

*INF-BIOx121 2017*

---

# Basic file formats

Arvind Sundaram

Sep 05, 2017

Norwegian Sequencing Centre  
OUS, Ullevål, Oslo

---

---

# File formats

---

- ❖ *FASTA* - sequence data
- ❖ *FASTQ* - sequence data including quality (**SRA**)
- ❖ GTF / GFF, GFF3 - Gene feature & transfer format
- ❖ *SAM* / **BAM** - aligned / mapped data (**CRAM**)
- ❖ *VCF* / **BCF** - variant calling

*Text format*

**Binary format**

---

# FASTA

---

- ❖ Represents a sequence - DNA / RNA / protein
- ❖ Two lines:
  - ❖ header starts '>'
  - ❖ Second line contains the sequence

header

```
>NM_001199335.1 Danio rerio toll-like receptor 21 (tlr21), mRNA  
CTAGACGAAGCTCTCACAGAGTGAAAGGCTTTGTAATGTTGGACTCAGTTTATAGGAATT  
ACATTTAAAAGTGGAAGTGAAAAACATGGCACACTCTGCATGCCACAACTGATACTAAA  
GGCCACATTCATCTGTCTCATAAACTTGCCTGCAGCTACAGTTTCAGAAGTTGCATAGA  
GATCCCAGATTCTAATCATACAATCTTTACATGTGTTAAAAGTTATGAACGAGACATAAC  
TGCGATTGTGAGTGATGTATTTCCCACTGCATTAAATCTTACAATCTCTCACA
```

Genomes  
Transcriptomes

# FASTQ

- ❖ Represents a sequence along with quality information
- ❖ Four lines:
  - ❖ header starts '>'
  - ❖ Second line contains the sequence
  - ❖ Quality header starts with '+'
  - ❖ Fourth line contains the quality for each base in the second line

Illumina read Phred 33

Instrument ID	Flowcell ID	tile#	Y axis	index used					
run#	lane#	X axis	read#						
@J00146	:32	:HJF3HBBXX	:8	:1101	:1387	:1121	1	:N:0	:GCGATA
NGTTCACCTTGGAGACCTGCTGCGGTTATGAGTACGACCTGGCGTGAAAACCTATTCCTTCCA									
+									
#AAA<<F<F<7A<-7<JFJJJJJJJFJF<FAJA<FJAAF<<<JFJJ7FFJJJJJJJAJJJJJ<									

# FASTQ quality encoding

[illegible]

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



---

# GFF/GTF

---

- ❖ GFF: General Feature Format
  - ❖ Current format is based on version 3 and called GFF3
  - ❖ format consists of one line per feature
  - ❖ each containing 9 columns of data, plus optional track definition lines.
- ❖ GTF: General Transfer Format is identical to GFF version 2.5.

# GFF3

seqid	source	chromosome coordinate system type	start	end	score	strand	phase	attributes	ID Name Parent Note .... *case-sensitive
##gff-version 3									
ctg123	example	gene	1050	9000	.	+	.	ID=EDEN;Name=EDEN;Note=protein kinase	
ctg123	example	mRNA	1050	9000	.	+	.	ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Index=1	
ctg123	example	five_prime_UTR	1050	1200	.	+	.	Parent=EDEN.1	
ctg123	example	CDS	1201	1500	.	+	0	Parent=EDEN.1	
ctg123	example	CDS	3000	3902	.	+	0	Parent=EDEN.1	
ctg123	example	CDS	5000	5500	.	+	0	Parent=EDEN.1	
ctg123	example	CDS	7000	7608	.	+	0	Parent=EDEN.1	
ctg123	example	three_prime_UTR	7609	9000	.	+	.	Parent=EDEN.1	
ctg123	example	mRNA	1050	9000	.	+	.	ID=EDEN.2;Parent=EDEN;Name=EDEN.2;Index=1	
ctg123	example	five_prime_UTR	1050	1200	.	+	.	Parent=EDEN.2	
ctg123	example	CDS	1201	1500	.	+	0	Parent=EDEN.2	
ctg123	example	CDS	5000	5500	.	+	0	Parent=EDEN.2	
ctg123	example	CDS	7000	7608	.	+	0	Parent=EDEN.2	
ctg123	example	three_prime_UTR	7609	9000	.	+	.	Parent=EDEN.2	

# BED format

- ❖ Browser Extensible Data
- ❖ 3-12 fields; but generally contains 6 or more

chromosome #	chromStart	chromEnd	name	score	strand
chr7	127471196	127472363	Pos1	0	+
Chr7	127472363	127473530	Pos2	0	+
chr7	127473530	127474697	Pos3	0	+
chr7	127474697	127475864	Pos4	0	+
chr7	127475864	127477031	Neg1	0	-
chr7	127477031	127478198	Neg2	0	-

There are far more annotation formats but will stop here!!!



