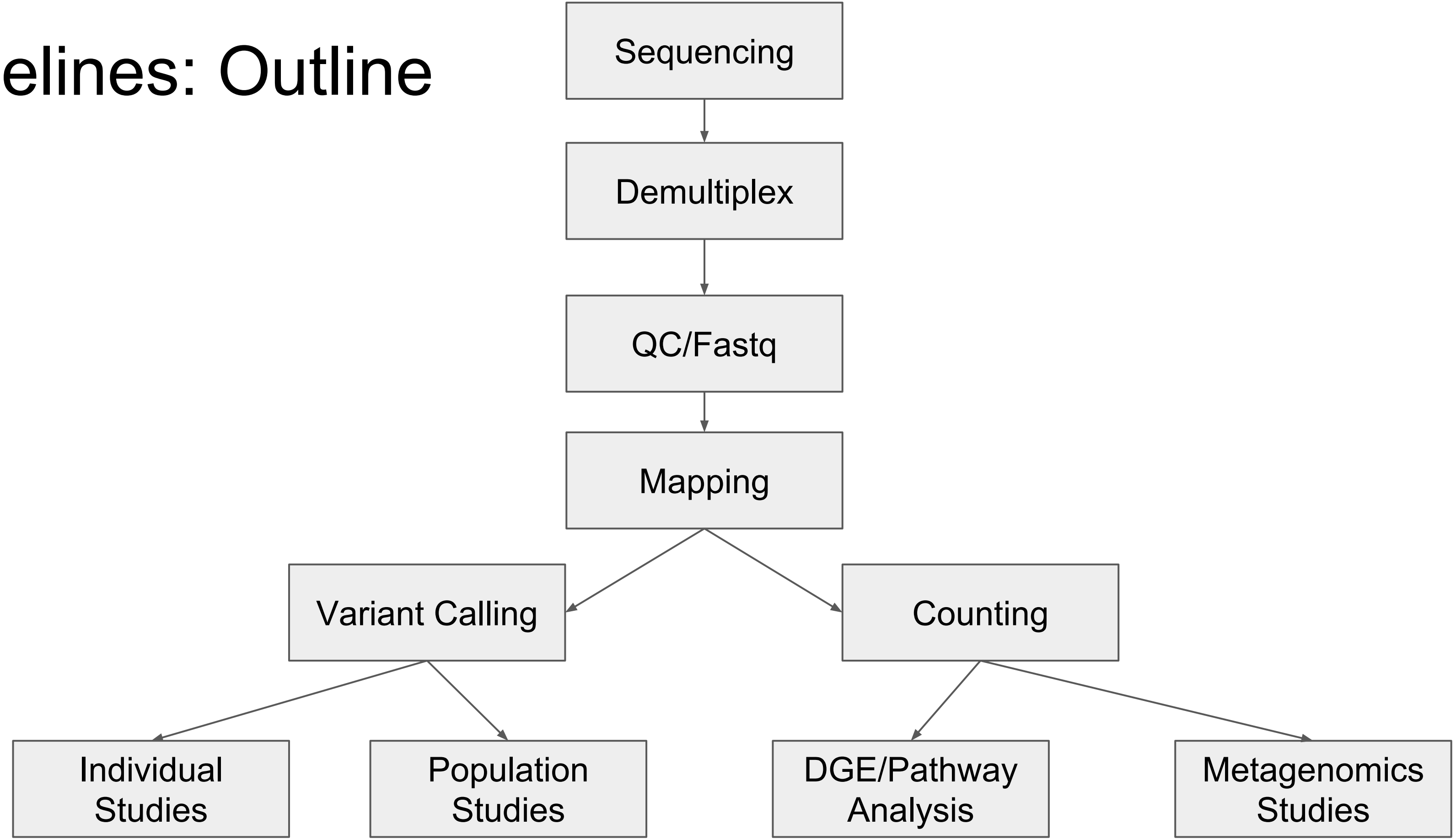
An introduction to

The Bits and Bytes to understand your Nucleotides

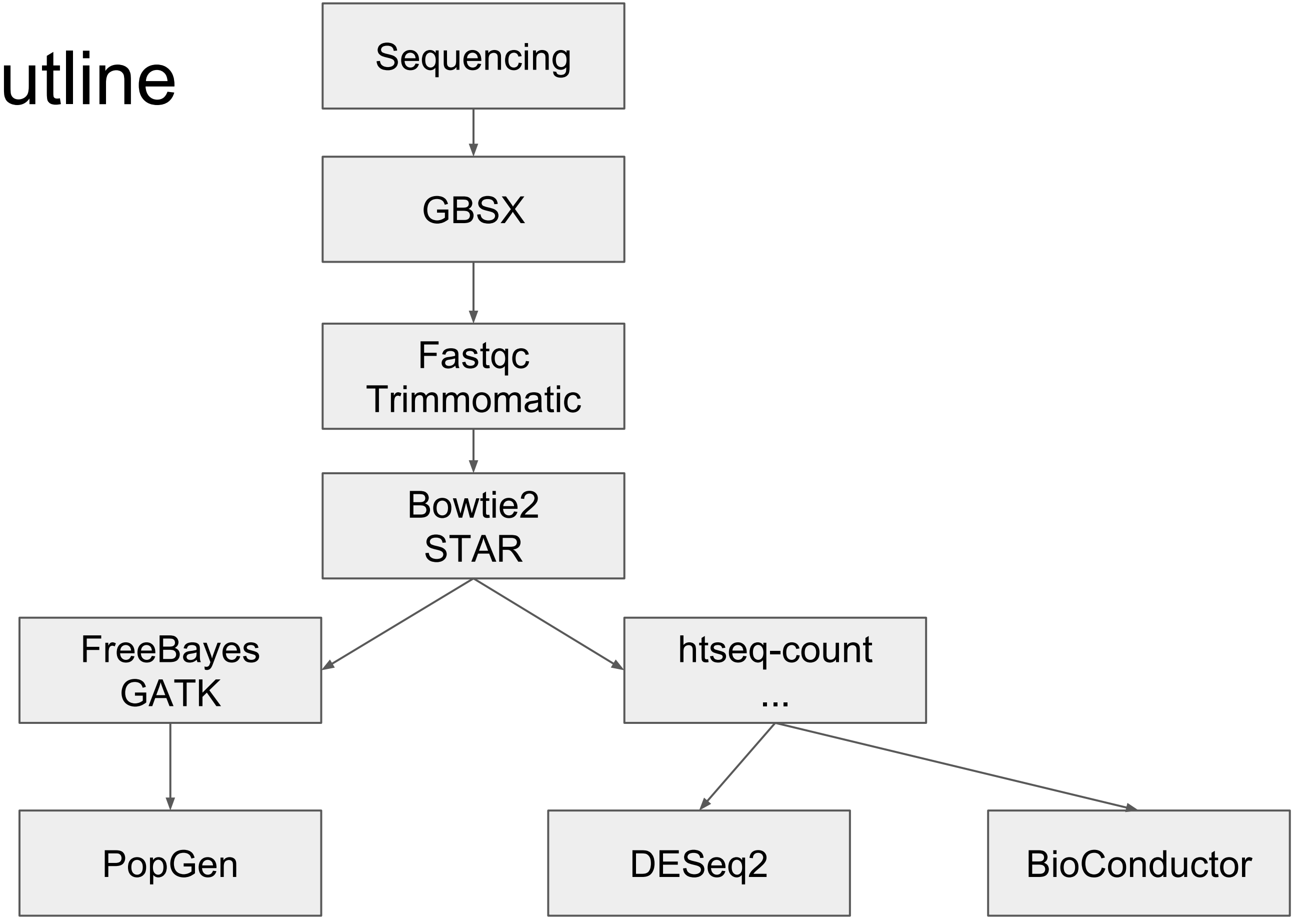
GENOMICS
CORELEUVEN

Koen Herten

Pipelines: Outline



Pipelines: Outline



Sequencing

GBSX

Fastqc
Trimmomatic

Bowtie2
STAR

FreeBayes
GATK

htseq-count
...

PopGen

DESeq2

BioConductor

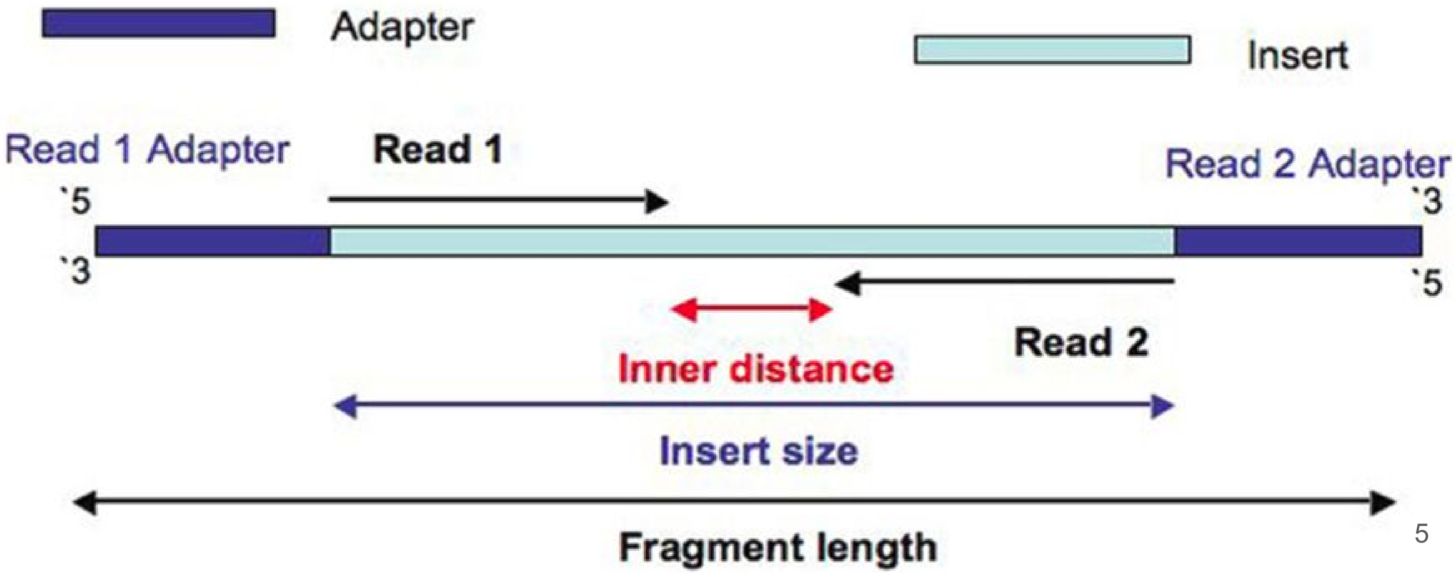
Basic Bioinformatic terms

- read
- fragment
- k-mer
- sequence depth
- coverage
- assembly
- alignment
- phred score

Basic Bioinformatic terms

- **read**
- **fragment**
- k-mer
- sequence depth
- coverage
- assembly
- alignment
- phred score

Read: a raw sequence originating from a sequencing machine
Fragment: the DNA template/amplicon that was loaded on the sequencing machine (is not always completely sequenced)



Basic Bioinformatic terms

- read
- fragment
- **k-mer**
- sequence depth
- coverage
- assembly
- alignment
- phred score

K-mer is a substring of length k
Usually the word seed is used in mapping context

```
AGCATACGATCAG
AGCAT
 GCATA
  CATAC
   ATACG
    TACGA
     ACGAT
      CGATC
       GATCA
        ATCAG
```

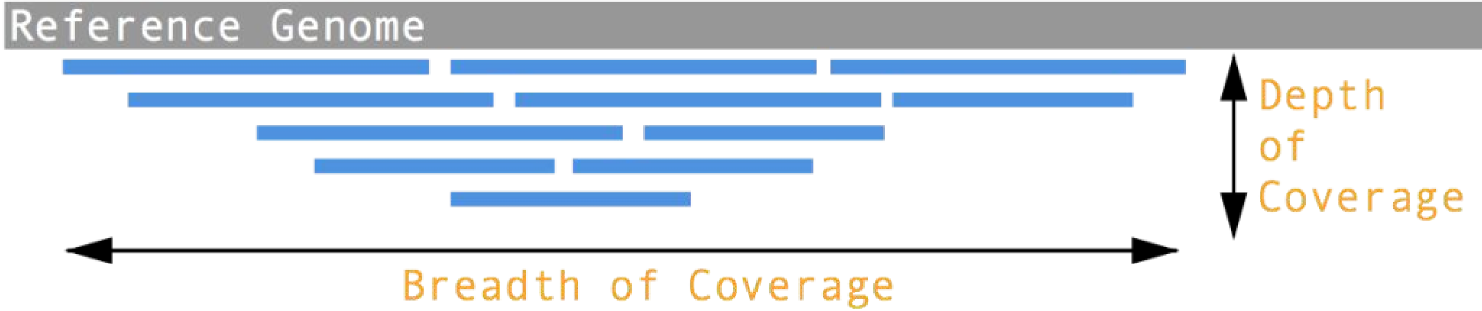
Basic Bioinformatic terms

- read
- fragment
- k-mer
- **sequence depth**
- coverage
- assembly
- alignment
- phred score

The total number of sequences generated for a sample.
Usually expressed in fragments or reads
(for Illumina Paired-end: #fragments*2=#reads)

Basic Bioinformatic terms

- read
- fragment
- k-mer
- sequence depth
- **coverage**
- assembly
- alignment
- phred score



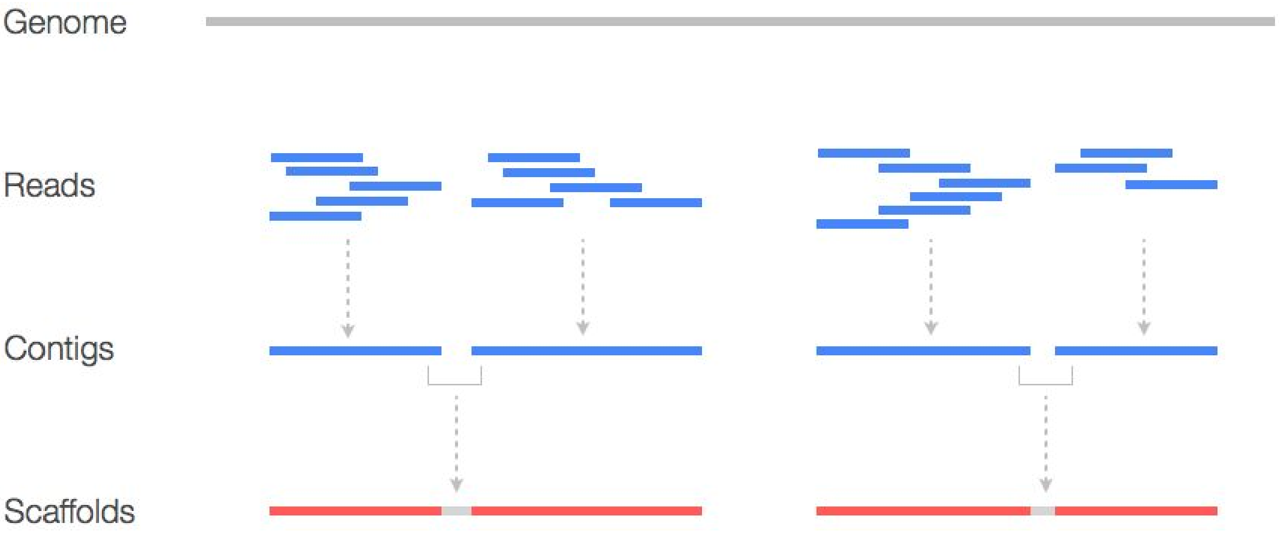
Coverage is a measure on how much of the target was seen.
Coverage can be:

per position (base)	The number of reads that are overlapping this position
per target (amplicon/exon)	The number of bases mapping to the target / the size of the target
over complete target or genome	The total number of bases generated by the sequencer / the size of the target or genome

