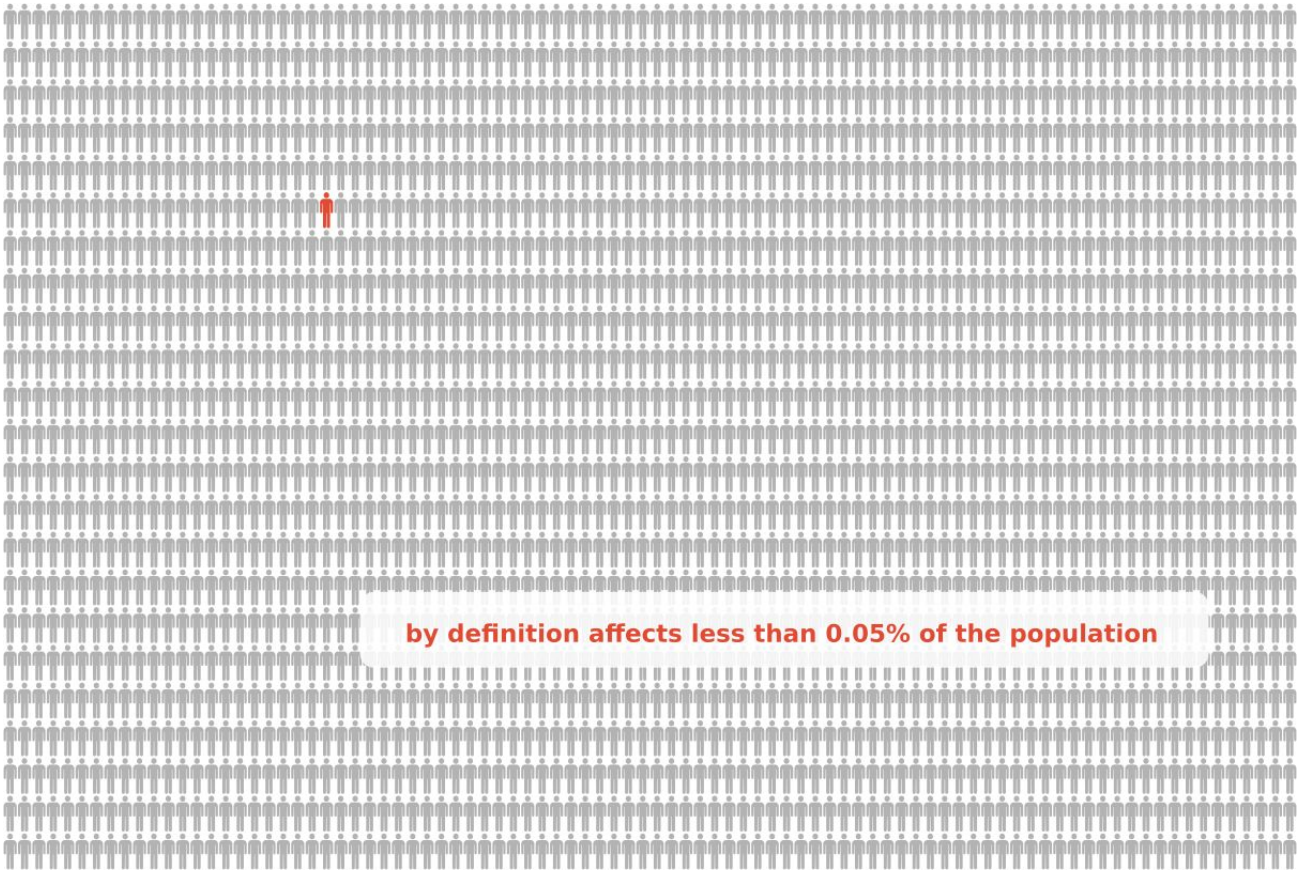
Bioinformatics and Genomics

Merriam-Webster

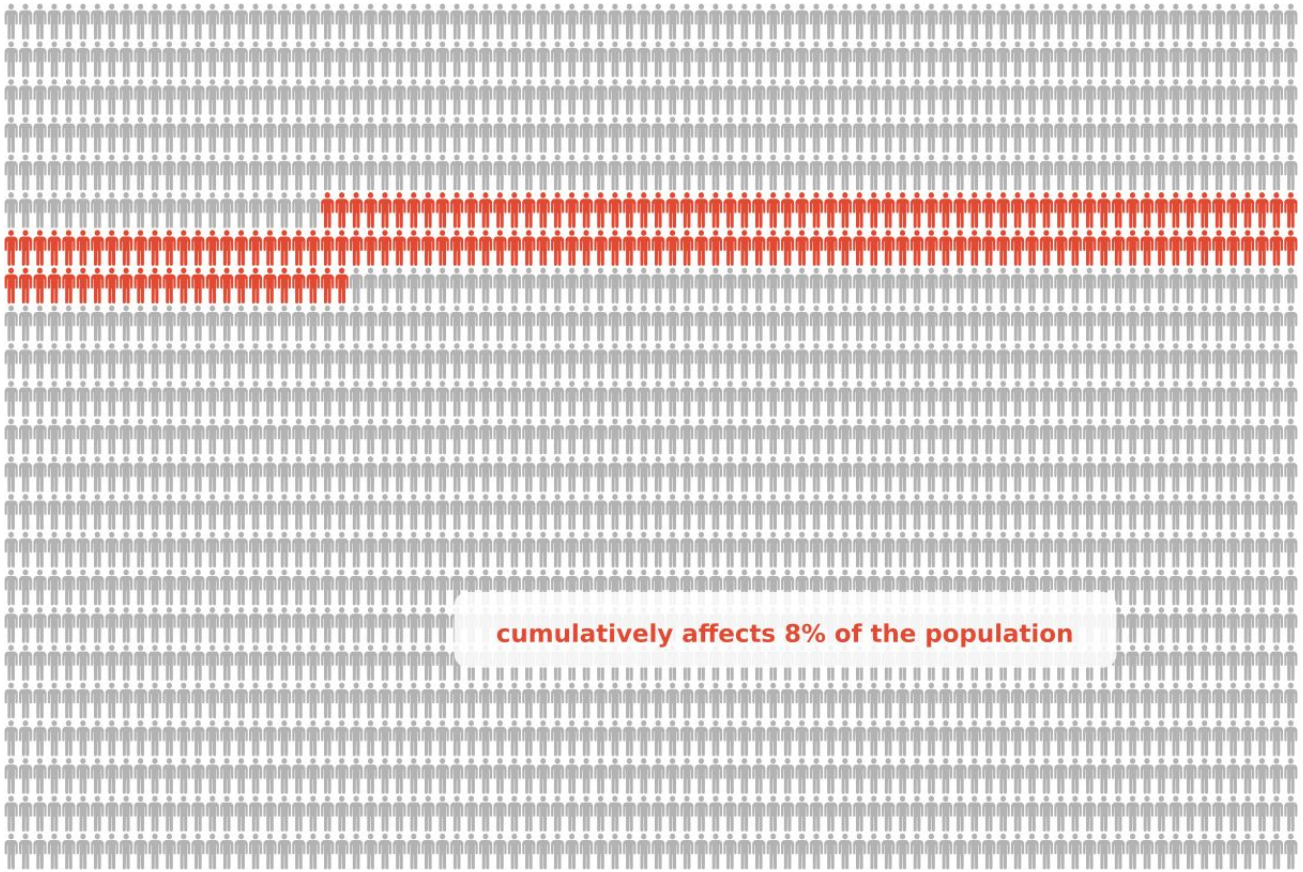
[The science of] *collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics*

