UMC St Radboud

Human Genetics Nijmegen

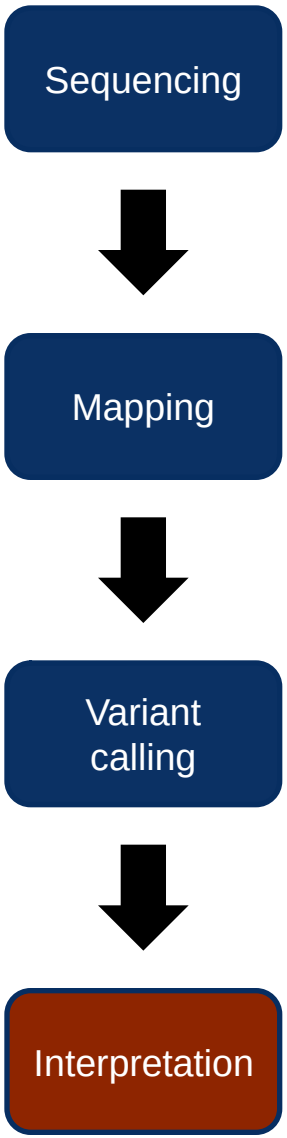
Variant Detection & Interpretation in a diagnostic context

Christian Gilissen

c.gilissen@gen.umcn.nl

28-05-2013

So far…



Johan den Dunnen
Marja Jakobs
Ewart de Bruijn

Victor Guryev

Laurent Francioli

What to interpret?

- Variants → SNVs and small indels
- 3 million SNVs per individual genome
- 20,000 to 50,000 variants per individual exome

How to identify variants that are involved in a patient's disease?

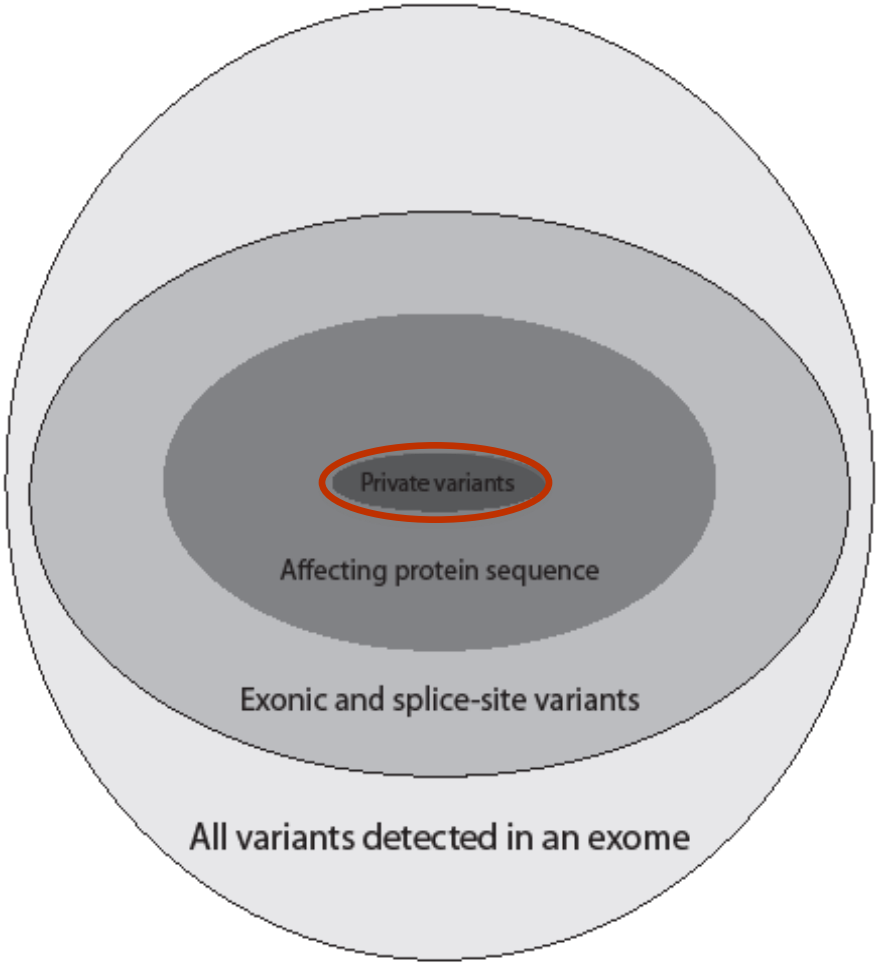
Variant interpretation

1. Annotation of variants
2. Strategies for prioritization
3. Computational prediction of pathogenicity

Part I - Interpretation of exome data

- An initial approach:

~150-500 private non-synonymous variants



Annotation

- Publicly available sources
  - SeattleSeq, Annovar, Vaast, Ensembl AP, SNPEff, dbNSFP
- Commercial packages:
  - CLC Bio
  - NextGene
  - Cartagenia
  - Ingenuity VA
- Home-made software
- **All tools:**
  - Effect of variant on protein coding gene
  - Overlap with databases of polymorphisms

What can you get?

- SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/)
  - Conservation scores, Polyphen predictions, on-line
  - No indels, input format is very specific
- Annovar (http://www.openbioinformatics.org/annovar/):
  - Pro: Sift (old) and polyphen predictions
  - Con: local install required → web interface now available: **wAnnovar**
- Vaast (http://www.yandell-lab.org/software/vaast.html):
  - Pro: statistic framework for candidate gene selection
  - Con: local install required, no indels (yet)
- Ensembl API (http://www.ensembl.org/info/docs/api/variation/index.html)
  - Pro: flexible
  - Con: requires installation and programming, not all data available
- SNPEff (http://snpeff.sourceforge.net/)
  - Pro: fast, indels, multiple species
  - Con: local install, only does effect on protein

Variant frequency sources

- **dbSNP:** largest dataset, but polluted
- **1000 genomes:** frequencies available but from cell-lines
- **ESP database:** no indels, patients, no validation
- **Published studies:** GONL, Complete genomics genomes
- **In house databases / DVD:** population/sequencing specific variants

ESP6500 variants for *ASXL1*

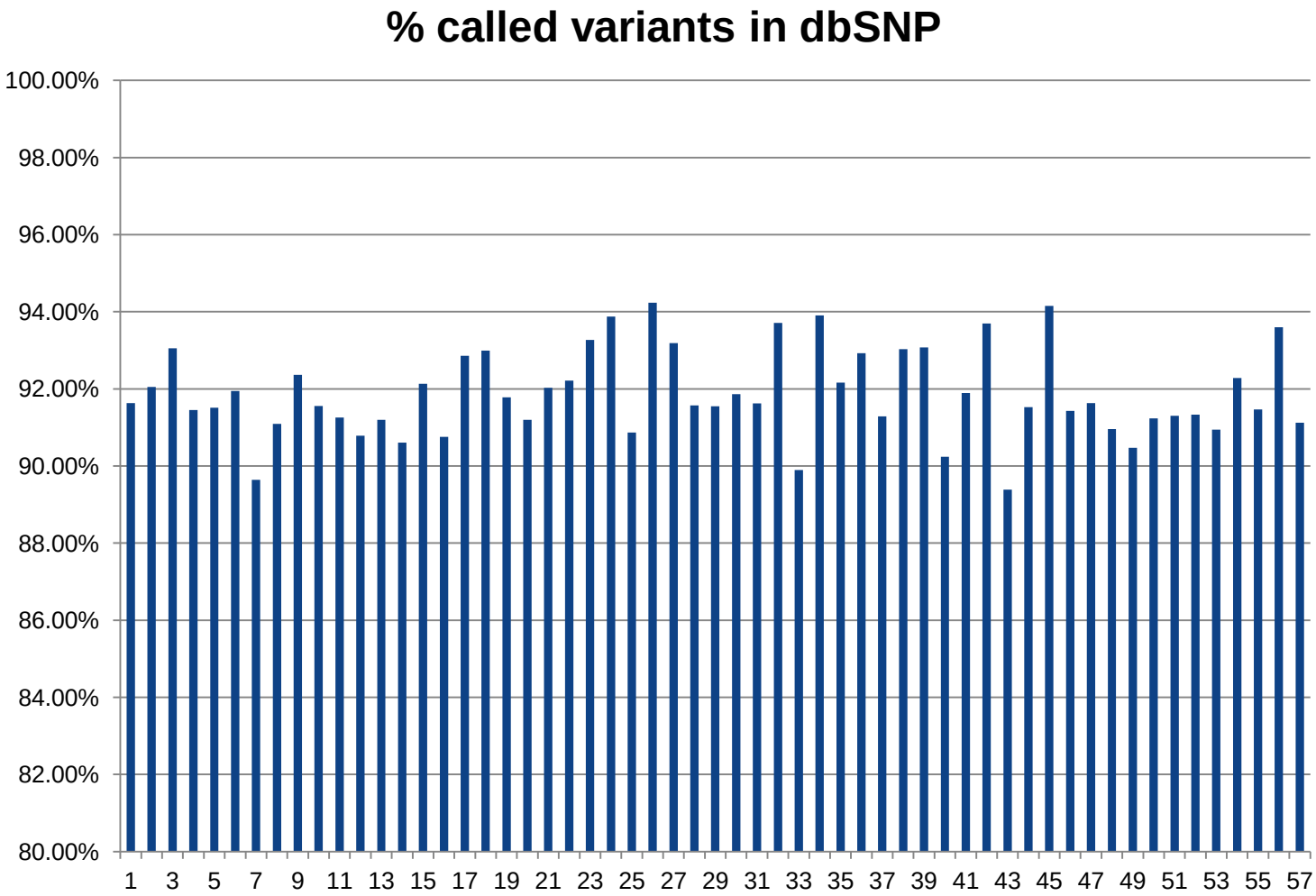
Variant Pos	rs ID	Alleles	EA Allele #	AA Allele #	All Allele #	Avg. Sample Read Depth	Genes	mRNA Accession #	GVS Function	Amino Acid
20:31021163	rs145699348	A/G	A=1/G=8599	A=0/G=4406	A=1/G=13005	81	ASXL1	NM_015338.5	missense	ILE,VAL
20:31021190	unknown	T/C	T=1/C=8599	T=0/C=4406	T=1/C=13005	90	ASXL1	NM_015338.5	missense	CYS,ARG