Basic Bioinformatic terms

- read
- fragment
- k-mer
- sequence depth
- coverage
- **assembly**
- alignment
- phred score

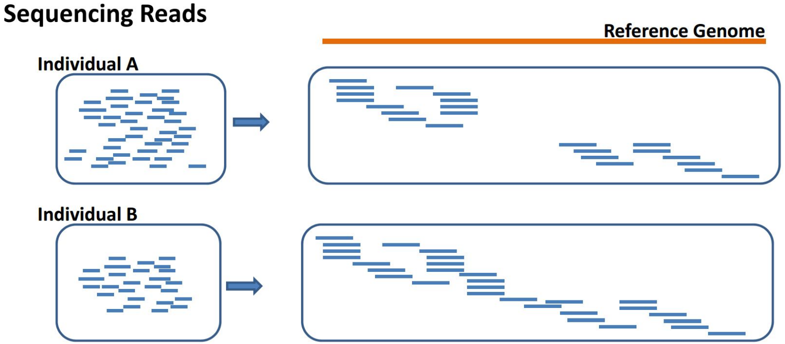
The **generation of a reference**, from scratch (*de novo*) or reference assisted.
Overlapping reads are merged to contigs (smallest unitable unit without unknown bases)
Contigs that belong together, but where the connecting sequence is unknown, can be connected to scaffolds, inserting N's for the unknown bases



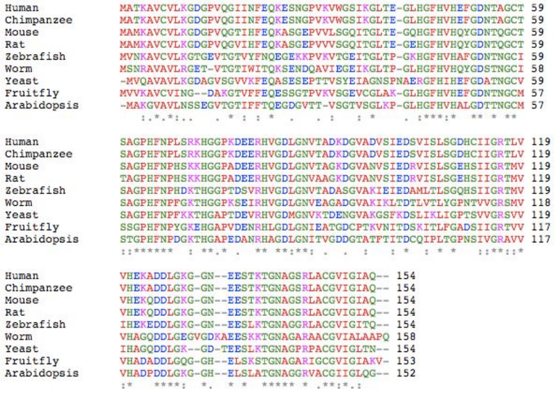
Basic Bioinformatic terms

- read
- fragment
- k-mer
- sequence depth
- coverage
- assembly
- **alignment**
- phred score

Mapping:
the process to find the position of the read on the given reference
Alignment:
the process to compare 2 sequences with each other



Mapping



```
Human            MATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTE-GLHGFHVHEFGDNTAGCT 59
Chimpanzee       MATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTE-GLHGFHVHEFGDNTAGCT 59
Mouse            MAMKAVCVLKGDGPVQGTIHFEQKASGEPVVLSGQITGLTE-GQHGFHVHQYGDNTQGCT 59
Rat              MAMKAVCVLKGDGPVQGVIHFEQKASGEPVVVSGQITGLTE-GEHGFHVHQYGDNTQGCT 59
Zebrafish        MVNKAVCVLKGTGEVTGTVYFNQEGEKKPVKVTGEITGLTP-GKHGFHVHAFGDNTNGCI 59
Worm             MSNRAVAVLRGET-VTGTIWITQKSENDQAVIEGEIKGLTP-GLHGFHVHQYGDSTNGCI 58
Yeast            -MVQAVAVLKGDAGVSGVVKFEQASESEPTTVSYEIAGNSPNAERGFHIHEFGDATNGCV 59
Fruitfly         MVVKAVCVING--DAKGTVFFEQESSGTPVKVSGEVCGLAK-GLHGFHVHEFGDNTNGCM 57
Arabidopsis      -MAKGVAVLNSSEGVTGTIFFTQEGDGVTT-VSGTVSGLKP-GLHGFHVHALGDTTNGCM 57
                    :.*.*:..   . * : : *  .   . :   : *    . :***:* ** * **

Human            SAGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHCIIGRTLV 119
Chimpanzee       SAGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHCIIGRTLV 119
Mouse            SAGPHFNPHSKKHGGPADEERHVGDLGNVTAGKDGVANVSIEDRVISLSGEHSIIGRTMV 119
Rat              TAGPHFNPHSKKHGGPADEERHVGDLGNVAAGKDGVANVSIEDRVISLSGEHSIIGRTMV 119
Zebrafish        SAGPHFNPHDKTHGGPTDSVRHVGDLGNVTADASGVAKIEIEDAMLTLSGQHSIIGRTMV 119
Worm             SAGPHFNPFGKTHGGPKSEIRHVGDLGNVEAGADGVAKIKLTDTLVTLYGPNTVVGRSMV 118
Yeast            SAGPHFNPFKKTHGAPTDEVRHVGDMGNVKTDENGVAKGSFKDSLIKLIGPTSVVGRSVV 119
Fruitfly         SSGPHFNPYGKEHGAPVDENRHLGDLGNIEATGDCPTKVNITDSKITLFGADSIIGRTVV 117
Arabidopsis      STGPHFNPDGKTHGAPEDANRHAGDLGNITVGDDGTATFTITDCQIPLTGPNSIVGRAVV 117
                 ::******  : **.* .  ** **:**: .  .  :   : *  : * *   ::**::*

Human            VHEKADDLGKG-GN--EESTKTGNAGSRLACGVIGIAQ-- 154
Chimpanzee       VHEKADDLGKG-GN--EESTKTGNAGSRLACGVIGIAQ-- 154
Mouse            VHEKQDDLGKG-GN--EESTKTGNAGSRLACGVIGIAQ-- 154
Rat              VHEKQDDLGKG-GN--EESTKTGNAGSRLACGVIGIAQ-- 154
Zebrafish        IHEKEDDLGKG-GN--EESLKTGNAGGRLACGVIGITQ-- 154
Worm             VHAGQDDLGEGVGDKAEESKKTGNAGARAACGVIALAAPQ 158
Yeast            IHAGQDDLGK--GD-TEESLKTGNAGPRPACGVIGLTN-- 154
Fruitfly         VHADADDLGQG-GH--ELSKSTGNAGARIGCGVIGIAKV- 153
Arabidopsis      VHADPDDLGKG-GH--ELSLATGNAGGRVACGIIGLQG-- 152
                 :*   ****:  *.  * *  ***** * .**:*.:
```

Alignment

Basic Bioinformatic terms

- read
- fragment
- k-mer
- sequence depth
- coverage
- assembly
- alignment
- **phred score**

Phred score:
universal score for the probability of:

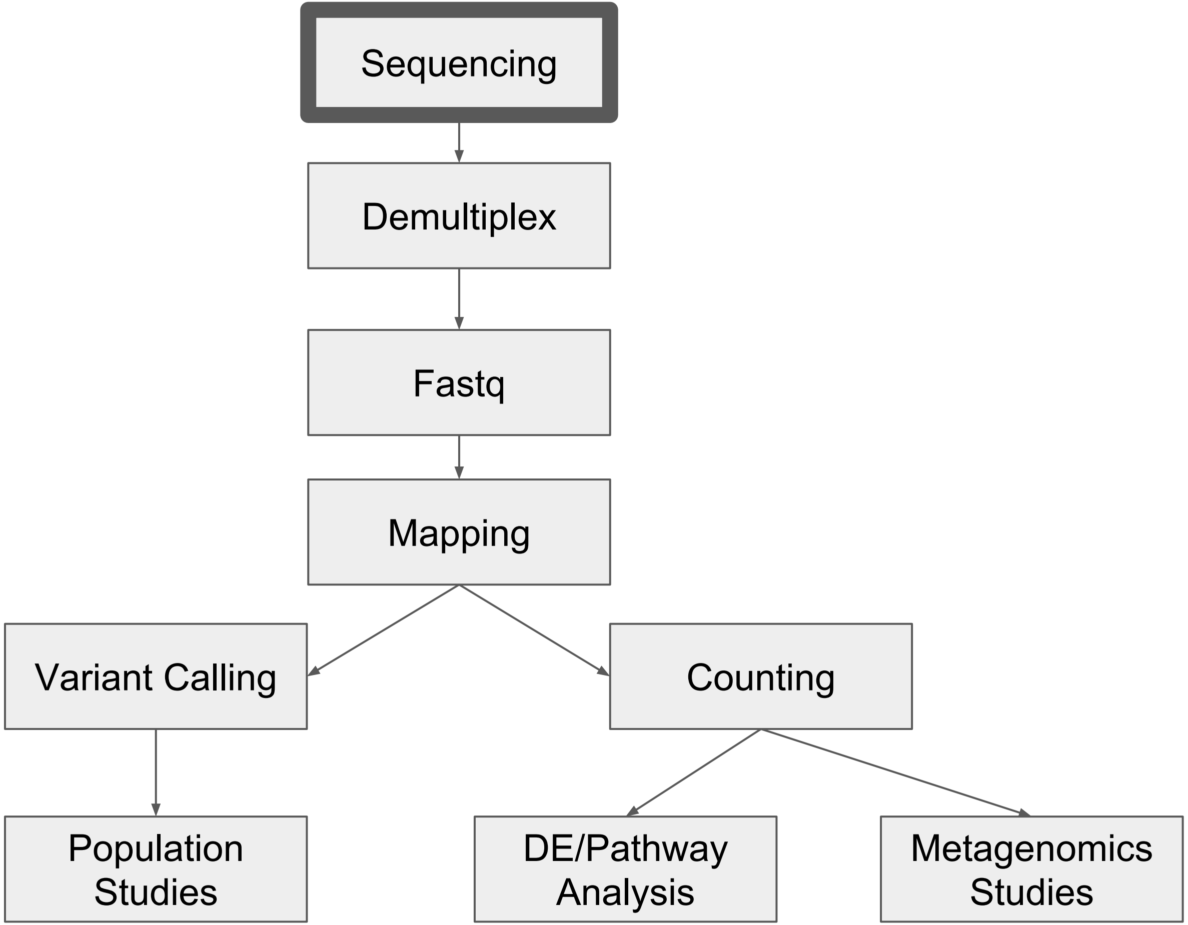
- base quality
- mapping quality
- genotype quality

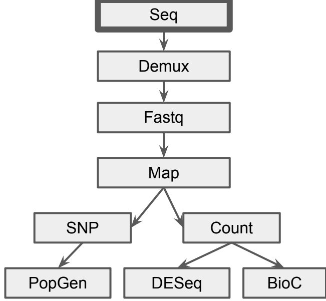
Formula:

$$Q = -10 \log_{10} P$$

Phred score	Probability	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%

Outline


Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies



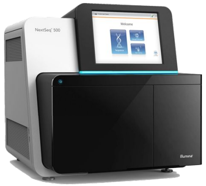
Sequencing Technologies

illumina®

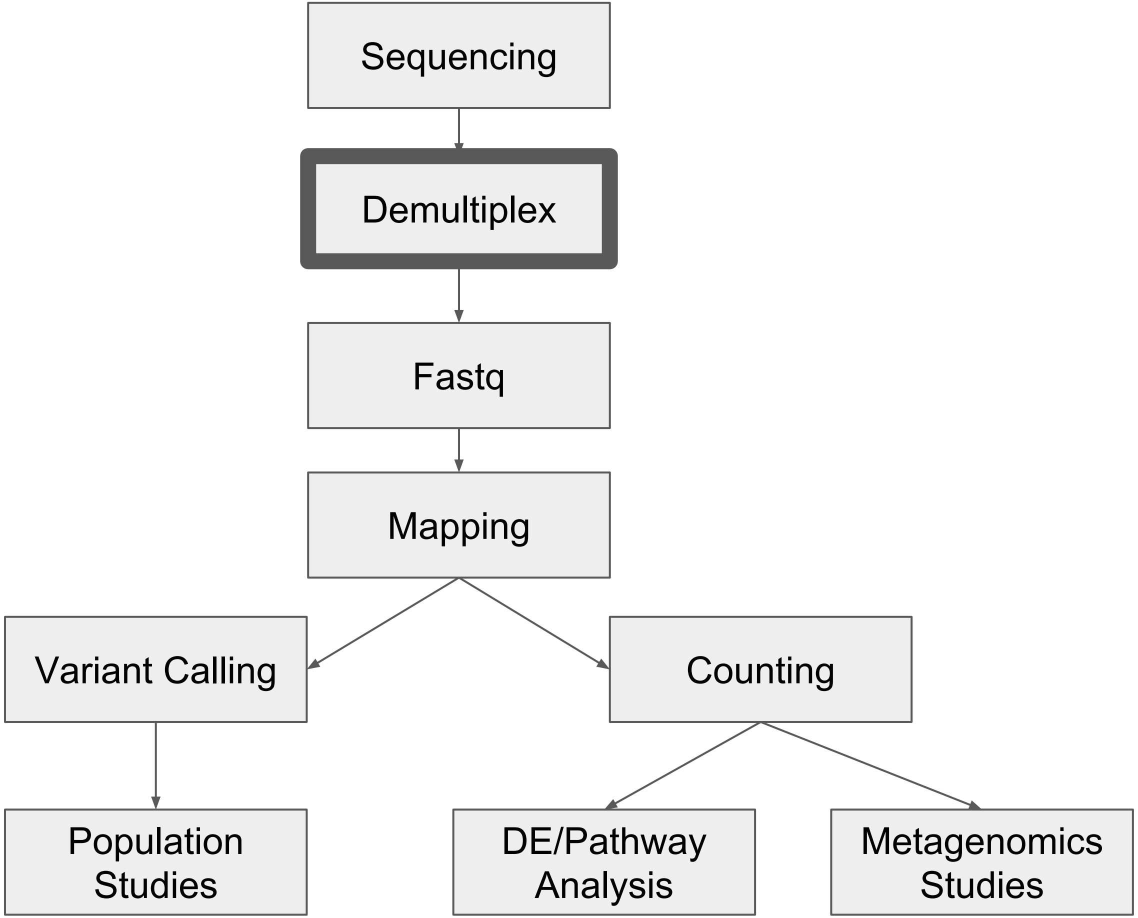
- Short reads (35-250bp)
- Single reads or Paired end
- Sequencing of clusters
  (groups of multiple same fragment)
- All raw reads have same length

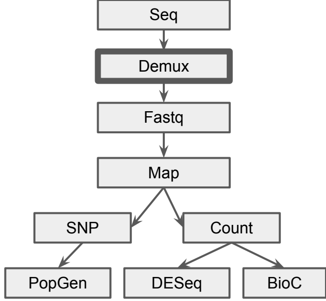
PACBIO®

- Long reads (avg >10,000bp)
- Subreads or CCS
  (same fragment, corrected with multiple passes)
- Single Molecule, Real Time (SMRT)
- Reads have different lengths

