LEIDEN UNIVERSITY MEDICAL CENTER

Prioritisation and variant effect prediction

Jeroen F.J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Variants

When we do a resequencing experiment, we find that about one in every 1,000 nucleotides differs from the reference sequence.

We expect roughly:

- 30,000 variants in an exome.
- 3,000,000 variants in a full genome.

We need some way to go through all of these variants.

Introduction

Exome sequencing

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

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

Introduction

Exome sequencing

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*.

Introduction

Illumina

Figure 1: HiSeq 2500.

Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length 2×125bp.
- Relatively long run time (6 days).
- Relatively expensive.

Introduction

Life Technologies

Figure 2: Ion proton.

Characteristics:

- Moderate throughput.
- Single end (for now).
- High accuracy.
- Read length $\pm$200bp.
- Short run time.
- Cheap runs.

Introduction

Pipelines

Figure 3: Scene from “Modern times”.

Introduction

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.

Introduction

Prioritisation

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

Introduction

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.

Introduction

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.

Inheritance based filtering

Trio analyses

Sequence the index patient and its parents.

Commonly used when:

- We expect a *de novo* variant.
- The disease is autosomal recessive.
- The disease is X-linked recessive.

These variants are relatively easy to find.

Inheritance based filtering

De novo

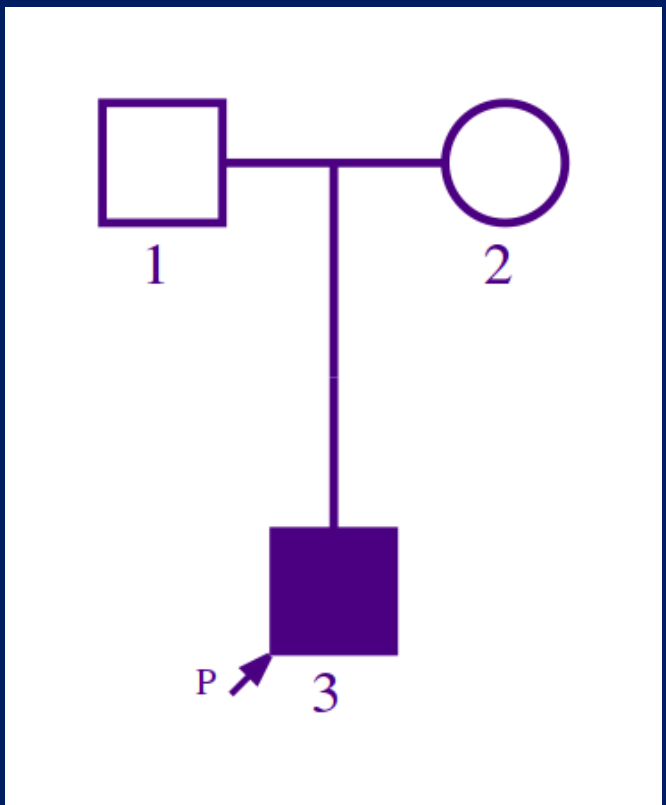


Figure 4: *De novo* variant.

Hypotheses:

- Both parents are unaffected.
- Both parents are not likely to be a *carrier*.
- The child is affected.

Filter all variants found in the parents from those of the patient.

