Leiden University
Medical Center

Whole genome sequencing of Dutch melanoma families

Experiences and challenges

Jeroen F.J. Laros
Leiden Genome Technology Center
Department of Human Genetics
Center for Human and Clinical Genetics

Full genome sequencing

In contrast to *exome sequencing*, where we only the *coding regions* are sequenced, we sequence everything.

Full genome sequencing

In contrast to *exome sequencing*, where we only the *coding regions* are sequenced, we sequence everything.

Advantages of *full genome sequencing*:

- Introns.
  - Branch points.
  - Intronic splicing enhancers.
- Promoters.
- Transcription factor binding sites, insulators, etc.
- Consistent coverage.
  - Copy number variation.

Full genome sequencing

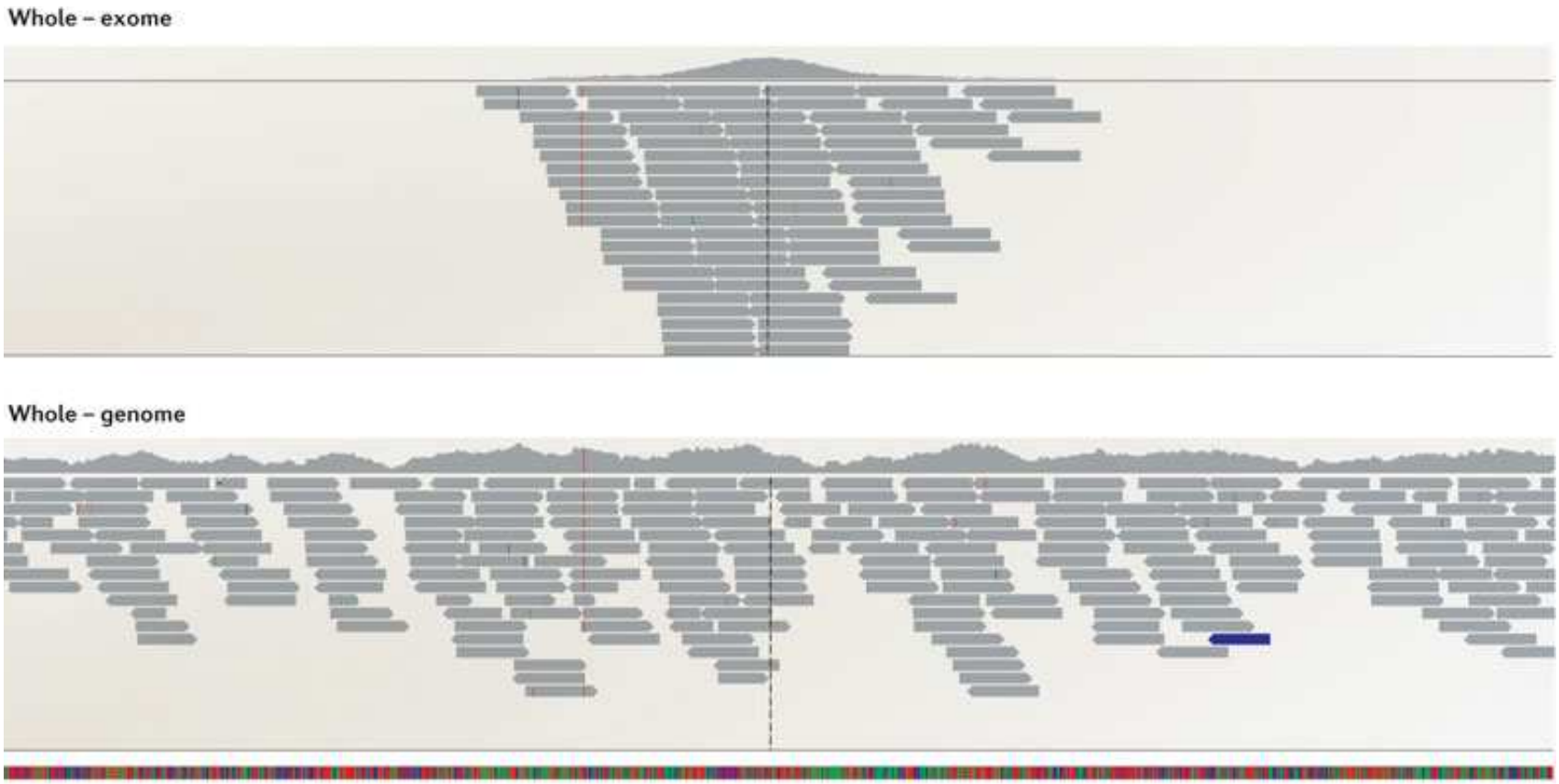

Figure 1: Exome- versus genome sequencing.

Copy number variation

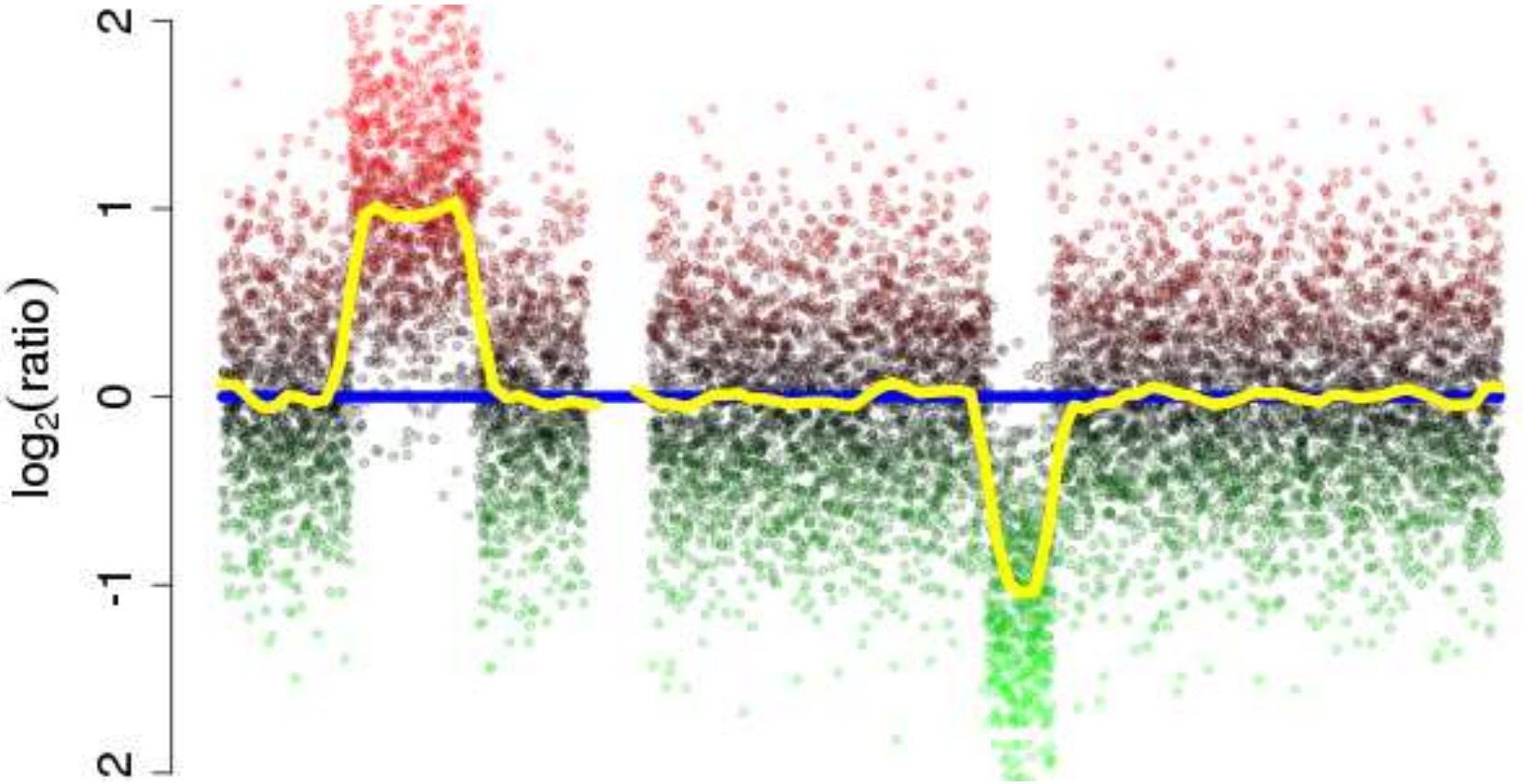

Figure 2: Coverage pattern over a whole chromosome.