Genomics = the study of the structure, function and evolution of genomes

Rare Diseases



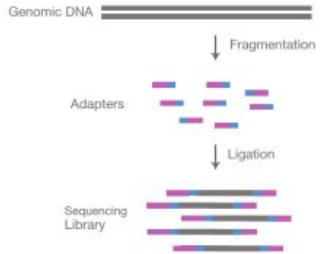
Rare Diseases



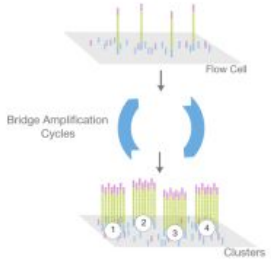
HTS data formats and Quality control

Illumina sequencing

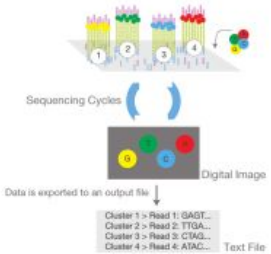
1. Library preparation



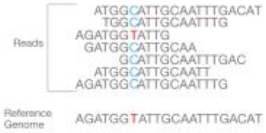
2. Cluster amplification



3. Sequencing

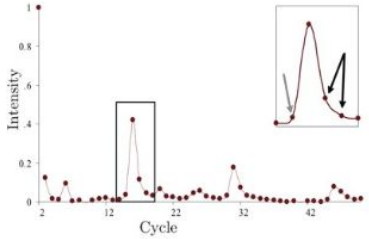


4. Alignment and data analysis

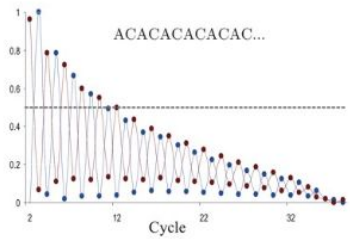


Base calling errors

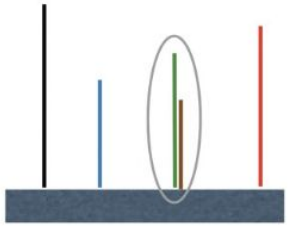
Phasing noise ϕ



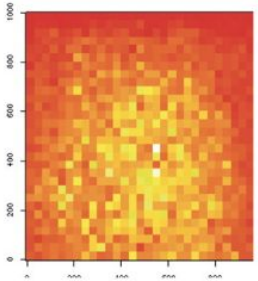
Signal Decay δ



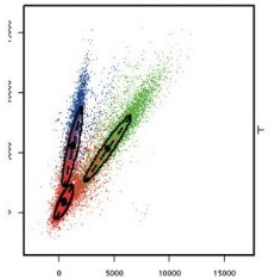
Mixed Cluster μ



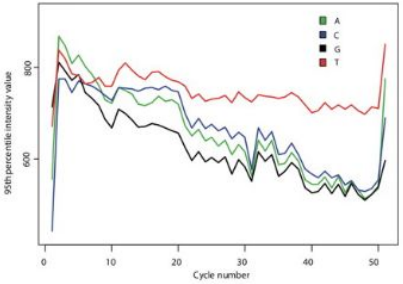
Boundary effects ω



Cross-talk Σ



T fluophore accumulation τ



Base-calling for next-generation sequencing platforms, doi: 10.1093/bib/bbq077

Quality = phred-scaled probability of an error

Quality	Probability of error	Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

$$Q = -10 \log_{10} P \qquad \ldots \qquad P = 10^{-Q/10}$$

FASTQ

Read 1

```
@ERR007731.739 IL16_2979:6:1:9:1684/1      ← Read name
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG...  ← Sequence
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBBABAAAABBBBB=@>BB...  ← Base qualities
```

Read 2

```
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC...
+
BBABB/ABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A...
```

- Simple format for raw unaligned sequencing reads
- Paired-end sequencing: two FASTQ files or one interleaved file

- Quality encoded in ASCII characters with decimal codes 33-126
  - ASCII code of "A" is 65, the corresponding quality is Q$= 65 - 33 = 32$

Base quality encoded as character

```
     ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J
```

Numeric ASCII value

```
     33 . . . . . . . . . . . . . 47 . . . . . . . . . . . . . . . . . . 65 . . . . . . . .
```

Base quality value

```
      0 . . . . . . . . . . . . . 14 . . . . . . . . . . . . . . . . . . 32 . . . . . . . .
```

(65-33 = 32)

  - Beware: multiple quality scores were in use!
    - Sanger, Solexa, Illumina 1.3+
    - See https://en.wikipedia.org/wiki/FASTQ_format for details

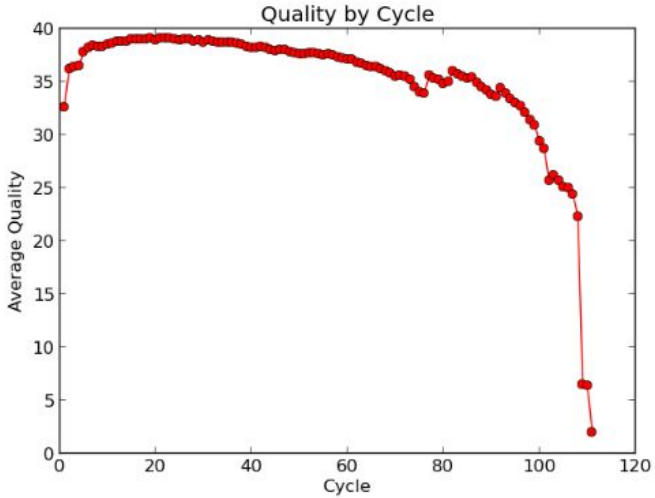
  - `perl -e 'printf "%d\n",ord("A")-33;'`

Base Quality

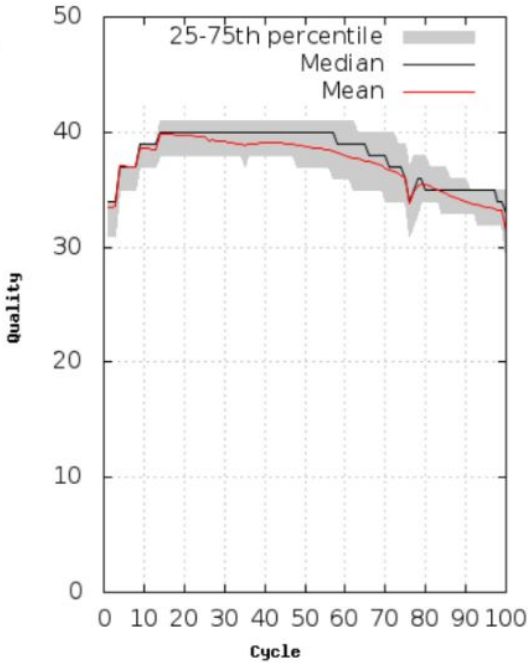
Sequencing by synthesis: dephasing

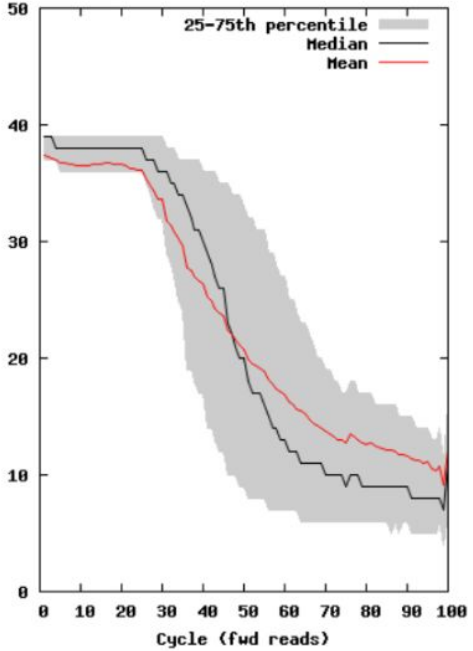
- growing sequences in a cluster gradually desynchronize
- error rate increases with read length

Calculate the average quality at each position across all reads



Base Quality





Read coverage

Read coverage / depth

- is every genomic position "covered" to a sufficient depth?
- average depth: number-of-reads / target-size
  - the whole human genome .. target-size = 3Gb
  - the exomes .. target-size = 50Mb

Exomes

- be careful to distinguish between the total sequencing yield and on-target bases