Inheritance based filtering

Autosomal recessive

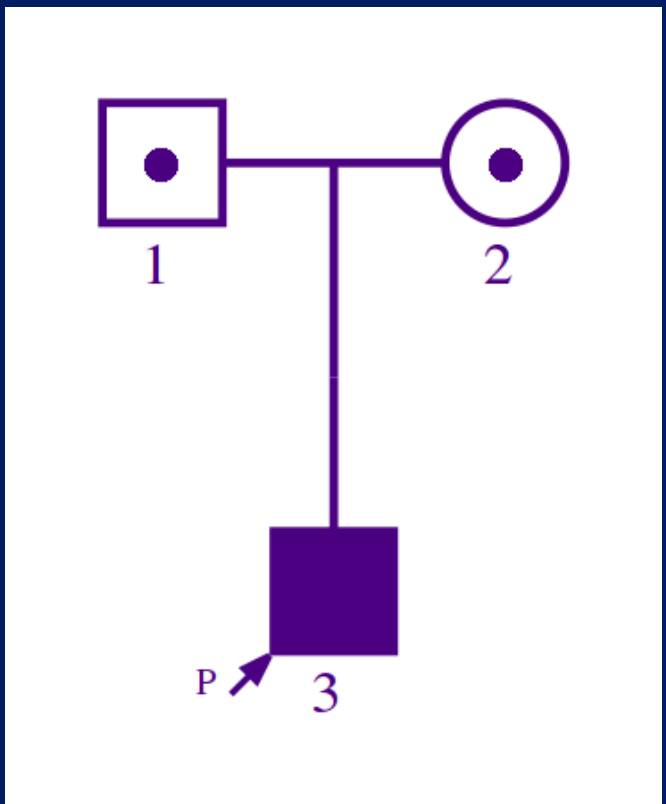


Figure 5: Recessive disease.

Hypotheses:

- Both parents are *carrier*.
- The child is affected.

Select all *homozygous* variants that are *heterozygous* in both parents.

Inheritance based filtering

X-linked recessive

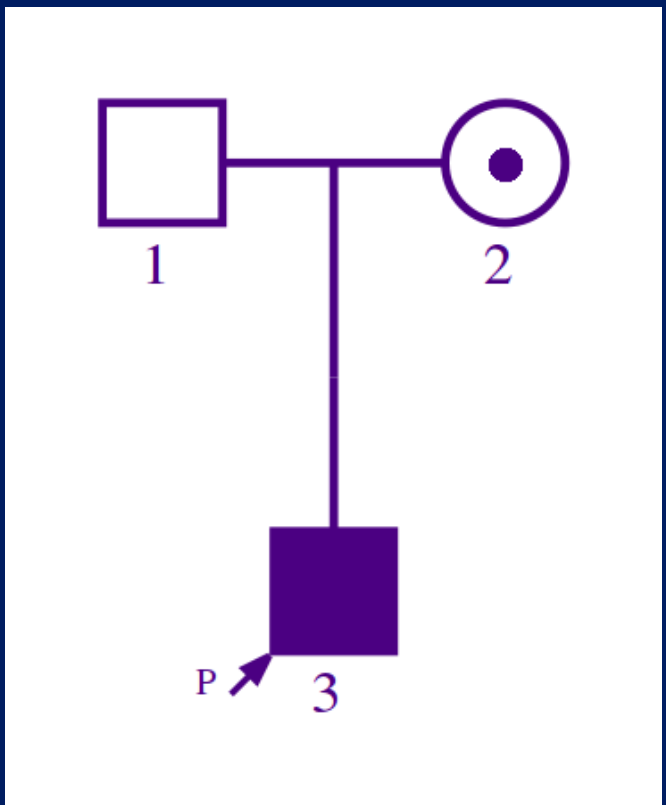


Figure 6: X-linked disease.

Hypotheses:

- The mother is a *carrier*.
- The child is male and affected.

Select the variants that are present on chromosome X which are *heterozygous* in the mother.

Annotation

Effect prediction

In most cases 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.
- Conservation.

Annotation

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`

Annotation

Gene panels

Sometimes we know which genes are associated with the disease.

- Online Mendelian Inheritance in Man (OMIM).

http://www.omim.org/

Annotation

Gene panels

Sometimes we know which genes are associated with the disease.

- Online Mendelian Inheritance in Man (OMIM).

To take it one step further, we can select the *transcripts* that are expressed in the tissue of interest.

Example: The DMD transcript expressed in brain cells is different from the one expressed in muscle cells.

`http://www.omim.org/`

Annotation

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.

