

Bioinformatics and Genome Analyses Course

Lecture 2 (Nov 13)

Stephan Pabinger @ait.ac.at

http://pabinger.site44.com

Recap



History

Gregor Mendel, Law of Heredity 1865 Johann Miescher, Purification of DNA 1866 1949 Sickle Cell Anemia Mutation 1953 Watson and Crick, Structure of DNA 1970 Recombinant DNA Technology 1977 **DNA** sequencing 1985 In Vitro Amplification of DNA (PCR) 2001 The Human Genome Project

Different sequencing technologies



SAM/BAM files

9PG ID: gatk target	interval cre	ator PN:GenomeAna	lysisTK	PP:sam to	fixed ban	VN:1.2-29	+gBacaf2d	CL: fava \$1	vm args -1a	r GenomeAnalysisTK.
dels file(s)										
gPG ID:bam realign	ment around k	nown indels PN:0	enomeAnalysi	STK P	Pigatk tan	get interval o	reator VN:1.	2-29-g8acaf2d	CL: ja	va \$jvm args -jar (
bam file -targetInters	als Sinterval	s file -known \$kr	own indels f	ile(s) -L	00 0.4 -no	del KNOWNS OM	Y -compress	disable b	em indexing	
BPG ID:ban count o	ovariates PN:	GenomeAnalysisTK	PP:bam n	ealignmen	it around k	nown indels \	N:1.2-29-g8a	caf2d CL	:java \$jvm	args -jar GenomeAna
recal data.csv -known5	ites \$known s	ites file(s) -l :	NFO -L '1;2;	3;4;5;6;7	;8;9;10;11	;12;13;14;15;1	6;17;18;19;2	3;21;22;X;Y;M	T' -cov Rea	dGroupCovariate -co
gPG ID:ban recalit	rate_quality_	scores PN:0	ienomeAnalysi	sTK P	P:bam_coun	t_covariates \	N:1.2-29-g8a	caf2d CL	ijava Sjvm_	args -jar GenomeAna
.csv -I sbam file -o s	recalibrated	bam_file -l INFO	-compress 0	disable	ban index	ing				
8PG ID:ban calcula	te bq PN:	samtools PP:I	am recalibra	te_qualit	y scores	VN:0.1.17	(r973:277)	CL:samtool	s calmd -Er	b \$bam_file \$refere
gPG ID:ban merge	PN:picard	PP:bam_calcu	late_bq '	M:1.53 C	L:java \$jv	m_args -jar Mo	rgeSamFiles.	ar IMPUT=\$ba	m_file(s) 0	UTPUT=\$merged_bam V
(IPG ID:ban mark du	plicates PN:	picard PP:	an merge	/N:1.53 C	L:java \$jvi	m args -jar Me	⊓kDuplicates	.jar INPUT=\$b	am_file OUT	PUT=\$markdup bam fi
										UTPUT=\$merged_bam \
<pre>9CO \$known_indels</pre>	file(s) = ftp	://ftp.1000genome	s.ebi.ac.uk/	rol1/ftp/	technical/	reference/phas	e2_mapping_r	esources/ALL.	wgs.indels_	mills_devine_hg19_1
										verage_vqsr.2010112
<pre>9CO \$known_sites_1</pre>	'lle(s) = ftp:	//ftp.1000genomes	.eb1.ac.uk/w	oll/ftp/t	echnical/n	eference/phase	i2_mapping_re	sources/ALL.w	gs.dbsnp.bu	ild135.snps.sites.v
SRR070823.24480225	163 11		100M		0720 355					CTTATCTGTGCCAAGTGGT
MVK3K9K9X3MNMGNDNNGNGC	KOKONKNIKKINHILL	LLLCKLLFKKLBLHJGI	HED X9:1:	3 X1:1:6	MD: Z: 100	PG:Z:5	FFR070323 AM	1:0 NM:1:0	SM:1:0 MQ	:1:0 XT:A:R BQ:Z:
30000000000000000I										
SRR070823.24480518	1187 11	69464 0			0720 355 MD:Z:188					CTTATCTGTGCCAAGTGGT
DJKBEIJBEJFIJCJGKEGKGO	HTGGHTHDHMFKC	DETAKELIKMEDER	7EB X0:1:	5 X1:1:6	MD: Z: 100	801:Z:S	PREFESZS AM	1:0 NM:1:0	SM:1:8 PQ	:1:0 XT:A:R BQ:Z:
9000000000000000 588070823, 24488225	83 11	60720 0	1008							TAACAGGTTTCTTTTTAAA
SHHOK1313631313GHH5631					0464 -35					:1:0 XT:A:R B0:Z:
2000000000000000	0360606360631	MODIFICATION	CLO WA:T:	A X11110	MD:2:100	1991211	PRIO/0923 PM	11:0 MM:1:0	5M:1:0 MU	THE MINNE BUILD
500000000000000000 500070023.24400510	1107 11	60720 0	100M		0464 -35					TAACAGGTTTCTTTTTAA
=EGJJF>7808AD00>888 <ae< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>:1:0 XT:A:R B0:Z:</td></ae<>										:1:0 XT:A:R B0:Z:
9090909090909090	: @ZBQBABBLPB;	::::egueuo;ec=+	udo yatri	VI:1:0	MD:2:100	M0:2::	PM070023 AT	11:0 MH:1:0	3H:1:0 HU	TIO NINK BUIL
5880705311, 23281260	99 11	61942 0	100H	. 6	2035 192	TECHTOCO	CTTTOATCACAG			TTCAGAAACTATATGACAT
J.MLKHLTJMJHJKNHFGKMM					MD: Z: 100					:1:0 XT:A:R B0:Z:
2000000000000000	MOLDINO TOTAL	OCHTA-HOHITI TOOM	HER MILES	NAILLIA	HD. E. 100	1991210	120070331 PE	11.0 Me1.11.0	301.1.10	TIO NINK DAIL
SRR070531.23281260	147 11	62035 0	1008		1942 -19	2 DOMOGRATIC	CTEAATTGACTG	MALATALOGAGO	TATTCCACAGA	CARTATAAATAAACATATA
MKK1K1KKC1KK3MK1111ML1					MD: Z: 199					11:0 XT:A:R B0:Z:
3000000000000000	TELEBOTTES?	13113111 1110101000			110727200	1991811			311.11.0	THE MILITING BRIDE
SBB070823.17243685	99 11	62388 0	100H		2452 163	CETTOCCA	TISTSTITCTCCT	TATTTCCTSCTS	GATATGACCAC	TOTCCTTCCATTGCATTGT

