LEIDEN UNIVERSITY MEDICAL CENTER

QC and accounting in GAPSS3

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

In *exome sequencing*, we select genomic regions of interest using a *target-enrichment strategy*.

- PCR.
- On array capture.
- In-solution capture.

Introduction

In *exome sequencing*, we select genomic regions of interest using a *target-enrichment strategy*.

- PCR.
- On array capture.
- In-solution capture.

Overview of an in-solution capture.

- Fragmentation.
- Size selection.
- Linker ligation.
- Capture.

Introduction

In *exome sequencing*, we select genomic regions of interest using a *target-enrichment strategy*.

- PCR.
- On array capture.
- In-solution capture.

Overview of an in-solution capture.

- Fragmentation.
- Size selection.
- Linker ligation.
- Capture.

These regions are then *sequenced*.

- Illumina Genome Analyser II (GAII).
- Illumina HiSeq 2000.

Introduction

The Illumina Genome Analyser II.

Introduction

The Illumina Genome Analyser II.

- Manufacturer: Illumina, Inc.
- Commercially available since 2005.
- Per cycle, one base is read.
- Reads up to 100×2 base pairs.
- Takes about 8 days.
- Produces about 40 Giga bases per run.

Introduction

The Illumina Genome Analyser II.

- Manufacturer: Illumina, Inc.
- Commercially available since 2005.
- Per cycle, one base is read.
- Reads up to 100 $\times$ 2 base pairs.
- Takes about 8 days.
- Produces about 40 Giga bases per run.

Pros:

- Does paired end sequencing.
- Cheap.

Introduction

The Illumina HiSeq 2000.

Introduction

The Illumina HiSeq 2000.

- Manufacturer: Illumina, Inc.
- Commercially available since 2010.
- Per cycle, one base is read.
- Reads up to 150×2 base pairs.
- Takes about 8 days.
- Produces about 150 Giga bases per run.

Introduction

The Illumina HiSeq 2000.

- Manufacturer: Illumina, Inc.
- Commercially available since 2010.
- Per cycle, one base is read.
- Reads up to 150×2 base pairs.
- Takes about 8 days.
- Produces about 150 Giga bases per run.

Pros:

- Even higher throughput.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.
4. Filtering.
    - Post-variant calling quality control.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.
4. Filtering.
    - Post-variant calling quality control.
5. Annotation.
    - Post-annotation quality control.

Pre-alignment

We use the FASTX toolkit for data cleaning.

- Remove linker sequences.
- Clip low quality reads at the end of the read.
- Judge the read that is left over.

Pre-alignment

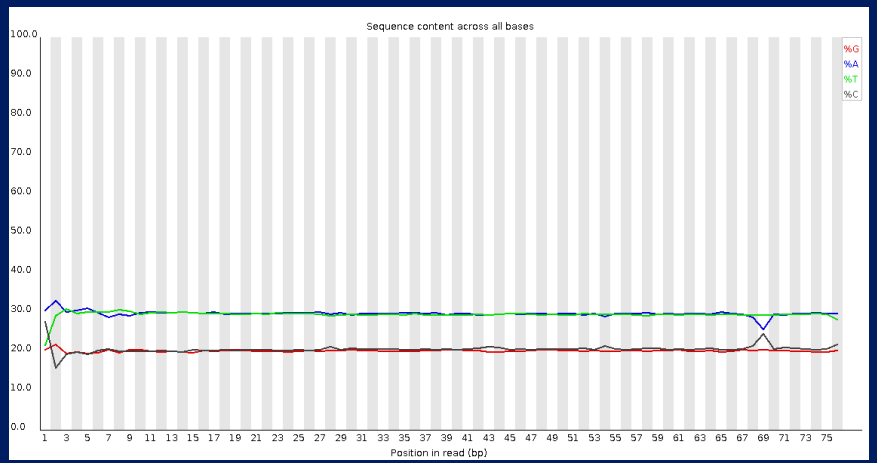
We use the FASTX toolkit for data cleaning.

