LEIDEN UNIVERSITY MEDICAL CENTER

Diagnostic Variant Database

an overview

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Share your results

Introduction

Share your results

There are a number of ways of storing and sharing your variants:

- LOVD.
- dbSNP.
- ...

Introduction

Share your results

There are a number of ways of storing and sharing your variants:

- LOVD.
- dbSNP.
- ...

Sharing is important:

- If no-one shared, you can not filter.
- Can lead to co-authorship when other people use your data.

Introduction

Figure 1: LOVD welcome screen.

Introduction

Variants in LOVD

Exon	DNA change	Var_pub_as	RNA change	Protein change
6	c.368dupC	-	r.(?)	p.(Met124Asnfs*14)
6	c.372delG	-	r.372del	p.Met124Ilefs*18
6	c.377delA	-	r.(?)	p.(Asn126Ilefs*16)
6	c.379_380del (Reported 4 times)	-	r.379_380del	p.Ile127Hisfs*10
6	c.386delC	-	r.(?)	p.(Ala129Valfs*13)
6	c.392T>A	-	r.(?)	p.(0)
6	c.397C>T	-	r.(?)	p.(Gln133*)
6	c.402_405del (Reported 2 times)	401_404delCCAA	r.(?)	p.(Asn135Valfs*6)
6	c.409G>T	-	r.(?)	p.(Glu137*)
6	c.419T>A (Reported 2 times)	-	r.(?)	p.(Leu140His)
6	c.429G>A	-	r.(?)	p.(Trp143*)
6	c.431T>A (Reported 2 times)	-	r.(?)	p.(Val144Asp)
6	c.433C>T (Reported 19 times)	-	r.(?)	p.(Arg145*)
6	c.433delC	-	r.(?)	p.(Arg145Aspfs*12)
6	c.440C>A	-	r.440c>a	p.Ser147*
6	c.440C>G (Reported 2 times)	-	r.(?)	p.(Ser147*)

Figure 2: Selection of variants.

Introduction

DVD

The *Diagnostic Variant Database*.

- Share variants found in exome sequencing experiments.
- Find functionally relevant variants.

Introduction

DVD

The *Diagnostic Variant Database*.

- Share variants found in exome sequencing experiments.
- Find functionally relevant variants.

Technical details:

- Annotation by sharing.
- Store coverage information to determine reference calls.
- Disambiguation of variant descriptions.
- Pooling without loss of information.
- Duplicate sample detection.
    - Allows for re-annotation without polluting the database.
- Encrypted connection with authentication.

Sequencing

Data generation

Figure 3: SOLiD 5500XL.

Figure 4: HiSeq 2000.

