

# Read Mapping and Variant Calling

# Whole Genome Resequencing

- Sequencing multiple individuals from the same species
- Reference genome is already available
- Discover variations in the genomes between and within samples
  - mutations
  - insertions
  - deletions
  - rearrangements
  - copy number changes

How long do the reads need to be?

For the human genome, estimates are:

25mers = 80% unique coverage

43mers = 90% unique coverage

But longer is better for certain applications.

Quality Assessment



Trimming



Quality Assessment



Mapping to a Reference



Visualization



Calling variants



Assessing functional impact of variants

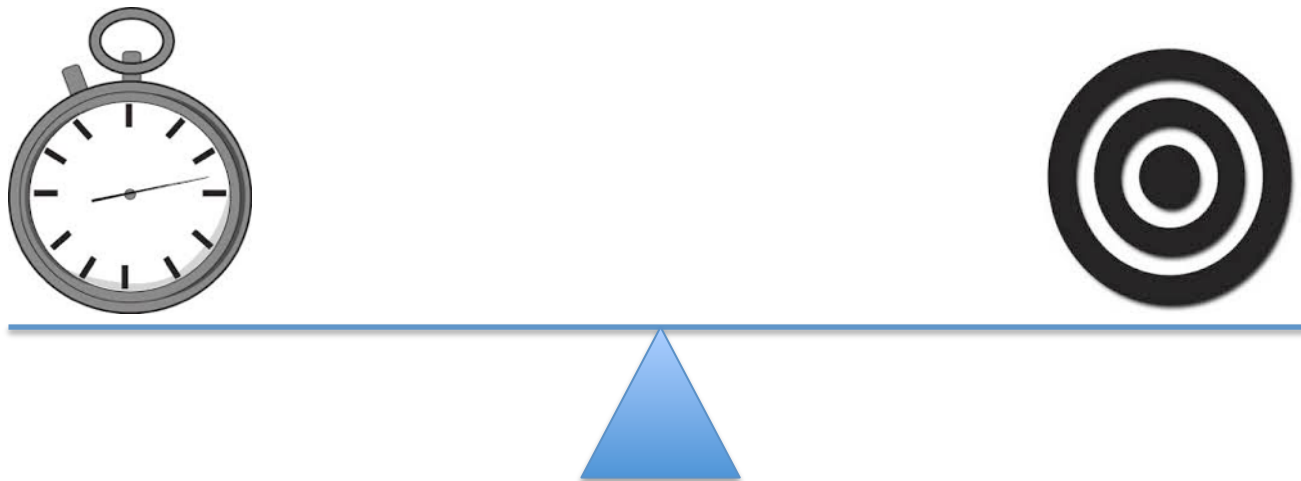


Submit to SRA

**Workflow**

# But we've already covered alignment with BLAST.

- Alignment methods must have tradeoffs for speed vs accuracy
- Depending on the application, may want to make different tradeoffs
- Global vs local
- Different types of alignment objectives lead to different categories of aligners



# Short Read Mappers

- BLAST is much faster than original algorithms (Smith Waterman for example)
- Still too slow for the amount of data produced by NGS technology
- Resequencing usually involves comparing very similar sequences (>90% identity in residues) to a reference genome
- Software that leverages this high percent identity can be faster
- Can generally utilize a global strategy

# Short Read Mappers

- Orders of magnitude faster than BLAST
- several tens of millions of reads mapped per hour per CPU
- Only matches of 95% identity or greater are found
- Indels are particularly problematic
- Usually only output the best hit or the set of hits all equivalently good
  - The point is usually to find the origin in the reference genome
  - Other genomic regions of lower identity are not considered useful

# Uniqueness

- Some reads can be mapped uniquely to the reference
- Some can't → Multiple alignment reads
- Multiple alignment reads are difficult to apply to downstream applications
  - RNASeq – which gene do they represent?
  - SNP – which location carries the polymorphism?
- How to deal with multiple alignment reads?
  - Throw them away
  - This introduces bias and ignores real genomic regions that may be biologically important

# Clever Tricks to find “Best” Alignments

- Use the quality values
  - Penalize mismatches at high quality bases more than mismatches at low quality bases
- Paired End information
  - If one read does not map uniquely, but the other does, use that information to place the non-unique one
  - Need to know your insert size



# Decisions for the end user

- How many mismatches are allowed for a read to be considered mapped?
  - Heterozygosity between sample and reference
  - Incomplete/low quality reference
- How many matches to report?
  - Does your downstream analysis need/want to include multiple matches?