Recap



Python programs

fasta_extractor.py

sam_examiner.py

```
| Diagram | Prince |
```

Go to https://github.com/spabinger/BioinformaticsAndGenomeAnalyses2018



Coordinate systems

Coordinate system



■ 0 based → 0, 1, 2, ... 9 | 1 based → 1, 2, 3, ... 10

- BED 0 based
- GFF 1 based
- Ensembl uses a one-based coordinate system UCSC use a zero-based coordinate system

	1 based	0 based
Third element	3	2
First ten	1, 10	0, 10
Second ten	11, 20	10, 20
One base long at 10	10,10	9,10
Interval	end – start + 1	end – start
Five elements at 100	100, 104	99, 104

Coordinate system



chr1		Т		Α		С		G		Т		С		Α	
	Ĩ	Ì	Ĩ	[1			ĵ	Ī		1	1		I	Ĩ
1-based		1		2		3		 4		5		6		7	
0-based	0		1		2		3		4		5		6		7

	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

							TTA								
chr1		Т		Α		С	1	G		Ŧ		C		Α	
		1	ľ	1	1	1	8	1		1	1	1	1	1	
1-based		1		2		3		4		5	ls	6		7	
0-based	0		1		2		3		4		5		6		7

	1-based	0-based		
Indicate a deletion	chr1:5-5 T/-	chr1:4-5 T/-		
Indicate an insertion	chr1:3-4 -/TTA	chr1:3-3 -/TTA		

Conversion tool



convert_zero_one_based

- Python CLI
- convert between zero and one based coordinate systems
- Use: convert_zero_one_based --help

https://github.com/griffithlab/convert_zero_one_based



Genome reference

Genome reference



Reference genome is a "consensus" across all chromosomes of DNA pooled from multiple individuals

UCSC and Genome Reference Consortium (GRCh) hg18, hg19, hg38 ←→ GRCh36, GRCh37, GRCh38

Newest version (hg38) release on Dez 24th 2013



Download (e.g.)

- http://hgdownload-test.cse.ucsc.edu/goldenPath/hg38/bigZips/
- ftp://gsapubftp-anonymous@ftp.broadinstitute.org

	HG38 (UCSC)	GRCh38
Prefix	Chr	-
Mitochondrial	chrM	MT
Order	chrM, chr1, chr2,chrX, chrY	1,2,, X, Y, MT

Genome reference

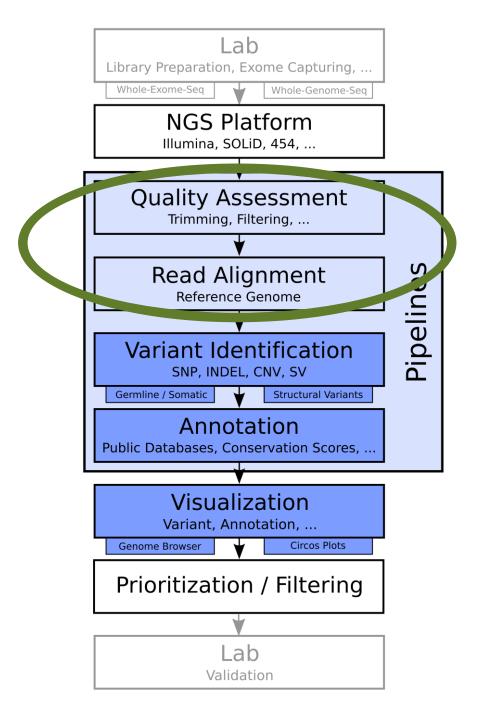


Indexing

- Fai file (created by samtools faidx)
 contig, size, location, bases-per-line and for efficient random
- Dict file (created by Picard CreateSequenceDictionary)
 SAM style header describing the contents of the fasta file
- Different mapping programs

Important

- Choose one reference genome (well sorted, indexed) and stick to it
- Be sure that previous variant calls use same reference otherwise convert coordinates (lift-over)





FASTQ

FASTQ



Storing and defining sequences from next-generation sequencing technologies

QSEQ_ID
Sequence ID GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
Sequence T
(separator) +
Quality scores !''*((((***+))%%%++)(%%%%).1***+*''))**55CCF>>>>>CCCCCCC65

Old format

@HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

New format

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

E40400	No continue to the continue of
EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Cleaning up FASTQ files