Sequencing

Data analysis

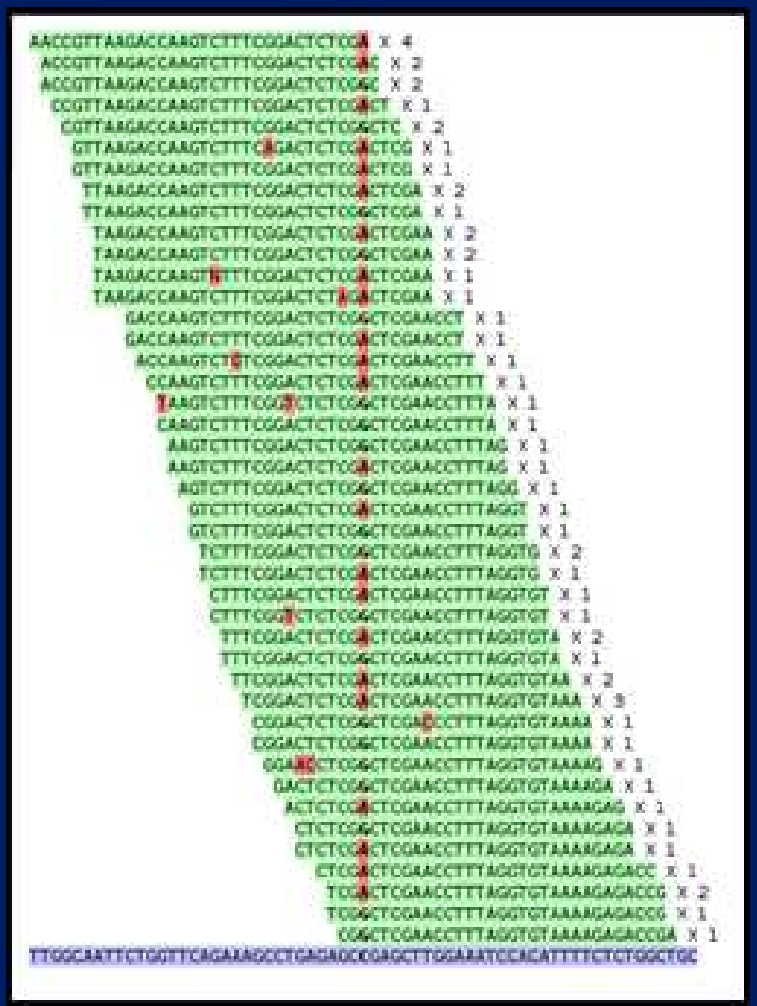

Figure 5: Variant calling.

Formats

Required fields in the VCF format

Field	Explanation.
CHROM	Name of the chromosome.
POS	Position on the chromosome.
ID	List of unique identifiers.
REF	Reference base(s).
ALT	List of alternate non-reference alleles.
QUAL	Phred-scaled quality score for the assertion made in ALT.
FILTER	PASS if this position has passed all filters.
INFO	Additional information.

Table 1: Required fields.

Formats

Disambiguation

One problem with the VCF format is that there are multiple ways of describing the *same* variant.

Original	Stored
`AT` ⇒ `A`	`AT` ⇒ `A`
`ATT` ⇒ `AT`	`AT` ⇒ `A`
`ATTC` ⇒ `ATC`	`AT` ⇒ `A`
`G` ⇒ `AG`	`G` ⇒ `AG`
`GG` ⇒ `AGG`	`G` ⇒ `AG`
`GGA` ⇒ `AGGA`	`G` ⇒ `AG`

Table 2: Resolving ambiguous variant descriptions.

Formats

Disambiguation

One problem with the VCF format is that there are multiple ways of describing the *same* variant.

Original	Stored
`AT` ⇒ `A`	`AT` ⇒ `A`
`ATT` ⇒ `AT`	`AT` ⇒ `A`
`ATTC` ⇒ `ATC`	`AT` ⇒ `A`
`G` ⇒ `AG`	`G` ⇒ `AG`
`GG` ⇒ `AGG`	`G` ⇒ `AG`
`GGA` ⇒ `AGGA`	`G` ⇒ `AG`

Table 2: Resolving ambiguous variant descriptions.

This is done automatically on the server.

Formats

Coverage

We need to store coverage information.

- Which variants *could* we have seen?

Formats

Coverage

We need to store coverage information.

- Which variants *could* we have seen?

Frequency calculation:

- A variant can be present only twice in the database.
- If it is only covered in two samples, the frequency is one.

Formats

Coverage

We need to store coverage information.

- Which variants *could* we have seen?

Frequency calculation:

- A variant can be present only twice in the database.
- If it is only covered in two samples, the frequency is one.

This enables us to store multiple experiment types.

- Different capture kits.
- Full genome sequencing.
- Amplicon resequencing.