Sequencing

Figure 3: HiSeq 2500.

Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length 2×150bp.
- Run time of 6 days.

Sequencing

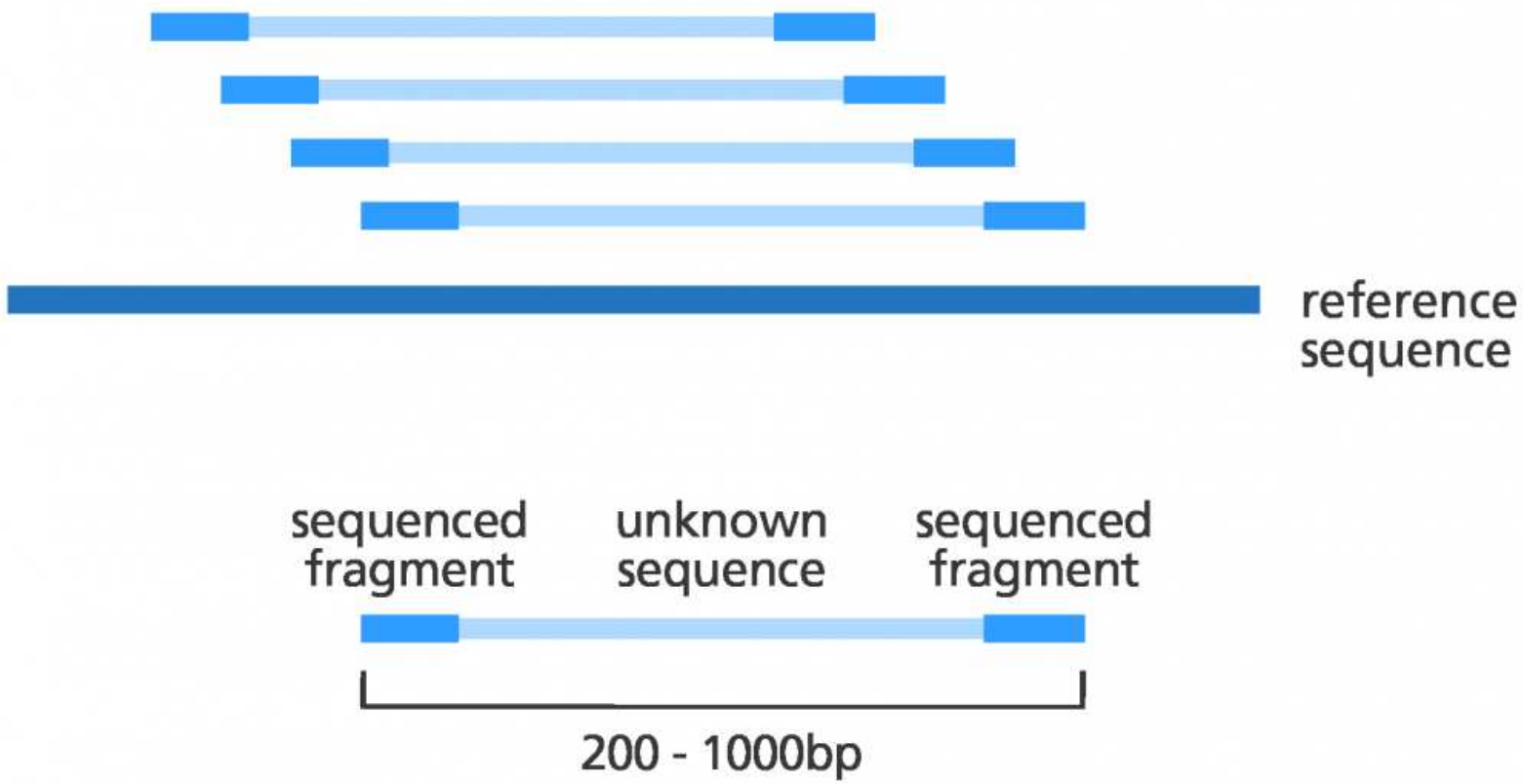

Figure 4: Paired end sequencing.