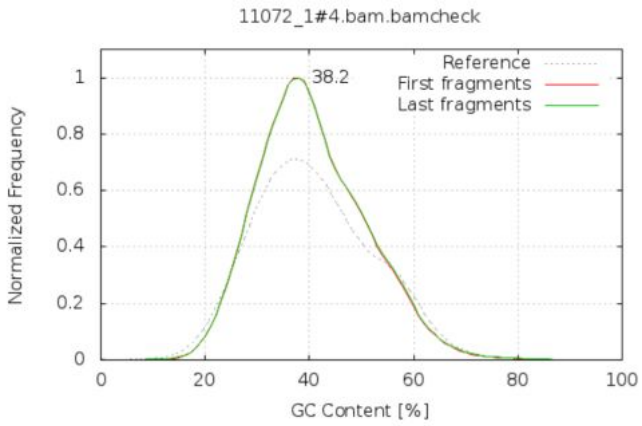
Useful coverage

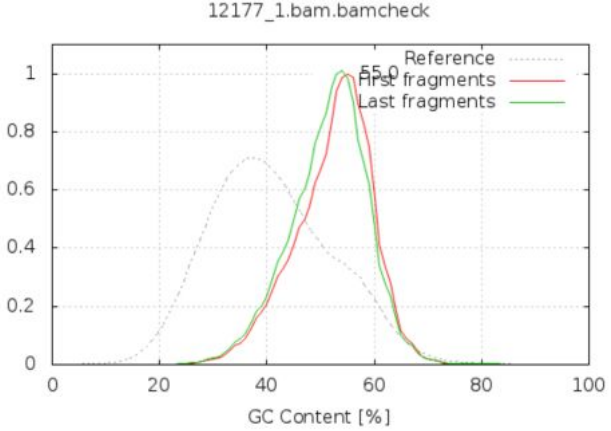
- 15x ok for common germline variants
- 30x ok for most things
- 100-200x for low VAF variants in tumors

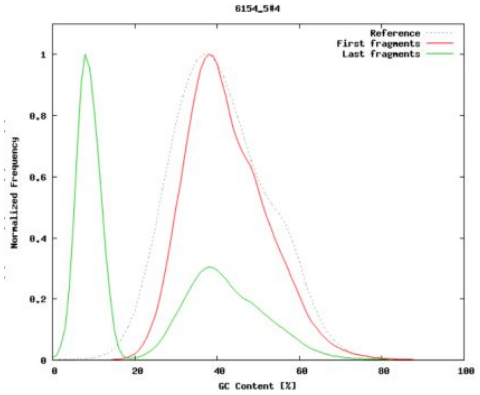
GC bias

GC- and AT-rich regions are more difficult to amplify

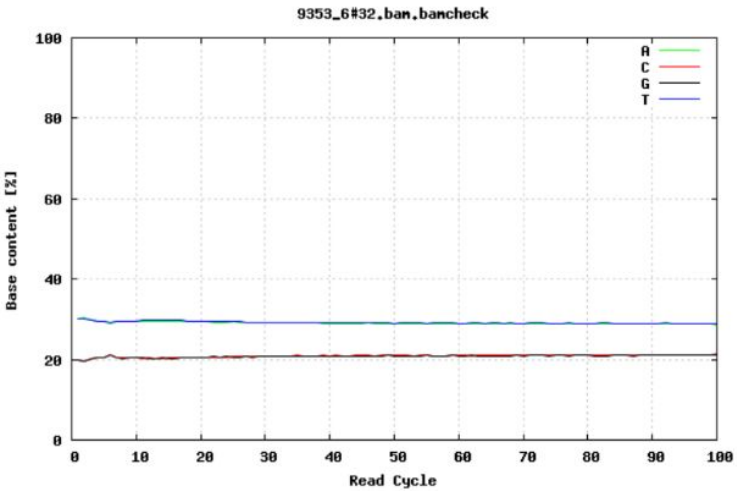
- compare the GC content against the expected distribution (reference sequence)

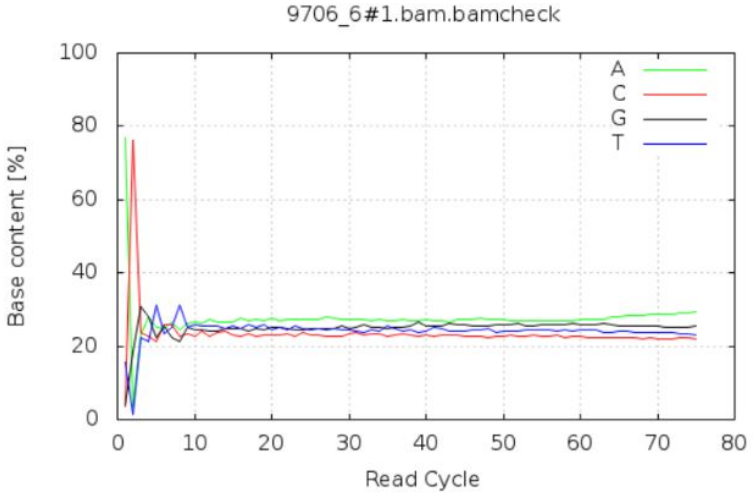






GC content by cycle

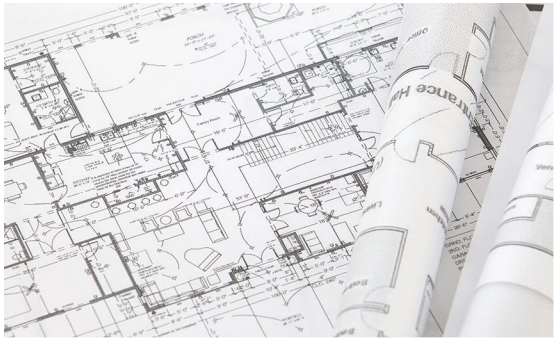




DNA is a computational problem

De novo assembly is very difficult

It is much easier to analyze known genomes



We need a reliable map...

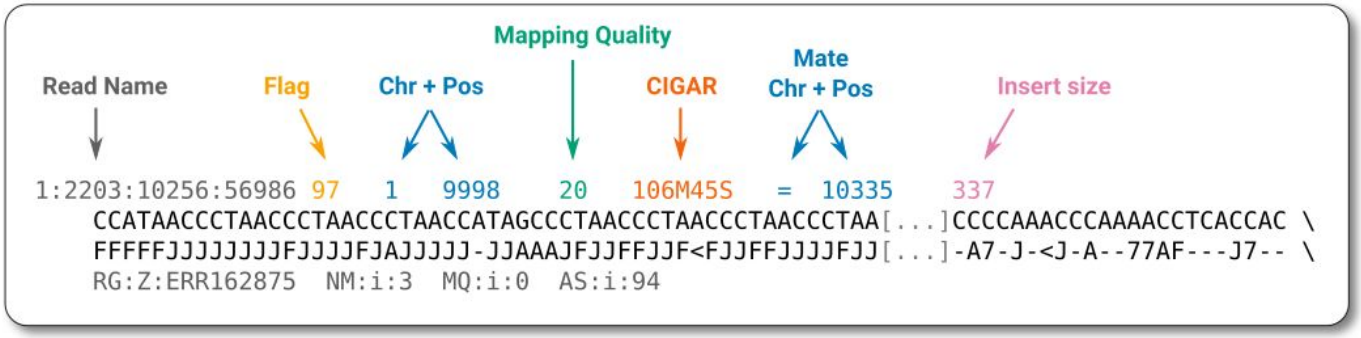
Reference Genome

```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TATTCAAAAAATTGAGAATTTCTGACCACTTAACAAACCCACAGAAAATCCACCCGAGTG
CACTGAGCACGCCAGAAATCAGGTGGCCTCAAAGAGCTGCTCCCACCTGAAGGAGACGCG
CTGCTGCTGCTGTCGTCCTGCCTGGCGCCTTGGCCTACAGGGGCCGCGGTTGAGGGTGGG
AGTGGGGGTGCACTGGCCAGCACCTCAGGAGCTGGGGGTGGTGGTGGGGGCGGTGGGGGT
GGTGTTAGTACCCCATCTTGTAGGTCTGAAACACAAAGTGTGGGGTGTCTAGGGAAGAAG
>2
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
AAAAGCATTTATGCTACAAATTACTATGGTAATTATGCTACAAATTTATGGTACCATAAA
TTACCATAGTAATTTGTAGCATAAATTTGTACTATGGTACAAATTACATGGGAGAGTGAA
GGTGGGTTAAAACATTCATATTAAAGAACTTCCACTCAGATTGCAAGAAAAGAGAGAGGA
ATGGAGATGGTAGCACAAGTCCCTACAATAAAAGTAGATGTTTTGAGATCAGTTCTATTT
```