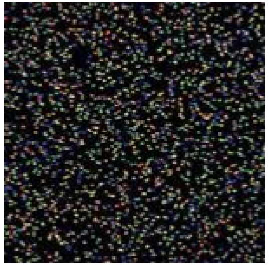


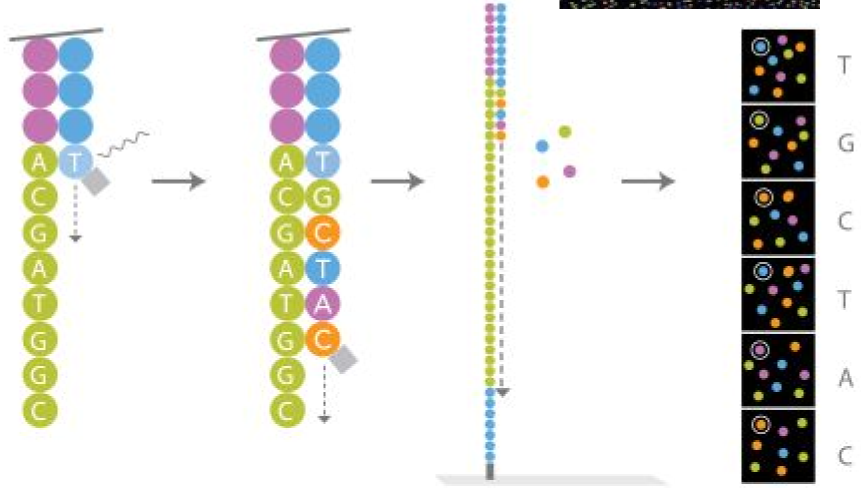
Outline

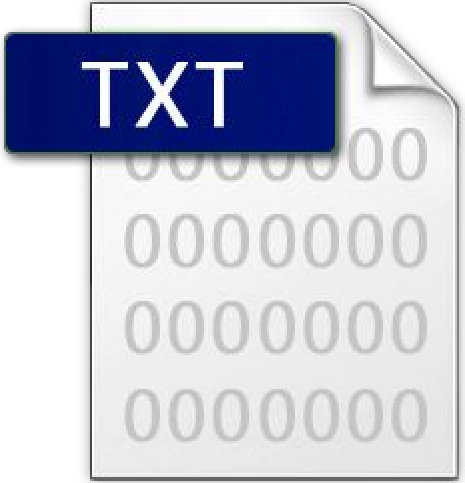

Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies



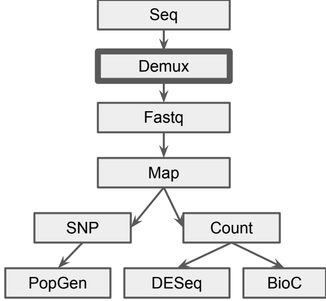
Demultiplexing: from image to fastq







Demultiplexing



Seq

Demux

Fastq

Map

SNP

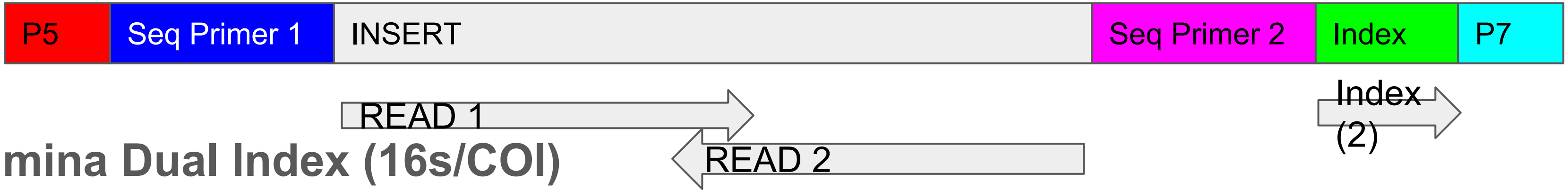
Count

PopGen

DESeq

BioC

Illumina Index (RNASeq/Shotgun Seq)



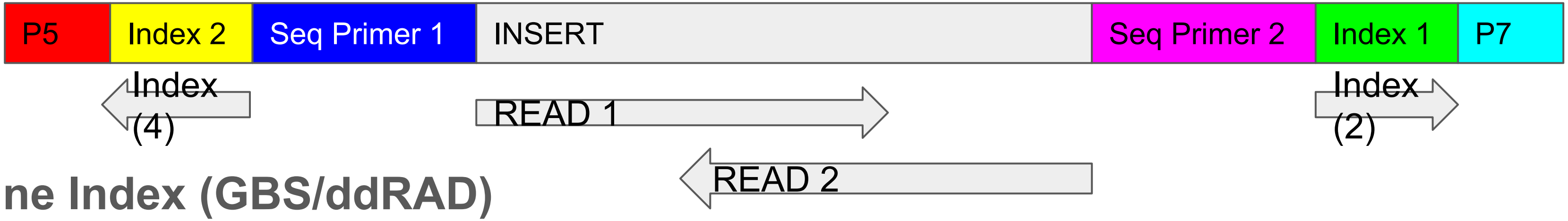
P5 | Seq Primer 1 | INSERT | Seq Primer 2 | Index | P7

READ 1

READ 2

Index (2)

Illumina Dual Index (16s/COI)



P5 | Index 2 | Seq Primer 1 | INSERT | Seq Primer 2 | Index 1 | P7

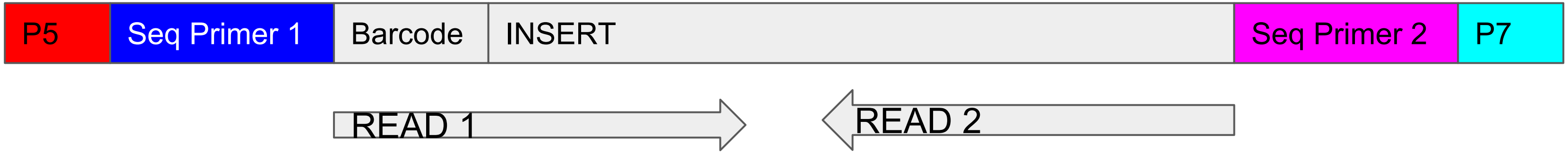
Index (4)

READ 1

Index (2)

READ 2

Inline Index (GBS/ddRAD)

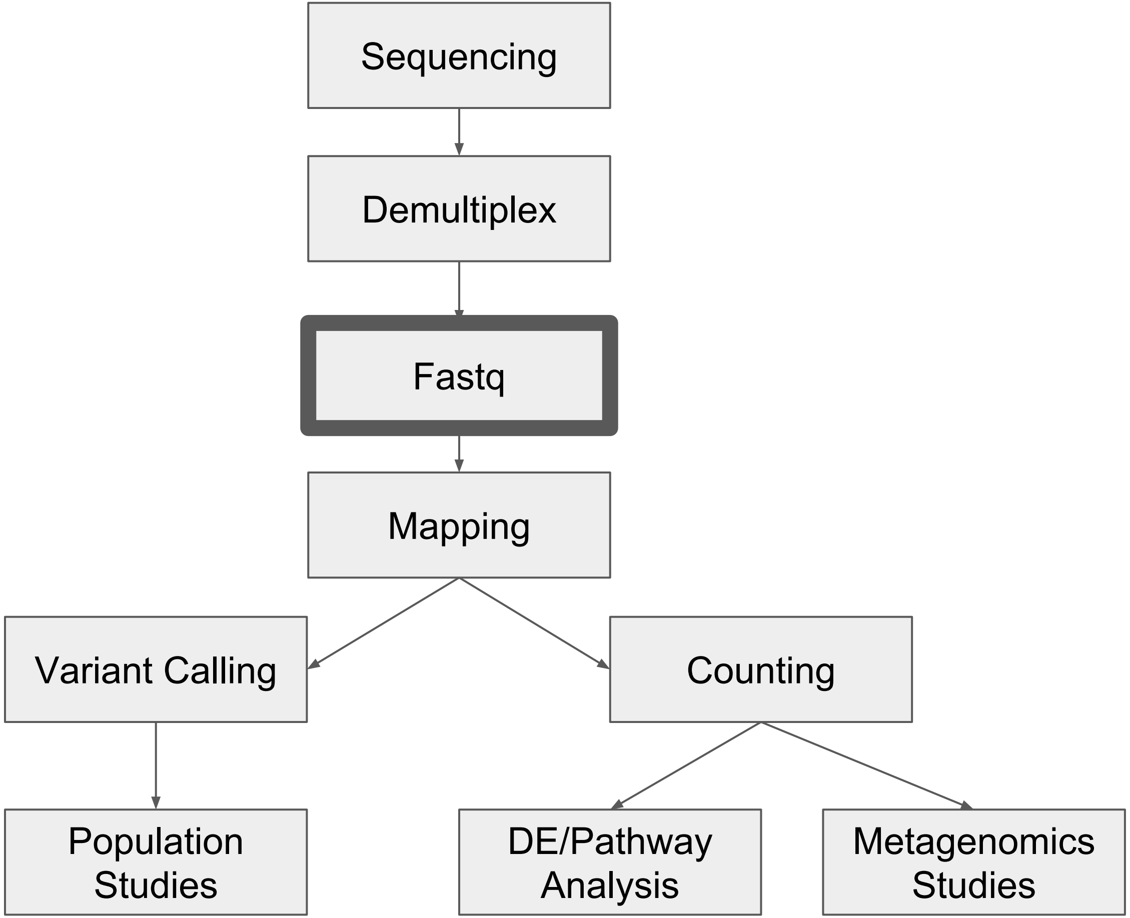


P5 | Seq Primer 1 | Barcode | INSERT | Seq Primer 2 | P7

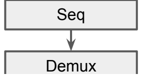
READ 1

READ 2

Outline



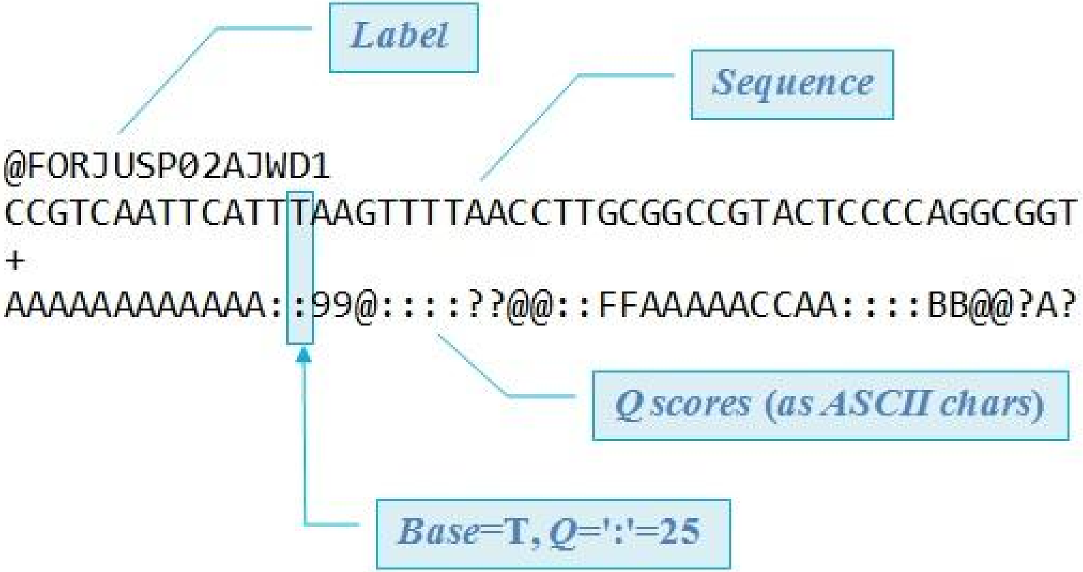
Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies



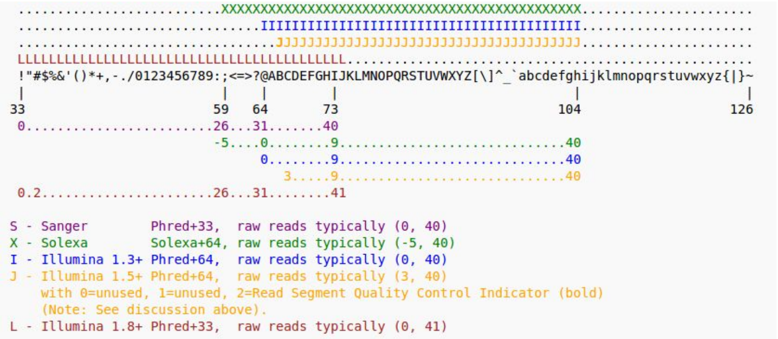
Fastq file

Fastq is a text based format containing the reads (sequences) that came from the machine.

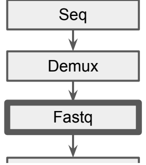
Often this file is several MB to GB in size, therefore it is gzipped (.fastq.gz).



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```



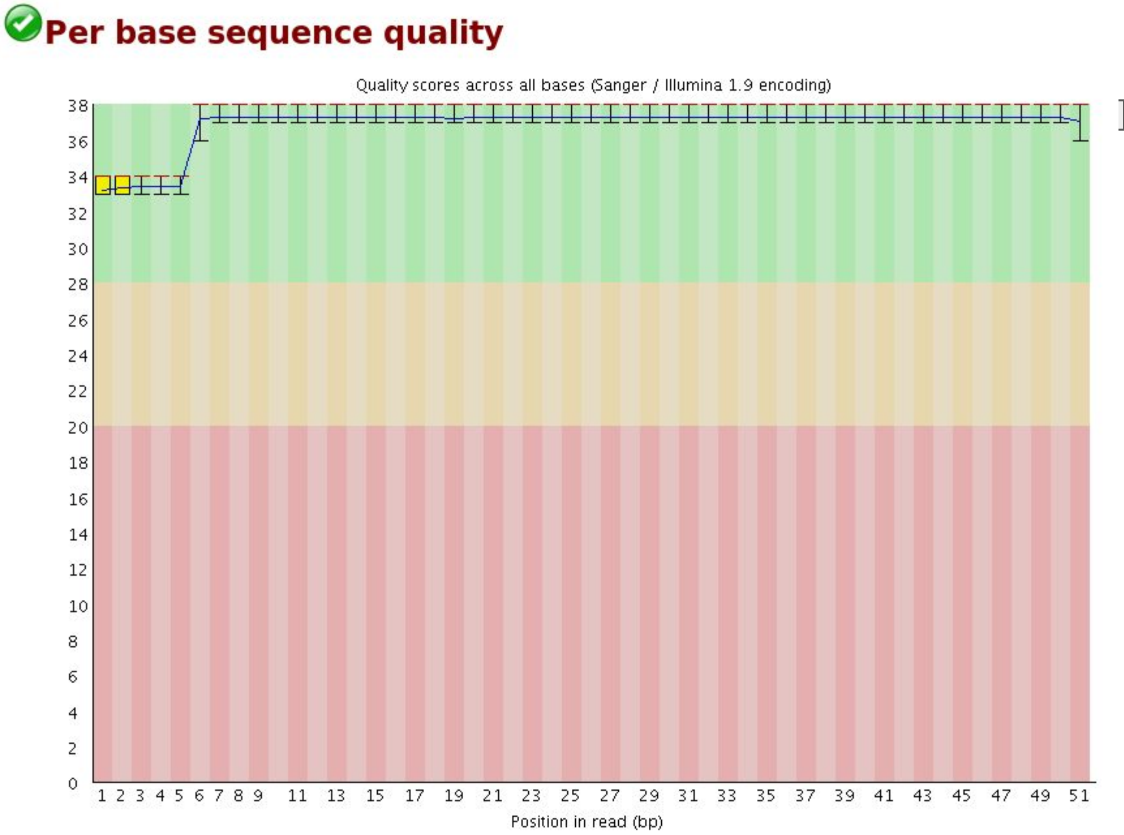
```
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

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, 40)
    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)
```