# Chromosome coordinate systems

## ❖ 0-based

ACTGACTG

012345678

To represent the TGAC:

0-based inclusive: 2-5

1-based inclusive: 3-6

1-based exclusive: 3-7

Ensembl: 1-based inclusive

UCSC: 0-based start

1-based end

1-based for display

## ❖ 1-based

ACTGACTG

123456789

SAM, VCF, GFF, GTF: 1-based

BAM, BCF: 0-based

BED: 0-based exclusive

Most tools are aware of these differences

---

# SAM/BAM

---

- ❖ SAM - Sequence Alignment/Map format
- ❖ BAM - same as/similar to SAM but in binary format
  - ❖ Most softwares and tools prefer BAM
- ❖ Header
  - (access using samtools view -H)
- ❖ Alignment records
  - (access using samtools view; use -h to include Header)



# SAM/BAM: header

```
@HD VN:1.0GO:none SO:coordinate ← sort order
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:249250621
...
@SQ SN:chrX LN:155270560
@SQ SN:chrY LN:59373566
...
@RG ID:86-191PL:ILLUMINA LB:IL500 SM:86-191-1 ← Read group with platform,
@RG ID:Bsk010PL:ILLUMINA LB:IL501 SM:Bsk010-1 library and sample information
...
@RG ID:SDH023PL:ILLUMINA LB:IL508 SM:SDH023 ↙ Program analysis history
@PG ID:GATK IndelRealigner VN:2.0-39-gd091f72CL:knownAlleles=[ ]
targetIntervals=tmp.intervals.list LODThresholdForCleaning=5.0
consensusDeterminationModel=USE_READS entropyThreshold=0.15
maxReadsInMemory=150000 maxIsSizeForMovement=3000 maxPositionalMoveAllowed=200
maxConsensuses=30 maxReadsForConsensuses=120 maxReadsForRealignment=20000
noOriginalAlignmentTags=false nWayOut=null generate_nWayOut_md5s=false
check_early=false noPGTag=false keepPGTags=false indelsFileForDebugging=null
statisticsFileForDebugging=null SNPsFileForDebugging=null
@PG ID:bwaPN:bwaVN:0.6.2-r126
```



# SAM/BAM: alignment records

1	2	3	4	5	6	7	8	9
HW-ST605:127:B0568ABXX:2:1201:10933:3739	147	chr1	27675	60	101M	=	27588	-188
TCATTTTATGGCCCCTTCTTCCTATATCTGGTAGCTTTTAAATGATGACCATGTAGATAATCTTTATTGTCCCTCTTTCAGCAGAC								
=7;::;<=?<=BCCEFFEJFCEGGEFFDF?E@E@ADCACB>CCDCBACDCDDAB@@BCADDCBC@BCBB8@ABCCDCBDA@>:/								
RG:Z:86-191 . . . . .								

Col	Field	Type	Regexp/Range	Brief Description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2 <sup>16</sup> -1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2 <sup>31</sup> -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 <sup>8</sup> -1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+	CIGAR string
7	RNEXT	String	\*  =  [!-( )+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 <sup>31</sup> -1]	Position of the mate/next read
9	TLEN	Int	[-2 <sup>31</sup> +1,2 <sup>31</sup> -1]	observed Template LENght
10	SEQ	String	\*  [A-Za-z=.]+	sequence SEQUENCE
11	QUAL	String	[!-~]+	ACSII of Phred-scaled base QUALity+33



---

# SAM/BAM

---

- ❖ To check FLAG info: <https://broadinstitute.github.io/picard/explain-flags.html>
- ❖ BAM/SAM files are produced by alignment/mapping softwares and other tools. You will hear more about these in Tim Hughes's *Algorithms* lecture
- ❖ QC of BAM/SAM files - Picard, Qualimap
- ❖ Manipulating and extracting information from varied file formats - SAMtools, BEDtools, BCFtools

---

# VCF/BCF

---

- ❖ Variant call format
- ❖ BCF - same as/similar to VCF but in binary format
  - ❖ Most softwares and tools prefer BCF or VCF in zip format
- ❖ header and records

header

```
##fileformat=VCFv4.2
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=xxxx,species="Homo sapiens",taxonomy=x>
...
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
...
##FILTER=<ID=q10,Description="Quality below 10">
...
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
...
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
```