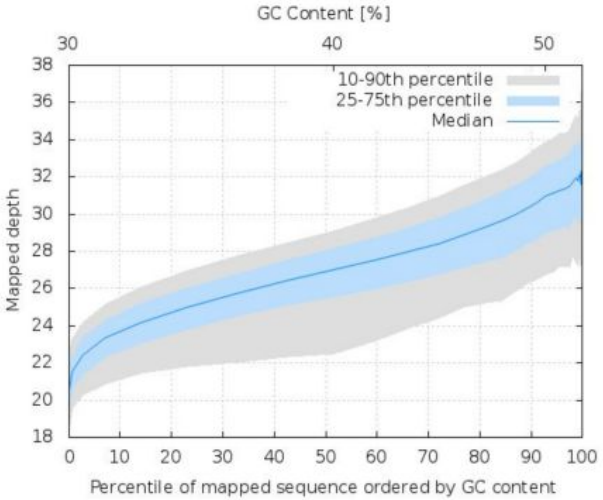
2003	NCBI Build 34	hg16
2004	NCBI Build 35	hg17
2006	NCBI Build 36.1	hg18
2009	GRCh37	hg19
2013	GRCh38	hg38

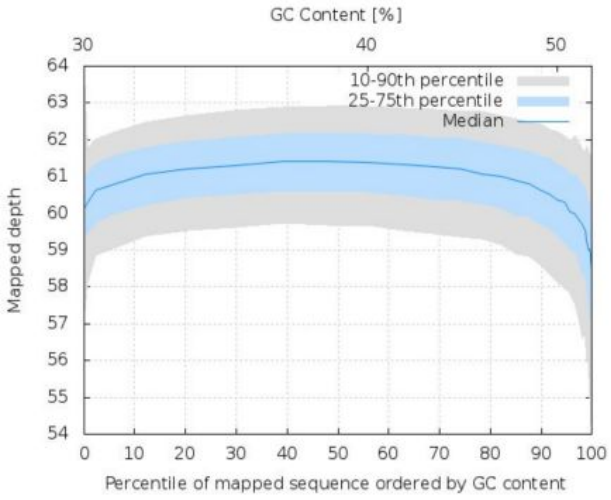
SAM / BAM: Sequence Alignment/Map format

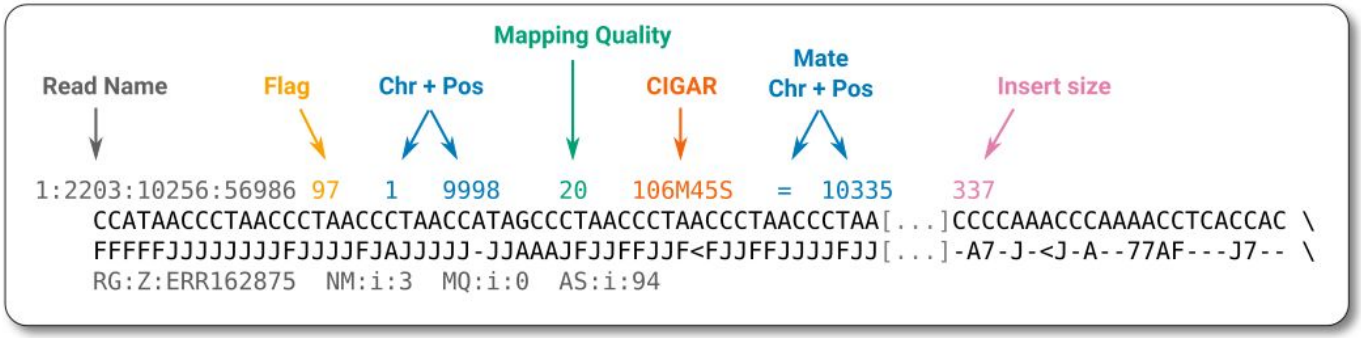
```
$ samtools view -h file.bam | less
@HD VN:1.0  GO:none SO:coordinate
@SQ SN:1    LN:249250621    UR:hs37d5.fa.gz AS:NCBI37   M5:1b22b98cdeb4a9304cb5d48026a85128 SP:Human
@SQ SN:2    LN:243199373    UR:hs37d5.fa.gz AS:NCBI37   M5:a0d9851da00400dec1098a9255ac712e SP:Human
@RG ID:1    PL:ILLUMINA PU:13350_1  LB:13350_1  SM:13350_1  CN:SC
@PG ID:bwa  PN:bwa  VN:0.7.10-r806    CL:bwa mem hs37d5.fa.gz 13350_1_1.fq 13350_1_1.fq
1:2203:10256:56986  97  1   9998    20   106M45S =  10335    0 \
    CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAACCCT[...]CAAACCCACCCCCAAACCCAAAACCTCACCAC \
    FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJJJFF[...]<---F-----A7-J-<J-A--77AF---J7-- \
    MD:Z:1G24C2A76  PG:Z:MarkDuplicates RG:Z:1  NM:i:3  MQ:i:0  AS:i:94 XS:i:94
```


Mapping Quality
Read Name
Flag
Chr + Pos
CIGAR
Mate
Chr + Pos
Insert size
1:2203:10256:56986 97 1 9998 20 106M45S = 10335 337
CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAA[...]CCCCAAACCCAAAACCTCACCAC \
FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJ[...]-A7-J-<J-A--77AF---J7-- \
RG:Z:ERR162875 NM:i:3 MQ:i:0 AS:i:94

GC content vs depth







Flag

Hex	Dec	Flag	Description
0x1	1	PAIRED	paired-end (or multiple-segment) sequencing technology
0x2	2	PROPER_PAIR	each segment properly aligned according to the aligner
0x4	4	UNMAP	segment unmapped
0x8	8	MUNMAP	next segment in the template unmapped
0x10	16	REVERSE	SEQ is reverse complemented
0x20	32	MREVERSE	SEQ of the next segment in the template is reversed
0x40	64	READ1	the first segment in the template
0x80	128	READ2	the last segment in the template
0x100	256	SECONDARY	secondary alignment
0x200	512	QCFAIL	not passing quality controls
0x400	1024	DUP	PCR or optical duplicate
0x800	2048	SUPPLEMENTARY	supplementary alignment

Bit operations made easy

- samtools flags
  0xa3 163 PAIRED,PROPER_PAIR,MREVERSE,READ2
- python
  0x1 | 0x2 | 0x20 | 0x80 .. 163
  bin(163) .. 10100011

PCR duplicates

Experiments start with small amounts of DNA

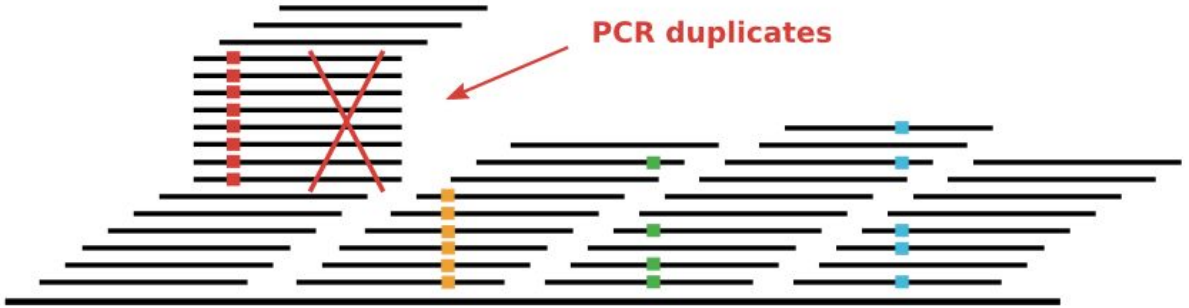
- a PCR amplification step is necessary for Illumina sequencing: one molecule => many identical molecules