Before mapping make sure

- Non genomic sequences are removed (barcodes, ...)
- Adapter sequences are removed
- Clean contaminations (PRINSEQ, DeconSeq)
- Trim Ns
- Trim bad quality reads

Skip cleaning → less read mapping; if not randomly distributed some areas wont get enough coverage

Tools for FASTQ manipulation



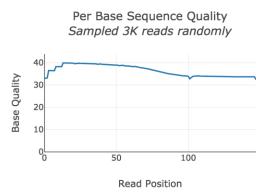
- FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ HTML output
- Fastx toolkit http://hannonlab.cshl.edu/fastx_toolkit/ Lots of tools, charts, trimming, clipping, filtering
- Cutadapt https://code.google.com/p/cutadapt/ Remove adapter sequences
- DeconSeq http://deconseq.sourceforge.net/
 User friendly interface, coverage plots, metagenomics datasets
- PRINSEQ: http://prinseq.sourceforge.net/ HTML output, trimming, filtering, contaminations
- Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic Paird-end trimming
- Atropos https://github.com/jdidion/atropos Multi-threading, paired-end, bisulfite-seq, ...

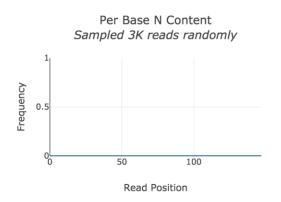
• ...

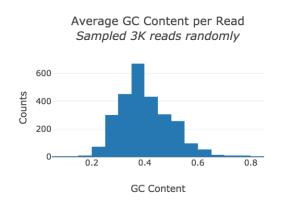
www.fastq.bio

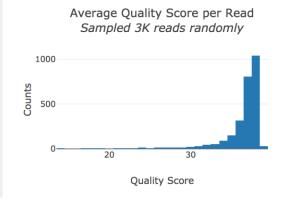


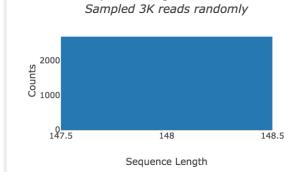












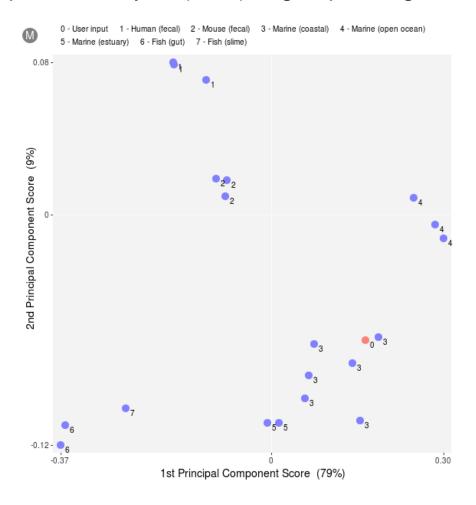
Sequence Length Distribution

Check for contamination



PRINSEQ

- Dinucleotide (e.g., TA, GC, ...) odds ratios
- Principal component analysis (PCA) to group metagenomes



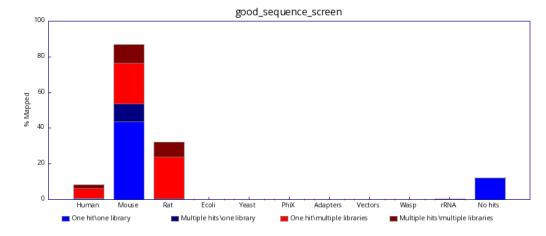
Check for contamination

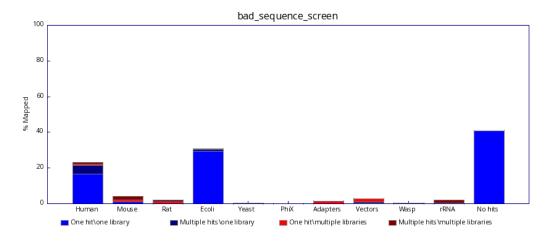


FastQ Screen

Screen against a set of sequence databases:

- genomes of all of the organisms you work on
- PhiX
- Vectors
- ...



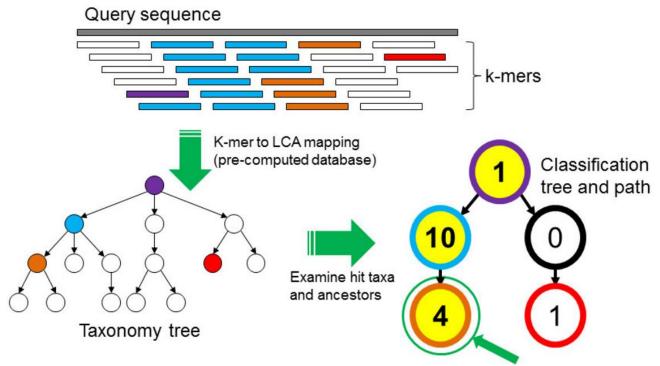


Check for contamination



Kraken (https://ccb.jhu.edu/software/kraken/)

- Assigning taxonomic labels to short DNA sequences
- Detect metagenomics contaminations



Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

Phred quality score



Characterize the quality of DNA sequences

$$q = -10\log_{10}(p)$$

p = error probability for the base

Phred quality score	Probability	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%