Data generation

Sequencing

Sequencing was done at the Sanger institute.

- Two times 100 nucleotides.
- 1,000,000,000 reads.
- 100,000,000,000 nucleotides.
- 150GB of data per sample (compressed).
- 35× coverage.

`http://www.sanger.ac.uk/`

Data generation

Sequencing

Sequencing was done at the Sanger institute.

- Two times 100 nucleotides.
- 1,000,000,000 reads.
- 100,000,000,000 nucleotides.
- 150GB of data per sample (compressed).
- 35× coverage.

A grand total of 4.5TB was generated, which completely filled up the storage at Sanger.

`http://www.sanger.ac.uk/`

Data generation

Data transfer

We need to make sure the data is transfered in a *secure* way.

Data carrier:

- Disks.
- Network.

Data generation

Data transfer

We need to make sure the data is transfered in a *secure* way.

Data carrier:

- Disks.
- Network.

Public server, GPG encryption.

- Encrypted with my *public key*.
- The encrypted data is public.
- Can only be decrypted with my *private key*.

Pipelines

Figure 5: Scene from "Modern times".

Data analysis

Resequencing 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.

Variant calling

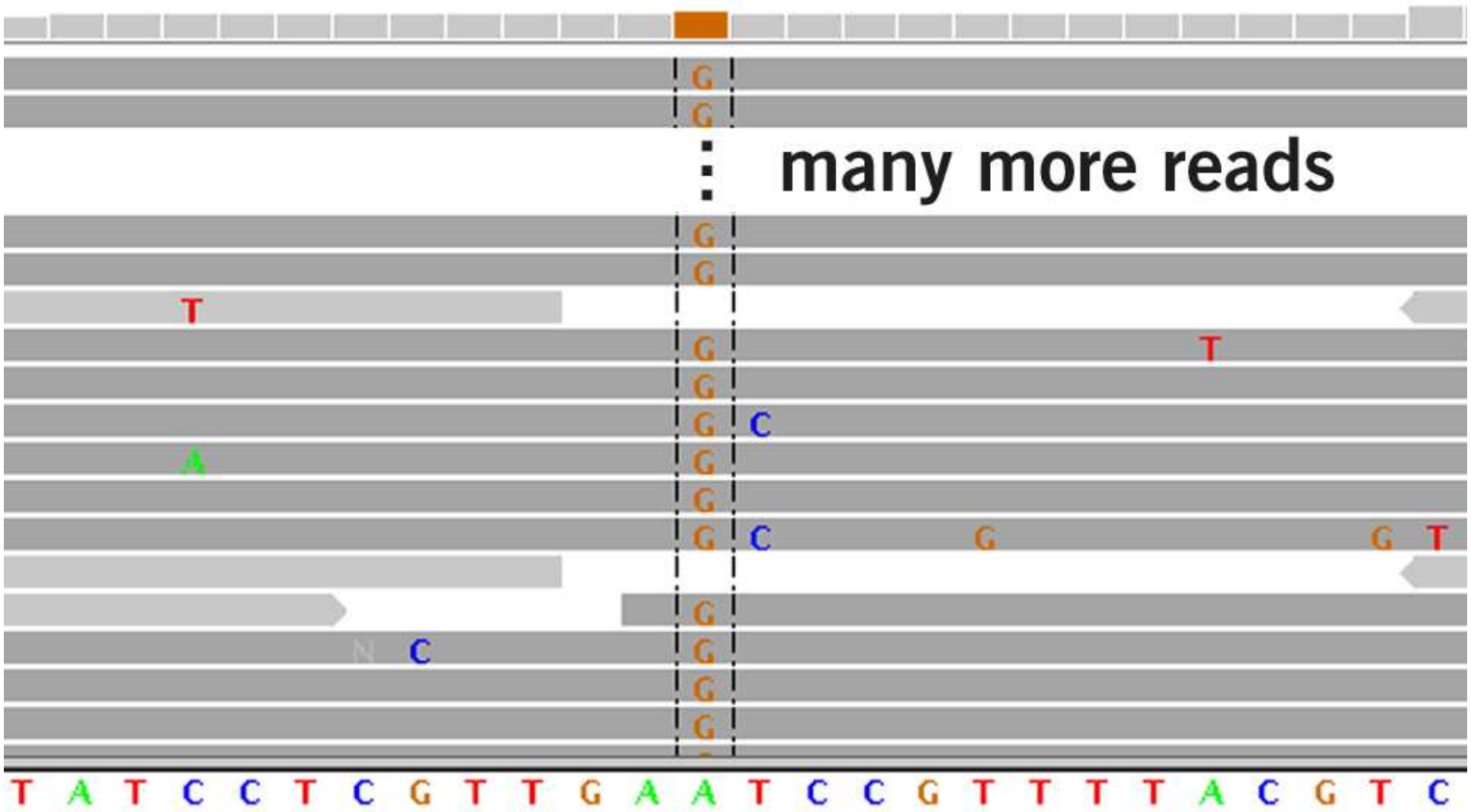

Figure 6: Result of an alignment.

Prioritisation

Prioritisation is mainly done by filtering variants that we expect to be irrelevant.

Prioritisation

Prioritisation is mainly done by filtering variants that we expect to be irrelevant.

This can be because the variant does not follow the *inheritance pattern* of the disease.

- The disease is recessive, but the variant is *homozygous* in an unaffected individual.

Prioritisation

Prioritisation is mainly done by filtering variants that we expect to be irrelevant.

This can be because the variant does not follow the *inheritance pattern* of the disease.