20:31021211	unknown	T/C	T=0/C=8600	T=2/C=4404	T=2/C=13004	92	ASXL1	NM_015338.5	stop-gained	stop,ARG
20:31021232	rs148964601	T/C	T=1/C=8599	T=0/C=4406	T=1/C=13005	90	ASXL1	NM_015338.5	missense	CYS,ARG
20:31021233	rs143719307	A/G	A=0/G=8600	A=1/G=4405	A=1/G=13005	90	ASXL1	NM_015338.5	missense	HIS,ARG
20:31021250	unknown	T/C	T=1/C=8599	T=0/C=4406	T=1/C=13005	87	ASXL1	NM_015338.5	stop-gained	stop,ARG
20:31021324	unknown	C/T	C=0/T=8600	C=1/T=4405	C=1/T=13005	98	ASXL1	NM_015338.5	coding-synonymous	none
20:31021332	unknown	G/C	G=1/C=8599	G=0/C=4406	G=1/C=13005	97	ASXL1	NM_015338.5	stop-gained	stop,SER
20:31021337	unknown	A/G	A=1/G=8599	A=0/G=4406	A=1/G=13005	97	ASXL1	NM_015338.5	mis[illegible]e	ILE,VAL
20:31021384	unknown	A/G	A=2/G=8598	A=0/G=4406	A=2/G=13004	102	ASXL1	NM_015338.5	coding[illegible]ous	none
20:31021389	unknown	A/G	A=1/G=8599	A=1/G=4405	A=2/G=13004	104	ASXL1	NM_015338.5	[illegible]	ASN,SER
20:31021430	rs141346625	C/G	C=2/G=8598	C=17/G=4389	C=19/G=12987	107	ASXL1	NM_015338.5	n[illegible]	GLN,GLU
20:31021466	rs142172134	G/C	G=0/C=8600	G=10/C=4396	G=10/C=12996	106	ASXL1	NM_015338.5	n[illegible]	GLY,ARG
20:31021475	rs145913172	C/G	C=0/G=8600	C=2/G=4404	C=2/G=13004	109	ASXL1	NM_015338.5	n[illegible]	PRO,ALA
20:31021521	rs138971201	A/T	A=0/T=8600	A=14/T=4392	A=14/T=12992	131	ASXL1	NM_015338.5	n[illegible]	ASN,ILE
20:31021544	unknown	A/G	A=0/G=8600	A=1/G=4405	A=1/G=13005	140	ASXL1	NM_015338.5	n[illegible]	MET,VAL

Variation Color Code:
splice or nonsense or frameshift
missense
coding-synonymous
coding
utr
codingComplex