Widely used tools



SAMtools

Utilities for manipulating SAM files

GATK

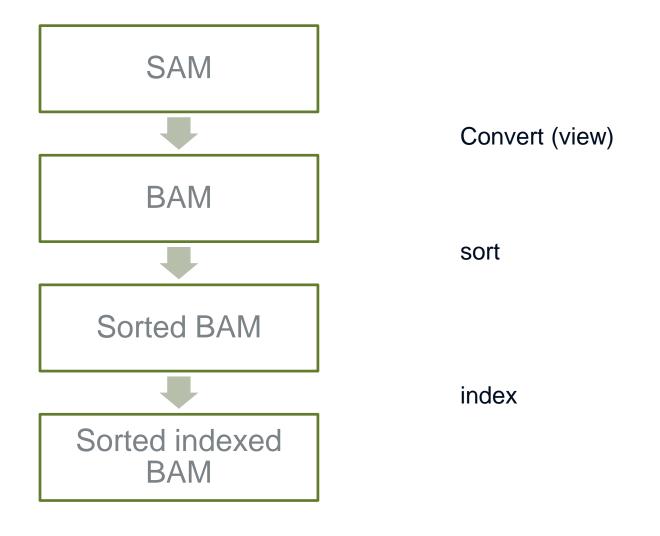
 Genome Analysis Toolkit - wide variety of tools, variant discovery and genotyping, quality assurance

Picard

Utilities for manipulating SAM files

SAM / BAM





Quality check of alignment



Based on SAM/BAM files

Detect biases in the sequencing and/or mapping

Metrics

- Coverage / nucleotide distribution
- Reads mapped outside of a target (e.g., Exome sequencing)
- Number of mapped reads (wrong reference genome?)
- Insert size statistics
- Mapping quality rule of thumb: Anything less than Q20 is not useful data

Tools

- Qualimap 2 http://qualimap.bioinfo.cipf.es/
- bamstats http://bamstats.sourceforge.net/

Read Group Tag (RG)



RG - Meta information

- ID: unique e.g., SRA number (Sequence read archive)
- PL: Sequencing platform
- PU: Platform unit (run name / flowcell-barcode-lane)
- LB: Library name
- PI: Insert size (Predicted mean insert size)
- SM: Sample (Individual)
- CN: Sequencing center

PICARD

java -jar picard/AddOrReplaceReadGroups.jar I=exampleWO_RG.bam O=
exampleW_RG.bam SORT_ORDER=coordinate RGID=superID RGLB=superLib
RGPL=illumina RGSM=superSample

Read Group Tag – why important?



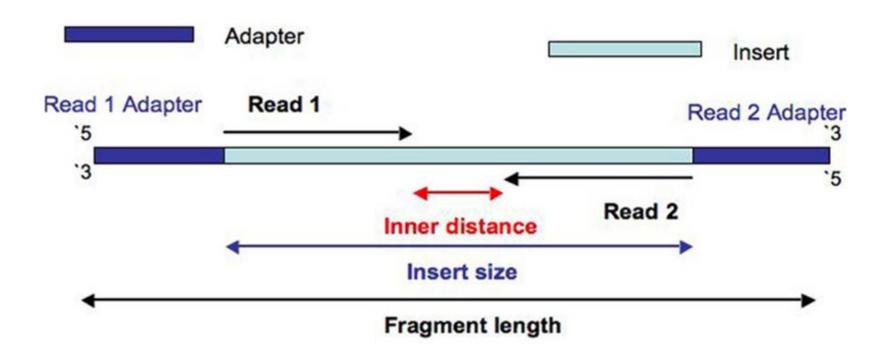
- It refers to a set of reads that were generated from a single run of a sequencing instrument.
- When multiplexing is involved, then each subset of reads originating from a separate library run on that lane will constitute a separate read group.
- To see the read group information for a BAM file, use the following command:

```
samtools view -H sample.bam | grep '@RG'
```

- Check that your FASTQ files have appropriate RQ when performing demultiplexing
- Important to have a correct RG tag → required by bioinformatics analysis tools (GATK, ...)

