LEIDEN UNIVERSITY MEDICAL CENTER

The Dutch Variant Database.

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

The Dutch Variant Database.

Goal: Share variants for annotation purposes.

Introduction

The Dutch Variant Database.

Goal: Share variants for annotation purposes.

Prototype by:

- Hubrecht / UMCU Utrecht.
- UMCN Nijmegen.
- LUMC Leiden.

Introduction

The Dutch Variant Database.

Goal: Share variants for annotation purposes.

Prototype by:

- Hubrecht / UMCU Utrecht.
- UMCN Nijmegen.
- LUMC Leiden.

Other participants:

- UMCG Groningen.
- AMC Amsterdam.
- EMC Rotterdam.
- VUMC Amsterdam.
- NBIC.

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*.

- ABI SOLiD 5500.
- Illumina HiSeq 2000.

Introduction

Figure 1: SOLiD 5500XL.

Figure 2: HiSeq 2000.

Paired end, high throughput, cheap.

Introduction

Raw data aquisition.

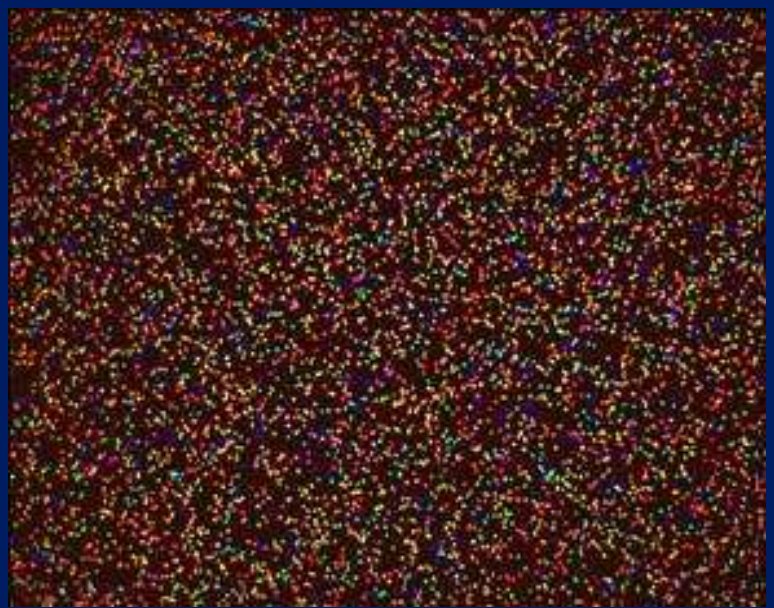


Figure 3: Raw image.

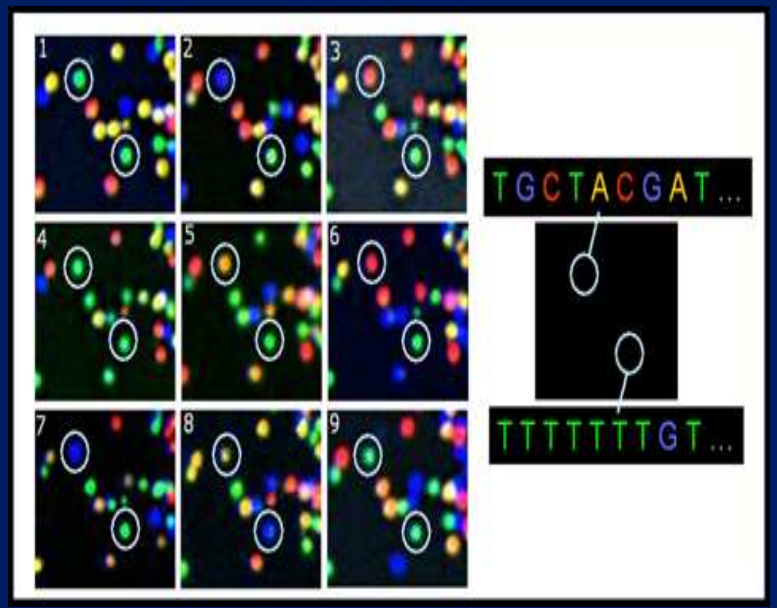


Figure 4: Base calling.

Introduction

Raw data aquisition.

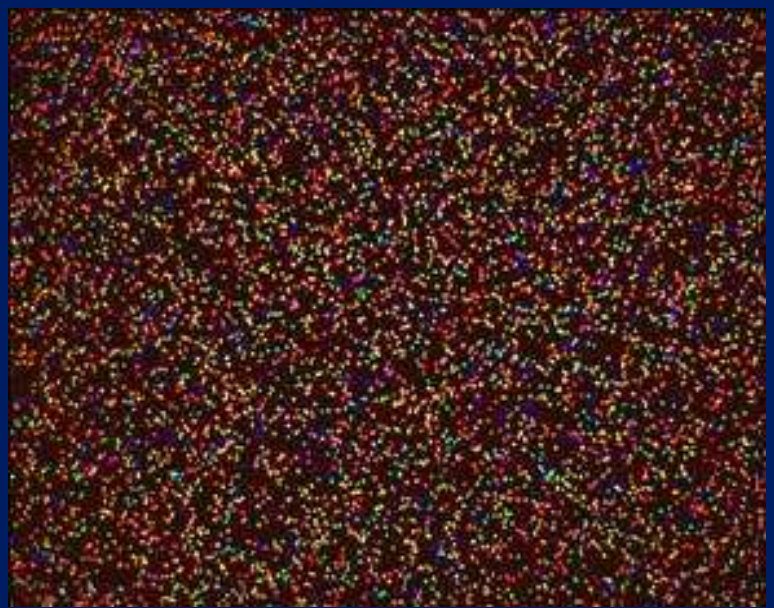


Figure 3: Raw image.