Coverage

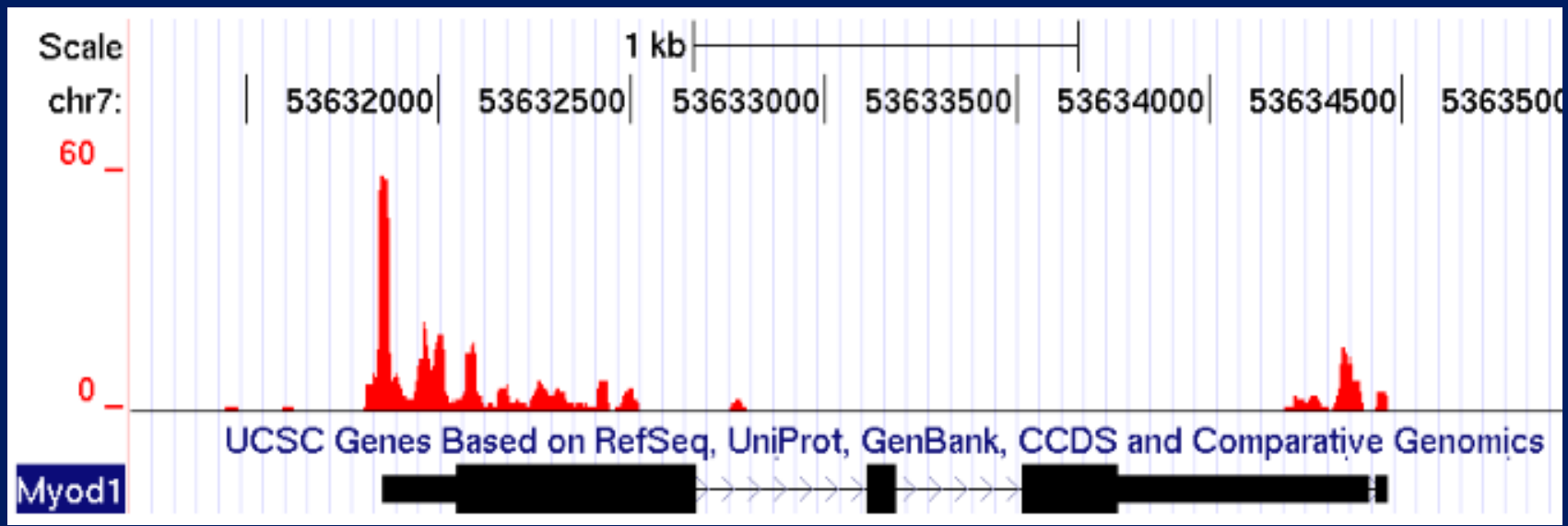

Figure 6: Coverage / depth histogram.

Coverage

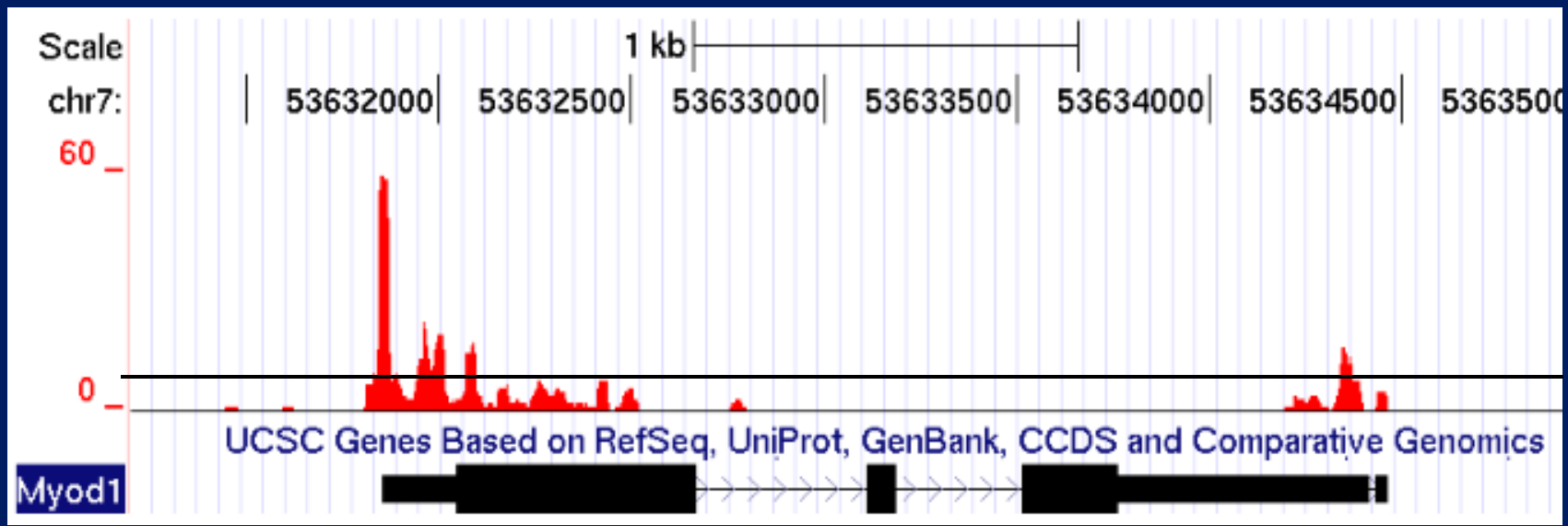

Figure 6: Coverage / depth histogram.