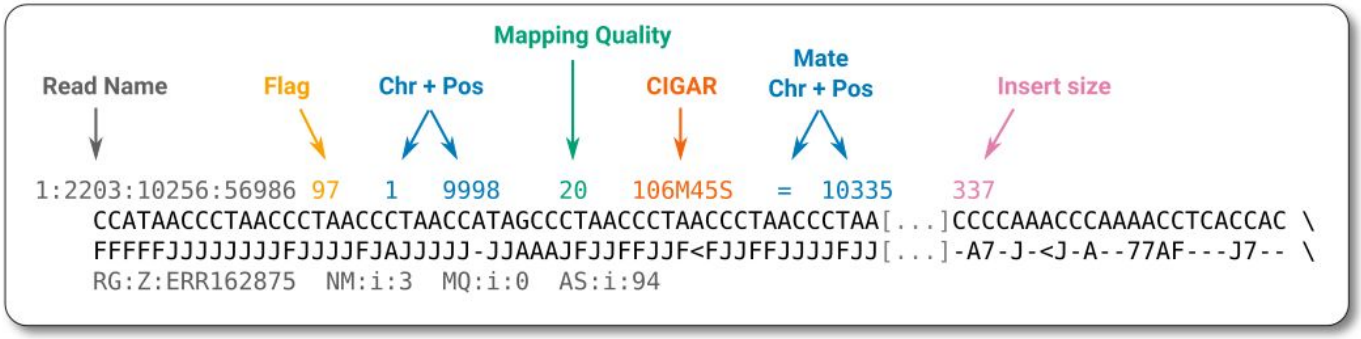
Problem:

- additional PCR-copy molecules are not informative

Solution:

- infer and mark PCR-dupliates, discount in later analysis
  - mark if reads and their mates start at the same position
- use `picard MarkDuplicates` or `samtools markdup`
- typical dup rates: Exomes ~ 15-20%, Genomes < 5%



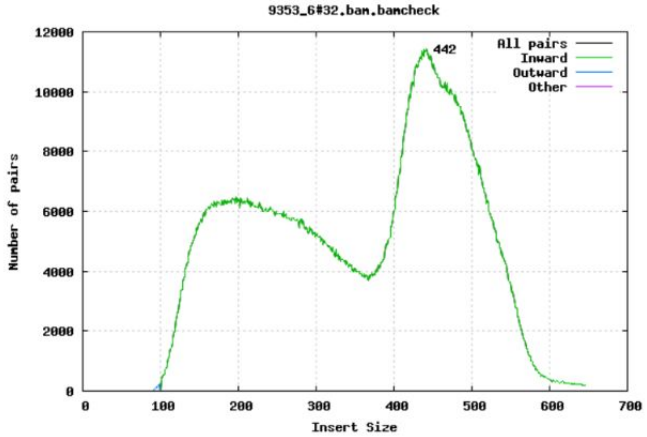


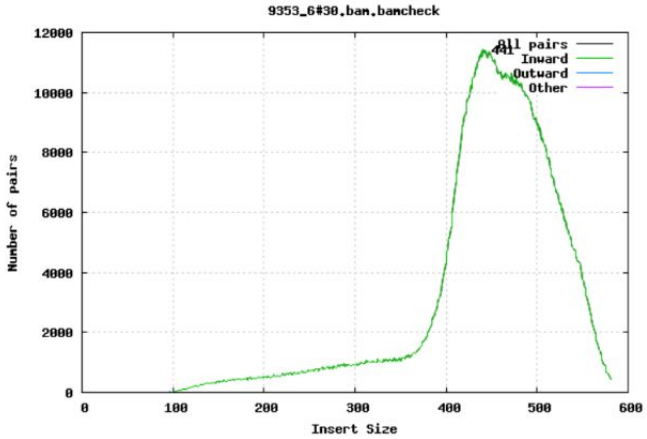
Insert size

length of the DNA fragment sequenced from both ends by paired-end sequencing:

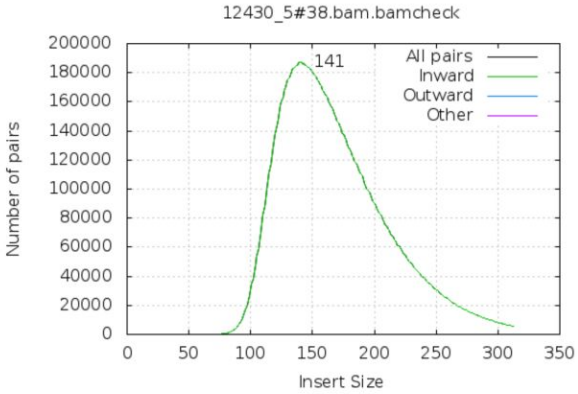
(How paired end sequencing works: https://www.youtube.com/watch?v=0vqajoP08Jg)

Insert Size

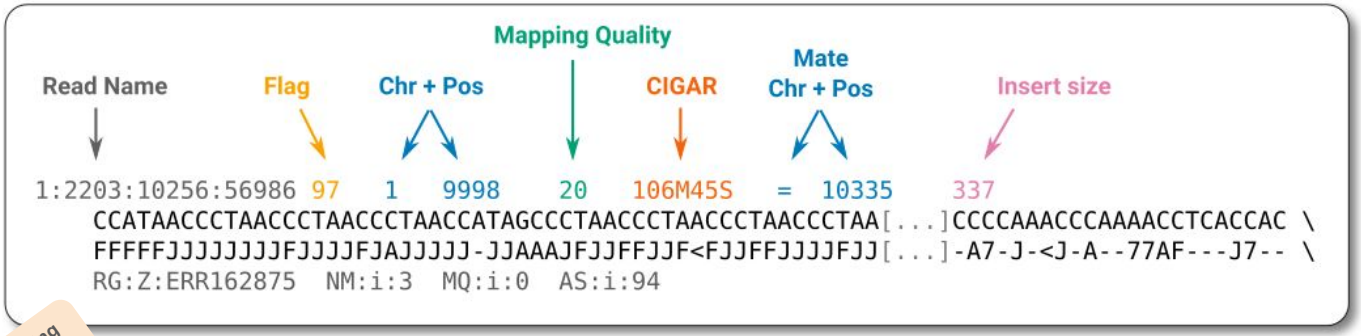

9353_6#32.bam.bamcheck
All pairs
Inward
Outward
Other
442
Number of pairs
Insert Size
0
2000
4000
6000
8000
10000
12000
0
100
200
300
400
500
600
700


9353_6#30.bam.bamcheck
All pairs
Inward
Outward
Other
441
Number of pairs
Insert Size
0
2000
4000
6000
8000
10000
12000
0
100
200
300
400
500
600

Insert Size



This is 100bp paired-end sequencing. Can you spot any problems??