- Remove linker sequences.
- Clip low quality reads at the end of the read.
- Judge the read that is left over.

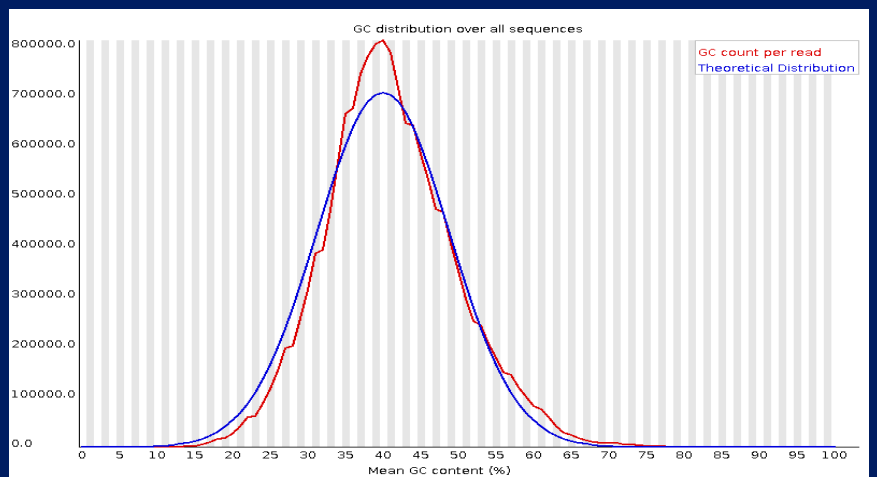
The FASTQC toolkit is used for quality control (both before and after the data cleaning step).

- GC content.
- GC distribution.
- Quality scores distribution.
- ...

Pre-alignment

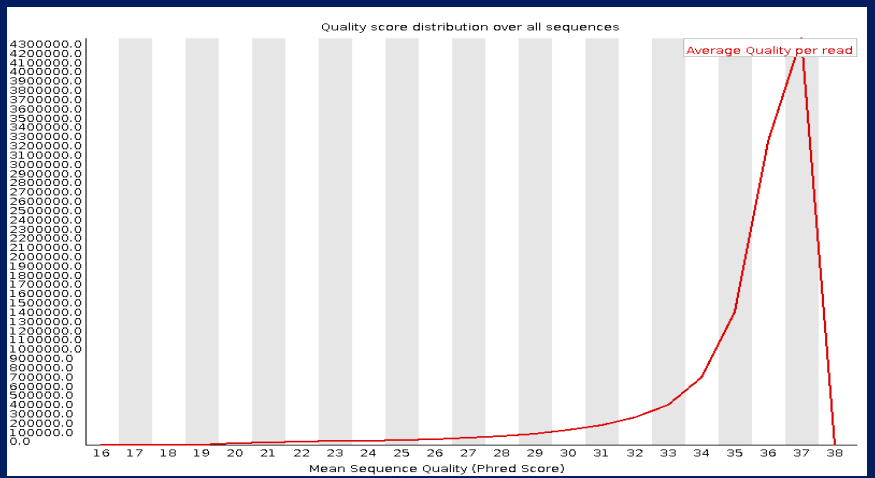


Per base sequence content.