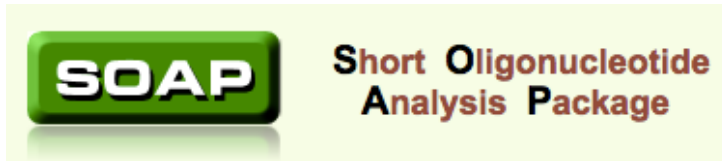
**Explore the documentation and parameters for your software of choice**

**Is it doing what you think its doing?**

# Lots of choices

65 software packages listed on the wikipedia page for alignment software

Blat  
BWA  
Eland  
GSNAP  
Maq  
RMAP  
Stampy  
SHRiMP



**mosaik**



Prize for best named:  
VelociMapper



What's in the box matters.

# How to choose?

- Memory efficient
- Good documentation
- Responsive mailing list or help forum
- Maintained and updated when bugs are found
  - BWA

# What mappers have in common:

## Indexing Strategies

- Usually, the first step is to transform part of the data into a more suitable form for fast searching
- Indexing – creating a glossary or look up table
- Without indexing you would have to scan everything each time you did a search
- Consider web search engines



# Burrows-Wheeler Aligner

- Has three algorithms
- Individual chromosomes cannot be longer than 2GB
- Output in SAM format



# Burrows-Wheeler Aligner

<http://bio-bwa.sourceforge.net/>

- three algorithms:
- **BWA-backtrack**
  - Meant for sequences of up to 100bp in length
  - Meant for reads with less than 2% error (can do some end trimming)
- **BWA-MEM**
  - Use for any sequences greater than 70bp up to 1Mb
  - Much more widely used now that sequencers output longer reads
  - Will work with reads with
    - 2% error for 100bp
    - 3% error for 200bp
    - etc
  - Has split read support
    - structural variations, gene fusion or reference misassembly
- **BWA-MEM is more accurate and faster**