Q: Mapping vs Alignment

CIGAR string

compact representation of sequence alignment:

M	alignment match or mismatch
=	sequence match
X	sequence mismatch
I	insertion to the reference
D	deletion from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
N	skipped region from the reference
P	padding (silent deletion from padded reference)

```
Ref:   ACGTACGTACTGT
Read:  ACGT----ACTGA
Cigar: 4M 4D 5M
```

```
Ref:   ACGT----ACGTA
Read:  ACGTACGTACGTA
Cigar: 4M 4I 5M
```

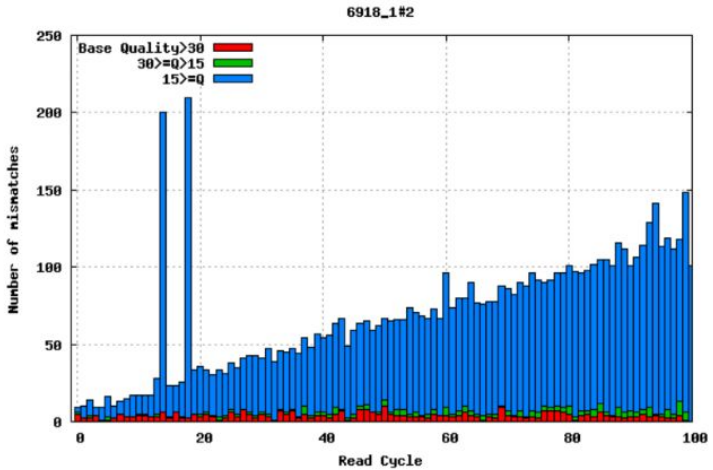
```
Ref:   CTCAGTG-GTCATCGTT
Read:  CGCA-TGAGTCTAGACG
Cigar: 4M 1D 2M 1I 3M 6S
```

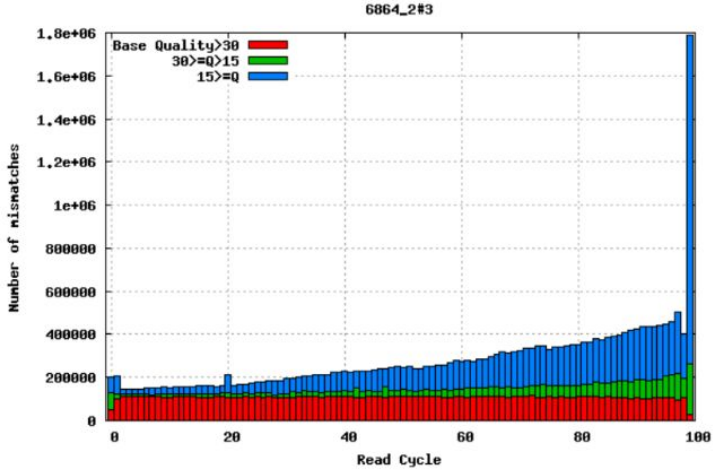
```
Ref:   ACG--CAGTAAAAT
Read:  AAGCGCAGTCCGTC
Cigar:     ???
```

Mismatches per cycle

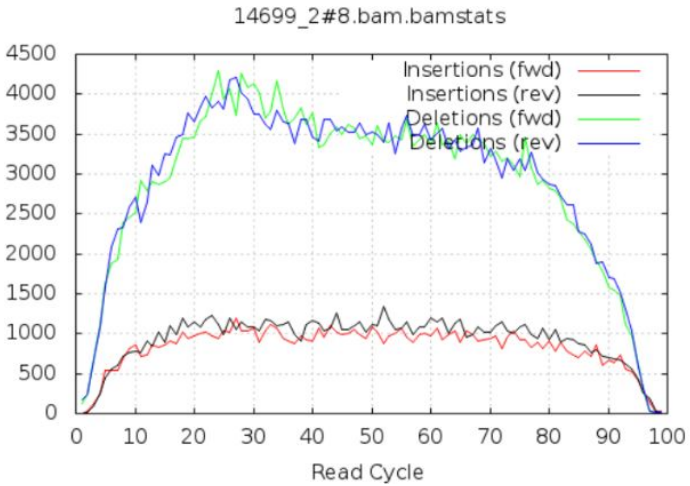
Mismatches in aligned reads (requires reference sequence)

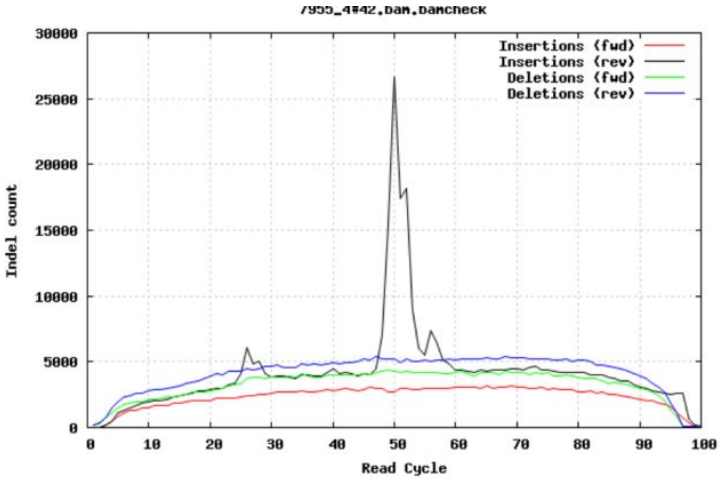
- detect cycle-specific errors
- base qualities are informative!

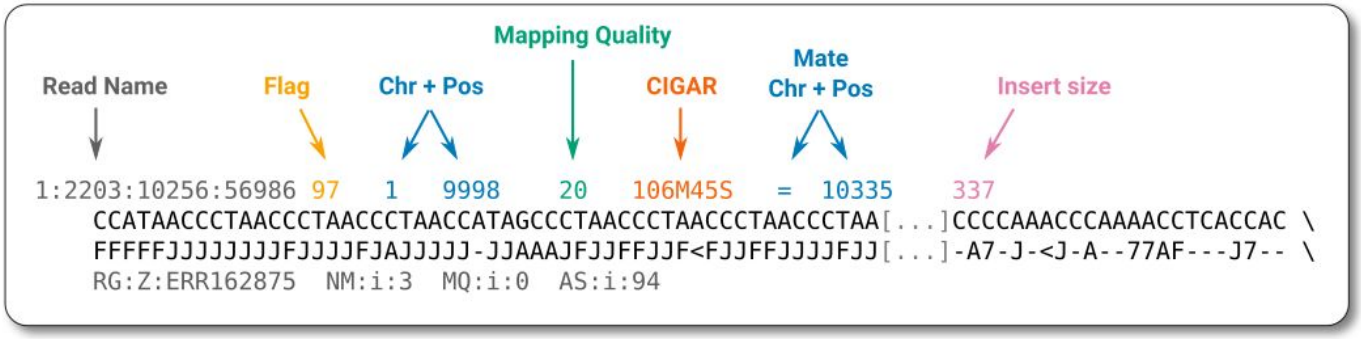




Indels per cycle







Optional tags

AS	Alignment score by the aligner
NM	Edit distance to the reference
MQ	Mapping quality of the mate
RG	Read group

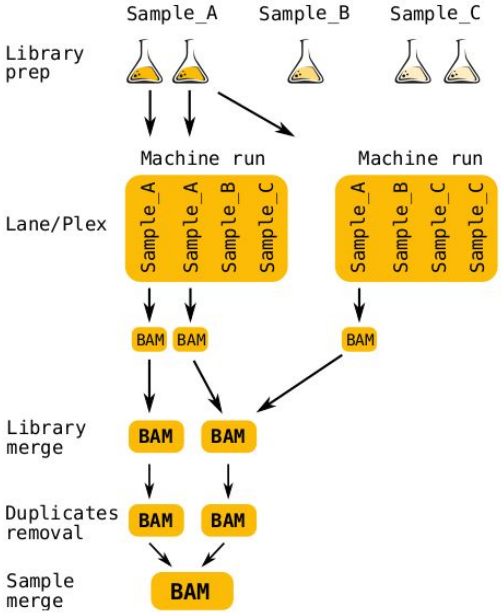
Read Group

ID	SRR/ERR number
PL	Sequencing platform
PU	Run name
LB	Library name
PI	Insert fragment size
SM	Individual
CN	Sequencing center

BAM specification

http://samtools.github.io/hts-specs/SAMv1.pdf

http://samtools.github.io/hts-specs/SAMtags.pdf

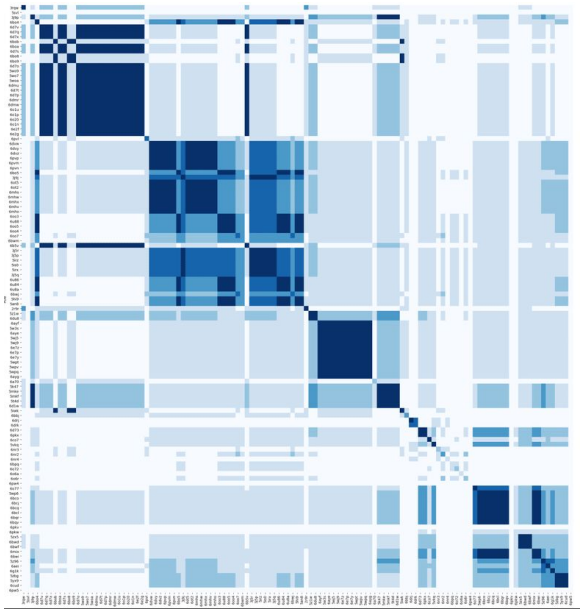


Sample swaps

How would you do that?