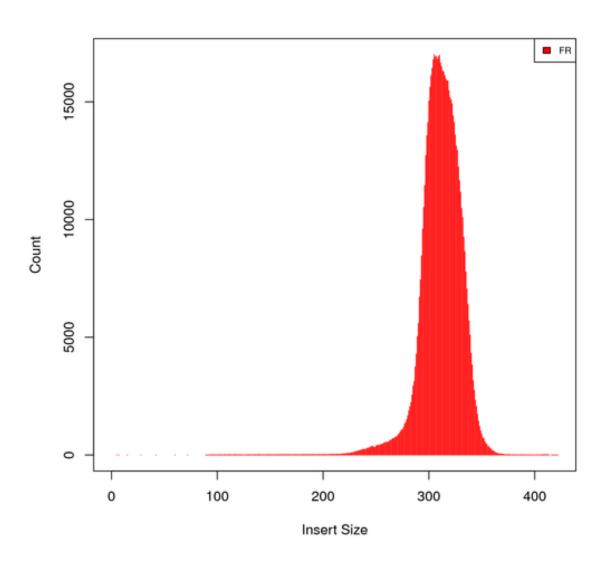
Insert size





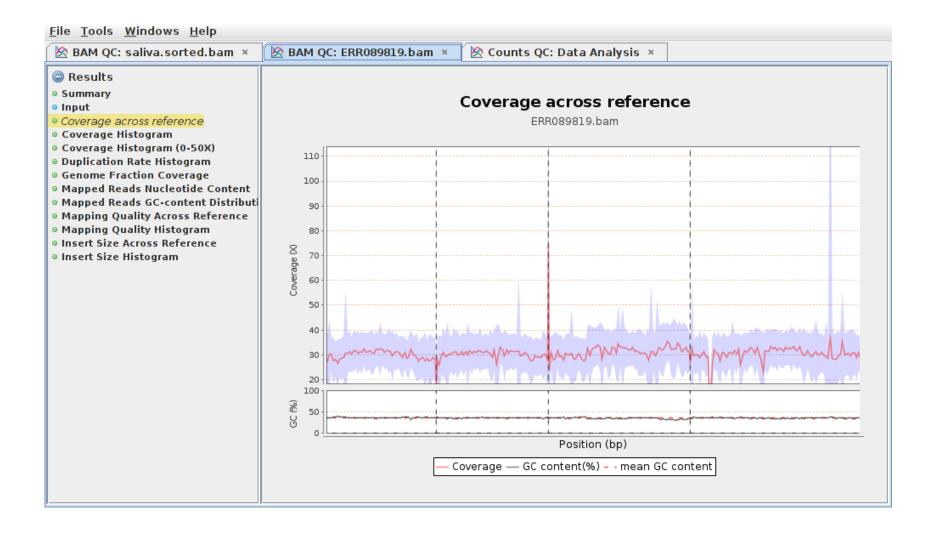
Insert size histogram





Qualimap 2





Read duplicates



Origin

- PCR amplification step in library preparation
 - Get DNA pieces (shatter / enrich DNA)
 - 2. Ligate adapters to both ends of the fragments
 - 3. PCR amplify the fragments with adapters
 - 4. Put fragments on heads or across flowcells
 - 5. Amplify fragments
 - 6. Sequence

Identification

Have the same starting position

Read duplicates

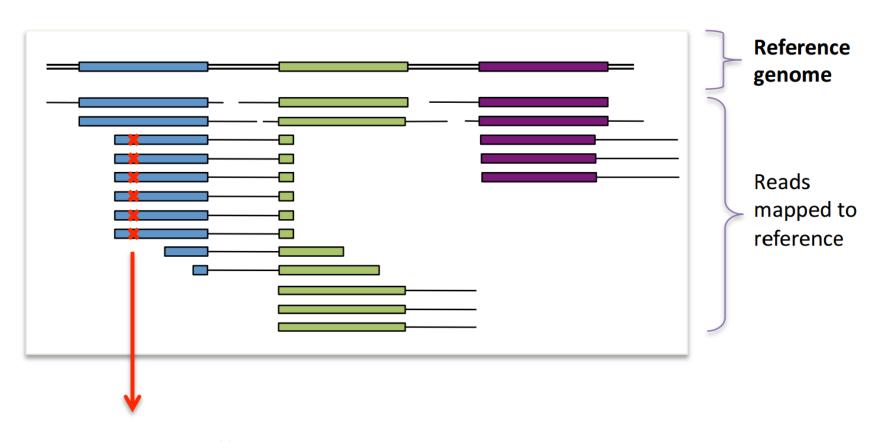


Problems/causes

- More steps during PCR amplification with little input material → more duplicates
 - This may result in duplicate DNA fragments in the final library
- Higher rates (~30%) arise when too little starting material is used
 → more amplification of the library is needed
- May result in false SNP calls (statistical model gets mixed up)

Read duplicates





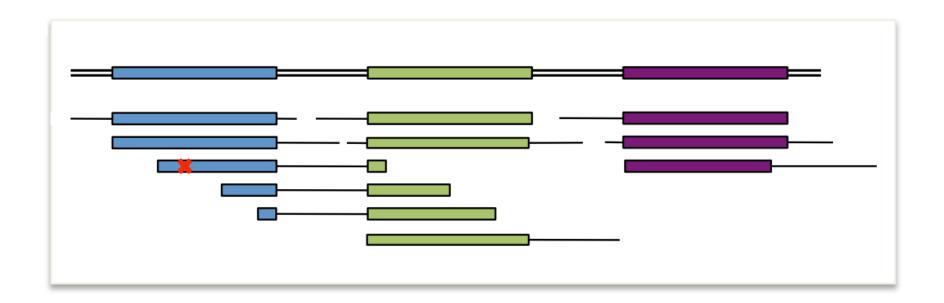
FP variant call (bad)

Read duplicates - removal



- → Identify reads that map to the same location
- → Remove all but one

After marking duplicates



Read duplicates - removal



Attention!

Do not remove for

- Haloplex enrichment (nonrandom fragmentation method)
- PCR based enrichment

You would remove results

Base Quality Score Recalibration



- Various sources of systematic error
 over / under estimated base quality scores
- Quality score assigned to single base in isolation (assigned by the sequencing machines)
- Variant calling algorithms rely heavily on base quality scores

Solution → Correct base quality scores

Base Quality Score Recalibration



Apply machine learning to model these errors empirically and adjust the quality scores accordingly

- First the program builds a model of covariation based on the data
 - Reported quality score
 - Position in the read (cycle)
 - Preceding and current base sequence context (homopolymer, ...)
- and a set of known variants (1000g, dbSNP, large private cohort)
 - discount most of the real genetic variation

First pass: calculate new QS based on the model

Second pass: adjust the base quality scores

→ Visual inspection with before/after plots

Base Quality Score Recalibration



WES

For WES restrict to capture targets
 off-target sites are likely to have higher error rates

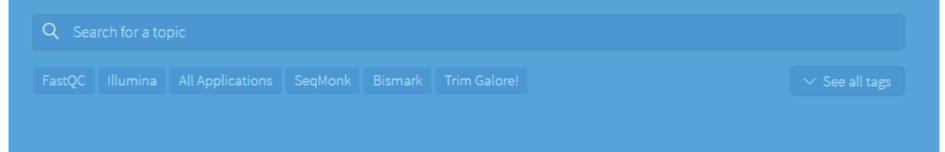
Organism with no known variants?