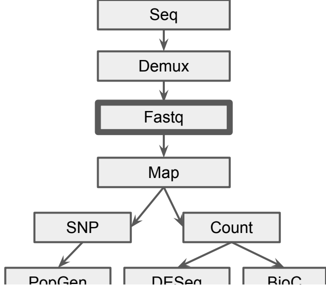
Fastq Quality Control

Quality scores across bases

- Indication for the quality of the run, and quality of input DNA for the sample
- First 8-10 bp always lower, since the sequencer (Illumina) needs to find the location of the reads
- Low end?
  => Quality trimming



Fastqc: Base Quality



Seq

Demux

Fastq

Map

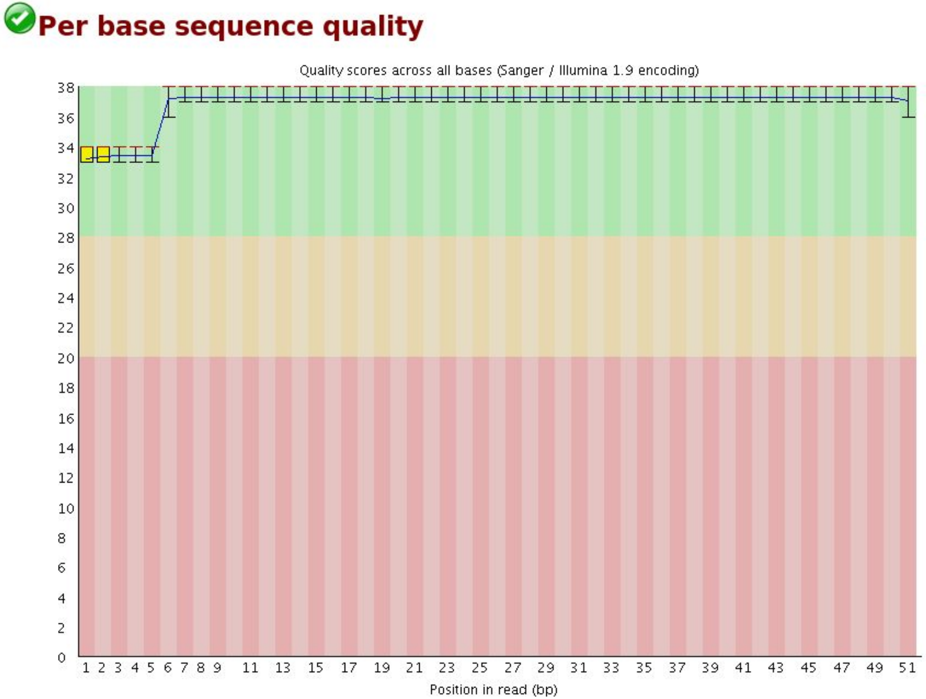
SNP

Count

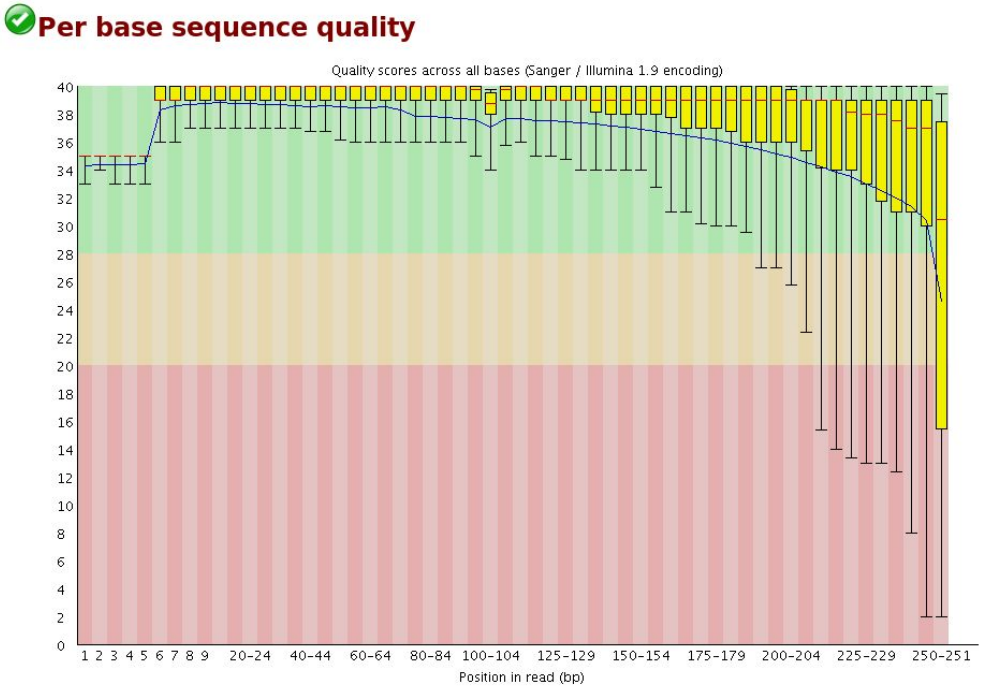
PopGen

DESeq

BioC

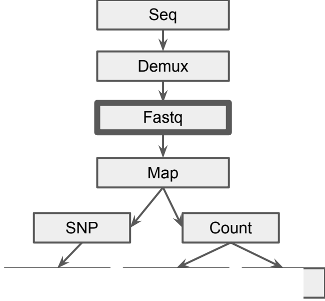


Per base sequence quality

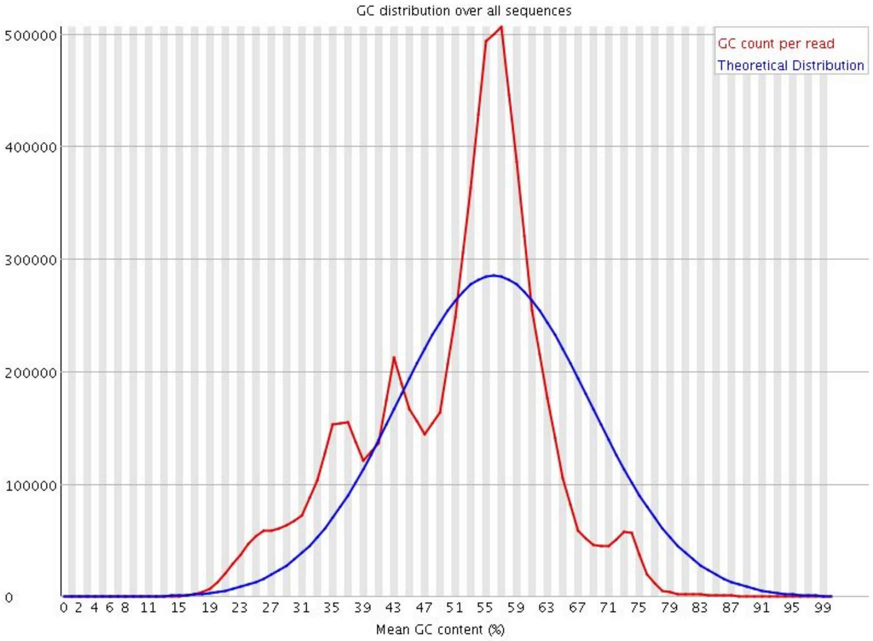


Per base sequence quality

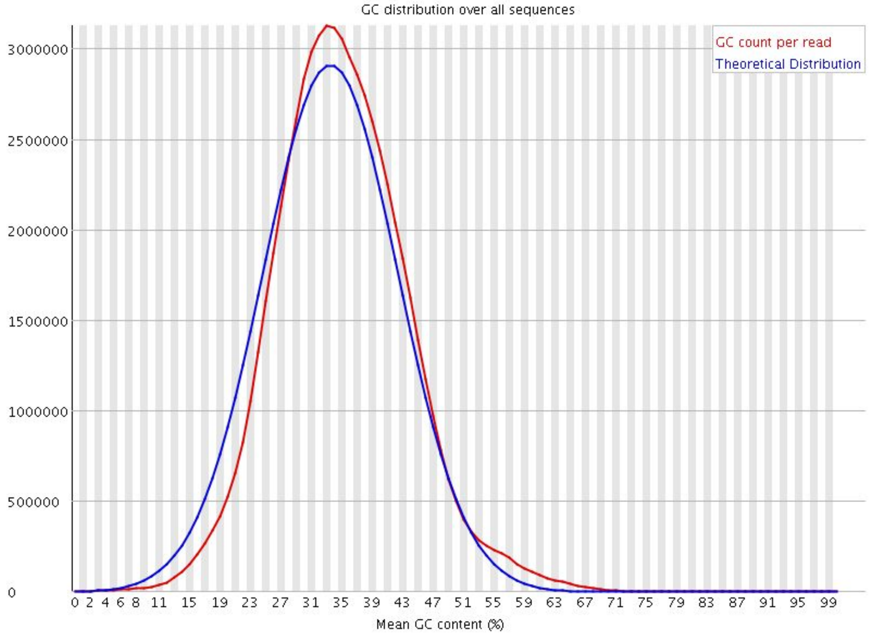
Fastqc: GC content



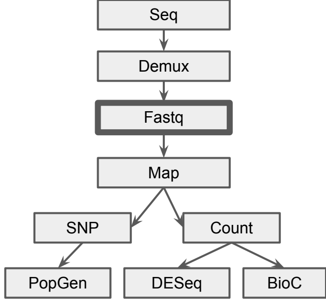
✗ **Per sequence GC content**



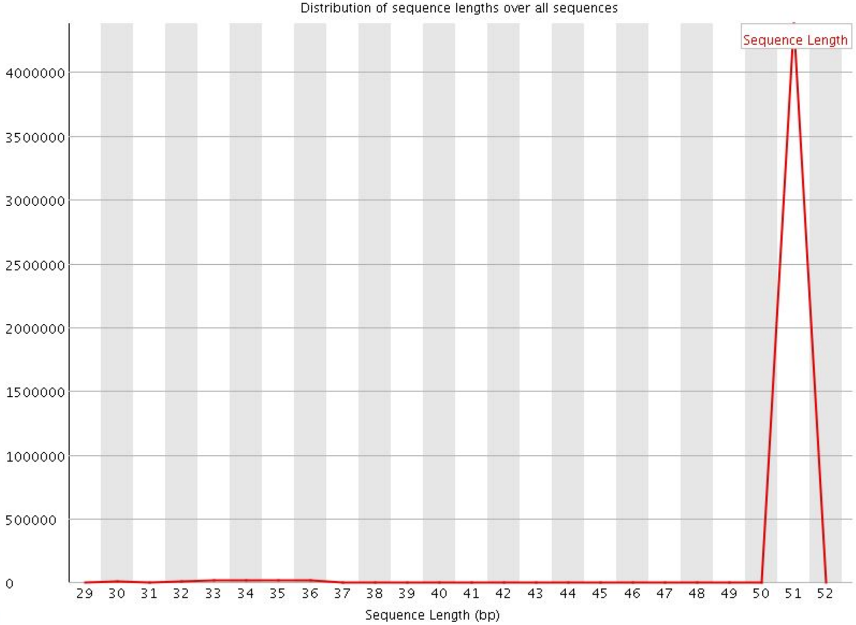
✓ **Per sequence GC content**



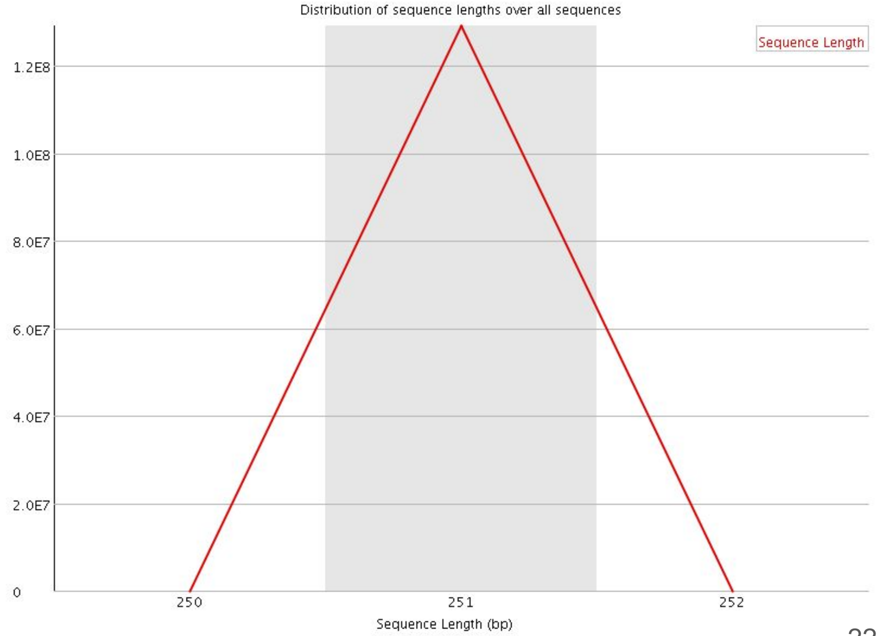
Fastqc: Read length



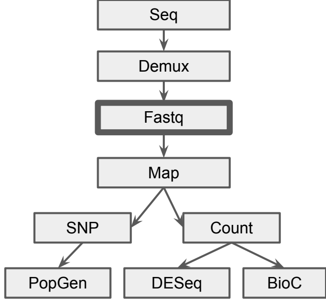
Sequence Length Distribution



Sequence Length Distribution



Fastqc: duplication levels



Seq

Demux

Fastq

Map

SNP

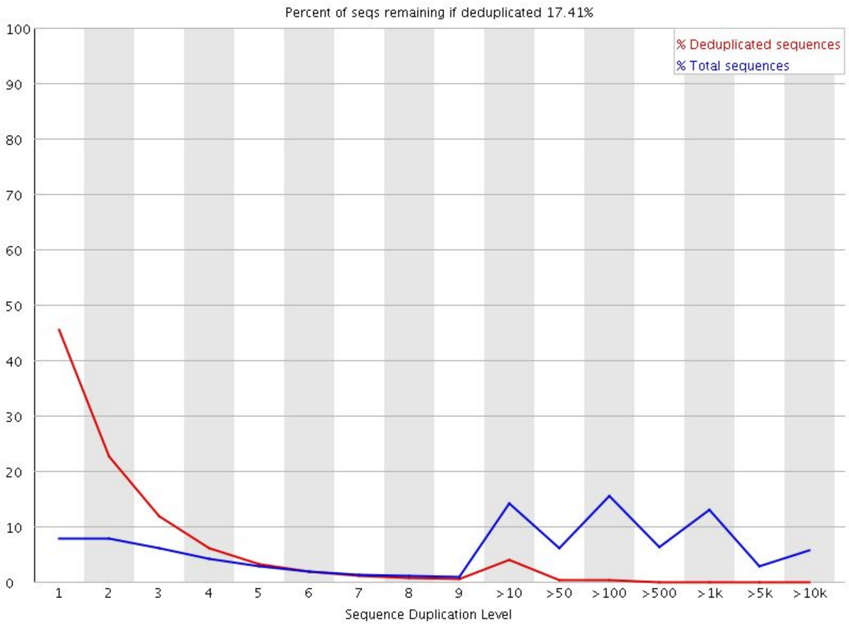
Count

PopGen

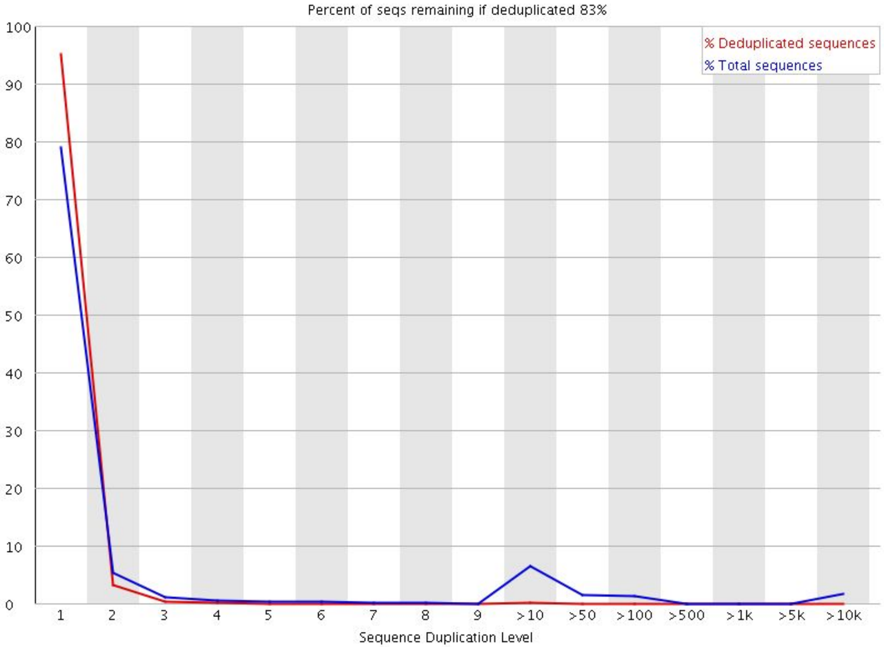
DESeq

BioC

✗ Sequence Duplication Levels



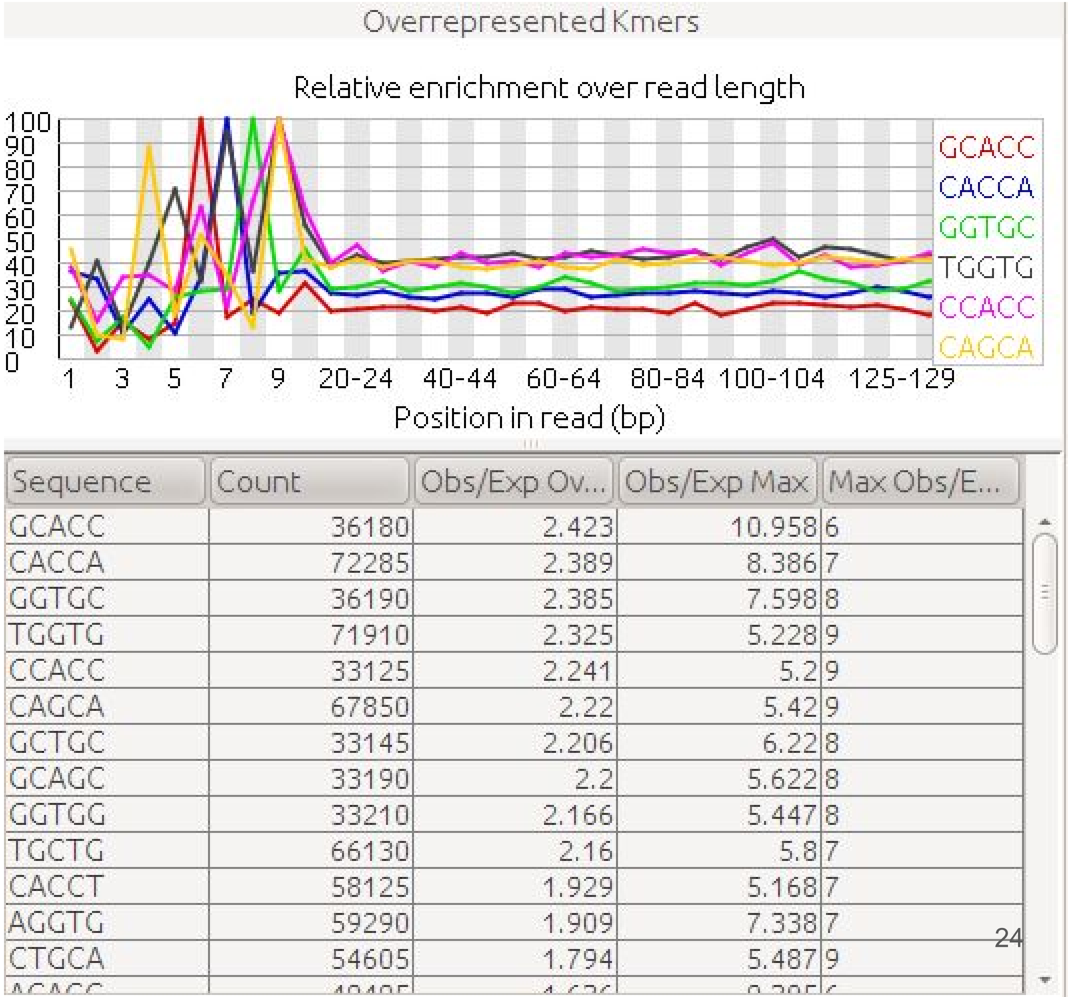
✓ Sequence Duplication Levels



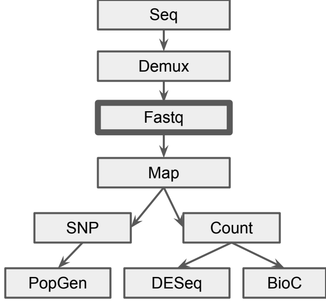
Fastq Quality Control

Overrepresented Kmers

- indicates frequent found sequence
- ALWAYS visible in amplicon based data (like 16S/GBS)