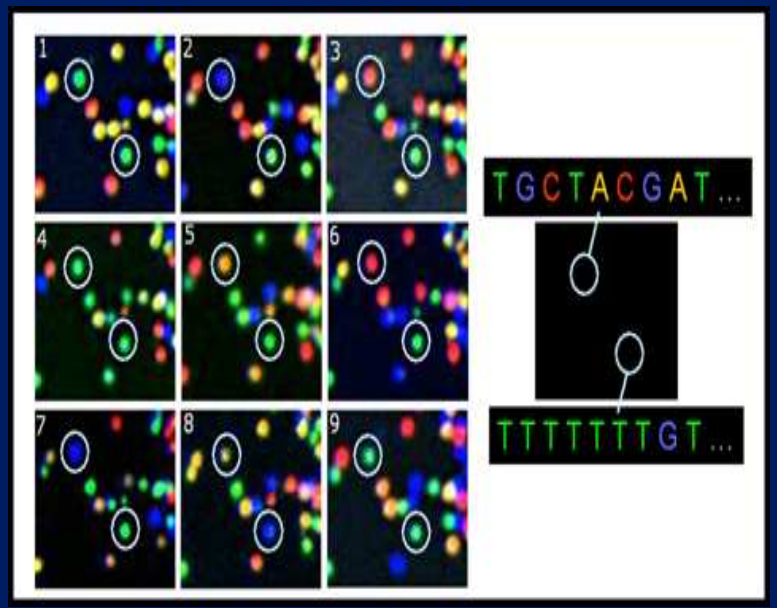


Figure 4: Base calling.

Output:

- Long list of short DNA sequences.
- Quality scores per base.

Analysis pipelines

Exome sequencing pipelines can roughly be divided in five steps.

Analysis pipelines

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.

Analysis pipelines

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.

Analysis pipelines

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.
3. Variant calling.
    - Description of difference with the reference.

Analysis pipelines

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.
3. Variant calling.
    - Description of difference with the reference.
4. Filtering.
    - Variant calling has a high false positive rate.

Analysis pipelines

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.
3. Variant calling.
    - Description of difference with the reference.
4. Filtering.
    - Variant calling has a high false positive rate.
5. Annotation.

Pre-alignment

We use the Trimmomatic / 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 Trimmomatic / 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

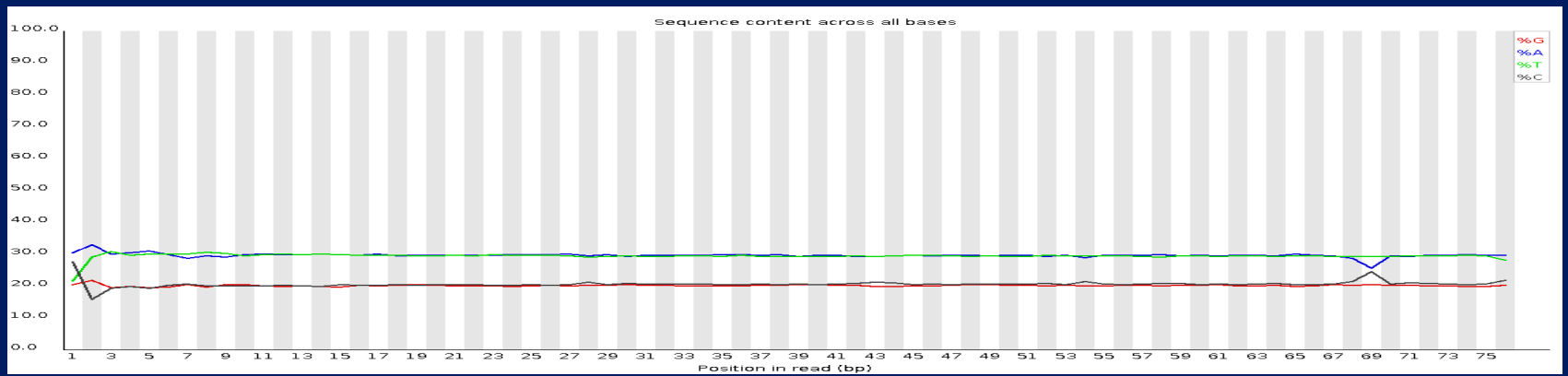


Figure 5: Per base sequence content.