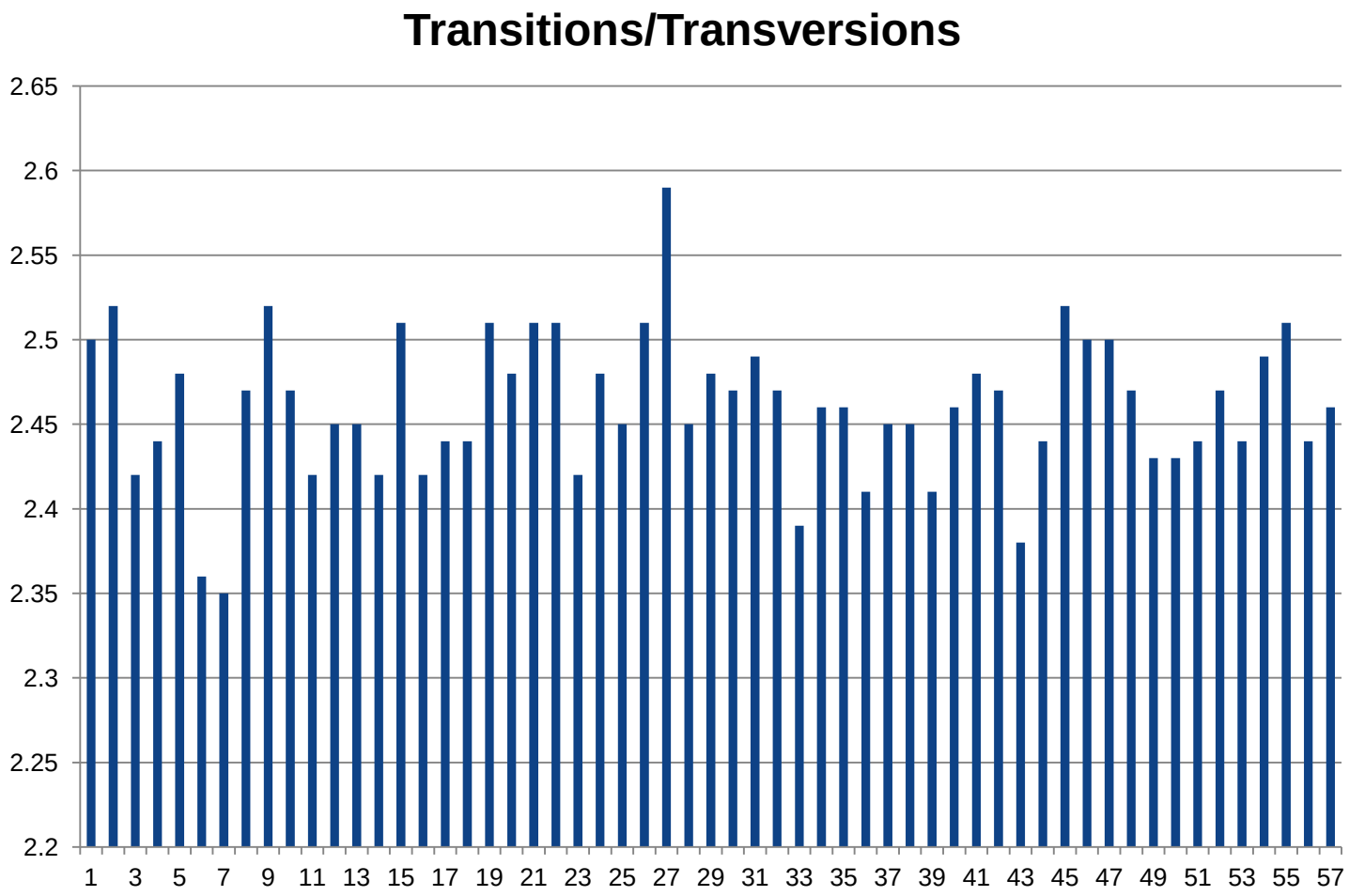
Bohring-Opitz syndrome is often fatal in early childhood.