- Can indicate sequenced adaptors or adaptor dimers

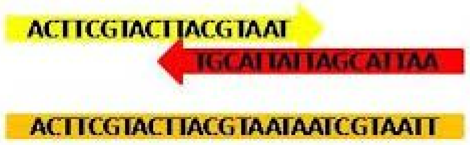


Sequence	Count	Obs/Exp Ov...	Obs/Exp Max	Max Obs/E...
GCACC	36180	2.423	10.958	6
CACCA	72285	2.389	8.386	7
GGTGC	36190	2.385	7.598	8
TGGTG	71910	2.325	5.228	9
CCACC	33125	2.241	5.2	9
CAGCA	67850	2.22	5.42	9
GCTGC	33145	2.206	6.22	8
GCAGC	33190	2.2	5.622	8
GGTGG	33210	2.166	5.447	8
TGCTG	66130	2.16	5.8	7
CACCT	58125	1.929	5.168	7
AGGTG	59290	1.909	7.338	7
CTGCA	54605	1.794	5.487	9
ACACC	40495	1.636	9.395	6

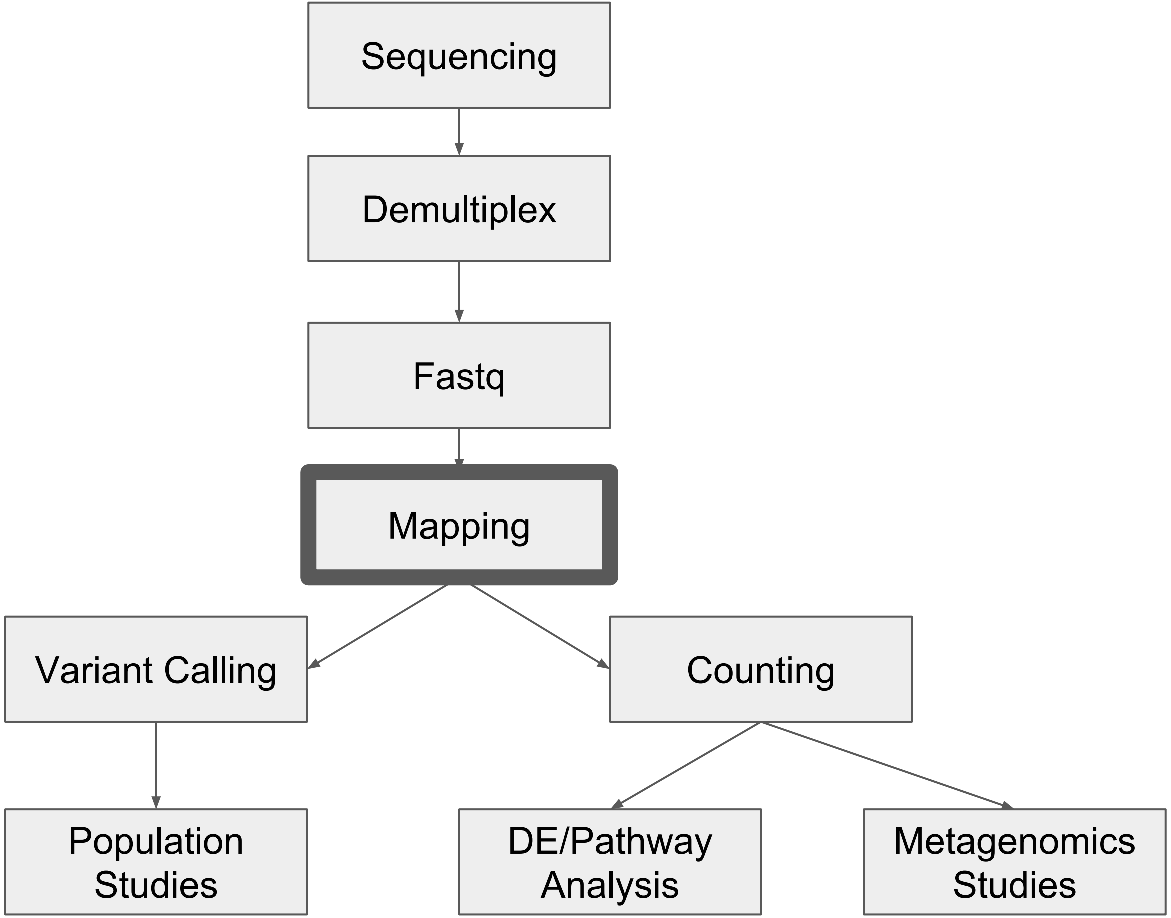


Fastq modifications

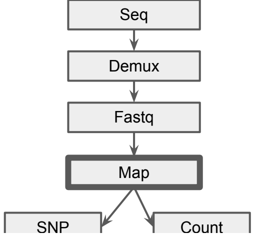
- QC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- adaptor and quality trimming
  - removal of adaptors (will affect mapping)
  - removal of low quality ends (can have an effect on mapping and downstream analysis such as variant calling)
  - Tools:
    - Trimmomatic, bbduk, ea-utils, fastx-toolkit, ...
- read merging
  - Merging of overlapping reads into the original fragments
  - Improvement on mapping and variant calling
  - Tools:
    - FLASH, PEAR, ...

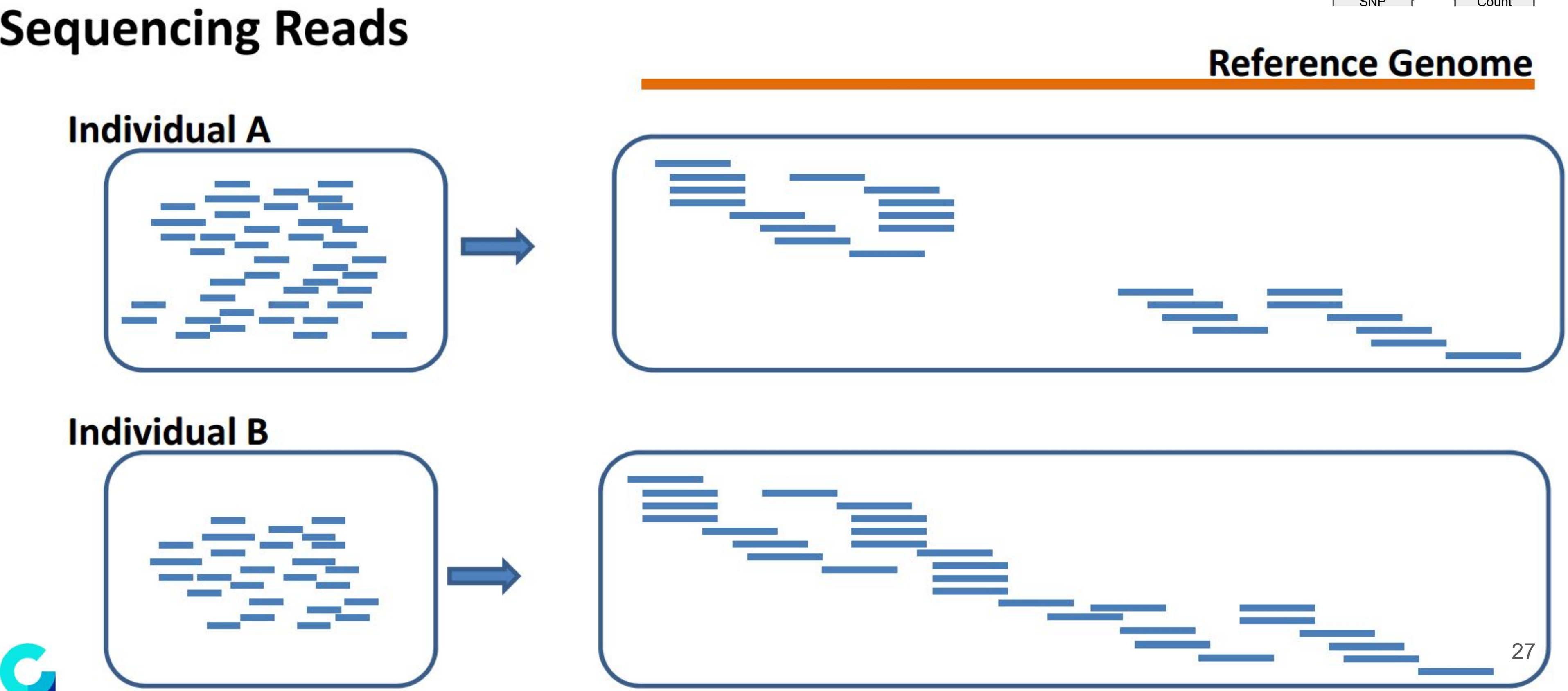


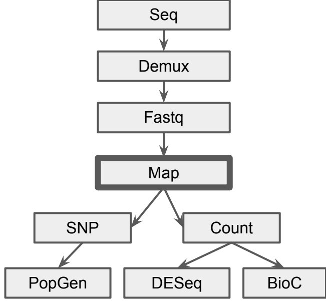
Outline


Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies

Mapping


Seq
Demux
Fastq
Map
SNP
Count


Sequencing Reads
Reference Genome
Individual A
Individual B

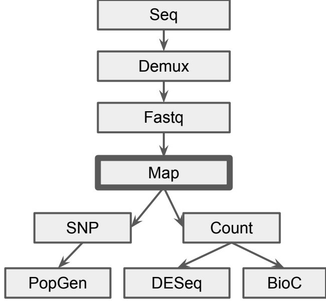


Mapping: Where to start?