# **SAM AND BAM FORMAT**

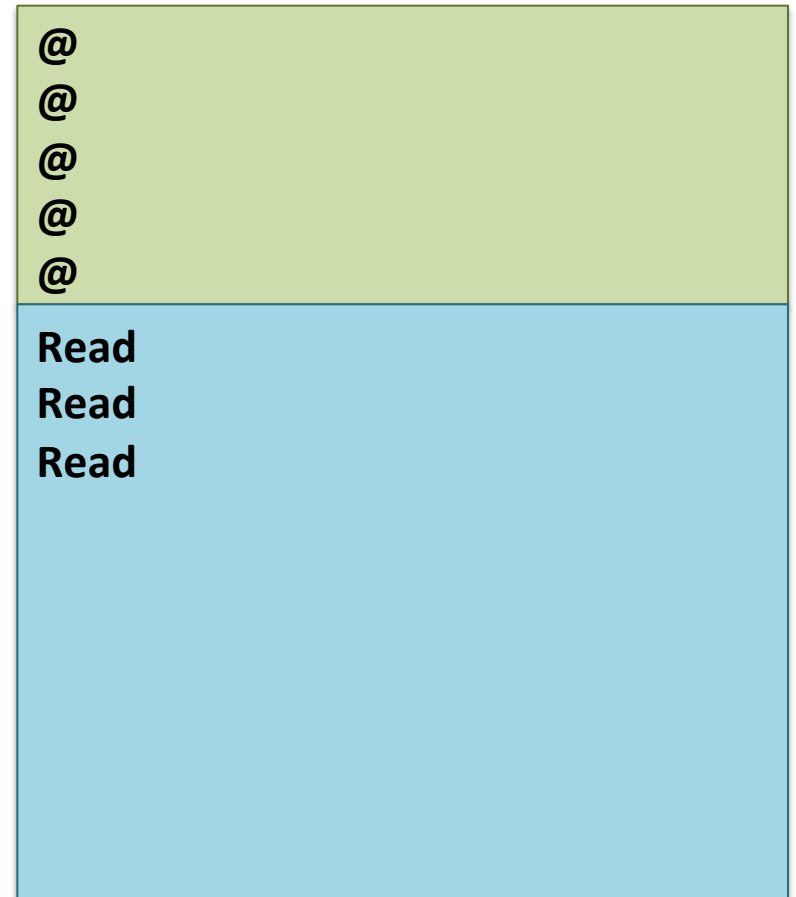
# SAM Format

- SAM = Sequence Alignment/Map format
- Tab delimited plain text
- Store large nucleotide sequence alignments
  - Alignment of every read
  - Including gaps and SNPs
  - Pairing of reads
  - Can record more than one alignment location in the genome
  - Stores quality values
  - Stores information about duplication
- Flexible
- Useful for operations on very large sequences
- Extremely detailed documentation
  - <https://samtools.github.io/hts-specs/SAMv1.pdf>
- Manipulations are primarily done with the software samtools
- Originally designed to store mapping information, now used as a primary storage format for unmapped sequences as well



# SAM - Header

- Structure
  - Optional Header at top of file
  - Alignment information



# SAM - Header

- Header lines start with @ symbol
- Always at top of file
- Contain lots of information about what was mapped, what it was mapped to, and how (metadata)
  - the version information for the SAM/BAM file
  - whether or not and how the file is sorted
  - information about the reference sequences
  - any processing that was used to generate the various reads in the file
  - software version

# Simple Header

@HD = first line

VN = version of SAM format

SO = sort order (this is sorted by coordinates)

@HD VN:1.5 SO:coordinate

@SQ SN:ref LN:45

@SQ = reference sequence

SN = Sequence reference Name

LN = sequence reference length

# Alignment Line

- Below the headers are the alignment records
  - Tab-delimited fields
- 
- 1 QNAME Query template/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 TLEN inferred Template LENgth (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:VALUE

Lets unpack this alignment line, taken from a SAM file:

```
SRR030257.2000020    83 gi|254160123|ref|
NC_012967.1|329575260 36M   =  3295706-82
TGCTGGCGGCGATATCGTCCGTGGTTCCGATCTGGT
?%<91<?>>??AAAAAAAAAAAAAAAAAAAAAAAAA
XT:A: NM:i:0  SM:i:37  AM:i:37  X0:i:1X1:i:0
XM:i:0  XO:i:0  XG:i:0  MD:Z:36
```

# SAM Field 1

Query name

SRR030257.2000020

2. Flag:

83

## Field 2: Flag

83

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

64 + 16 + 2 + 1

1 = Read is paired

2 = Read mapped in proper pair

16 = Read mapped to reverse strand

64 = First in pair

Look up aSAM flag: <https://broadinstitute.github.io/picard/explain-flags.html>

# SAM Field 3

Reference sequence name (useful especially if you have multiple chromosomes)

```
gi | 254160123 | ref | NC_012967.1 |
```



# SAM Field 4

Position- 1-based leftmost mapping POSition of the first matching base

3295752

# SAM Field 5

## Mapping Quality

- equals  $-10 \log_{10} \text{Pr}\{\text{mapping position is wrong}\}$ , rounded to the nearest integer
- Probability of 99.9% = map quality of 30
- Probability of 0% = map quality of 0
- value 255 indicates that the mapping quality is not available.

60

.000001% probability wrong

# SAM Field 6

## CIGAR String

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

36M 36 nucleotides match (perfect match)

8S28M 8 nucleotides clipped, 28 match

# More CIGAR

Aligning these two:

RefPos:	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T		G	A	A	C	T	G	A	C	T	A	A	C
Read:					A	C	T	A	G	A	A	C	T	G	G	C	T			

Position:

5


CIGAR:

3M1I3M1D5M

# SAM Field 7

Reference sequence for the next read in the template

- For a forward read, this is the reference where the reverse read maps
- For a reverse read, this is the reference where the forward read maps

 = reverse read maps on the same reference

# SAM Field 8

Position where the next read maps

3295706

(Forward read mapped at 3295752.  
Remember the forward read mapped  
to the reverse strand)