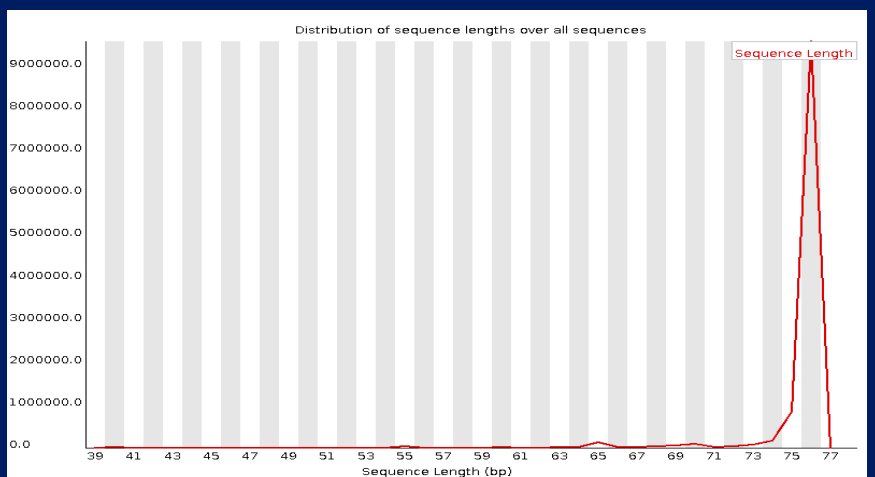


Per sequence GC content.

Pre-alignment



Per sequence quality.



Sequence length distribution.

Alignment

Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.

Alignment

Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.

Some features:

- Base quality recalibration.
  - First map 1% of the input.
  - Recalibrate the Fastq quality scores.
  - Redo the alignment with the recalibrated scores.

Alignment

Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.

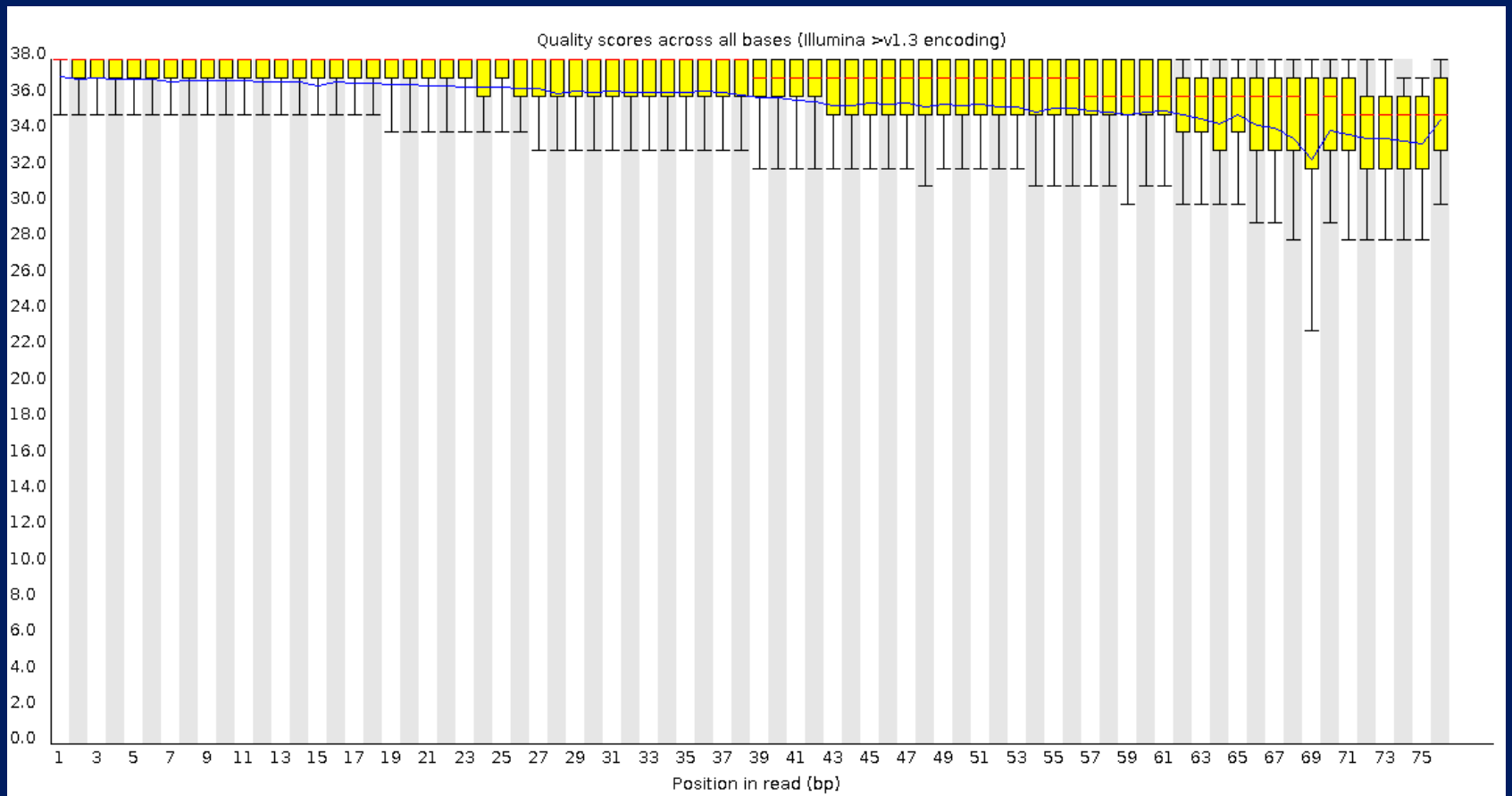
Some features:

- Base quality recalibration.
  - First map 1% of the input.
  - Recalibrate the Fastq quality scores.
  - Redo the alignment with the recalibrated scores.
- Uses BWA for the hard work.
  - Switches to its accurate built in aligner when BWA fails.

Burrows-Wheeler Aligner (BWA) is a short read aligner that allows small insertions and deletions.

Alignment

Base quality recalibration.



Variant calling

Variant calling is done by Samtools, BCFtools / VCFutils.

The output of most modern aligners is in *Sequence Alignment / Map* (SAM) format.

Variant calling

Variant calling is done by Samtools, BCFtools / VCFutils.

The output of most modern aligners is in *Sequence Alignment / Map* (SAM) format.

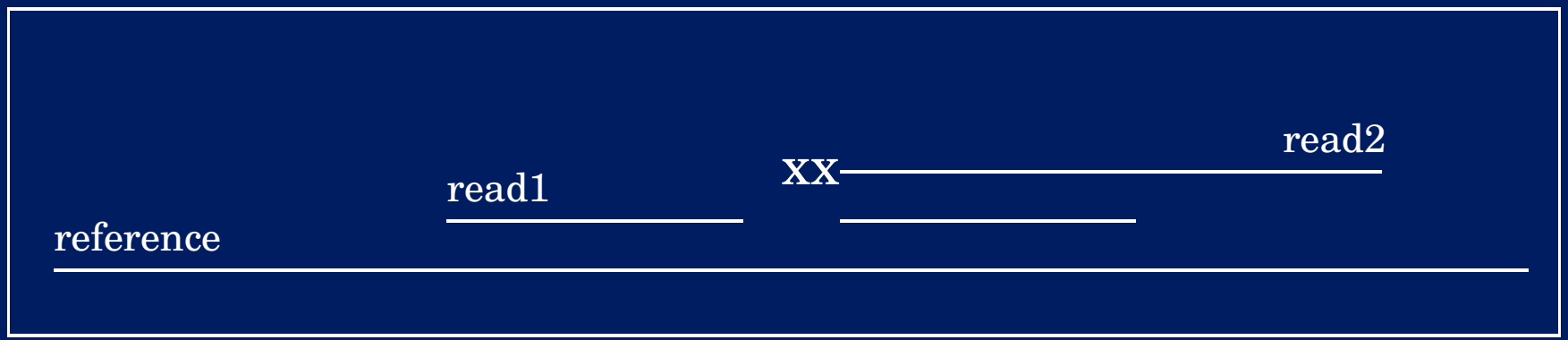
Mainly file format conversions.