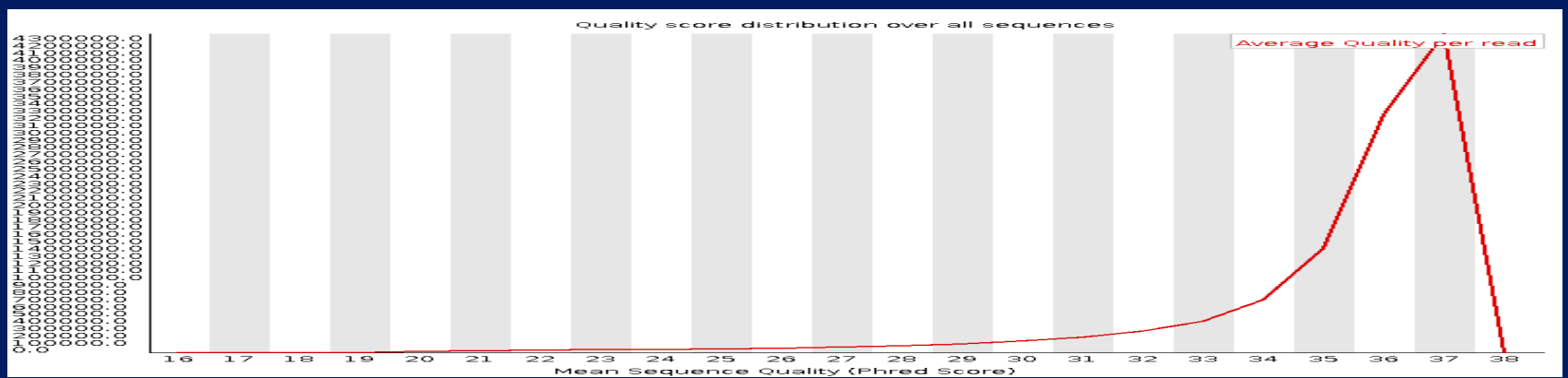


Figure 6: Per sequence quality.

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

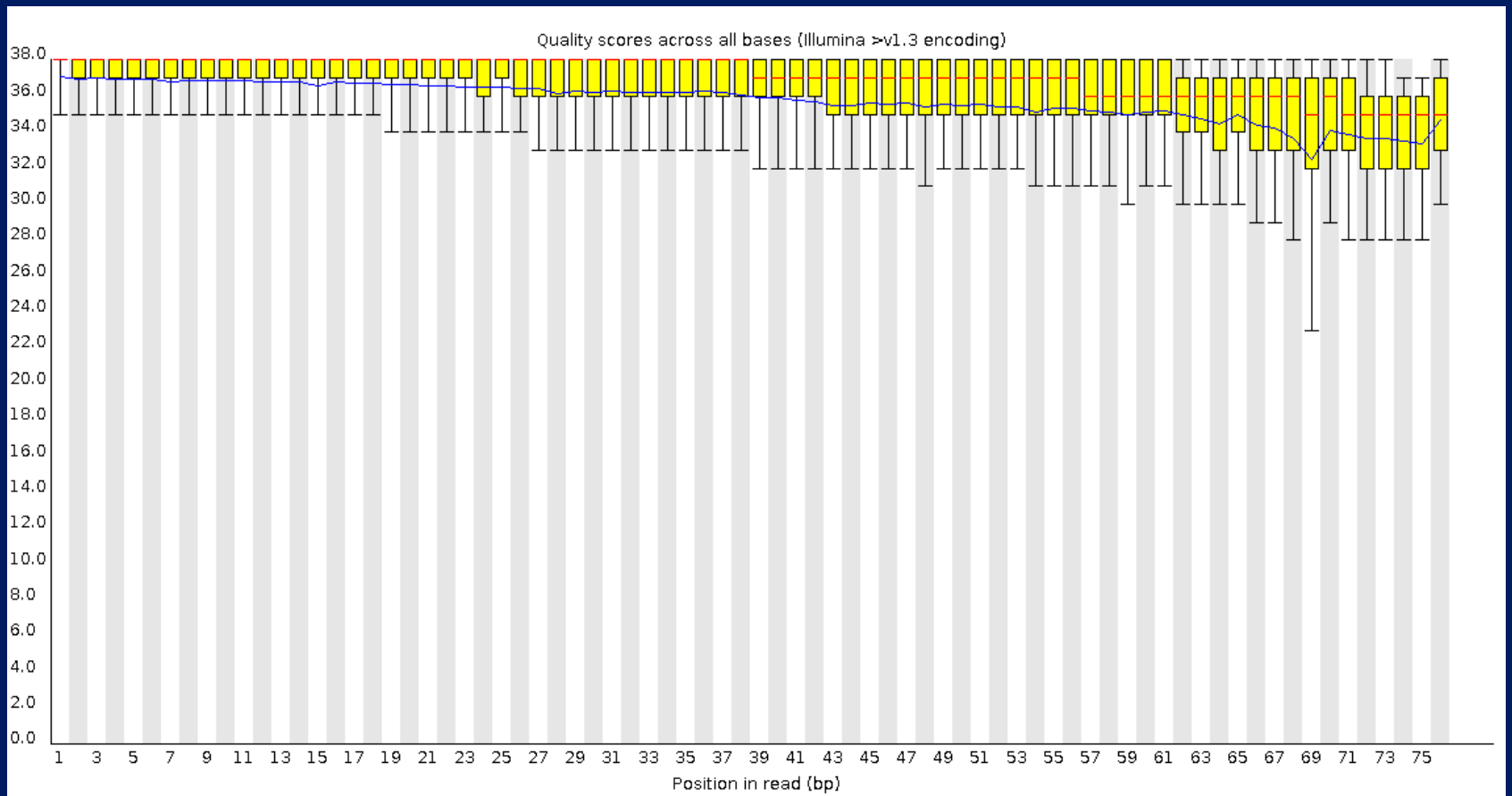


Figure 7: Base quality recalibration.