Annotation

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/`

Annotation

Conservation

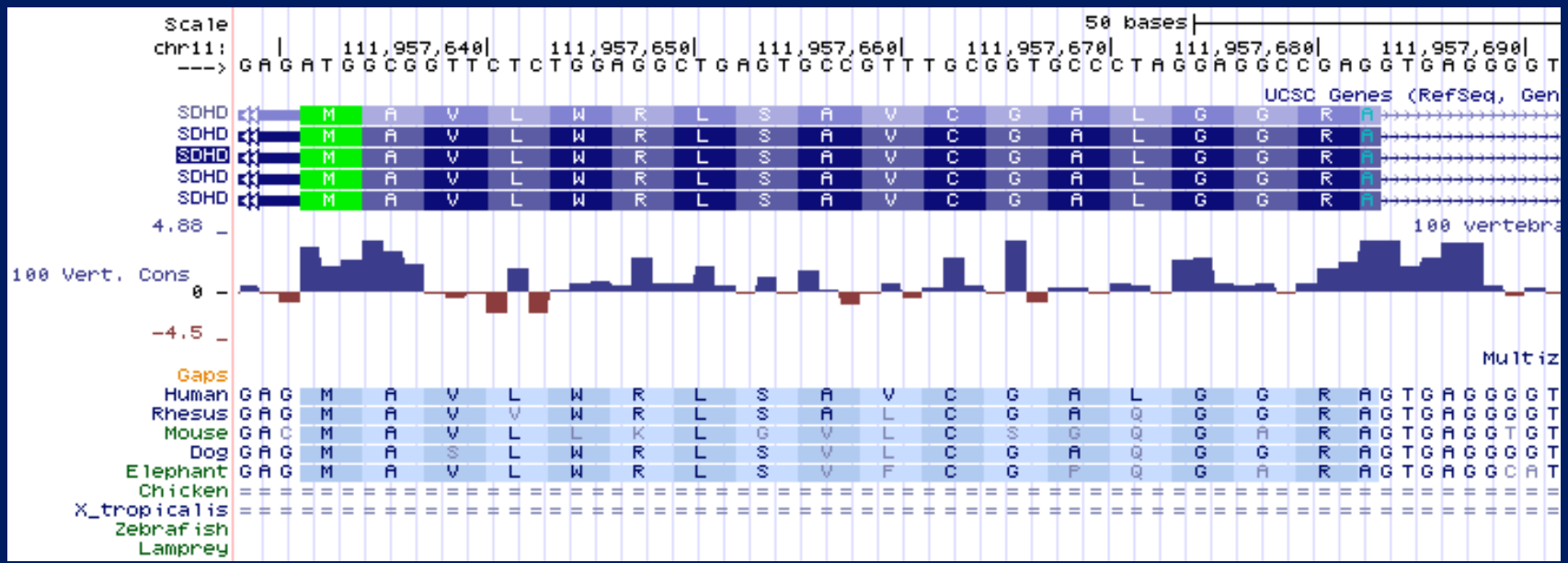


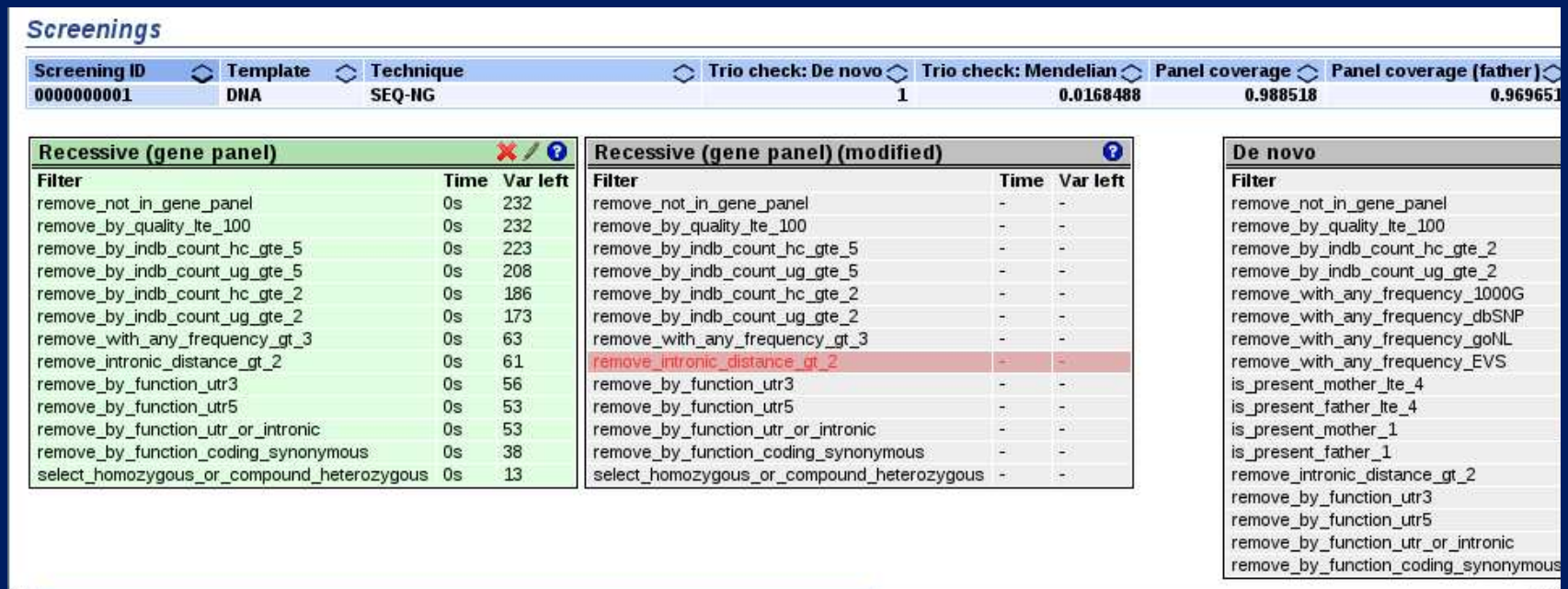
Figure 7: PhyloP scores based on 100 vertebrate species.

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

Putting it all together

Interactive filters

Prioritise your variants with interactive software like CLC bio, NextGENe, Cartagania, . . .



Screenings

Screening ID	Template	Technique	Trio check: De novo	Trio check: Mendelian	Panel coverage	Panel coverage (father)
0000000001	DNA	SEQ-NG	1	0.0168488	0.988518	0.969651

Recessive (gene panel)

Filter	Time	Var left
remove_not_in_gene_panel	0s	232
remove_by_quality_lte_100	0s	232
remove_by_indb_count_hc_gte_5	0s	223
remove_by_indb_count_ug_gte_5	0s	208
remove_by_indb_count_hc_gte_2	0s	186
remove_by_indb_count_ug_gte_2	0s	173
remove_with_any_frequency_gt_3	0s	63
remove_intronic_distance_gt_2	0s	61
remove_by_function_utr3	0s	56
remove_by_function_utr5	0s	53
remove_by_function_utr_or_intronic	0s	53
remove_by_function_coding_synonymous	0s	38