- SAM → BAM.
- BAM → BAM.sorted.
- BAM.sorted → BAM.sorted.index.
- BAM.sorted → mpileup (BAQ realignment).
- BAM.sorted → BCF.
- BCF → VCF.

We end up with a list in *Variant Call Format* (VCF).

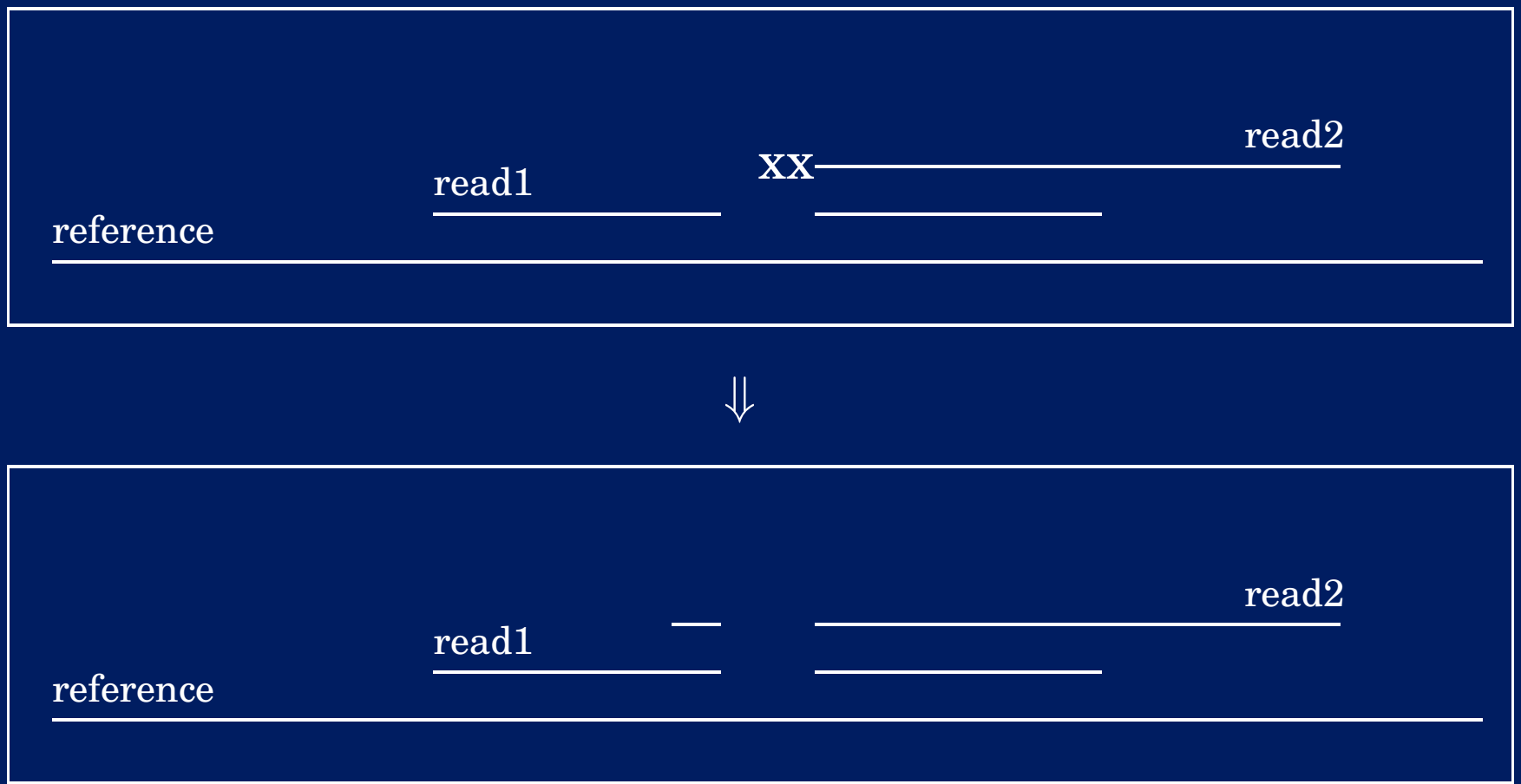
Variant calling

Base Alignment Quality (BAQ) realignment:
Remove SNPs around indels.