Required information:

1. The dataset in fastq format (1 or 2 files, depends on single reads or paired end data)
2. A reference
   a. The reference is species dependent.
   b. It can have any quality (from 1k+ contigs, to chromosome level)
   c. It can be the genome or the transcriptome
   d. As in *de novo* RAD/GBS mode, a set of micro-contigs assembled

Mapping



- Burrow wheeler transformation
  - Most used, low RAM needed
- Hashmap
  - Needs a lot of RAM

AC**GTAC**GCA

Find seed in table (usually multiple seeds)

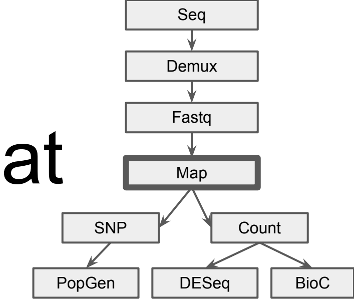
Seed	Position
AGGT	1:50
GGTC	5:1563
GTAC	3:1563
CGGA	3:1566
...	...

Note: seed can be found on multiple positions

Check surrounding bases (bases between matching seeds)

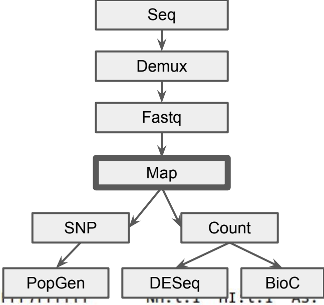
```
REF:    ...ACGTACGGA...
SEQ:       ACGTACCGA
POS:    3:1561
```

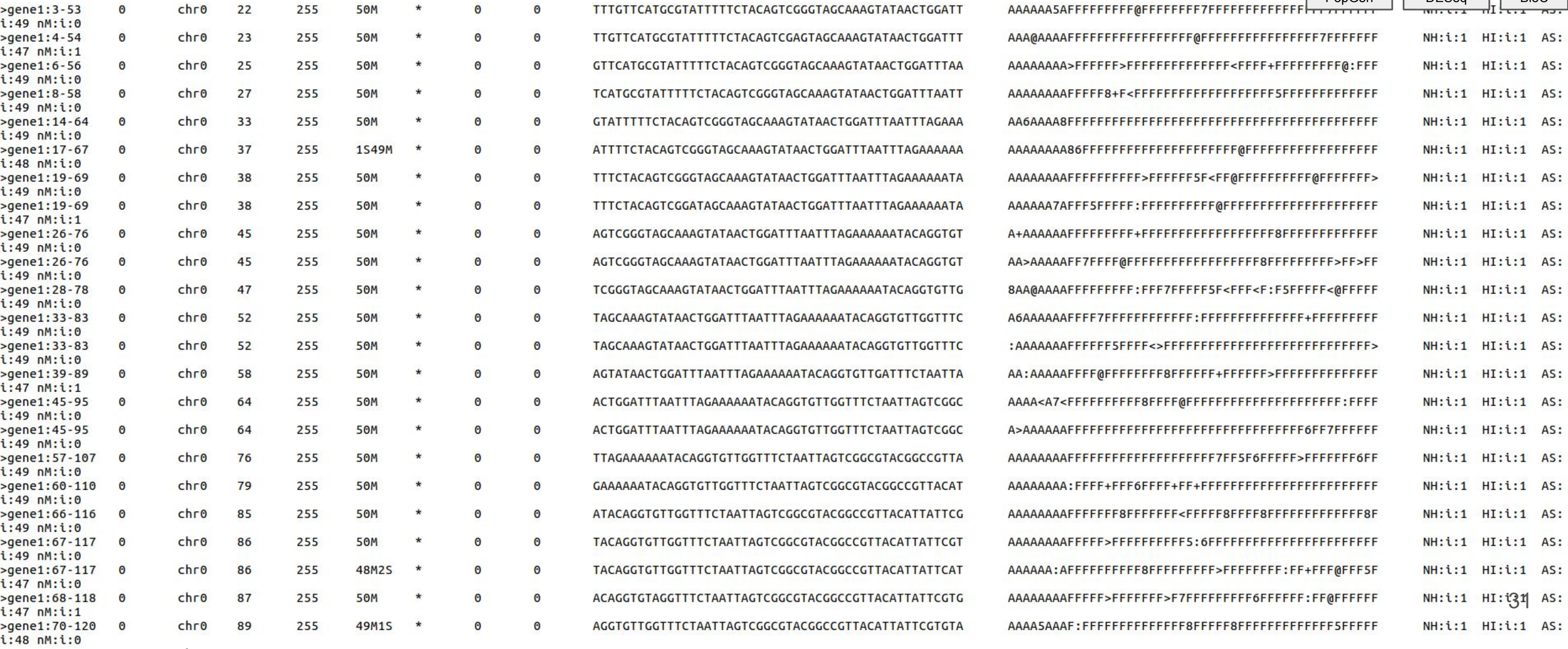
Mapping: Output format: SAM and BAM format



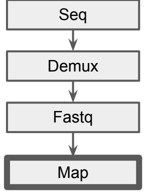
- SAM
  Sequence Alignment/Map format
  Header starts with @
  Mapped Reads are tab-delimited lines
  1-based system (includes SAM, **VCF**, GFF, GTF)
- BAM
  Binary Alignment/Map format
  Is the binary form of the SAM format, so reduces storage
  Contains exact the same information as the SAM format
  0-based system (includes BAM, BED)

Mapping: Output format: SAM format





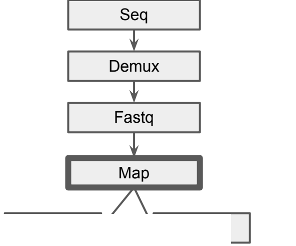
```
>gene1:3-53     0   chr0   22   255   50M     *   0   0   TTTGTTCATGCGTATTTTTCTACAGTCGGGTAGCAAAGTATAACTGGATT   AAAAAA5AFFFFFFFFF@FFFFFFFF7FFFFFFFFFFFFF              NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:4-54     0   chr0   23   255   50M     *   0   0   TTGTTCATGCGTATTTTTCTACAGTCGAGTAGCAAAGTATAACTGGATTT   AAA@AAAAFFFFFFFFFFFFFFFFF@FFFFFFFFFFFFFFF7FFFFFFF   NH:i:1   HI:i:1   AS:
i:47 nM:i:1
>gene1:6-56     0   chr0   25   255   50M     *   0   0   GTTCATGCGTATTTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTTAA   AAAAAAAA>FFFFFF>FFFFFFFFFFFFFF<FFFF+FFFFFFFFF@:FFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:8-58     0   chr0   27   255   50M     *   0   0   TCATGCGTATTTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTTAATT   AAAAAAAAFFFFF8+F<FFFFFFFFFFFFFFFFFFF5FFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:14-64    0   chr0   33   255   50M     *   0   0   GTATTTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTTAATTTAGAAA   AA6AAAA8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:17-67    0   chr0   37   255   1S49M   *   0   0   ATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTTAATTTAGAAAAAA   AAAAAAAA86FFFFFFFFFFFFFFFFFFFFFF@FFFFFFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:48 nM:i:0
>gene1:19-69    0   chr0   38   255   50M     *   0   0   TTTCTACAGTCGGGTAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATA   AAAAAAAAFFFFFFFFFF>FFFFFF5F<FF@FFFFFFFFFF@FFFFFFF>   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:19-69    0   chr0   38   255   50M     *   0   0   TTTCTACAGTCGGATAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATA   AAAAAA7AFFF5FFFFF:FFFFFFFFFF@FFFFFFFFFFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:47 nM:i:1
>gene1:26-76    0   chr0   45   255   50M     *   0   0   AGTCGGGTAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATACAGGTGT   A+AAAAAAFFFFFFFFF+FFFFFFFFFFFFFFFFFFF8FFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:26-76    0   chr0   45   255   50M     *   0   0   AGTCGGGTAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATACAGGTGT   AA>AAAAAFF7FFFF@FFFFFFFFFFFFFFFFFFF8FFFFFFFFF>FF>FF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:28-78    0   chr0   47   255   50M     *   0   0   TCGGGTAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATACAGGTGTTG   8AA@AAAAFFFFFFFFFF:FFF7FFFFF5F<FFF<F:F5FFFFF<@FFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:33-83    0   chr0   52   255   50M     *   0   0   TAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATACAGGTGTTGGTTTC   A6AAAAAAFFFF7FFFFFFFFFFFF:FFFFFFFFFFFFFF+FFFFFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:33-83    0   chr0   52   255   50M     *   0   0   TAGCAAAGTATAACTGGATTTAATTTAGAAAAAAATACAGGTGTTGGTTTC   :AAAAAAAFFFFFF5FFFF<>FFFFFFFFFFFFFFFFFFFFFFFFFFFFF>   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:39-89    0   chr0   58   255   50M     *   0   0   AGTATAACTGGATTTAATTTAGAAAAAAATACAGGTGTTGATTTCTAATTA   AA:AAAAAFFFF@FFFFFFFF8FFFFFF+FFFFFF>FFFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:47 nM:i:1
>gene1:45-95    0   chr0   64   255   50M     *   0   0   ACTGGATTTAATTTAGAAAAAAATACAGGTGTTGGTTTCTAATTAGTCGGC   AAAA<A7<FFFFFFFFFF8FFFF@FFFFFFFFFFFFFFFFFFFFF:FFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:45-95    0   chr0   64   255   50M     *   0   0   ACTGGATTTAATTTAGAAAAAAATACAGGTGTTGGTTTCTAATTAGTCGGC   A>AAAAAAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF6FF7FFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:57-107   0   chr0   76   255   50M     *   0   0   TTAGAAAAAAATACAGGTGTTGGTTTCTAATTAGTCGGCGTACGGCCGTTA   AAAAAAAAFFFFFFFFFFFFFFFFFFFF7FF5F6FFFFF>FFFFFFF6FF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:60-110   0   chr0   79   255   50M     *   0   0   GAAAAAAATACAGGTGTTGGTTTCTAATTAGTCGGCGTACGGCCGTTACAT   AAAAAAAA:FFFF+FFF6FFFF+FF+FFFFFFFFFFFFFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:66-116   0   chr0   85   255   50M     *   0   0   ATACAGGTGTTGGTTTCTAATTAGTCGGCGTACGGCCGTTACATTATTCG   AAAAAAAAFFFFFFF8FFFFFFF<FFFFF8FFFF8FFFFFFFFFFFFF8F   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:67-117   0   chr0   86   255   50M     *   0   0   TACAGGTGTTGGTTTCTAATTAGTCGGCGTACGGCCGTTACATTATTCGT   AAAAAAAAFFFFF>FFFFFFFFFF5:6FFFFFFFFFFFFFFFFFFFFFFF   NH:i:1   HI:i:1   AS:
i:49 nM:i:0
>gene1:67-117   0   chr0   86   255   48M2S   *   0   0   TACAGGTGTTGGTTTCTAATTAGTCGGCGTACGGCCGTTACATTATTCAT   AAAAAA:AFFFFFFFFFF8FFFFFFFFF>FFFFFFFF:FF+FFF@FFF5F   NH:i:1   HI:i:1   AS:
i:47 nM:i:0
>gene1:68-118   0   chr0   87   255   50M     *   0   0   ACAGGTGTAGGTTTCTAATTAGTCGGCGTACGGCCGTTACATTATTCGTG   AAAAAAAAFFFFF>FFFFFFF>F7FFFFFFFFF6FFFFFF:FF@FFFFFF   NH:i:1   HI:i:1   AS:
i:47 nM:i:1
>gene1:70-120   0   chr0   89   255   49M1S   *   0   0   AGGTGTTGGTTTCTAATTAGTCGGCGTACGGCCGTTACATTATTCGTGTA   AAAA5AAAF:FFFFFFFFFFFFFF8FFFFF8FFFFFFFFFFFFF5FFFFF   NH:i:1   HI:i:1   AS:
i:48 nM:i:0
```


Seq
Demux
Fastq
Map

Mapping: Output format: SAM format

```
>gene1:3-53     0       chr0    22      255     50M     *       0       0       TTTGTTCATGCGTATTTTTCTACAGTCGGGTAGCAAAGTATAACTGGATT
AAAAAA5AFFFFFFFFF@FFFFFFFF7FFFFFFFFFFFFFFFF7FFFFFF      NH:i:1  HI:i:1  AS:i:49 nM:i:0
```

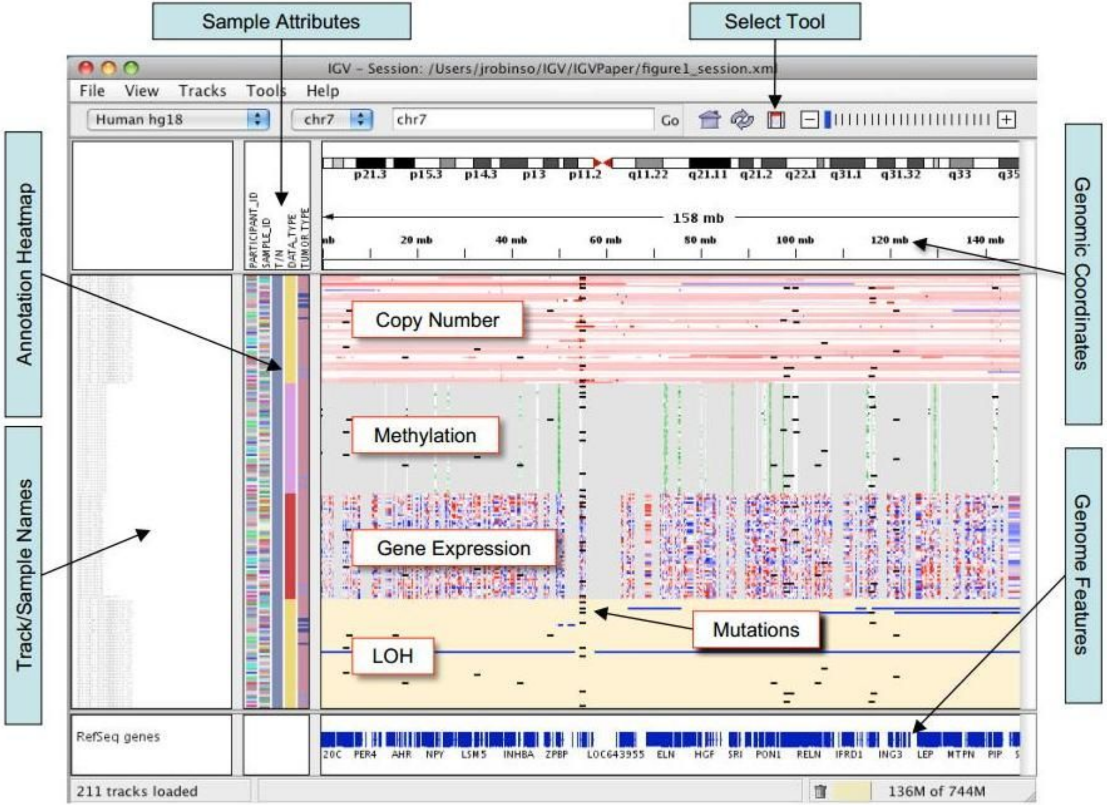
>gene1:3-53	Name of the sequence
0	The flag (containing information about mapped/unmapped, forward/reverse, paired, info of the paired read)
chr0	The chromosome
22	Start position
255	Mapping quality (phred based, 0 always unmapped, below 10 multi mapping. Old scheme: 255: uniquely mapped)
50M	CIGAR string (info about the alignment)
*	Chromosome of the paired read (no pair here)
0	Start position of the paired read
0	Template length or insert size (calculated from the start from the first read, to the end of the second read)
TTTGTTCA….	Sequence
AAAAAA5A….	Quality