Variant calling

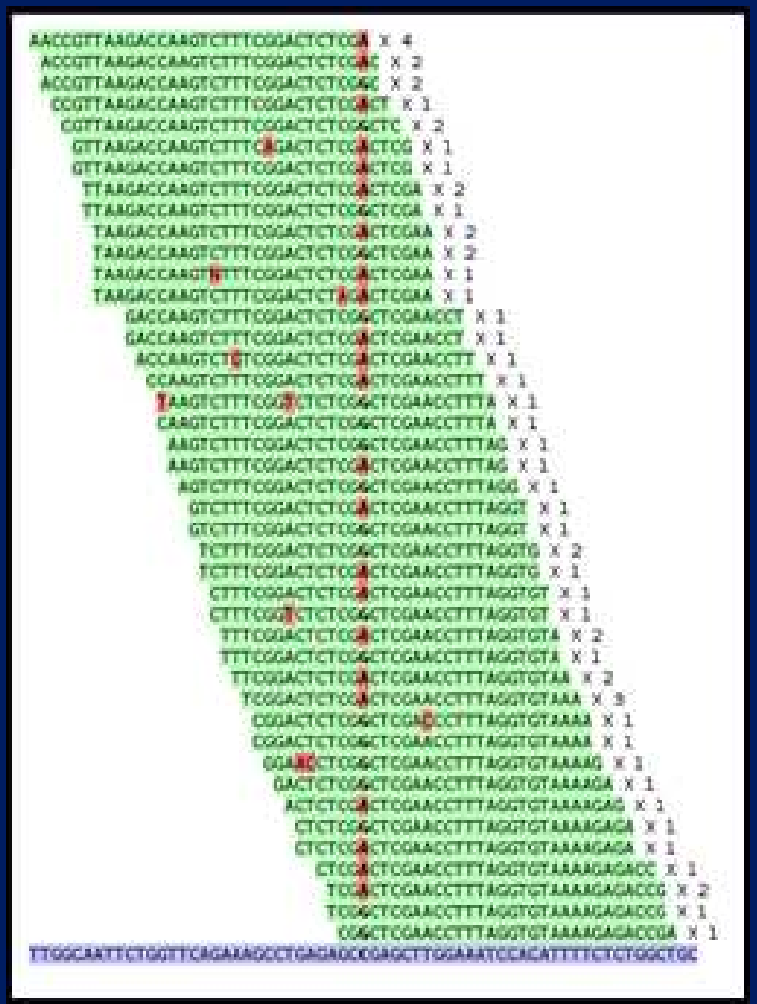


Figure 8: Variant calling.

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.

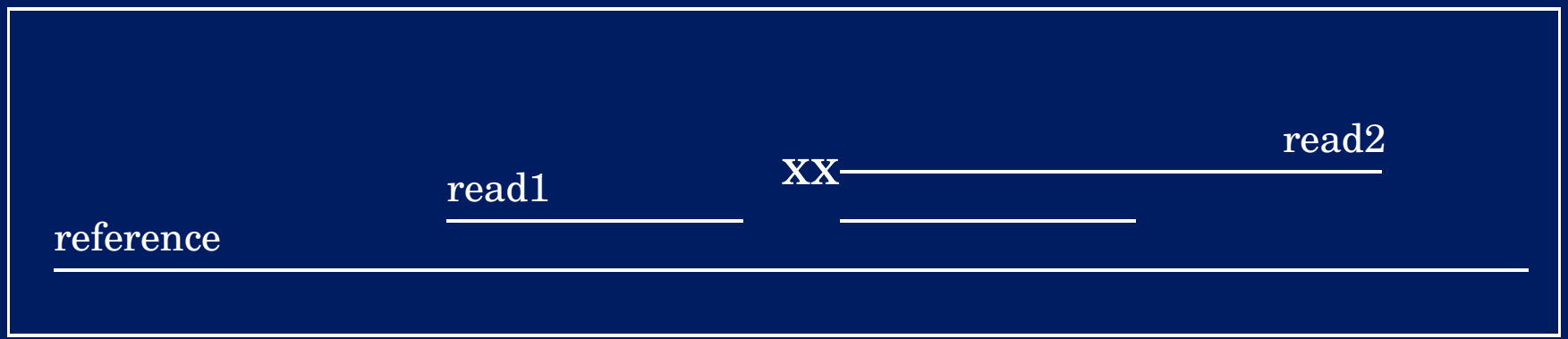
Mainly file format conversions.

- 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).