# SAM Field 9

Observed template length

-82

# SAM Field 10

Sequence of the read

TGCTGGCGGCGATATCGTCCGTGGTTCCGATCTGGT



# SAM Field 11

Quality of the read

```
?%<91<?>>??AAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

# SAM Field 12

Optional MORE information

TAG:TYPE:VALUE format

```
XT:A:U   NM:i:0   SM:i:37  AM:i:37  X0:i:1  
X1:i:0   XM:i:0   XO:i:0   XG:i:0   MD:Z:36
```

Anything with an X is specified by the user or by the mapping software, and is not part of the SAM spec.

# Decipher the last fields

XT:A:U	One of Unique/Repeat/N/Mate-sw
NM:i:0	Edit distance to the reference
SM:i:37	Template-independent mapping quality
AM:i:37	Smallest template-independent mapping quality of other segments
X0:i:1	Number of best hits
X1:i:0	Number of suboptimal hits found by BWA
XM:i:0	Number of mismatches in the alignment
XO:i:0	Number of gap opens
XG:i:0	Number of gap extensions
MD:Z:36	String for mismatching positions

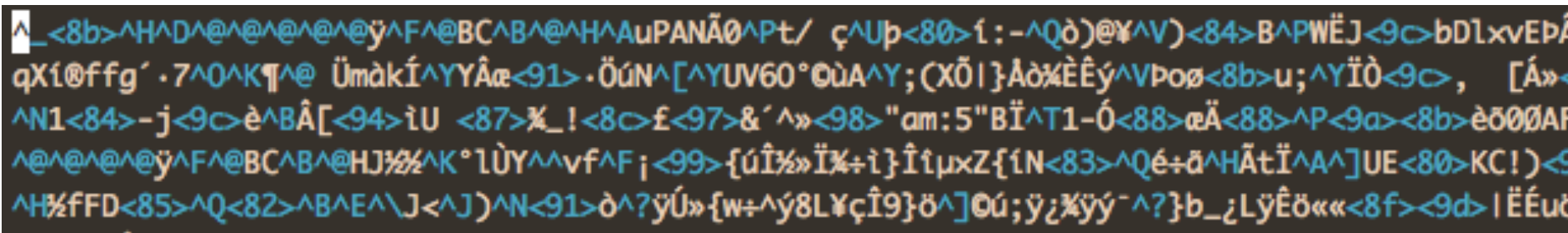
# SAM

The first three lines of a sam file:

```
@SQ  SN:gi|254160123|ref|NC_012967.1|  LN:4629812
@PG  ID:bwaPN:bwa  VN:0.7.12-r1039  CL:/lustre/projects/
rnaseq_ws/apps/bwa-0.7.12/bwa sampe ../raw_data/
NC_012967.1.fasta aln_SRR030257_1.sai
aln_SRR030257_2.sai ../raw_data/SRR030257_1.fastq ../
raw_data/SRR030257_2.fastq
SRR030257.1 99 gi|254160123|ref|NC_012967.1|
950180 60 36M = 950295 151
TTACACTCCTGTTAATCCATACAGCAACAGTATTGG
AAA;A;AA?A?AAAAA?;?A?1A;;????566)=*1 XT:A:U
NM:i:1SM:i:37 AM:i:25 X0:i:1 X1:i:0 XM:i:1 XO:i:0
XG:i:0 MD:Z:32C3
```

# BAM Format

- Sister format to SAM
- BAM – Binary version of SAM
- compressed **BGZF** (Blocked GNU Zip Format) - a variant of GZIP (GNU ZIP),
- files are bigger than GZIP files, but they are much faster for random access
- Can index and then look up information embedded in the file with decompressing the whole file
- up to 75% smaller in size
- Not readable by people