Bam file information

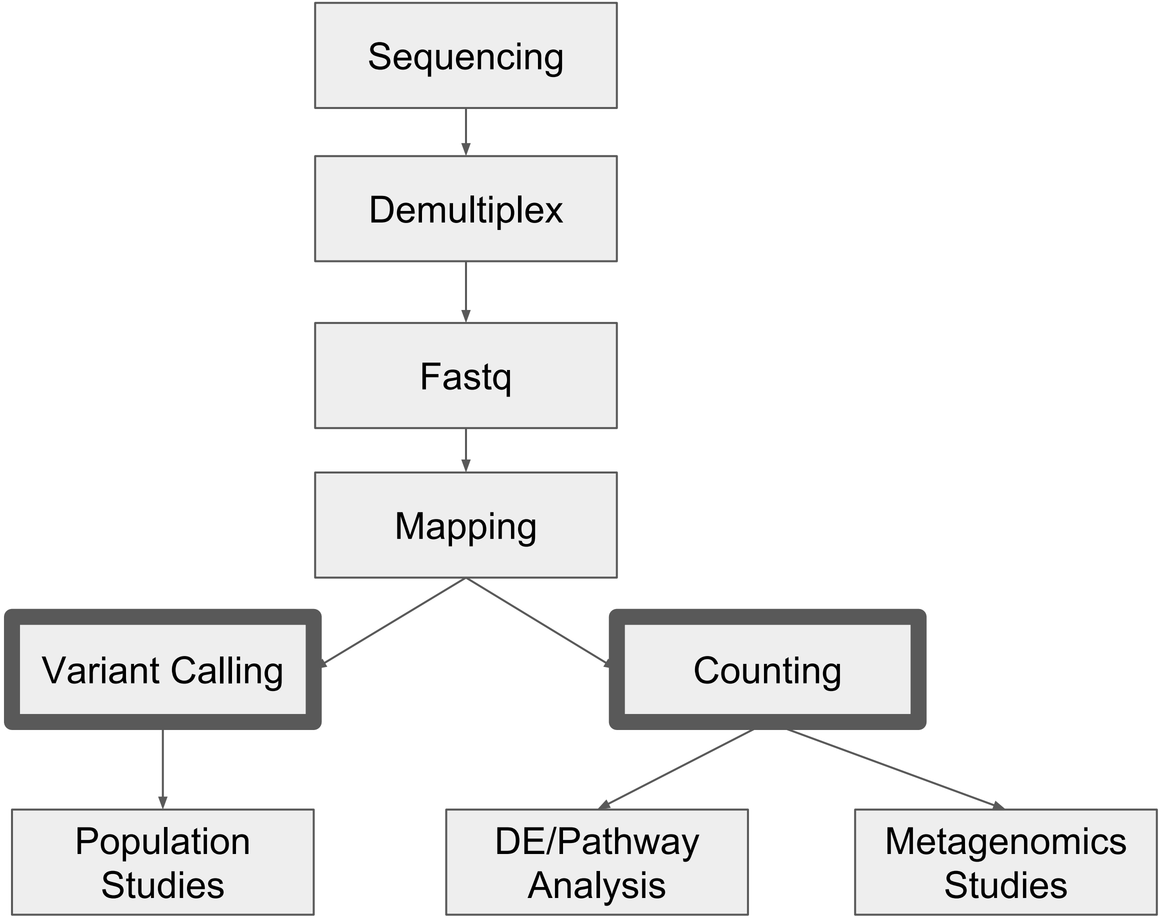
- Visible in IGV

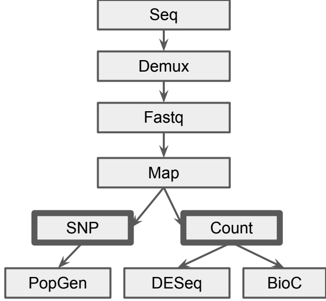
http://software.broadinstitute.org/software/igv/


Sample Attributes
Select Tool
Annotation Heatmap
Track/Sample Names
Genomic Coordinates
Genome Features
Copy Number
Methylation
Gene Expression
Mutations
LOH
IGV - Session: /Users/jrobinso/IGV/IGVPaper/figure1_session.xml
File View Tracks Tools Help
Human hg18
chr7
chr7
Go
PARTICIPANT_ID
SAMPLE_ID
T/N
DATA_TYPE
TUMOR TYPE
p21.3 p15.3 p14.3 p13 p11.2 q11.22 q21.11 q21.2 q22.1 q31.1 q31.32 q33 q35
158 mb
mb 20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb
RefSeq genes
20C PER4 AHR NPY LSM5 INHBA ZPBP LOC643955 ELN HGF SRI PON1 RELN IFRD1 ING3 LEP MTPN PIP
211 tracks loaded
136M of 744M

Downstream analysis

Outline


Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies

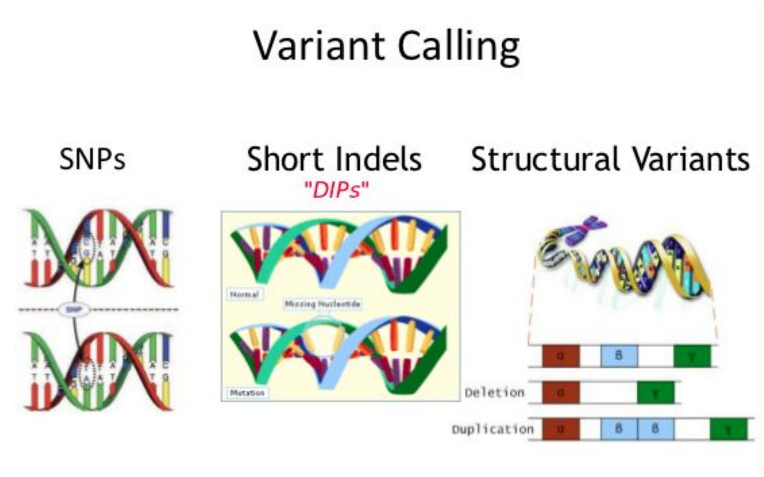


Downstream analysis

Variant Analysis

Based on the alignment of the read

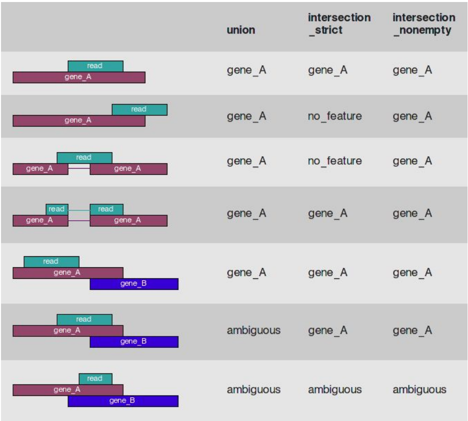
- Variant Calling
  - SNPs, Small Indels, Small polymorphisms
- Structural Variants
  - CNVs, Large Indels



Counting methods

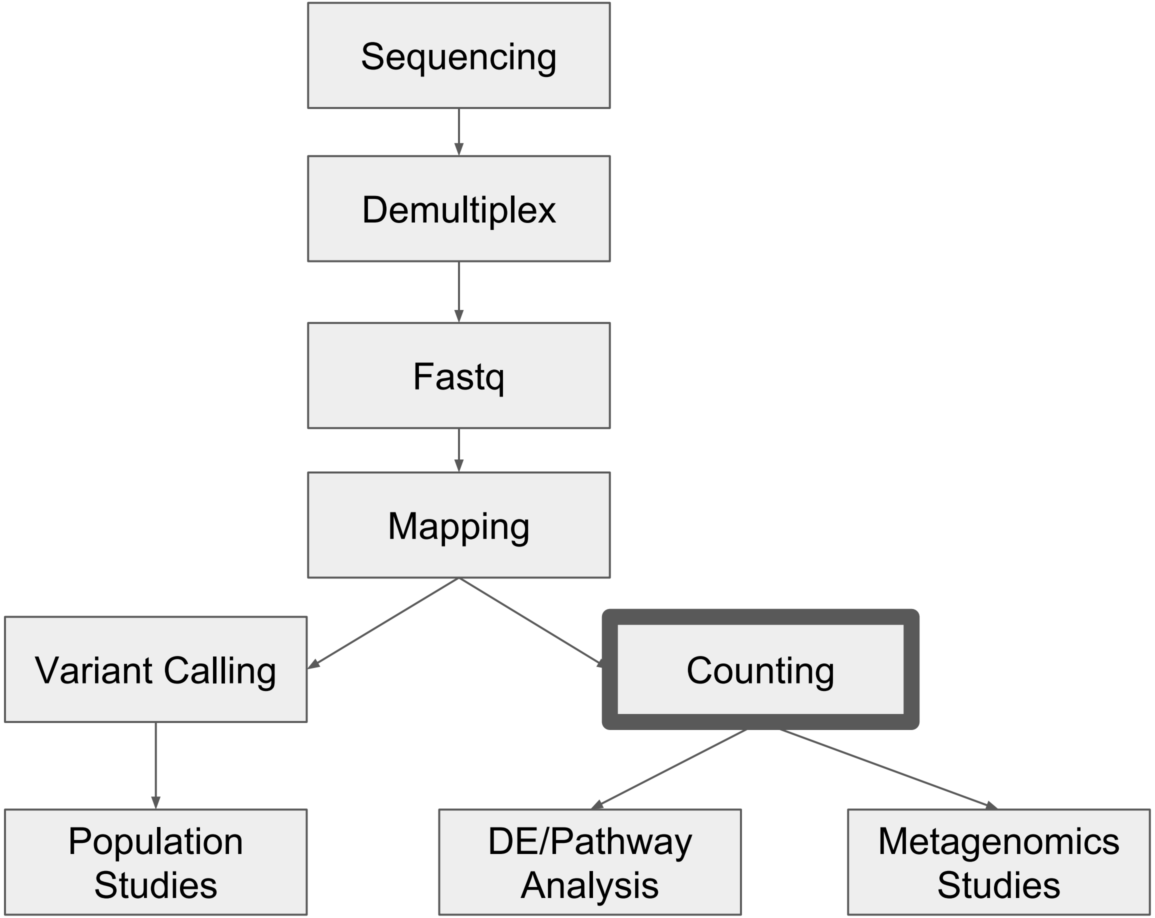
Based on the position of the read

- Differential expression
- Structural Variants
- 16S community analysis

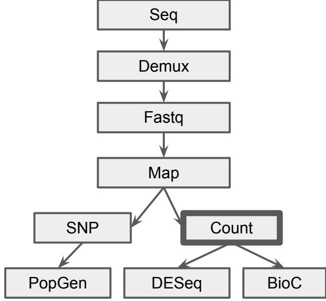


	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

Outline


Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies

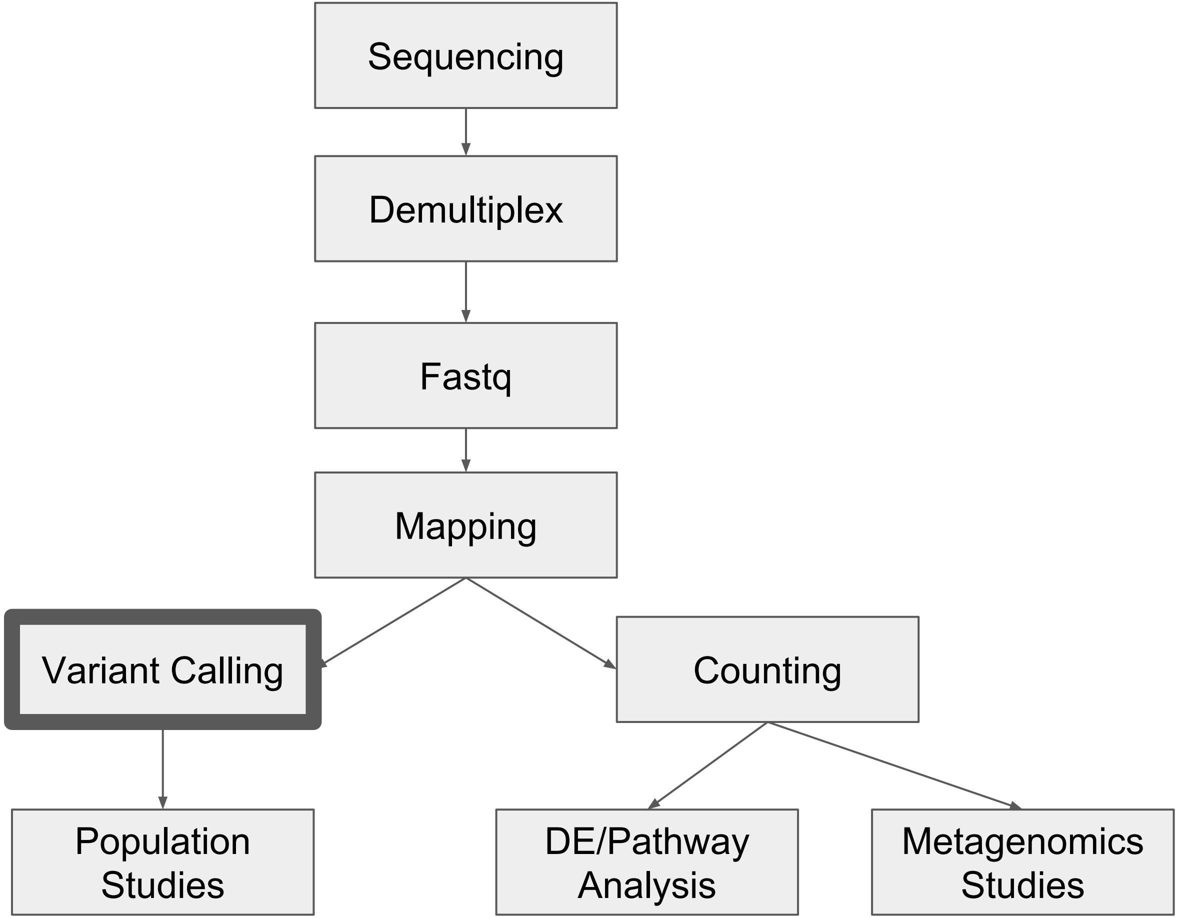
Counting Methods: Where to start?



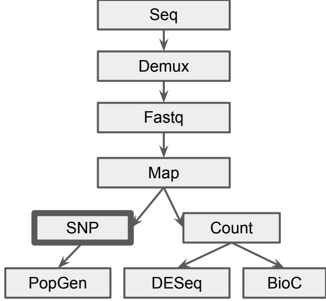
Needed information:

1. The mapped dataset in bam format
2. The definitions of your locations of interest on the reference
   a. must be the same reference as used for the mapping
   b. usually in gff or bed format
      i. gff is used for annotation (location of the genes, exons, ...)
      ii. bed format is used for locations, often more “self” defined

Outline


Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies

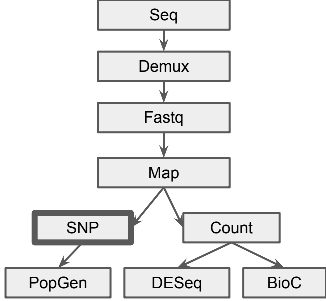
Variant Calling: Where to start?