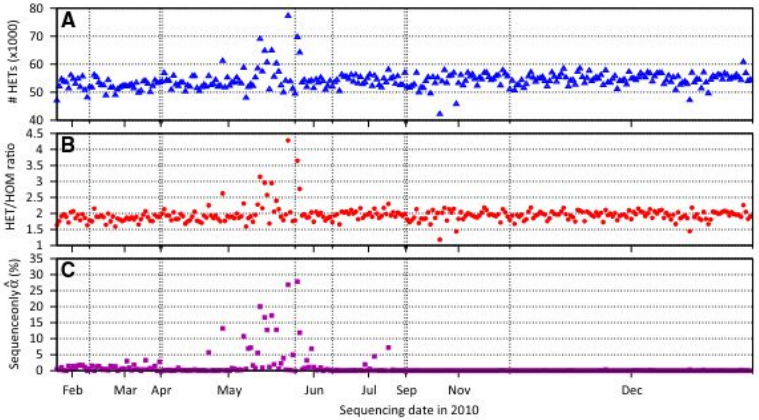
Check

- expected sex
- sample identity against a known set of variants ("barcodes")



Contaminations

- Detect sample mixture from population allele frequency



nhet

nhet/nhom

seq-only

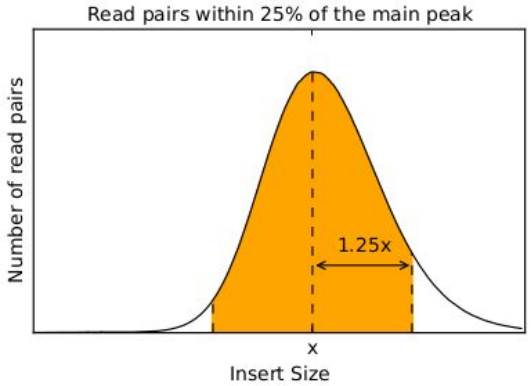
$$\mathcal{L}(\alpha) = \prod_{i=1}^{M} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{R_i} \sum_{e_{ij}} ((1-\alpha) P(b_{ij} \mid g_i^1, e_{ij}) + \alpha P(b_{ij} \mid g_i^2, e_{ij})) P(e_{ij}) \right\} P(g_i^2) P(g_i^1)$$

verifyBamID: Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data, American journal of human genetics doi:10.1016/j.ajhg.2012.09.004

Auto QC

A suggestion for human data:

Minimum number of mapped bases	90%
Maximum error rate	0.02%
Maximum number of duplicate reads	5%
Minimum number of mapped reads which are properly paired	80%
Maximum number of duplicated bases due to overlapping read pairs	4%
Maximum in/del ratio	0.82
Minimum in/del ratio	0.68
Maximimum indels per cycle, factor above median	8
Minimum number of reads within 25% of the main peak	80%

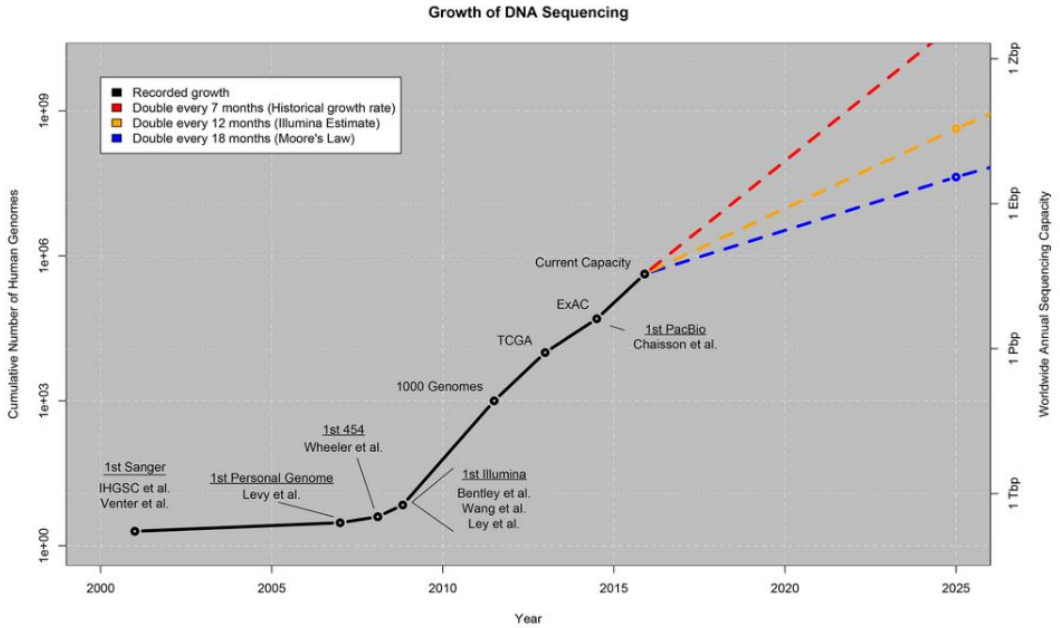


CRAM

BAM files are too large

- ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies



Zachary D. Stephens, *et al*, Big Data: Astronomical or Genomical? DOI: 10.1371/journal.pbio.1002195

CRAM

BAM stores all of the data

- Every read base
- Every base quality
- Using a single conventional compression technique for all types of data

```
Reference sequence: ACGTACGTACGTACGTACGTACGTACGTAC
read 1:             ACGTACGTACGTACGTACGTGC
read 2:                TACGTACGCACGTACGTGCGTA
read 3:                  CGTACGCACGTACGTACGTACG
read 4:                    TACGTACGTACGTGCGTACGTA
read 5:                      CGCACGTACGTACGTACGTACG
read 6:                            TACGTGCGTACGTACGTAC
```

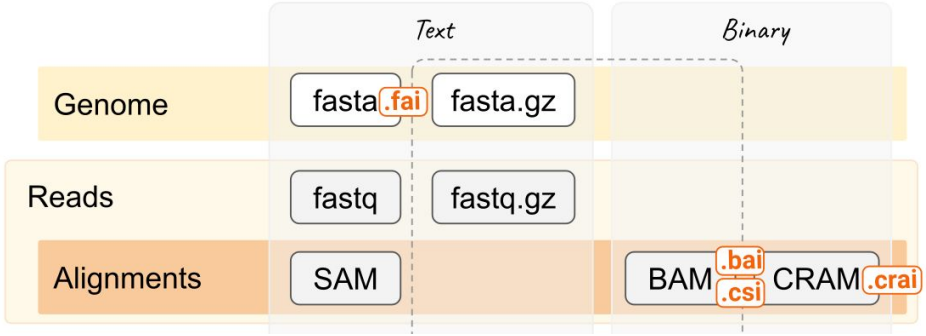
```
Reference sequence: ACGTACGTACGTACGTACGTACGTACGTAC
read 1:             ..................G.
read 2:                ........C.............
read 3:                  ......C...............
read 4:                    ............G........
read 5:                      ..C...................
read 6:                            .....G............
```