```
^_<8b>^H^D^@^@^@^@^@y^F^@BC^B^@^H^A^uPAN^0^P^t/  ç^Up<80>í:-^Qð)^@Y^V)^<84>B^PWËJ<9C>bDl^xvE^P^
qXí@ffg'.7^0^K^T^@ ÜmàkÍ^YY^Â^æ<91>·ÖúN^[^YUV60°0ùA^Y;(XÖI}Àð%ÈËý^V^oø<8b>u;^YÏÒ<9C>, [Á»
^N1<84>-j<9C>è^B^Â[<94>iU <87>%_!<8C>f<97>&'^»<98>"am:5"BÏ^T1-Ó<88>æÄ<88>^P<9a><8b>èðððAF
^@^@^@^@^@y^F^@BC^B^@HJ^%K^lÙY^^vf^F; <99>{úÎ%»Ï%+i}ÎîµxZ{îN<83>^Qé+ð^H^tÏ^A^]UE<80>KC!)<9
^H%FFD<85>^Q<82>^B^E^^\J<^J)^N<91>ð^?ýÚ»{w+^ý8L^çÎ9}ð^]0ú;ÿ¿%ÿý^-^?}b_¿LÿÊö««<8f><9d>|ËÉuð
```

# samtools

- View – print alignments to your screen or convert between formats. Can reduce files to a particular region only
- Tview - text alignment viewer, nifty for quick viewing of files
- Mpileup – generates a special mpileup formatted file needed for calling variants
- Sort – sort the alignments (by default, sorts by coordinate). Sorting is needed for most downstream applications.
- Merge – concatenate bam files together, while maintaining sorting order
- Index - index a bam or cram file, needed for most downstream applications
- Idxstats – get some stats about your bam file
- Faidx - index a fasta file, need for most downstream applications using a bam file
- Bam2fq – convert a bam file to a fastq file
- More...

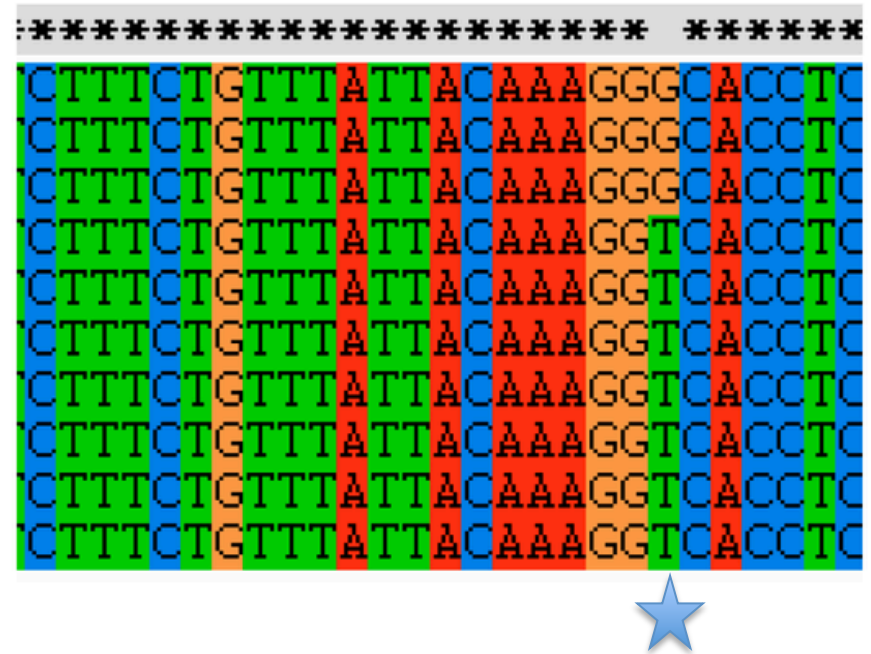
Always the format

Samtools subcommand –flags –moreflags

# SNP CALLING

# SNP Calling

- SNP = single nucleotide polymorphism
- Indel = insertion/deletion
- Examine the alignments of reads and look for differences between the reference and the individual(s) being sequenced





# SNP Calling

- Difficulties:
  - Cloning process (PCR) artifacts
  - Errors in the sequencing reads
  - Incorrect mapping
  - Errors in the reference genome
- Heng Li, developer of BWA, looked at major sources of errors in variant calls\*:
  - erroneous realignment in low-complexity regions
  - the incomplete reference genome with respect to the sample

\* Li 2014 Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics.

# Indel