select_homozygous_or_compound_heterozygous	0s	13

Recessive (gene panel) (modified)

Filter	Time	Var left
remove_not_in_gene_panel	-	-
remove_by_quality_lte_100	-	-
remove_by_indb_count_hc_gte_5	-	-
remove_by_indb_count_ug_gte_5	-	-
remove_by_indb_count_hc_gte_2	-	-
remove_by_indb_count_ug_gte_2	-	-
remove_with_any_frequency_gt_3	-	-
remove_intronic_distance_gt_2	-	-
remove_by_function_utr3	-	-
remove_by_function_utr5	-	-
remove_by_function_utr_or_intronic	-	-
remove_by_function_coding_synonymous	-	-
select_homozygous_or_compound_heterozygous	-	-

De novo

Filter
remove_not_in_gene_panel
remove_by_quality_lte_100
remove_by_indb_count_hc_gte_2
remove_by_indb_count_ug_gte_2
remove_with_any_frequency_1000G
remove_with_any_frequency_dbSNP
remove_with_any_frequency_goNL
remove_with_any_frequency_EVS
is_present_mother_lte_4
is_present_father_lte_4
is_present_mother_1
is_present_father_1
remove_intronic_distance_gt_2
remove_by_function_utr3
remove_by_function_utr5
remove_by_function_utr_or_intronic
remove_by_function_coding_synonymous

Figure 8: Filtering with LOVD$^+$.