- The disease is recessive, but the variant is *homozygous* in an unaffected individual.

It can be because the *predicted effect* of a variant does not fit in the phenotype.

- A variant found in an unrelated gene.
- A variant that does not alter the protein.

Databases

In most cases we are not interested in common variants.

- dbSNP.
- 1000 Genomes.
- Exome Variant Server (EVS).

A cut-off of 1% is usually fine.

Databases

In most cases we are not interested in common variants.

- dbSNP.
- 1000 Genomes.
- Exome Variant Server (EVS).

A cut-off of 1% is usually fine.

Databases containing detailed information about variants:

- *Locus specific* databases.
  - LOVD.
- Human Gene Mutation Database (HGMD).

`http://www.lovd.nl/`

`http://www.hgmd.cf.ac.uk/`

Inheritance based filtering

We used the following intersection method.

"A variant should be present in all affected members of a family, but it may not occur in part of the affected members of an other family."

Family one	Family two	Filter result
5/5	4/4	Pass
0/5	4/4	Pass
3/5	4/4	Filtered
3/5	3/4	Filtered

Table 1: Advanced intersection.

Filters

We first did some general filtering based on allele frequencies found in databases and inheritance patterns.

Filter	Variants left
None	12,820,660
Unaffected member	7,354,674
EU MAF below 1%	5,504,165
GoNL MAF below 1%	4,681,268
Intersection	5,973,169
Advanced intersection	1,549,550

Table 2: Single filters.

Filters

By combining filters, we are left with a reasonable amount of variants.

Filter	Variants left
None	12,820,660
EU MAF below 1%	5,504,165
GoNL MAF below 1%	4,327,913
Unaffected member	2,471,569
Intersection	479,494
Advanced intersection	40,944

Table 3: Combining filters.

Note: 25,127 variants are present in every individual.

Effect prediction

We are still left with a lot of variants.

Variant annotation.