http://evs.gs.washington.edu/EVS/

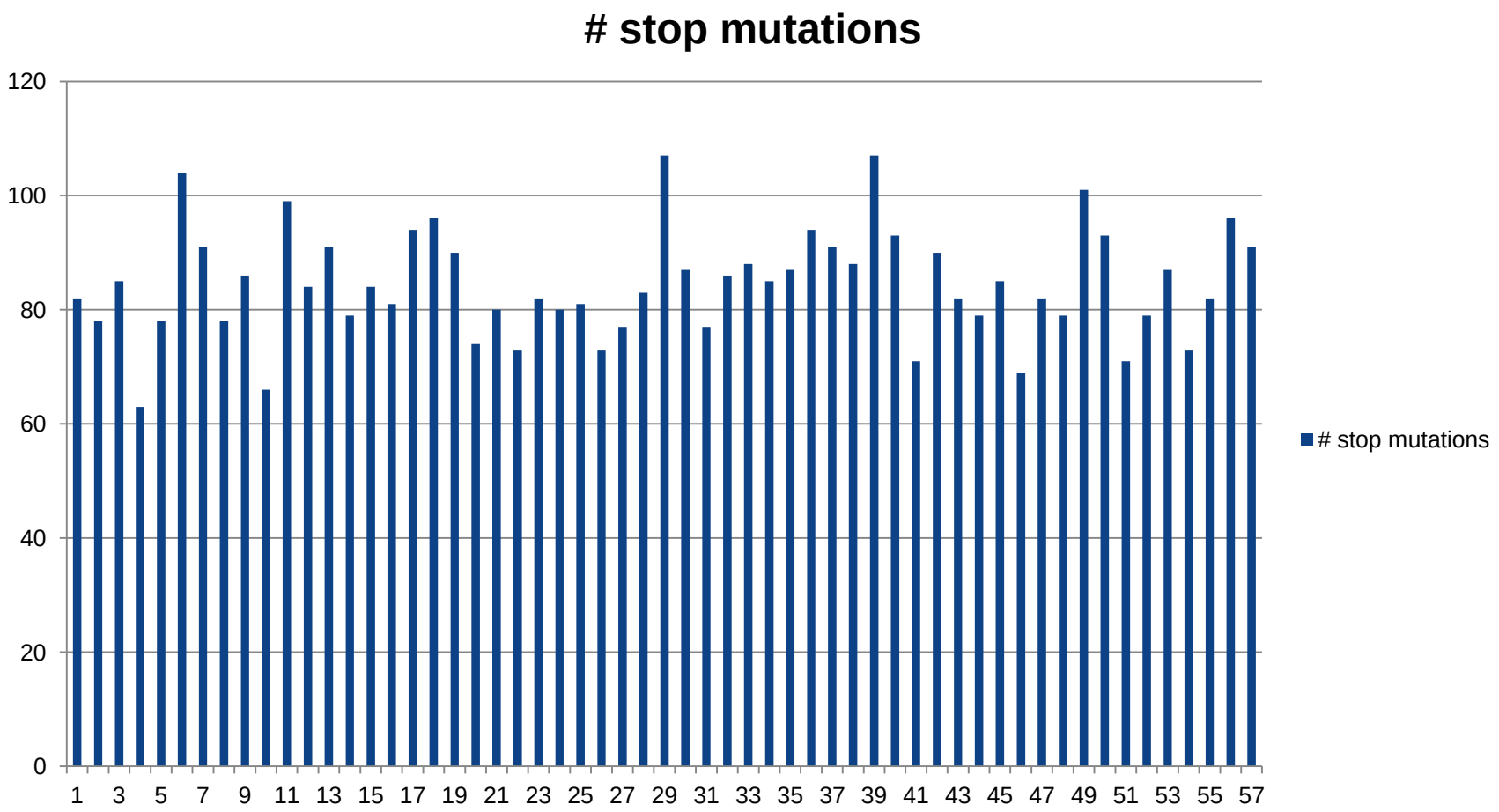
Not just interpretation: also QC



QC from annotation: Tr/Ti



QC from annotation: stop mutations



Other (common) annotations:

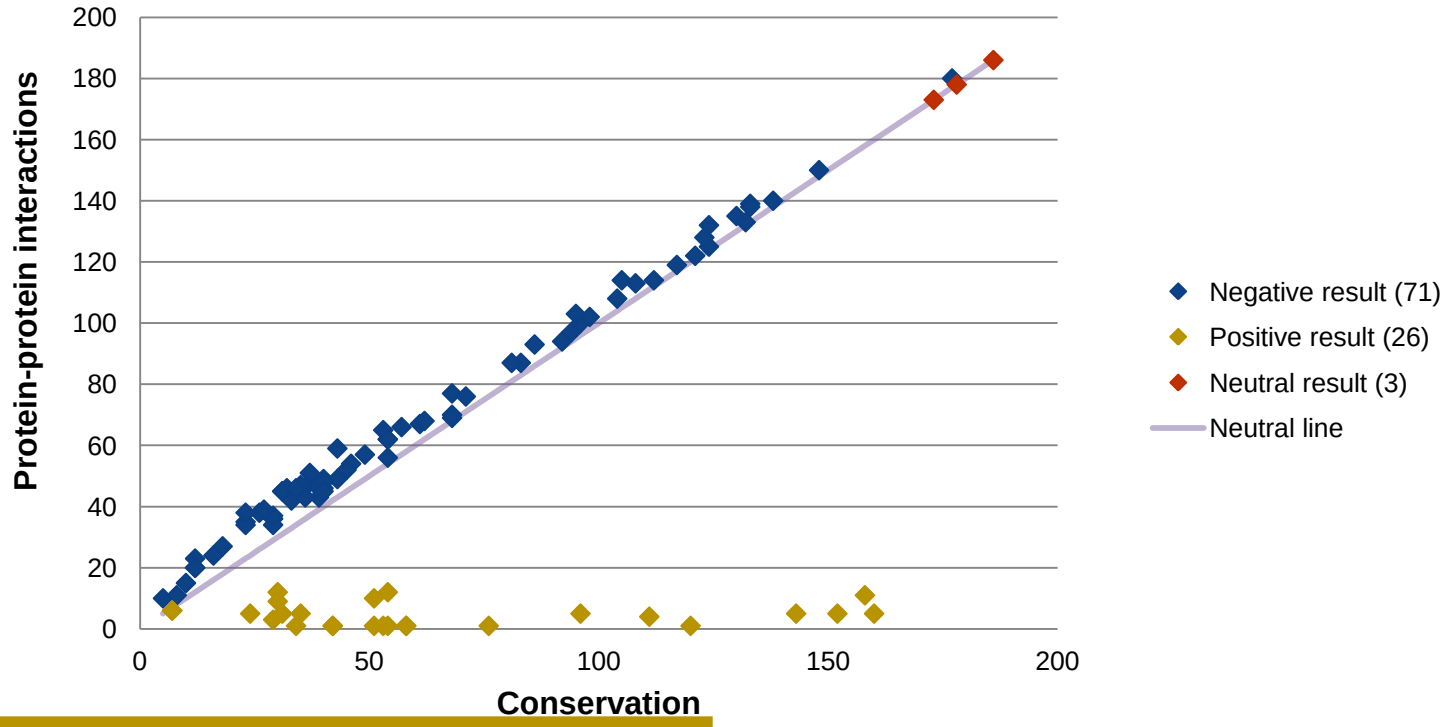
- **Variant based:**
  - Grantham / substitution scores
  - HGMD
  - Protein domains
  - Protein level conservation
  - Repeat
- **Gene based**
  - OMIM (disease gene),
  - MGI: Mouse knock-out phenotypes / zebrafish knock-out
  - Kegg pathways and GO biological processes
  - Loss off function gene