Putting it all together

Interactive filters

13 entries on 1 page. Showing entries 1 - 13.

25 per page ▼ **Legend**

	Chr	Effect	DNA change (genomic)	Alamut link	PhyloP conservation	HGMD association	Read depth Alt (fraction)	GATKcal
		!-						
☐	3	?/?	g.142215124_142215136del	Alamut	-0.305	-	0	u,n,k,n,
☐	3	?/?	g.142215367_142215373del	Alamut	0.187	-	0	u,n,k,n,
☑	4	?/?	g.5624370C>G	Alamut	0.399	-	0	u,n,k,n,
☑	4	?/?	g.5642503C>T	Alamut	3.469	-	0	u,n,k,n,
☑	4	?/?	g.5691037T>C	Alamut	0.382	-	0	u,n,k,n,
☐	7	?/?	g.103155949_103155950insAAA	Alamut	-0.959	-	0	u,n,k,n,
☐	10	?/?	g.27822618_27822620del	Alamut	0.535	-	0	u,n,k,n,
☐	10	?/?	g.27822652_27822653insT	Alamut	0.433	-	0	u,n,k,n,
☐	11	?/?	g.93534930_93534932del	Alamut	-0.021	-	0	u,n,k,n,
☐	11	?/?	g.134014672_134014673insCCT	Alamut	0.31	-	0	u,n,k,n,
☐	13	?/?	g.31821083_31821084insT	Alamut	-0.086	-	0	u,n,k,n,
☐	13	?/?	g.31821091_31821092insTC	Alamut	-0.587	-	0	u,n,k,n,
☐	15	?/?	g.90173732_90173733insAC	Alamut	-0.016	-	0	u,n,k,n,

25 per page ▼ **Legend**

Powered by LOVD v.
©2004-2013 Leiden Univers

Figure 9: Results of an analysis.

Phasing

Unphased variants

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Reference protein:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT LHGHWGLGQV VTDYVHGDAL
121  QKAAKAGLLA LSALTFAGLC YFNYHDVGIC KAVAMLWKL*
```

Protein predicted from variant coding sequence:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA RTIPWLQPSL FMVTGALDKL LLTMFMGMPC
121  RKLPRQGFWH FQL*
```

Figure 10: Predicted frameshift.

NM_003002.2(SDHD_v001):c.272del

Phasing

Unphased variants

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Reference protein:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT LHGHWGLGQV VTDYVHGDAL
121  QKAAKAGLLA LSALTFAGLC YFNYHDVGIC KAVAMLWKL*
```

Protein predicted from variant coding sequence:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT SWSLGPWTSC Y*
```

Figure 11: Predicted frameshift.

NM_003002.2(SDHD_v001):c.301_302del

Phasing

Phased variants

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Reference protein:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA MDYSLAAALT LHGHWGLGQV VTDYVHGDAL
121  QKAAKAGLLA LSALTFAGLC YFNYHDVGIC KAVAMLWKL*
```

Protein predicted from variant coding sequence:

```
  1  MAVLWRLSAV CGALGGRALL LRTPVVRPAH ISAFLQDRPI PEWCGVQHIH LSPSHHSGSK
 61  AASLHWTSER VVSVLLLGLL PAAYLNPCSA RTIPWLQPSL HGHWGLGQVV TDYVHGDALQ
121  KAAKAGLLAL SALTFAGLCY FNYHDVGICK AVAMLWKL*
```

Figure 12: Predicted indel.

NM_003002.2(SDHD_v001):c.[272del;301_302del]

Questions?

Acknowledgements:

Ivo Fokkema
Sander Bollen
Gijs Santen
Claudia Ruivenkamp
Mariëtte Hoffer
Johan den Dunnen