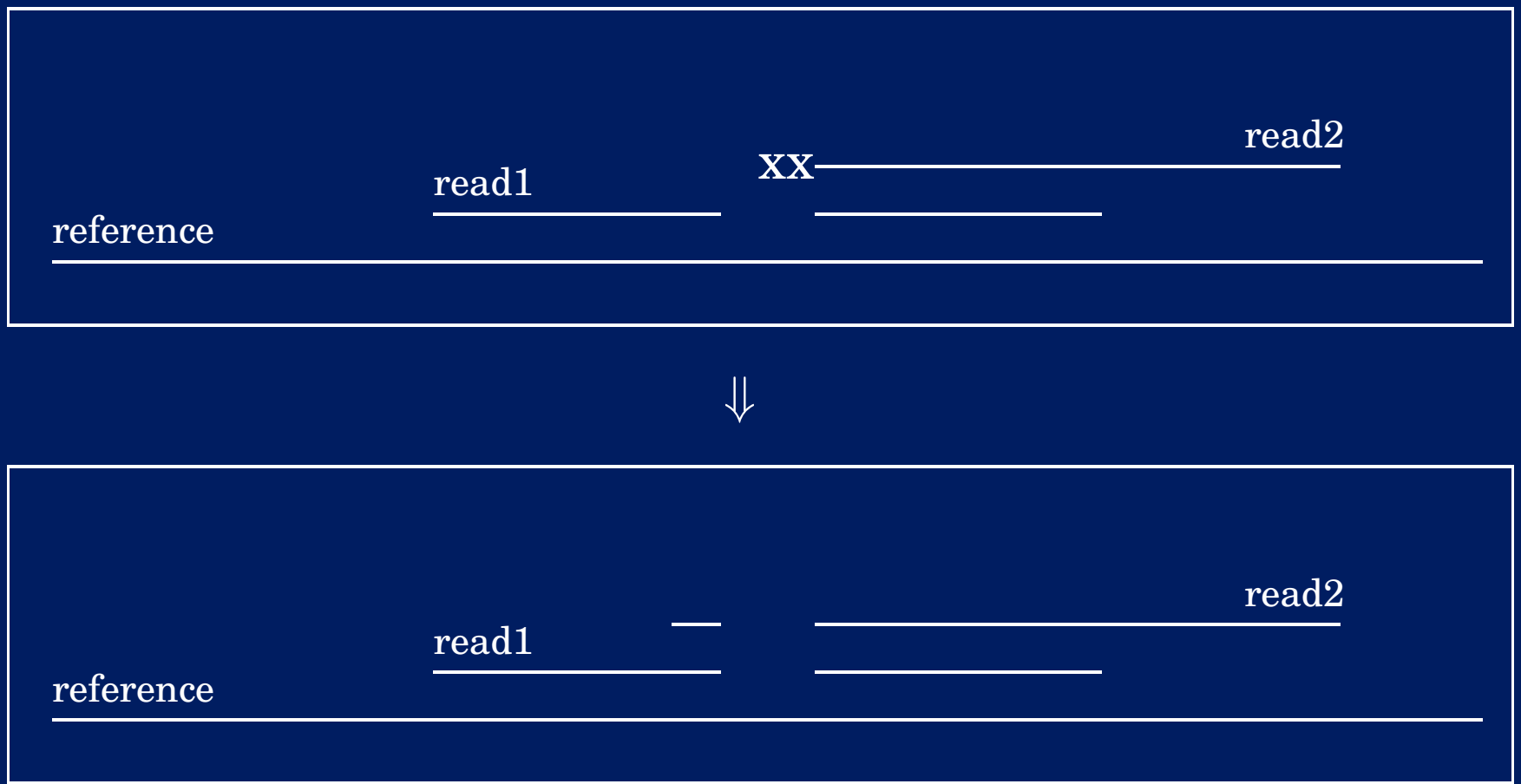
Variant calling

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:

- Transitions: **A** $\Leftrightarrow$ **G** or **C** $\Leftrightarrow$ **T**.
- Transversions: **A** $\Leftrightarrow$ **C** or **G** $\Leftrightarrow$ **T** or **A** $\Leftrightarrow$ **T** or **C** $\Leftrightarrow$ **G**.
- Transitions are more frequent.

Filtering

Samtools varfilter.

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

Transition transversion rates:

- Transitions: **A** $\Leftrightarrow$ **G** or **C** $\Leftrightarrow$ **T**.
- Transversions: **A** $\Leftrightarrow$ **C** or **G** $\Leftrightarrow$ **T** or **A** $\Leftrightarrow$ **T** or **C** $\Leftrightarrow$ **G**.
- Transitions are more frequent.

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

Annotation

We use different annotation sources and an in-house database.

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - . . .

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - …
  - Is it in the 5'/3' UTR of a gene?
  - …

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - …
  - Is it in the 5'/3' UTR of a gene?
  - …
- Is it in a regulatory region?
- …

Pipelines

Combining all these tools in a pipeline:

```
1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1: Shell script

Pipelines

Combining all these tools in a pipeline:

```
1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1: Shell script

```
1  %.sai: %.fq
2    $(BWA) aln -t $(THREADS) $(call MKREF, $@) $< > $@
3
4  %.sam: %.sai %.fq
5    $(BWA) samse $(call MKREF, $@) $^ > $@
6
7  %.bam: %.sam
8    $(SAMTOOLS) view -bt $(call MKREF, $@) -o $@ $<
```

Listing 2: Makefile

Advanced annotation and filtering.

Description of the data.

Exome sequencing output:

- List of variants (VCF format).
- Regions with sufficient coverage (BED format).