Protein-protein interactions

How to use?

•Simulate 100 exomes with a “spiked-in” mutation in a deafness gene

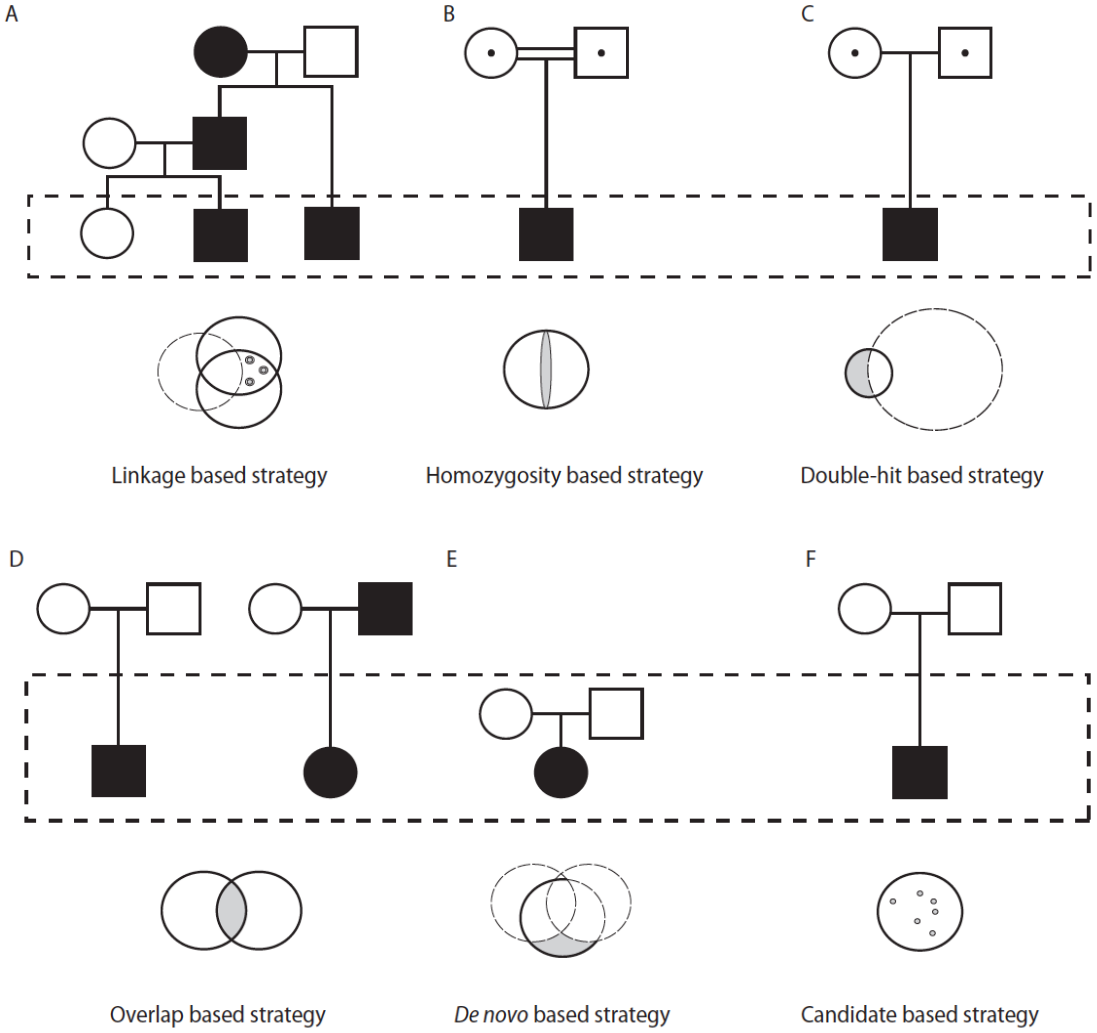
•Raking of variants using PPI and conservation compared to only on conservation