Needed information:

1. The mapped dataset in bam format
2. The reference
    a. Must be the same as used for the mapping

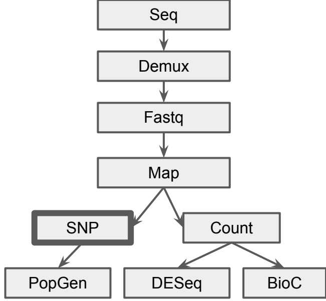
Variant Calling



Detecting variation at 1 position:

- Depth
- Frequency
- Genotype Quality

Variant Calling



Seq

Demux

Fastq

Map

SNP

Count

PopGen

DESeq

BioC

Detecting variation on 1 position:

- Depth
  - Low depth:
    - impossible to distinguish between:
      - error and true variant
      - homozygous or heterozygous

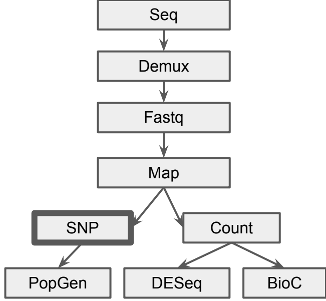
AGCTGAG
AGC**G**GAG
AGCTGAG

Depth: 3
Frequency: ⅓
error, variant?

AGCTGAG
AGC**G**GAG
AGCTGAG
AGCTGAG
AGC**G**GAG

Depth: 5
Frequency: 2/5
=> variant

- Frequency
- Genotype Quality

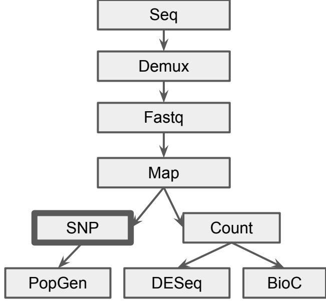


Variant Calling

Detecting variation on 1 position:

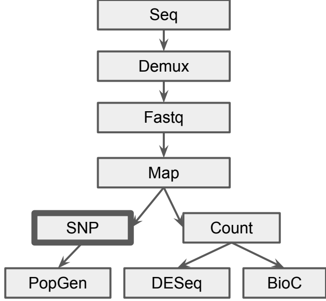
- Depth
- Frequency
  - Within sample:
    - low or high frequency (<5 or >95%) can be an error or mosaik
    - heterozygous genotypes (A/T) are never distributed 50-50!!!!
  - Within and between population:
    - very low frequency can be a sequencing/PCR error in a single sample
    - very high frequency of alternative allele can be an error in the reference
- Genotype Quality

Variant Calling



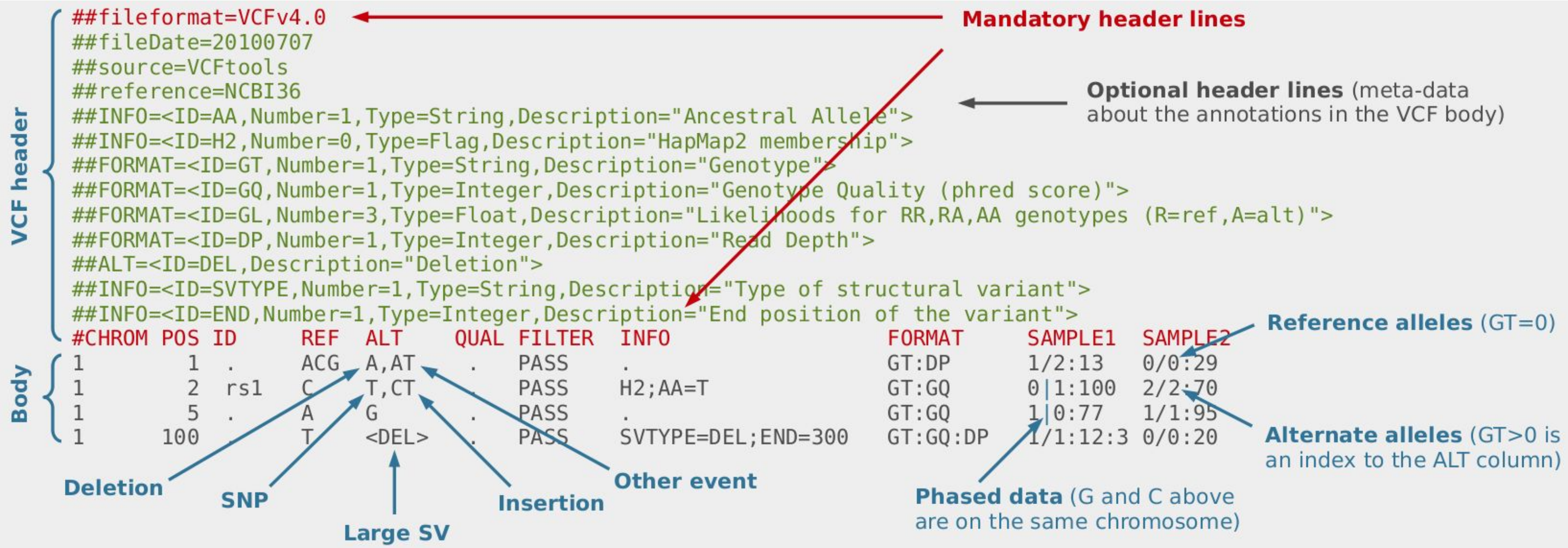
Detecting variation on 1 position:

- Depth
- Frequency
- Genotype Quality
  - per sample:
    - combined metric of depth, frequency, base and mapping quality
    - usually preferred over depth filter
  - multiple sample:
    - combination of GQ per sample, and frequency between samples
    - some bad samples can have a big influence



Variant Calling

Format: vcf (variant calling file)



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT     QUAL FILTER  INFO                  FORMAT     SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT    .    PASS    .                     GT:DP      1/2:13    0/0:29
1        2 rs1    C    T,CT    .    PASS    H2;AA=T               GT:GQ      0|1:100   2/2:70
1        5 .      A    G       .    PASS    .                     GT:GQ      1|0:77    1/1:95
1      100 .      T    <DEL>   .    PASS    SVTYPE=DEL;END=300    GT:GQ:DP   1/1:12:3  0/0:20
```

VCF header

Body

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

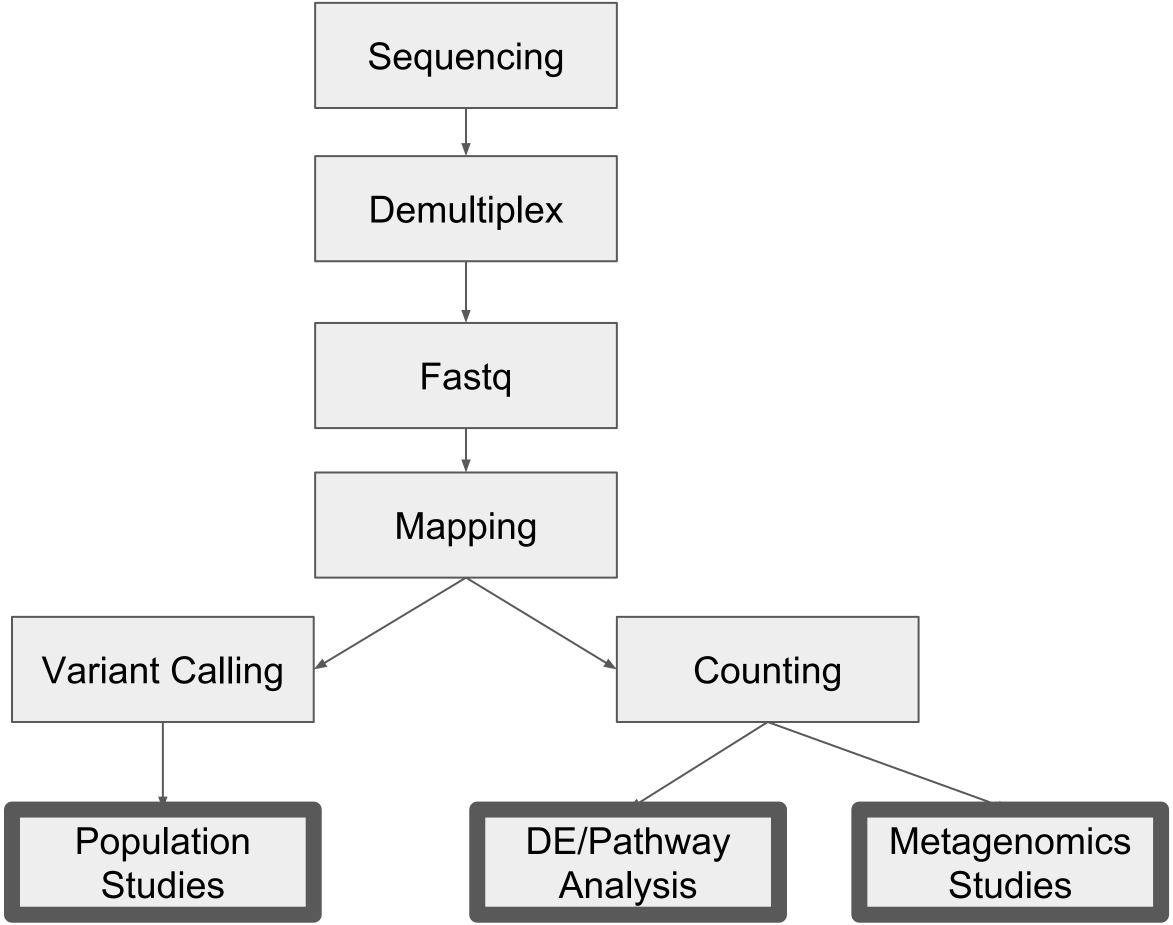
Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

Outline


Sequencing
Demultiplex
Fastq
Mapping
Variant Calling
Counting
Population Studies
DE/Pathway Analysis
Metagenomics Studies

Introduction to NGS Bioinformatics

Are you a biologist, a medical doctor or perhaps a bio-engineer interested on learning the basic bioinformatics skills and techniques that will open the door for analyzing your own NGS data?

The Genomics Core Leuven organizes its first workshop of the year on **March 31, 2017** on the topic of NGS bioinformatics. NGS data formats, reads mapping and variant calling will be among the covered topics.

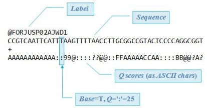
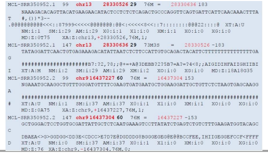
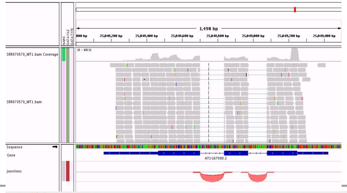
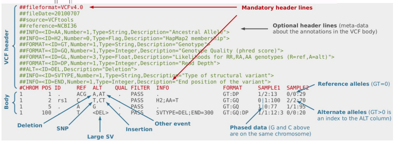
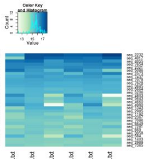
Interested on joining the NGS informatics revolution?

More information and **registration as from March 5, 2017** in: www.genomicscore.be

GENOMICS
CORELEUVEN

File formats and Tool Overview

File Formats, a small overview

	File Format	PipeLine	Tools	Previous step	Interesting for	
Fastq	Fastq	All (Fastq)	bcl2fastq (demultiplexing)	SEQUENCING	raw data	
Mapped Data DNA	BAM	exome, target, amplicon	BWA Bowtie2	Fastq	mapped vs reference genome	
Mapped Data RNAseq	BAM	rnaseq, mirnaseq, quantseq	STAR (split read) Bowtie2 (short read)	Fastq	mapped vs reference genome	
Genotyping	VCF Excel	exome, target, amplicon	GATK	Mapped Data (DNA or RNA)	SNP and variants	
Differential Expression	csv Excel	rnaseq, mirnaseq, quantseq	htseq-count, EdgeR, DESeq2	Mapped Data RNAseq	Different expression of genes, compared between 2 conditions	

```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

```
MCL-SRR350952.1   99   chr13    28330526 29   76M =    28330636 183
    NAAAGACACAGTTACATGAAGAACATACTCCTCTCTCAGACTGCCCAGGTTCAGTGATTCATTCAACAAACTTTA
T   #,())*3--
.@@@@@@@@@@<<<::87999<<<<<@@@@@@@:@@<:<<<<<8<<::7::::::::@@@22::::@  XT:A:U
    NM:i:1   SM:i:29  AM:i:29  X0:i:1   X1:i:0   XM:i:1   XO:i:0   XG:i:0
    MD:Z:0T75    XA:Z:chr13,+28330526,76M,1;
MCL-SRR350952.1   147  chr13    28330636 29   73M3S    =   28330526 -183
    TATAGGATTCAACTGTGAGAAAGACATATTAATCTCTTCCATTGTGCAGACTACATTCTTTTTTTTTTTTTTTGA
G
    #####################B7:?2,?8;;@+=+A@3DEBB?2?5B7=A?=?4<8;;AIGIDIHFAIIGHIIBI
I   XT:A:M   NM:i:2   SM:i:29  AM:i:29  XM:i:2   XO:i:0   XG:i:0   MD:Z:18A18G35
MCL-SRR350952.2   99   chr9 16437227 60   76M =    16437304 153
    NGAAATGCAAGGCTGTTTGGGATGTTTTCGAAGTGATGAATGCTGGAAGGATTGCTGTTCTCTAAGTGAGCAAGG
A
    ###########################################################################
#   XT:A:U   NM:i:1   SM:i:37  AM:i:37  X0:i:1   X1:i:0   XM:i:1   XO:i:0   XG:i:0
    MD:Z:0A75    XA:Z:chr9,+16437227,76M,1;
MCL-SRR350952.2   147  chr9 16437304 60   76M =    16437227 -153
    GCTGGGACTCCTGGTGCGATTATTGCTCTCAATGAAAGTCCTTATATCTGAGTCTGTCTTTGAAGATGGTACAGC
C
    DBAEA<>G>GGDGG<DD3E<CDCC>E?D?E@DGDDDG8BGGGEGEG@E@@BCCFEE,IHIIGEGGEFCCF<FFFF
D   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:76  XA:Z:chr9,-16437304,76M,0;
```

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT     QUAL FILTER INFO                FORMAT      SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT    .    PASS   .                   GT:DP       1/2:13    0/0:29
1        2 rs1    C    T,CT    .    PASS   H2;AA=T             GT:GQ       0|1:100   2/2:70
1        5 .      A    G       .    PASS   .                   GT:GQ       1|0:77    1/1:95
1      100 .      T    <DEL>   .    PASS   SVTYPE=DEL;END=300  GT:GQ:DP    1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Tools: fastq manipulation

- Quality Control
  - fastqc
- Inline barcode demultiplexing
  - GBSX
- Adaptor and quality trimming
  - trimmomatic
  - bbduk
  - ea-utils
  - fastx-toolkit
- Contamination check
  - fastq-screen

Tools: Mapping

- Some reads (like 20)
  - BLAST
  - BLAT
- Illumina DNA
  - BWA-mem
  - Bowtie2
  - bbmap
- Illumina RNA (split read mapping)
  - STAR
  - HiSat2
- PacBio
  - BLASR
  - BWA-mem
  - BWA-sw

Tools: Mapping manipulations

- BAM/SAM manipulations (sorting, duplicate marking)
  - elPrep
  - SAMtools
  - BAMtools
  - Picard
  - CRAMtools
- Quality control
  - Picard
  - BAMtools
  - QualiMap
- Visualisation
  - IGV

Tools: Counting methods

- RNAseq
  - htseq-count
  - bbcount
- CNV
  - seqCBS

Tools: Variant Calling

- Variant calling:
  - GATK haplotype caller
  - FreeBayes
  - SAMtools
  - Varscan
- Haplotyping
  - FreeBayes
- Filtering and manipulations:
  - vcflib
  - vcftools