Advanced annotation and filtering.

Description of the data.

Exome sequencing output:

- List of variants (VCF format).
- Regions with sufficient coverage (BED format).

We need the coverage information to calculate the frequencies correctly.

- We need to see who *could* have seen a variant.

Advanced annotation and filtering.

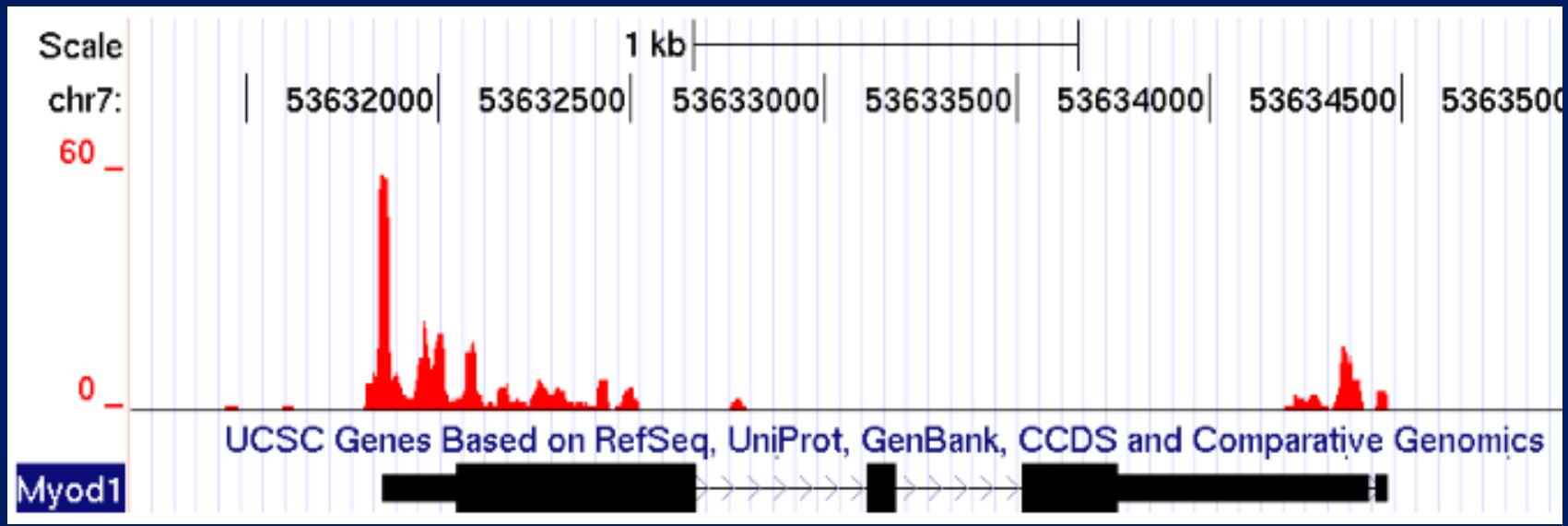


Figure 9: Coverage / depth histogram.

Advanced annotation and filtering.

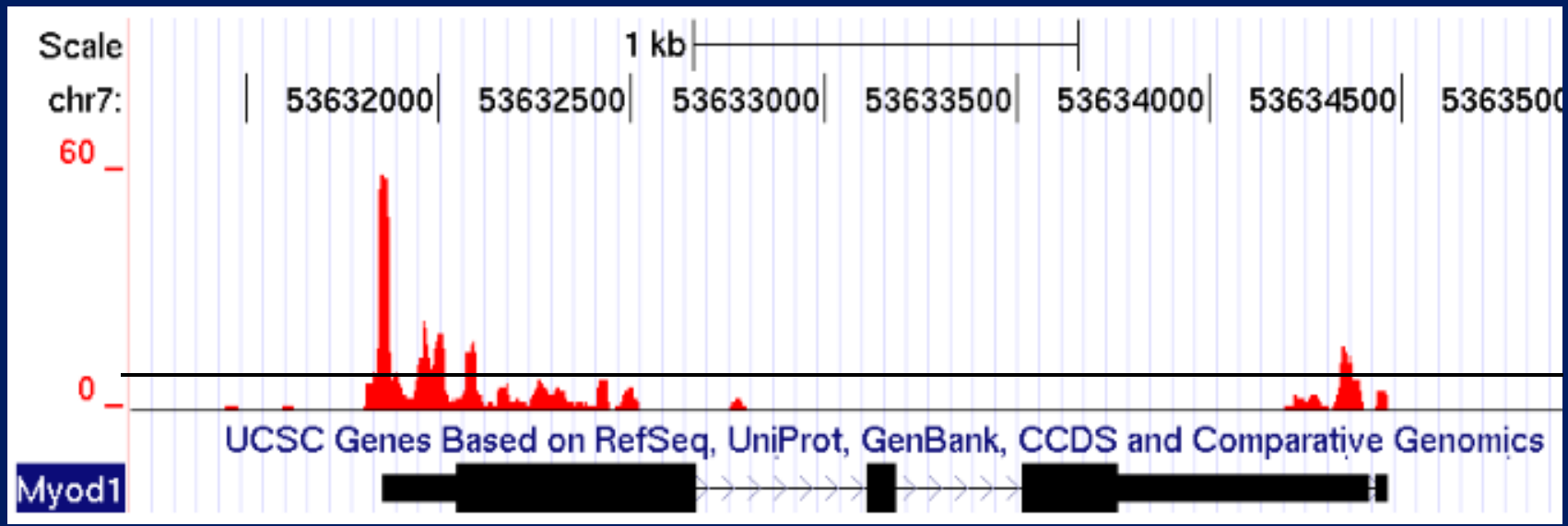


Figure 9: Coverage / depth histogram.