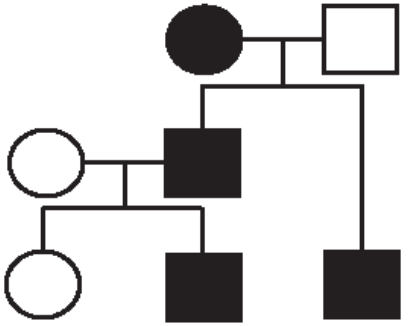
Interpretation of non-coding variants

- Many more variants, much less information
- What can you use?
  - Evolutionary conservation
  - Overlap with regulator regions (Encode)
  - Proximity to known genes
- Similar ways of reducing the candidates as exome analysis: *de novo* variants, family analysis

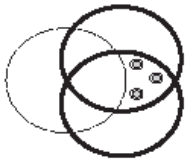
Part II –
Strategies to prioritize variants from exome studies



Linkage strategy



- Select variants that segregate with the disease or lie within a region that segregates with the disease



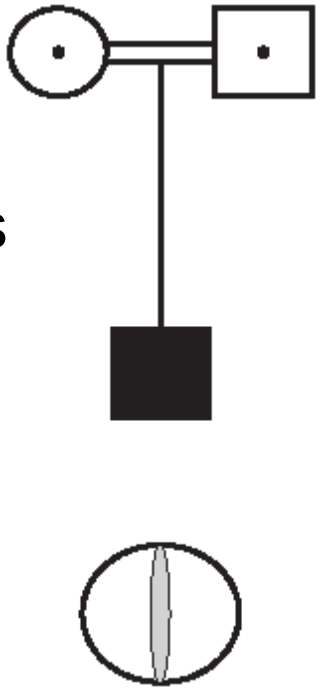
- Applies to both dominant and recessive disorders

1. Overlap / exclude variants from family members
   Two affected siblings, reducing the number of candidates to 9 genes.[1]

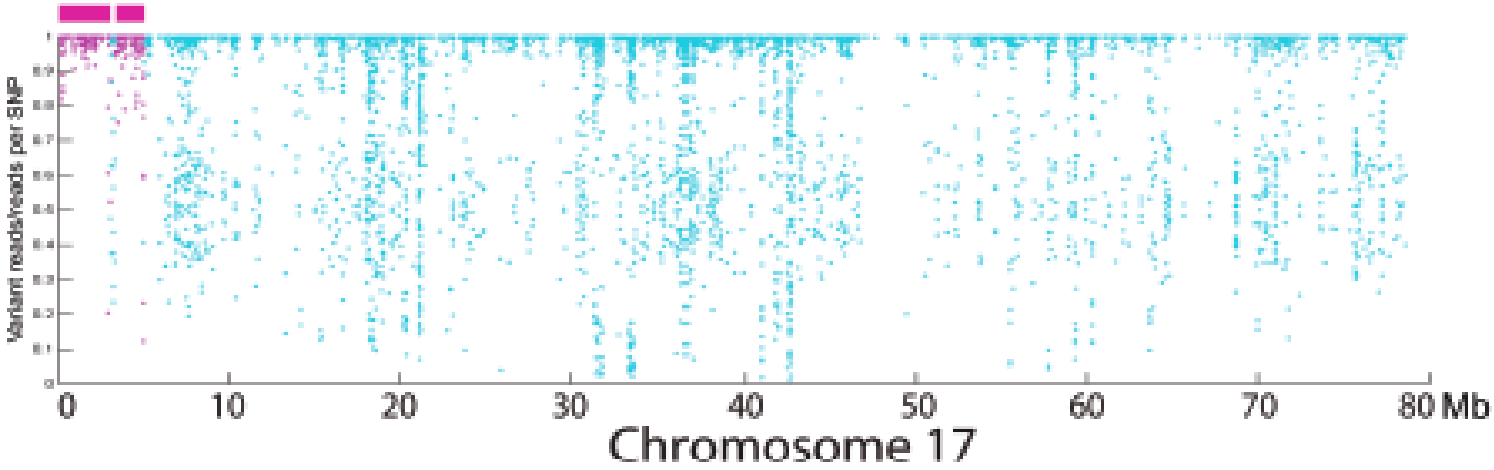
2. Determine regions of Identity By Descent
   Three affected siblings, reducing the number from 14 to 2 genes.[2]