Variant calling

Base Alignment Quality (BAQ) realignment:
Remove SNPs around indels.



Filtering

Samtools varfilter.

- Minimum coverage threshold.
- Strand bias.
- Quality scores.
- Maximum coverage threshold.
  - Copy number variation.
  - Alignment artefacts.

Filtering

Samtools varfilter.

- Minimum coverage threshold.
- Strand bias.
- Quality scores.
- Maximum coverage threshold.
  - Copy number variation.
  - Alignment artefacts.

Transition transversion rates:

- Around 2.1 human full genome.
- Around 2.8 to 3.0 human exome.

Filtering

Samtools varfilter.

- Minimum coverage threshold.
- Strand bias.
- Quality scores.
- Maximum coverage threshold.
    - Copy number variation.
    - Alignment artefacts.

Transition transversion rates:

- Around 2.1 human full genome.
- Around 2.8 to 3.0 human exome.

Still working on:

- Maximum coverage per region.
    - Probe affinity can vary greatly.

Annotation

We use five annotation sources.

- Seattle Seq.
- Ensembl.
- Mutalyzer / SVEP.
- LOVD.
- In house database.

Annotation

We use five annotation sources.

- Seattle Seq.
- Ensembl.
- Mutalyzer / SVEP.
- LOVD.
- In house database.
  - HGMD data.
  - 1000 genomes project.
  - Genome of the Netherlands (250 triplets).
  - All variants called by this pipeline.

Annotation

We use five annotation sources.

- Seattle Seq.
- Ensembl.
- Mutalyzer / SVEP.
- LOVD.
- In house database.
  - HGMD data.
  - 1000 genomes project.
  - Genome of the Netherlands (250 triplets).
  - All variants called by this pipeline.
    - Coverage per variant.
    - Number of reads supporting the variant.
    - Horizontal coverage per sample.

Annotation

We use five annotation sources.

- Seattle Seq.
- Ensembl.
- Mutalyzer / SVEP.
- LOVD.
- In house database.
  - HGMD data.
  - 1000 genomes project.
  - Genome of the Netherlands (250 triplets).
  - All variants called by this pipeline.
    - Coverage per variant.
    - Number of reads supporting the variant.
    - Horizontal coverage per sample.

dbSNP rate should be around 90% (human exome).

Technical details

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).

Technical details

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.

Technical details

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.
- Database to keep track of the versions of all used tools and custom scripts.
  - If one or more tools are upgraded, the new versions are stored.
  - The version number of the pipeline is incremented.
  - The versions of all tools of all pipeline versions can be retrieved from this database.

Technical details

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.
- Database to keep track of the versions of all used tools and custom scripts.
  - If one or more tools are upgraded, the new versions are stored.
  - The version number of the pipeline is incremented.
  - The versions of all tools of all pipeline versions can be retrieved from this database.
- LaTeX documentation is automatically generated.
  - Compiled to pdf that can be handed over to the customer.

Technical details

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.
- Database to keep track of the versions of all used tools and custom scripts.
  - If one or more tools are upgraded, the new versions are stored.
  - The version number of the pipeline is incremented.
  - The versions of all tools of all pipeline versions can be retrieved from this database.
- LaTeX documentation is automatically generated.
  - Compiled to pdf that can be handed over to the customer.
- All individual commands are logged.

Questions?

Acknowledgements

Michiel van Galen
Yu-Ching Lai
Martijn Vermaat
Bradley ten Broeke
Jaap van der Heijden
Michel Villerius
Matthew Hestand
Johan den Dunnen

http://www.lgtc.nl