Advanced annotation and filtering.

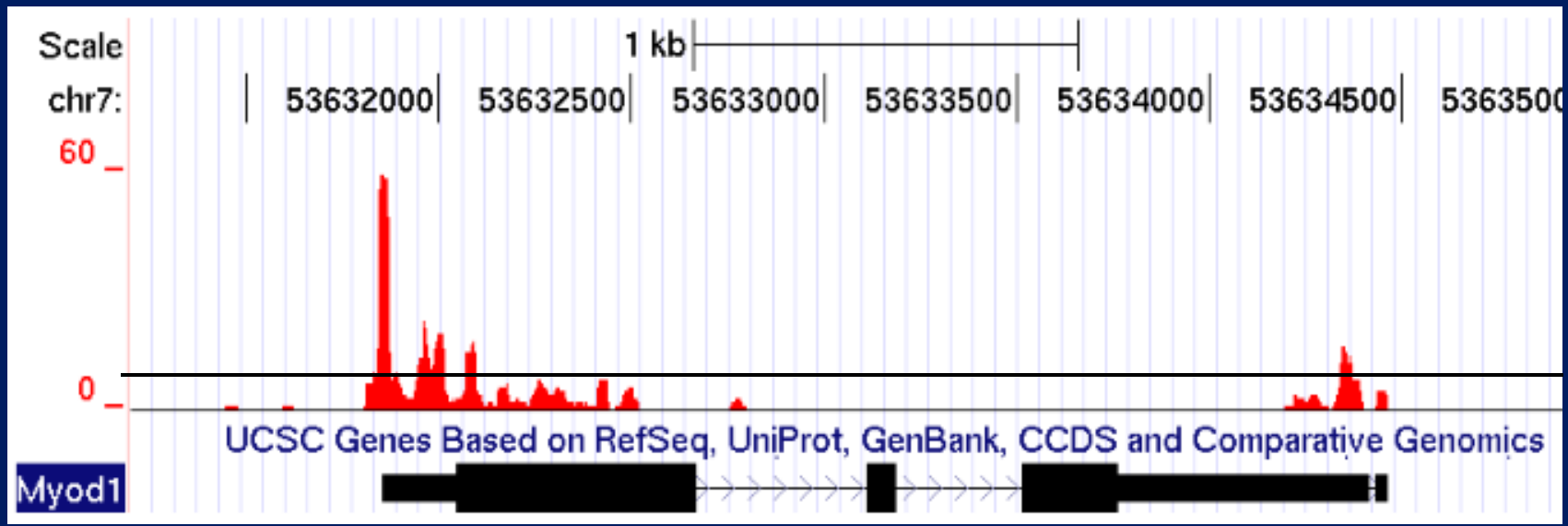


Figure 9: Coverage / depth histogram.

Figure 10: Coverage summary.

National variant database

Proposed setup:

For privacy reasons:

- No browsing.
- Pre-defined interface.
- Encrypted connection.
- Authentication.

National variant database

Proposed setup:

For privacy reasons:

- No browsing.
- Pre-defined interface.
- Encrypted connection.
- Authentication.

General design:

- A SOAP interface for interaction with pipelines.
- SOAP over HTTPS (SSL) for encryption.
- Trivial authentication.
    - Send username / password in each call.

National variant database

Prototype running in Utrecht:

- MySQL database.
- SOAP interface.

National variant database

Prototype running in Utrecht:

- MySQL database.
- SOAP interface.

The exome data can be enriched with other datasets, not available anywhere else:

- Latest 1000 Genomes Project results.
- Genome of the Netherlands (GoNL).
- RefSeq false positives.