```
coord      12345678901234      5678901234567890123456
ref        aggtttttataaaac----aattaagtctacagagcaacta
sample     aggtttttataaaacAAATaattaagtctacagagcaacta
read1      aggtttttataaaac****aaAtaa
read2      gggtttttataaaac****aaAtaaTt
read3      ttataaaacAAATaattaagtctaca
read4      CaaaT****aattaagtctacagagcaac
read5      aaT****aattaagtctacagagcaact
read6      T****aattaagtctacagagcaacta
```

Can be difficult to decide where the best alignment actually is

# SNP Calling Software

- Samtool's mpileup -> bcftools
- GATK (Genome Analysis Toolkit)
- FreeBayes



- Take into account the quality value of the individual base and the quality value for the alignment of the read
- Can utilize information about previously called/confirmed SNPs

# Steps

- Preprocessing read alignments\*
  - Picard MarkDuplicates
  - GATKbase quality score recalibration
  - GATK realignment around indels
- Call the SNPs with software of your choice
- Filter the SNPs
  - Goal to remove false positives

May or may not be worth preprocessing: <https://bcbio.wordpress.com/2013/10/21/updated-comparison-of-variant-detection-methods-ensemble-freebayes-and-minimal-bam-preparation-pipelines/>

# Filtering

- Depth
  - How many reads do you need to sample to confidently call a SNP?
  - $> 20X$  = very good
  - $5-20X$  = okay
  - $< 5X$  = missing many heterozygous calls
  - Low depth may be overcome using statistics – see overview of genotype likelihoods here:
  - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3593722/>
- Low quality – use the quality estimates provided by the software
- Number of alleles (monomorphic vs biallelic)
- High coverage – can indicate a duplicated region in the genome
- Highly variable region – can also indicate a duplicated region
- Low complexity regions

# Your mileage may vary

- Different decisions about how to align reads and identify variants can yield very different results

Low concordance of multiple variant-calling pipelines:  
practical implications for exome and genome sequencing

Jason O'Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson,  
W Evan Johnson, Zhi Wei, Kai Wang ✉ and Gholson J Lyon ✉

*Genome Medicine* 2013 5:28 | DOI: 10.1186/gm432 | © O'Rawe et al.; licensee BioMed Central Ltd. 2013

- 5 pipelines
- “SNV concordance between five Illumina pipelines across all 15 exomes was 57.4%, while 0.5 to 5.1% of variants were called as unique to each pipeline. Indel concordance was only 26.8% between three indel-calling pipelines”

# bcftools

- BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF.
- Ack, more formats!!!

# VCF

- Variant Call Format
- Official spec:  
[http://  
samtools.github.io/hts-  
specs/VCFv4.2.pdf](http://samtools.github.io/hts-specs/VCFv4.2.pdf)
- Header lines starting  
with # signs
- Lines with variants  
afterward

#	
#	
#	
#	
#	
Read	
Read	
Read	



# VCF (cont)

- Tab delimited fields
  - Chromosome
  - Location
  - ID (if this is a named variant)
  - Reference sequence
  - Alternate sequence
  - Quality score
  - Filter (true/false – whether or not it passed filtering)
  - Info – lots of additional info such as CIGAR string, depth across different samples, etc.
  - Columns follow for each genotype if available
- BCF is the compressed binary format
  - SAM <-> BAM
  - VCF <-> BCF

# VCF Example

#CHROM	20
POS	14370
ID	rs6054257
REF	G
ALT	A
QUAL	29
FILTER	PASS
INFO	NS=3;DP=14;AF=0.5;DB;H2
FORMAT	GT:GQ:DP:HQ
NA00001	0 0:48:1:51,51
NA00002	1 0:48:8:51,51
NA00003	1/1:43:5:.,.

# bcftools

- Bcftools calls SNPs from mpileup file and manipulates vcf/bcf formatted files
- Call – SNP/indel calling
- Filter – filter the variants by quality
- Merge – merge VCF files together
- Consensus – resequenced an individual and generate the reference sequence for that individual
- Stats - statistics
- Convert – convert between formats

# Overview

Samtools

- Works with SAM/BAM files
- Produces mpileup

Alignment  
Data

Bcftools

- Call SNPs from mpileup
- Works with VCF/BCF files

Variant Data

# IGV

- high-performance visualization tool for interactive exploration of large, integrated genomic datasets
- Free but requires one time registration
- Visualizes lots of data types
  - NGS read alignments
  - Gene annotation
  - Variants
  - Etc.

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

**Integrative  
Genomics  
Viewer**