Coverage

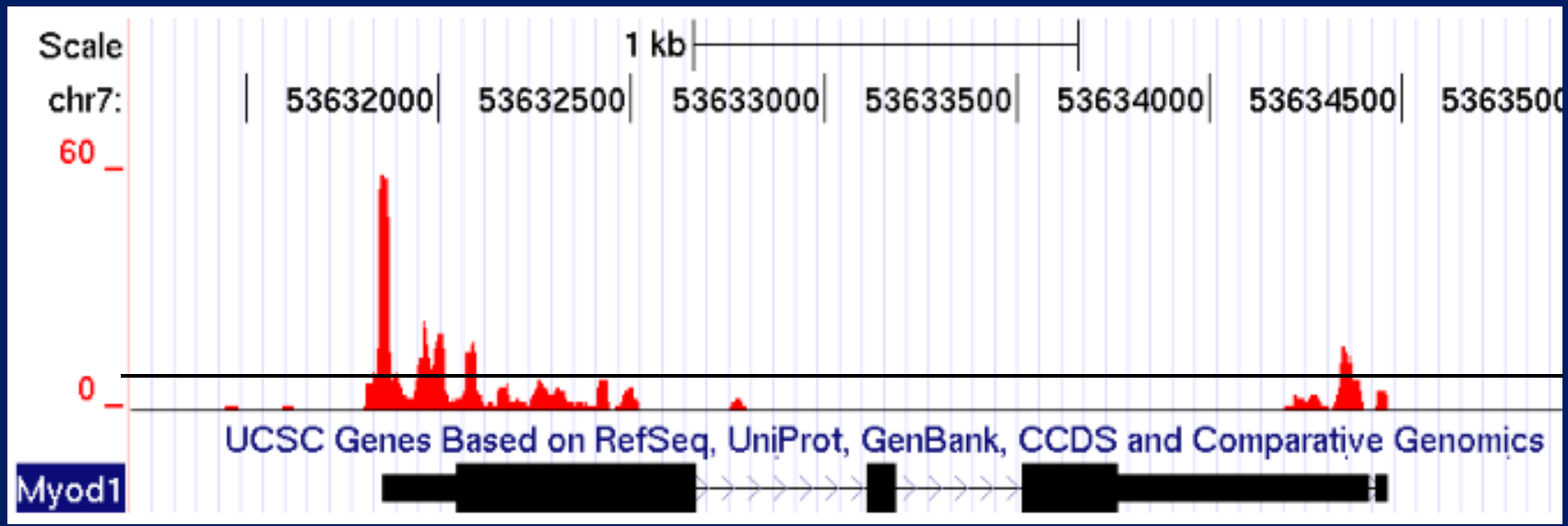

Figure 6: Coverage / depth histogram.

Figure 7: Coverage summary.

Formats

Minimal BED format

We only store minimal information about coverage.

Field	Explanation.
CHROM	Name of the chromosome.
BEGIN	First position of the region.
END	Last position of the region.

Table 3: Three column BED format.

Formats

Pooling

We provide a script to merge multiple VCF files into one.

- VCF files are merged and sorted on the *client* side.
  - The server has no knowledge of the individual variant calls.
- This makes identification of a person hard if the pool size is large enough.

Formats

Pooling

We provide a script to merge multiple VCF files into one.

- VCF files are merged and sorted on the *client* side.
  - The server has no knowledge of the individual variant calls.
- This makes identification of a person hard if the pool size is large enough.

As long as we calculate the *coverage* files separately, we can still calculate the frequencies correctly.

- One variant file and multiple coverage files are sent to the server.

Formats

Pooling

We provide a script to merge multiple VCF files into one.

- VCF files are merged and sorted on the *client* side.
  - The server has no knowledge of the individual variant calls.
- This makes identification of a person hard if the pool size is large enough.

As long as we calculate the *coverage* files separately, we can still calculate the frequencies correctly.

- One variant file and multiple coverage files are sent to the server.

Anonymity with no loss of functionality.