- Call variants -> apply stringent filter -> use these for recalibration
- Repeat previous steps





Articles about common next-generation sequencing problems





Working with SAM/BAM files



http://samtools.sourceforge.net

View

```
samtools view -h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view -f 0x02 <file.bam> > <only_proper_paired.sam>
```



http://samtools.sourceforge.net

View

```
samtools view -h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view -f 0x02 <file.bam> > <only_proper_paired.sam>
```



http://samtools.sourceforge.net

View

```
samtools view -h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view -f 0x02 <file.bam> > <only_proper_paired.sam>
```

Sort

```
samtools sort -o <sorted.bam> <aln.bam>
```

Index

```
samtools index <sorted.bam>
(only for BAM)
```



Simple stats (reads mapped, reads paired, ...)

samtools flagstat <file.bam>

Stats for each chr/contig - reads mapped and unmapped

samtools idxstats <sorted.bam> (and indexed)

Convert SAM to BAM

samtools view -S -h -b <aln.sam> > <aln.bam>

Converting SAM to a sorted BAM file (without intermediate file)

samtools view -Shb <file.sam> | samtools sort -o
file sorted.bam -

http://davetang.org/wiki/tiki-index.php?page=SAMTools

Other tools



SAMBAMBA - "Multithreaded" SAMtools - https://github.com/lomereiter/sambamba

- view
- sort
- index
- merge
- flagstat
- markdup

elprep - https://github.com/exascience/elprep

- High-performance tool for preparing .sam / .bam / .cram files
- In-memory and multi-threaded application
- Requires lots of memory (WGS ~256GB)
- Replacement for SAMTOOLS & Picard

PICARD & iobio



PICARD

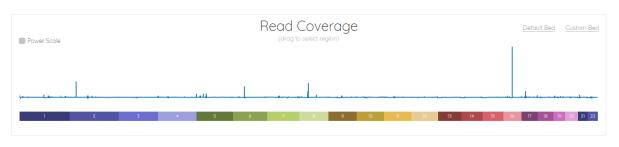
- JAVA based tool (http://picard.sourceforge.net/)
- BuildBamIndex, FastqToSam, MergeSamFiles, ...

bam.iobio.io

- Web-based (http://bam.iobio.io)
- Coverage overview
- Mapping overview

bam.iobio.io







an iobio project









Multiple mapping

Multiple mapping - MAPQ



Bowtie 1

- 255 for uniquely mapped reads (default)
- 0 for multiply mapped reads

Bowtie 2

- MAPQ filter of >=40 to get reads which had only 1 convincing alignment
- Lower filter to allow multi-mapped reads; secondary alignment with varying degrees of difference to the primary

BWA

- Phred scores as MAPQ values
- 0 if read maps to multiple positions



Genetic variations

Genetic variation



A regular diploid human cell contains 46 chromosomes (23 pairs)

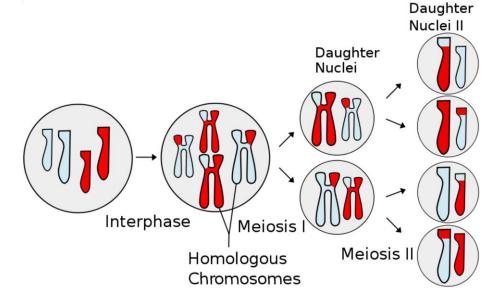
- 22 pairs + sex chromosomes XX(female) XY(male)
- One set of chromosomes inherited from each parent

Germline

 Meiosis → four genetically unique haploid gametes that each contain a unique mixture of the genetic code of the maternal and paternal chromosomes of the cell

Somatic mutation

- Not inherited from parent
- Acquired from spontaneous mutations during DNA replication



Types of genetic variations



SNV / SNP

- A single nucleotide A, T, C or G in the genome differs between members of a population or chromosome pairs
- Originally defined as occurring at least in one individual of the population (these definitions may shift in time)
- SNV (single nucleotide variant) if observed very rarely
- SNP, SNV → may fall within
 - coding sequences of genes
 - non-coding regions of genes
 - intergenic regions

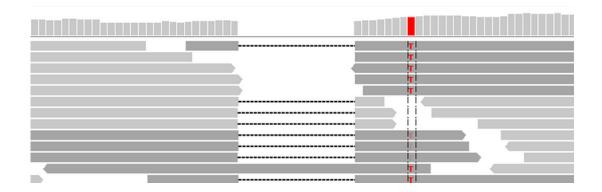
Types of genetic variations



INDEL

- Insertion / deletion of bases
- Coding regions of the genome produce a frameshift mutation (unless multiple of 3)
- There are approximately 190-280 frameshifting INDELs in each person.