# VCF/BCF: records

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

20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50

20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
```

	#CHRO	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLEs
<b>SNP</b>	20	3	.	C	G	.	PASS	DP=10		
<b>Deletion</b>	20	2	.	TC	C	.	PASS	DP=10		
<b>Insertion</b>	20	2	.	TC	TCA	.	PASS	DP=10		
<b>Alleles</b>	20	2	.	TC	TG,T	.	PASS	DP=10		
<b>Alleles</b>	20	2	.	TCG	TG,T,TCA	.	PASS	DP=10		

+ Structural variants

---

# Qualimap

---

- ❖ GUI, command line based
  - ❖ examines sequencing alignment data in SAM/BAM format
- ❖ requires Java, R
- ❖ Uses:
  - ❖ FastQC, Picard, SAMTools, NOISeq, Repitools and JProfiler
  - ❖ BAM QC
  - ❖ RNA-seq QC
  - ❖ Counts QC
  - ❖ Multi-sample BAM QC



---

# SAMtools/BCFtools

---

## ❖ SAMtools

- ❖ manipulating SAM/BAM

## ❖ BCFtools

- ❖ manipulating VCF/BCF

## ❖ Part of HTSlib tool kit

```
-- indexing
  faidx    index/extract FASTA
  index    index alignment

-- editing
  calmd    recalculate MD/NM tags and '=' bases
  fixmate  fix mate information
  reheader replace BAM header
  rmdup    remove PCR duplicates
  targetcut cut fosmid regions (for fosmid pool only)

-- file operations
  bamshuf  shuffle and group alignments by name
  cat      concatenate BAMs
  merge    merge sorted alignments
  mpileup  multi-way pileup
  sort     sort alignment file
  split    splits a file by read group
  bam2fq   converts a BAM to a FASTQ

-- stats
  bedcov   read depth per BED region
  depth    compute the depth
  flagstat simple stats
  idxstats BAM index stats
  phase    phase heterozygotes
  stats    generate stats (former bamcheck)