- Frequency within a population.
- Location of the variant.
  - Gene panels.
  - Location within a gene.
  - Regulatory regions.
- Conservation.

Variant Effect Predictor

A selection of VEP annotation:

- Affected genes and transcripts.
- Location of the variant.
  - Upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory region.
- Consequence on the protein sequence.
  - Stop gained, missense, stop lost, frameshift.
- Minor allele frequencies from the 1000 Genomes Project.
- SIFT and PolyPhen scores for changes to protein sequence.

`http://www.ensembl.org/info/docs/tools/vep/index.html`

Conservation

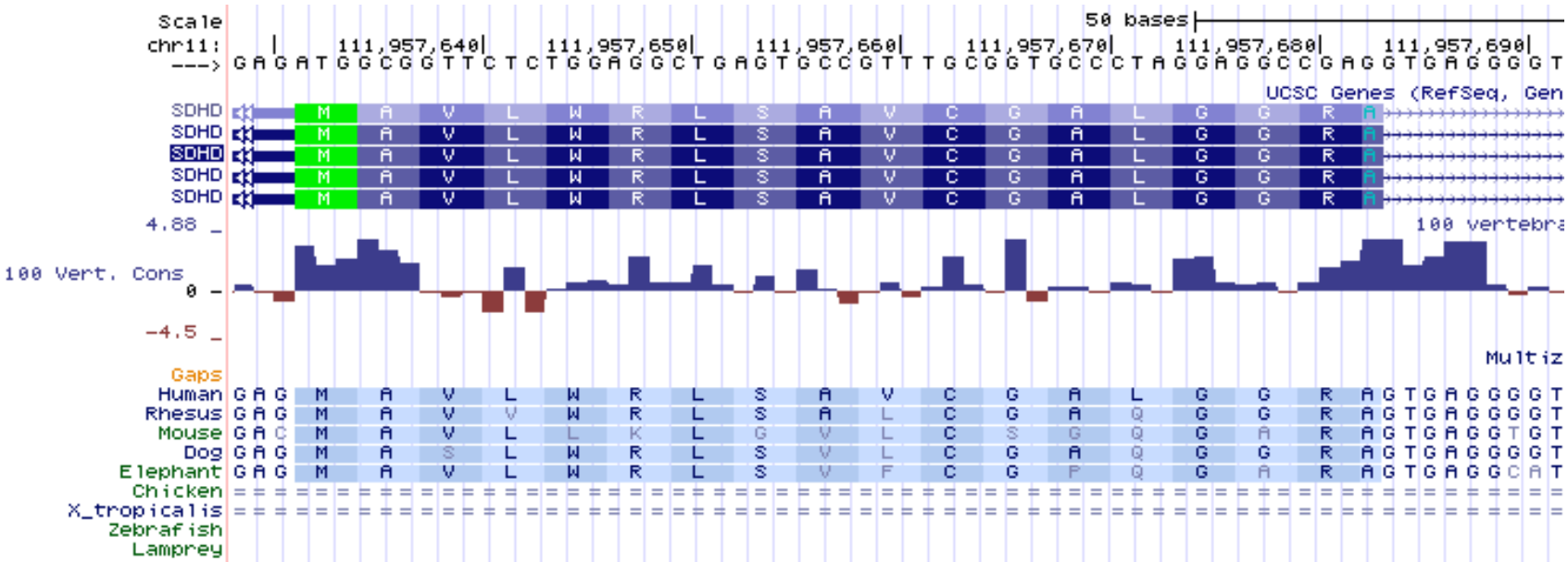

Figure 7: PhyloP scores based on 100 vertebrate species.

http://compgen.bscb.cornell.edu/phast/

Leiden University
Medical Center

Acknowledgements

Dermatology
Mijke Visser
Nelleke Gruis
Remco van Doorn
Remco van Doorn

SASC
Peter van 't Hof
Sander van der Zeeuw
Leon Mei

Sanger
Thomas Keane
Kim Wong
Daniela Espinoza
David Adams

Clinical Genetics
Nienke van der Stoep

Medical Statistics
Ramin Monajemi
Jeanine Houwing