National variant database

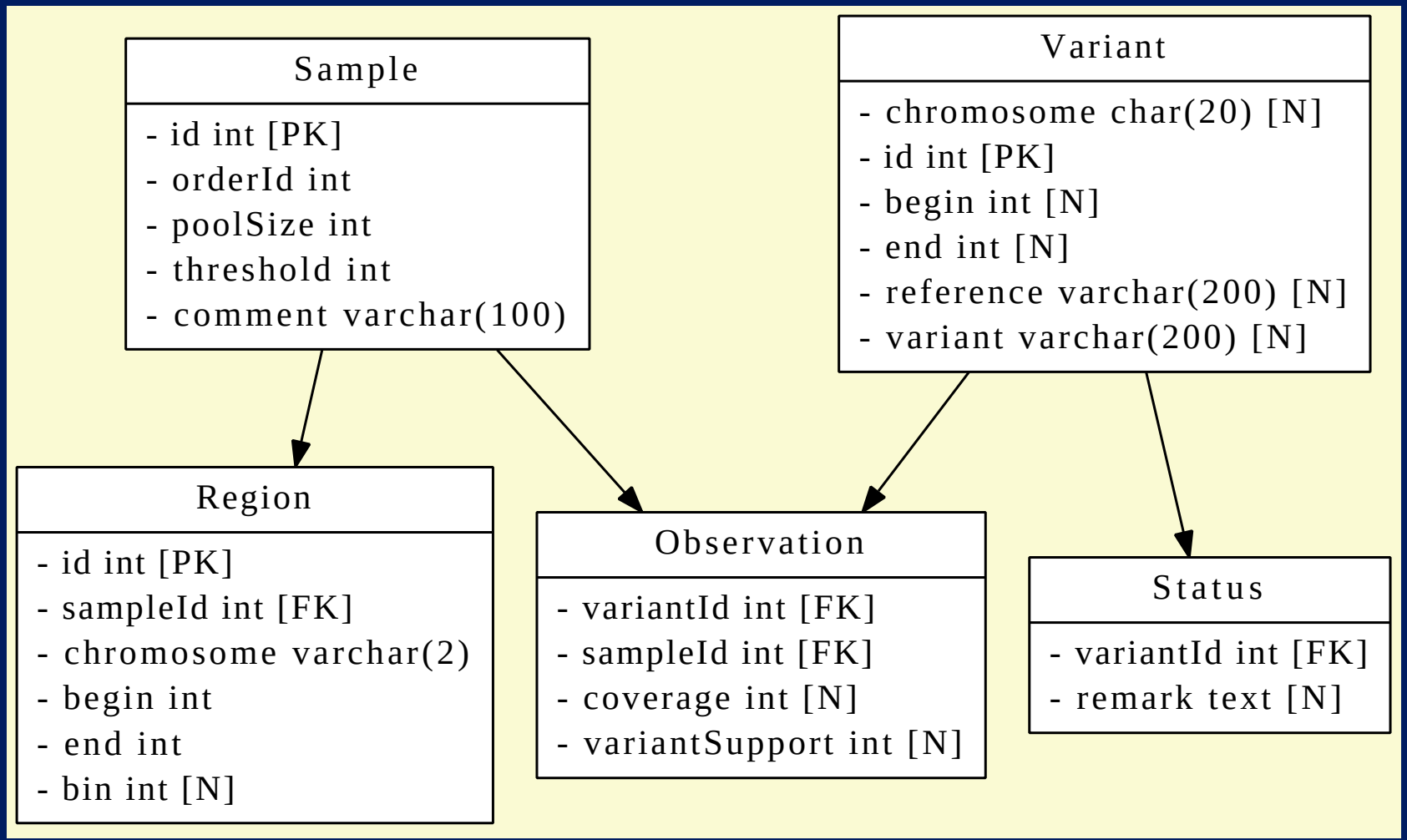


Figure 11: Database schema.

National variant database

Client perspective:

In principle, one program that handles all:

- Upload a VCF and a BED.
- Server responds with an ID.
- Use the ID to poll for the results.

National variant database

Client perspective:

In principle, one program that handles all:

- Upload a VCF and a BED.
- Server responds with an ID.
- Use the ID to poll for the results.

Client characteristics:

- Around 30 lines of code.
- Programming language independent.
- Operating system independent.

National variant database

Server perspective:

Uploading:

- Receive a VCF and a BED.
- Respond with an ID.
- A process is forked that starts annotation and importing.

National variant database

Server perspective:

Uploading:

- Receive a VCF and a BED.
- Respond with an ID.
- A process is forked that starts annotation and importing.

Annotation:

- Receive an ID.
- If the annotation for this ID is finished, return the result.

National variant database

Server perspective:

Uploading:

- Receive a VCF and a BED.
- Respond with an ID.
- A process is forked that starts annotation and importing.

Annotation:

- Receive an ID.
- If the annotation for this ID is finished, return the result.

Note that if the annotation part is finished, the result can already be returned.

National variant database

Column	Name	Description
1	CHROM	Name of the chromosome.
2	POS	Position on the chromosome.
3	REF	Reference allele.
4	ALT	Genotype.
5	Frequency	Frequency of this genotype.
6	Occurrences	Number of hits.
7	IDs ($>$ 100)	If the frequency is low: a list of sample IDs.
8	IDs ($<$ 100)	A list with hits in reserved samples; 1000 genomes, GoNL, HGMD, etc.

Table 1: Output format.

Discussion

Stakeholder meeting.

Discussion

Stakeholder meeting.

- The proposed interface.

Discussion

Stakeholder meeting.

- The proposed interface.
- More sensitive issues.
  - Who gets access?

Discussion

Stakeholder meeting.

- The proposed interface.
- More sensitive issues.
  - Who gets access?
  - What if someone submits the same data twice?

Discussion

Stakeholder meeting.

- The proposed interface.
- More sensitive issues.
  - Who gets access?
  - What if someone submits the same data twice?
  - Can we get all variants in a gene?
    - Maybe only for public samples.

Questions?

Acknowledgements:

Christian Gilissen
Victor Guryev
Frans-Paul Ruzius
Nienke Wieskamp
Hailiang Mei
Johan den Dunnen

https://www.mutalyzer.nl/svn/NGSdata/branches/dvd/