 A map of human genome variation from population-scale sequencing. Nature 467 (7319)



Types of genetic variations



Structural variations (SV)

- Variation in structure of an organism's chromosome
- InsertionsDeletions

CNV

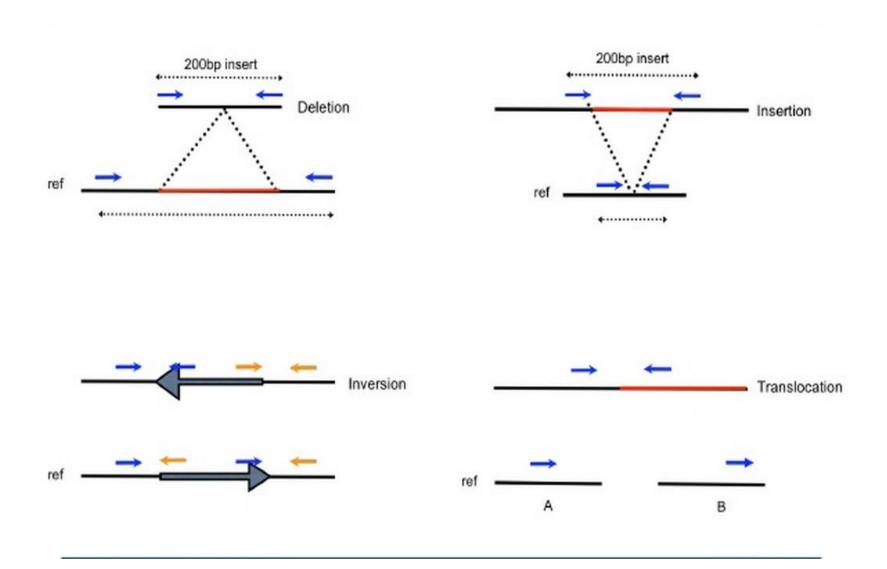
Inversions

Translocations

12.11.2018

Types of structural genetic variations

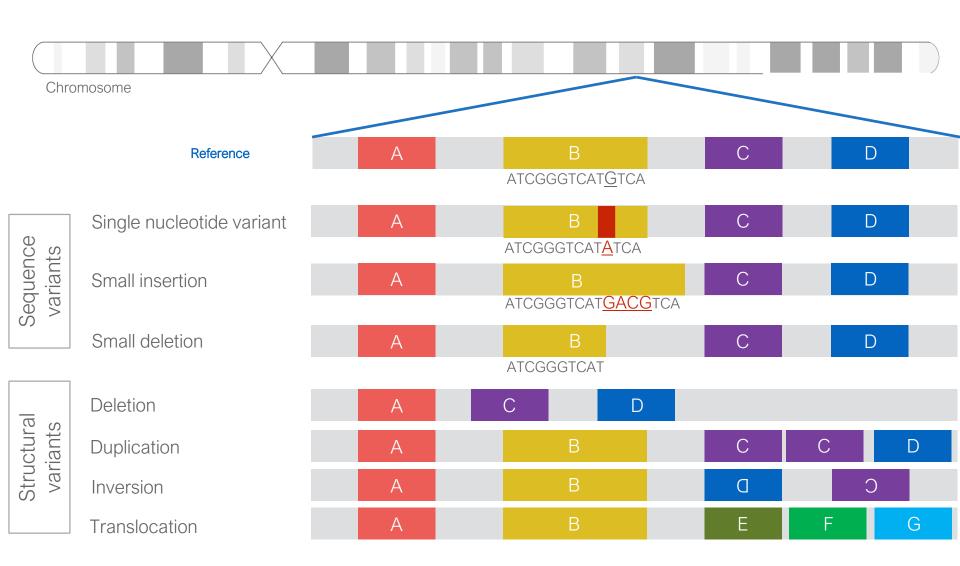




12.11.2018

Summary DNA Variants





Credit: Elizabeth Ruzzo, PhD, CHGV

Human genome



- Typical genome differs from the reference human genome at
 4.1 million to 5.0 million sites.
- ~99.9% of variants consist of SNVs and INDELs
- Structural variants affect more bases
 - 2,100 to 2,500 structural variants
 (~1,000 large deletions, ~160 copy-number variants, ~1100 insertions)
 - Affecting ~20 million bases of sequence
- African ancestry populations harbor the greatest numbers of variant sites (as predicted by the out-of-Africa model of human origins)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/
The 1000 genome project; A global reference for human genetic variation; Nature 2015

What kind of data do you have?



Genotyping

- Single samples
- Family → finding causing mutation of rare disease
- Time series

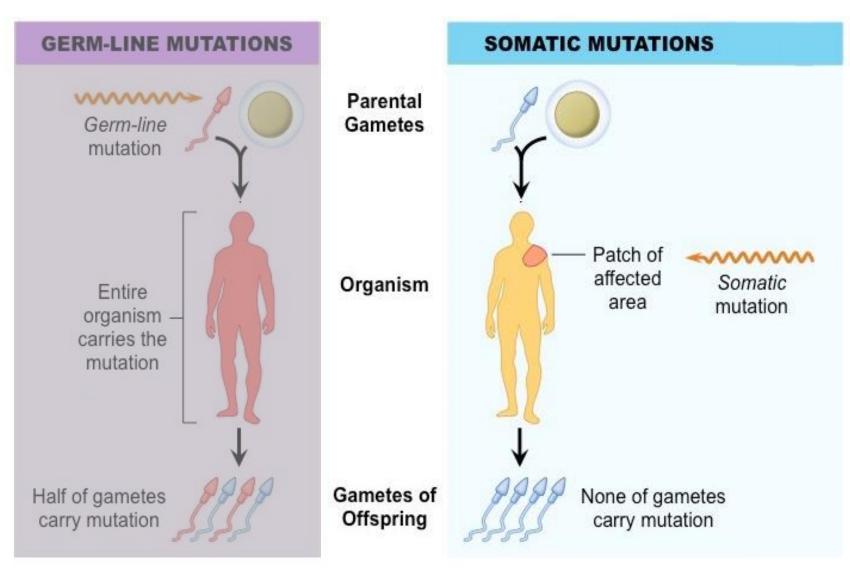
Somatic mutations

- Primary tumor tissue
- Blood samples

→ Different callers / strategies for respective samples

Somatic mutations





Somatic mutations



- Change in genetic structure **not** inherited from a parent
- Constantly happening in living organisms
- Can be used for detection of disease
- Detected by comparison to "normal" cells

