalog genetic variation at unprecedented depth.

There is rapid growth in the technologies used for NGS. There are also vast numbers of software solutions for quality control, sequence alignment, genome assembly, variant calling (including single nucleotide variants, indels, and structural variants), and variant prioritization.

Key file formats include FASTQ ("raw" reads), BAM/SAM (aligned reads), and VCF (variant calls). Many tools are available for the generation, analysis, and visualization of these types of files.