CRAM: in lossless mode 60% of BAM size

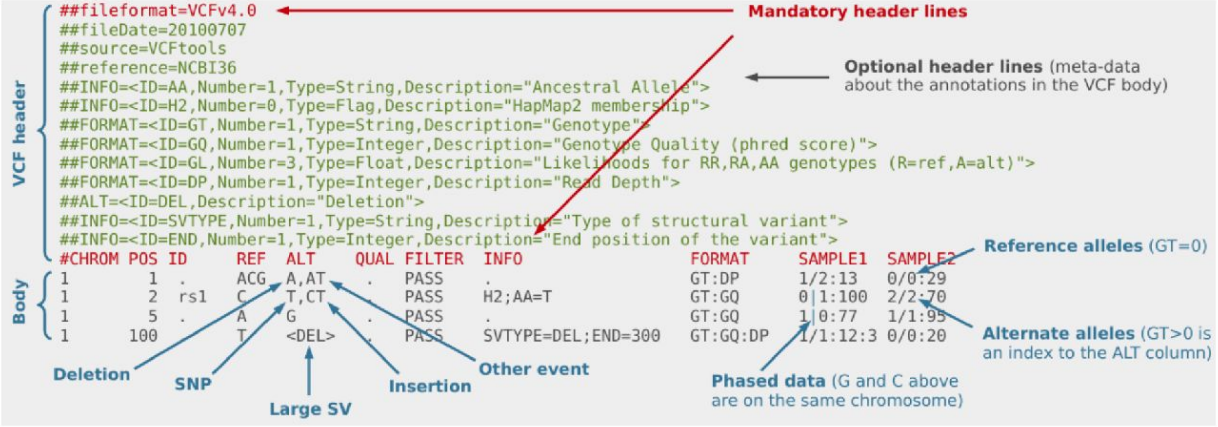
- Reference based compression
- Controlled loss of quality information
- Different compression methods for different type of data

Support for CRAM

- added to Samtools/HTSlib in 2014, to GATK in 2015
- CRAM is now mature and used in production pipelines
  - all sequencing data by default in CRAM format
  - 40% disk space saving immediately


Text
Binary
Genome
fasta
.fai
fasta.gz
Reads
fastq
fastq.gz
Alignments
SAM
BAM
.bai
.csi
CRAM
.crai

VCF: Variant Call Format



File format for storing variation data

- tab-delimited text, parsable by standard UNIX commands
- flexible and user-extensible
- compressed with BGZF (bgzip), indexed with TBI or CSI (tabix)

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Row-oriented, tab-delimited file with eight mandatory columns (CHROM-INFO)

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">

#CHROM POS  ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535  .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Genomic coordinates

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535 .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Arbitrary string, typically a dbSNP RefSNP id. Dot for missing value.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS    ID  REF ALT   QUAL FILTER INFO          FORMAT  SAMPLE1   SAMPLE2   SAMPLE3
11    24535    .   G   A     243  PASS   DP=221;AF=0.5 GT:AD   0/1:73,15 0/0:48,0  0/1:71,14
12    153927   .   C   CA,T  15   LowQ   AF=0,0.1      GT      2/2       1/2       0/1
```

All variation types can be represented:

	POS:	12345678	POS	REF	ALT
MNP	REF:	ACGTACGT	3	GT	TA
	ALT:	ACTAACGT			
Deletion		ACGTACGT	2	CGT	C
		AC--ACGT			
Insertion		AC--ACGT	2	C	CGT
		ACGTACGT			
Structural variation			2	C	<DEL>
			2	C	<DUP>

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT  QUAL FILTER INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535  .   G   A    243  PASS   DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Although in theory phred-scaled probability, don't expect truly probabilistic interpretation in practice.

QUAL = probability that the ALT call is wrong

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT  QUAL FILTER INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535 .   G   A    243  PASS   DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Soft-filter variants with e.g. low quality, low depth, etc.

VCF anatomy

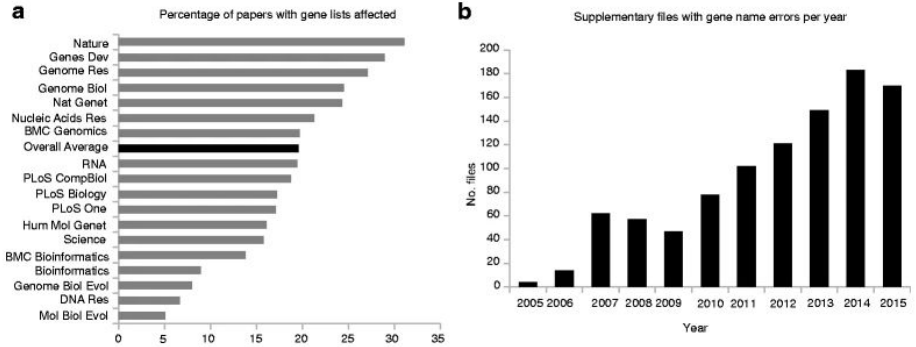
```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS    ID  REF ALT    QUAL FILTER  INFO              FORMAT  SAMPLE1    SAMPLE2    SAMPLE3
11    24535    .   G   A      243   PASS    DP=221;AF=0.5     GT:AD   0/1:73,15  0/0:48,0   0/1:71,14
12    153927   .   C   CA,T    15   LowQ    AF=0,0.1          GT      2/2        1/2        0/1
                   0   1  2                                          T/T        CA/T       C/CA
```

Genotype (GT) is represented as a 0-based index into the array of REF and ALT alleles

One file can contain zero, one or many samples

Don't use XLS

From: Gene name errors are widespread in the scientific literature



Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

5 October 2020

By Leo Kelion, Technology desk editor

BBC

https://eusprig.org/research-info/horror-stories/

Don't use CSV

Impractical format

- quoting of special characters makes it difficult to parse

```
chr1,10000,"gene X, Y, and Y"
chr1,10000,""gene "X""
```

Use UNIX escaping style instead, easy to parse

```
chr1,10000,gene X\, Y\, and Y
chr1,10000,gene X\\Y
```

TSV, Tab-Separated Values

- TAB separators are used frequently in bioinformatics pipelines
- if compressed with `bgzip`, `tabix` can be used to index by genomic coordinate
- `annot-tsv` from HTSlib package can be used to intersect files and transfer annotations from one file to another

General purpose formats

- genomic text files often BGZF-compressed

```
/* BGZF/GZIP header (specialized from RFC 1952; little endian):
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 | 31|139|  8|  4|               0|  0|255|       6| 66| 67|       2|BLK_LEN|
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
  BGZF extension:
                ^                                 ^   ^   ^
                |                                 |   |   |
                FLG.EXTRA                         XLEN  B   C

  BGZF format is compatible with GZIP. It limits the size of each compressed
  block to 2^16 bytes and adds and an extra "BC" field in the gzip header which
  records the size.
*/
```

and indexed with tabix

https://samtools.github.io/hts-specs/tabix.pdf

Field	Description	Type	Value
magic	Magic string	char[4]	TBI\1
n_ref	# sequences	int32_t	
format	Format (0: generic; 1: SAM; 2: VCF)	int32_t	
col_seq	Column for the sequence name	int32_t	
col_beg	Column for the start of a region	int32_t	
col_end	Column for the end of a region	int32_t	
meta	Leading character for comment lines	int32_t	
skip	# lines to skip at the beginning	int32_t	
l_nm	Length of concatenated sequence names	int32_t	
names	Concatenated names, each zero terminated	char[l_nm]	
	List of indices (n=n_ref):		
n_bin	# distinct bins (for the binning index)	int32_t	
	List of distinct bins (n=n_bin)		
bin	Distinct bin number	uint32_t	
n_chunk	# chunks	int32_t	
	List of chunks (n=n_chunk)		
cnk_beg	Virtual file offset of the start of the chunk	uint64_t	
cnk_end	Virtual file offset of the end of the chunk	uint64_t	
n_intv	# 16kb intervals (for the linear index)	int32_t	
	List of distinct intervals (n=n_intv)		
ioff	File offset of the first record in the interval	uint64_t	
n_no_coor (optional)	# unmapped reads without coordinates set	uint64_t	

Cheat sheet

File formats specifications

http://samtools.github.io/hts-specs

Index FASTA file

```
samtools faidx ref.fa
```

View a SAM/BAM/CRAM or a slice of it

```
samtools view file.bam | less
samtools view file.bam chr1:300000-310000 | less
```

Generate and plot stats

```
samtools stats file.bam > file.txt
plot-bamstats -p plots/ file.txt
```