Technical description

Database layout

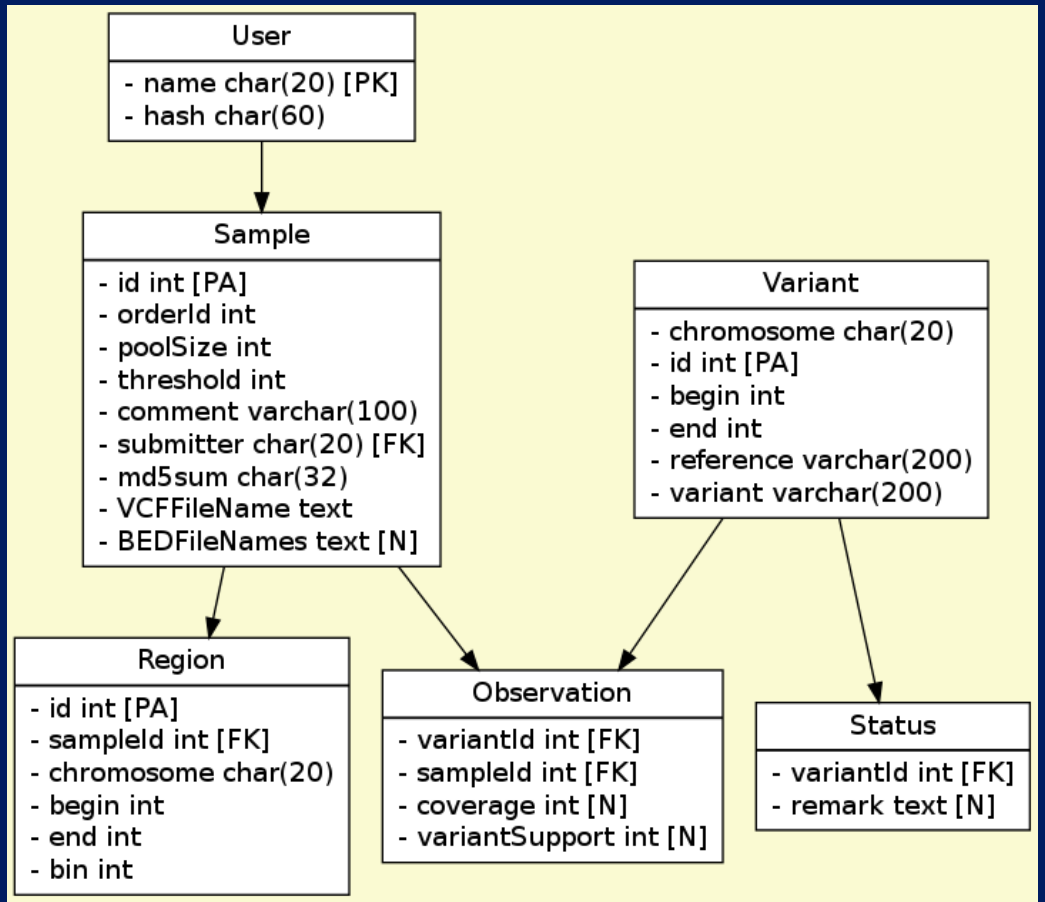

Figure 8: DVD SQL schema.

Technical description

Set up

For privacy reasons:

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

Technical description

Set up

For privacy reasons:

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

Protocols:

- A SOAP interface for interaction with pipelines.
- SOAP over HTTPS (SSL) for encryption.
- Authentication.
    - Send username / password in each call.
    - Only the password hash is stored in the database.

Technical description

Client perspective

In principle, one program that handles all:

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

Technical description

Client perspective

In principle, one program that handles all:

- Upload a VCF and one or more BED tracks.
- 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.

Technical description

Server perspective

Uploading:

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

Technical description

Server perspective

Uploading:

- Receive a VCF and one or more BED tracks.
- 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.

Technical description

Server perspective

Uploading:

- Receive a VCF and one or more BED tracks.
- 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 once the annotation part is finished, the result will be returned (before the import needs to be finished).

Technical description

Miscellaneous

The uploaded files are cached and *md5sums* are stored in the database.

- Prevents re-uploading the same data.
- Allows for re-annotation without polluting the database.
- Can be used to re-populate the database if needed.

Technical description

Miscellaneous

The uploaded files are cached and *md5sums* are stored in the database.

- Prevents re-uploading the same data.
- Allows for re-annotation without polluting the database.
- Can be used to re-populate the database if needed.

Highly optimised for genomic data.

- Binning to optimise region (spacial) data.
- Thoroughly tested and benchmarked.

Technical description

Output

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 4: The output format.

Current status

Production

Currently setting up a server at the UMCG.

- Nijmegen.
- Utrecht.
- Leiden.
- Amsterdam.
- Groningen.

https://trac.nbic.nl/dvd/

Current status

Production

Currently setting up a server at the UMCG.

- Nijmegen.
- Utrecht.
- Leiden.
- Amsterdam.
- Groningen.

Open source.

The core can be modified easily.

- Drop restrictions.
- Interface for browsing.
- ...

https://trac.nbic.nl/dvd/

Questions?

Acknowledgements:

Martijn Vermaat
Christian Gilissen
Victor Guryev
Frans-Paul Ruzius
Nienke Wieskamp
Wim Spee
Morris Swertz
Pieter Neerincx
Rob Hooft
David van Enckevort
Hailiang Mei
Johan den Dunnen

https://trac.nbic.nl/dvd/