-- viewing
  flags    explain BAM flags
  tview    text alignment viewer
  view     SAM<->BAM<->CRAM conversion
```

---

# BEDtools

---

- ❖ Awesome tool to compare dataset in BED / GFF / VCF and BAM formats
- ❖ Very well documented and Highly recommended!!
- ❖ Example:

<http://bedtools.readthedocs.io/en/latest/content/tools/intersect.html>



# IGV - Integrative Genomics Viewer

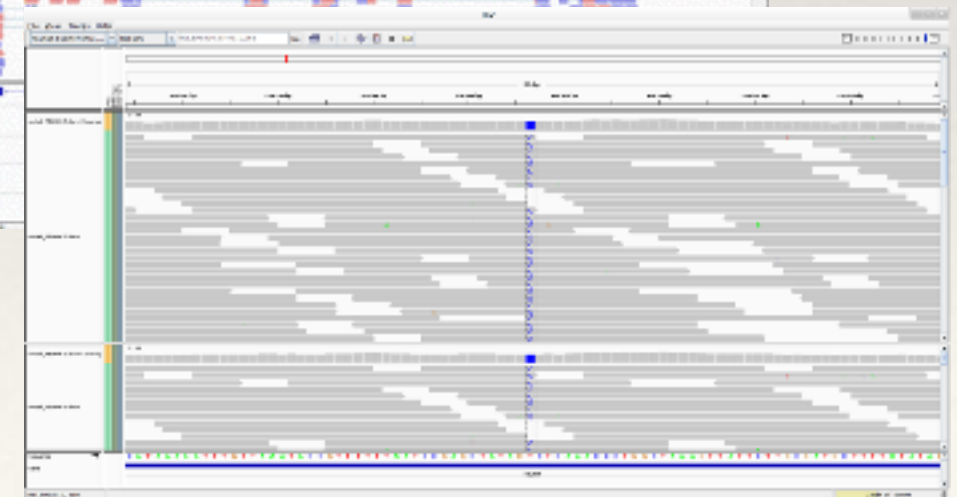
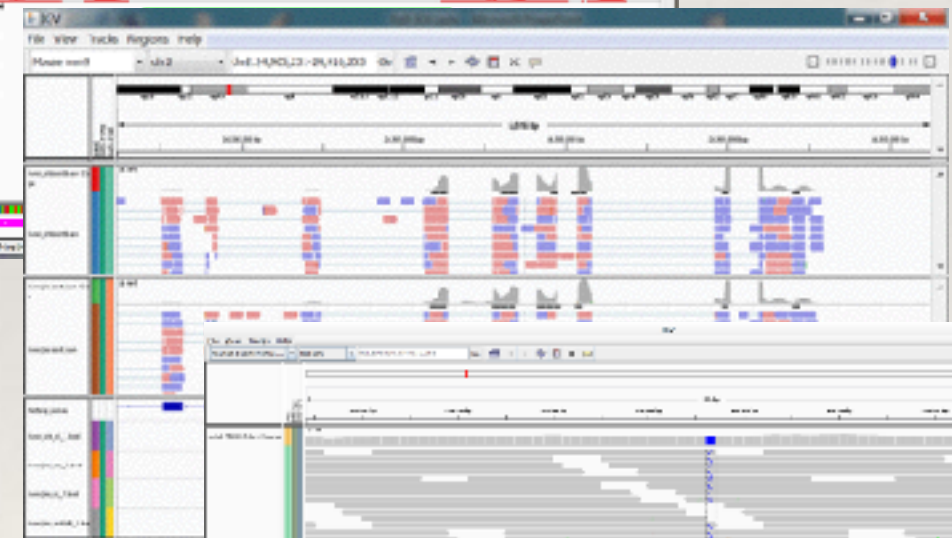
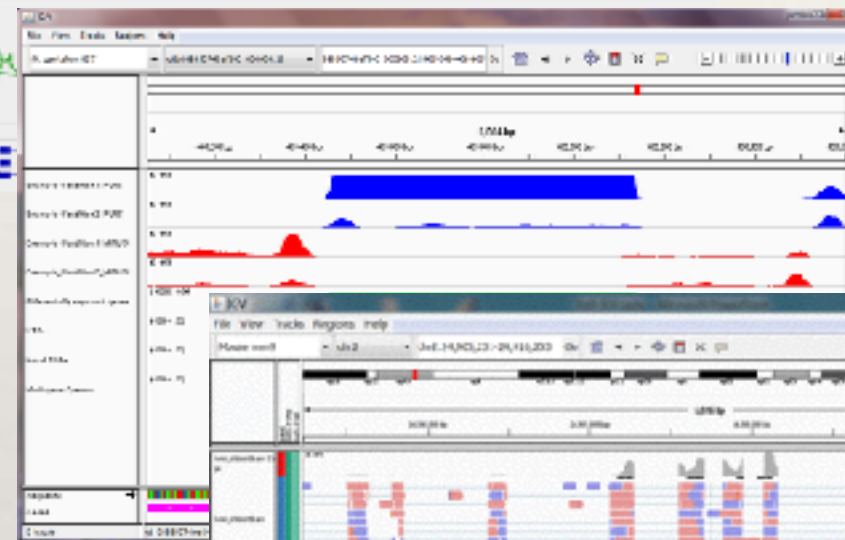
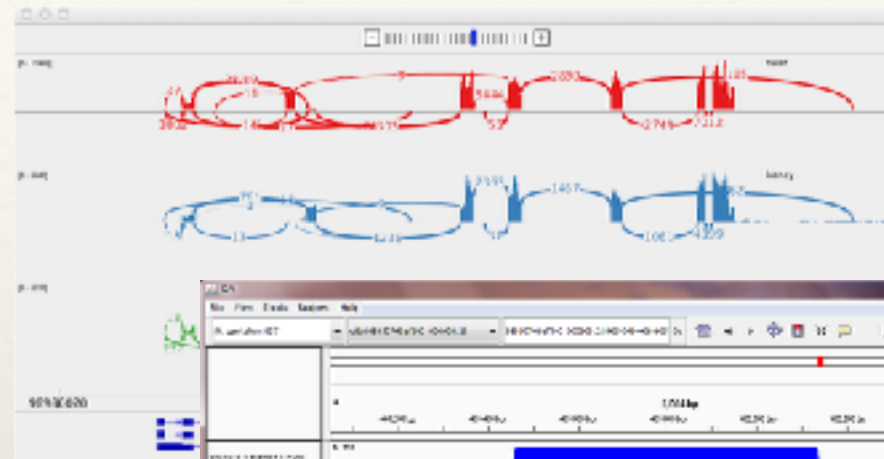
- ❖ View BAM, VCF, GTF files and more

- ❖ Ensembl browser

<http://ensembl.org/>

- ❖ UCSC browser

<http://genome.ucsc.edu/>



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

---

# Further reading

---

- ❖ GFF3, GTF - <http://gmod.org/wiki/GFF3>
- ❖ SAM/BAM format - <http://samtools.github.io/hts-specs/SAMv1.pdf>
- ❖ VCF/BCF format - <https://samtools.github.io/hts-specs/VCFv4.2.pdf>  
- <http://vcftools.sourceforge.net/VCF-poster.pdf>
- ❖ IGV - <http://software.broadinstitute.org/software/igv/>



---

# Databases

---

- ❖ Ensembl - <http://ensembl.org/>
- ❖ UCSC - <http://genome.ucsc.edu/>
- ❖ NCBI - <http://ncbi.nlm.nih.gov/>
  - ❖ RefSeq - <https://www.ncbi.nlm.nih.gov/refseq/>

There are too many databases containing specific information!!!