Cancer

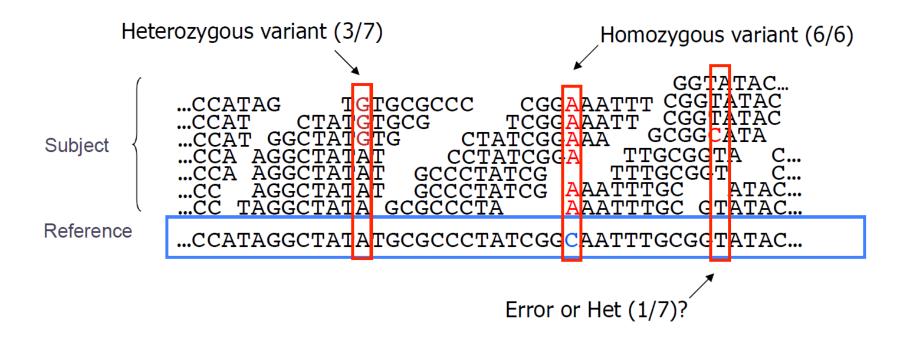
Used to characterize tumor cells



Variant Calling

Genotyping theory





- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!
- Sequencing instruments make mistakes
 - Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

What are variant calls?



Find differences to a reference (hg19, GRCh38)

Naive variant calling

- Check all the reads that cover base chr11:1234567
- Add up the bases at chr11:1234567
- e.g. 15 A's, 4 G's
- Is this an A/G heterozygous site or four sequencing errors?

Actual variant callers

- Estimate likelihood of a variant site vs a sequencing error
- Sequencing error rate
- Quality scores

(Other) reasons for a mismatch



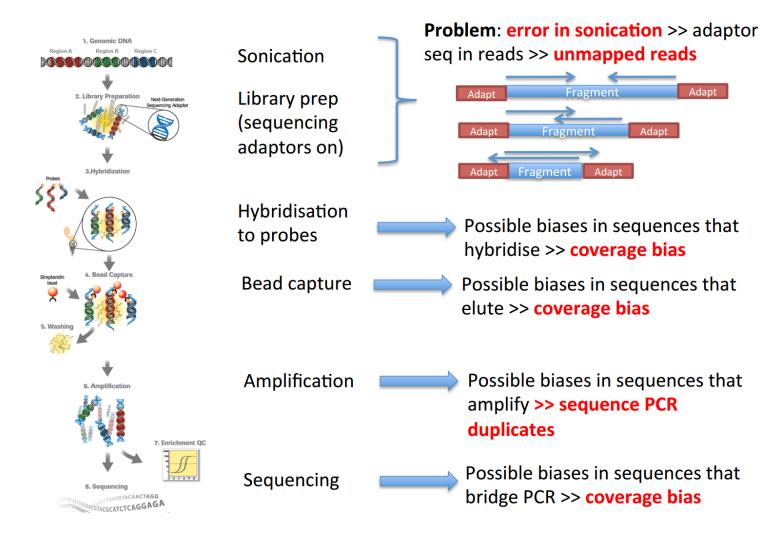
True SNP

OR

- Error generated in library preparation
- Base calling error
 - May be reduced by improved base calling methods, but cannot be eliminated
- Misalignment (mapping error)
 - Local realignment to improve mapping
- Error in reference genome sequence

Difficulties



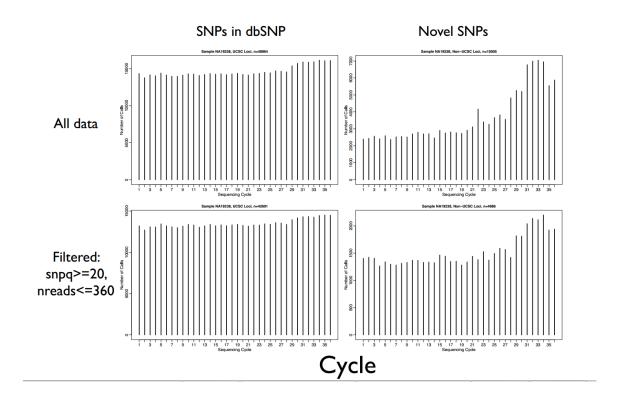


https://wiki.uio.no/projects/clsi/images/8/80/VariantCallingCourse_sept2013_Part1.pdf

Difficulties



Base calling gets more difficult the longer the read gets
 I000 Genomes Data



http://www.biostat.jhsph.edu/~khansen/LecIntro1.pdf

More difficulties



- High depth low quality regions → likely due to copy number or other larger structural events
- Repetitive regions → artefactual variants
- Regions of low complexity (~ 2% of genome)
 polypurine (AG), AT-rich regions, simple tandem repeats

Linderman et al. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Medical Genomics 2014, 7:20

Heng Li. Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. http://arxiv.org/pdf/1404.0929v1.pdf



Practicals – Day 2