[1]Ng *et al.* Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010
[2]Krawitz *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet. 2010

Homozygosity strategy



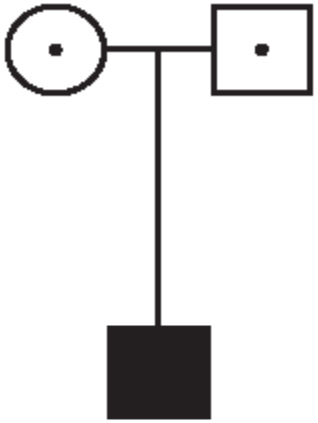
- Select variants that lie within a large homozygous region of the patient



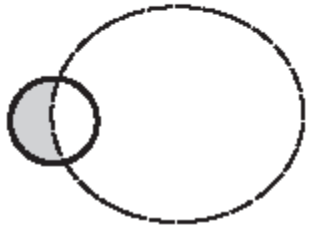
- Reduced the number of homozygous candidate variants from 17 to 3.[3]

[3]Becker *et al.* Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. Am J Hum Genet. 2011

Double hit strategy



- Select variants that are homozygous or compound-heterozygous in the patient

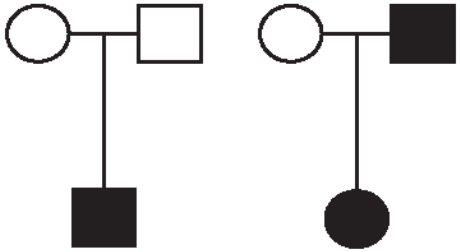


- Applies only to recessive disorders (with no consanguinity)

- A single exome can be sufficient, [4,5] reducing the number of candidates from 139 and 158 to 3 and 4 respectively.

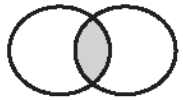
[4]Pierce *et al.* Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. Am J Hum Genet. 2010

[5]Gilissen *et al.* Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet. 2010



Overlap strategy

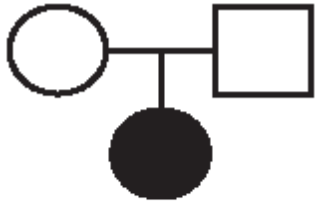
- Select unrelated patients and determine variants in multiple patients in the same gene[6,7]



- Used for rare sporadic dominant disorders
- Depends crucially on good phenotyping
- Disorder must be monogenic

- Three individuals can be enough to pinpoint a single gene.[8]

[6]Hoischen *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010
[7]Ng *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010
[8]Hoischen *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet. 2011



De novo strategy

- Exome sequencing an affected patient and his unaffected parents and select variants that are not inherited.[9,10,11]

- Applies to sporadic disorders with large genetic heterogeneity

- Methods for detecting *de novo* mutations enrich for sequencing and analysis errors.

[9]Vissers *et al*. A *de novo* paradigm for mental retardation. Nat Genet. 2010
[10]O'roak *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. Nat Genet. 2011
[11]Xu et al. Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. Nat Genet. 2011

Prioritization of candidate *de novo* variants

MR trio	1	2	3	4	5	6	7	8	9	10	average
High confidence variant calls	20,810	21,658	21,338	22,647	17,694	22,333	21,369	22,658	24,085	22,962	21,755

n=51

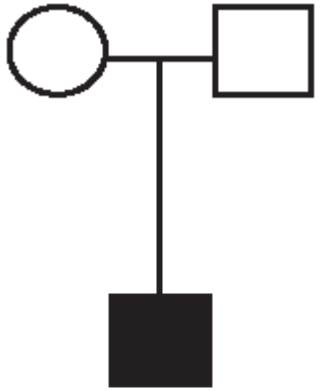
Systematic validation using Sanger sequencing

- 38 not validated in proband

  → *Median variant reads: 5*

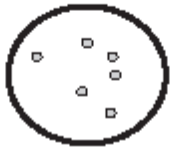
- 13 validated: **9 *de novo*!!!**

  → *Median variant reads: 17*



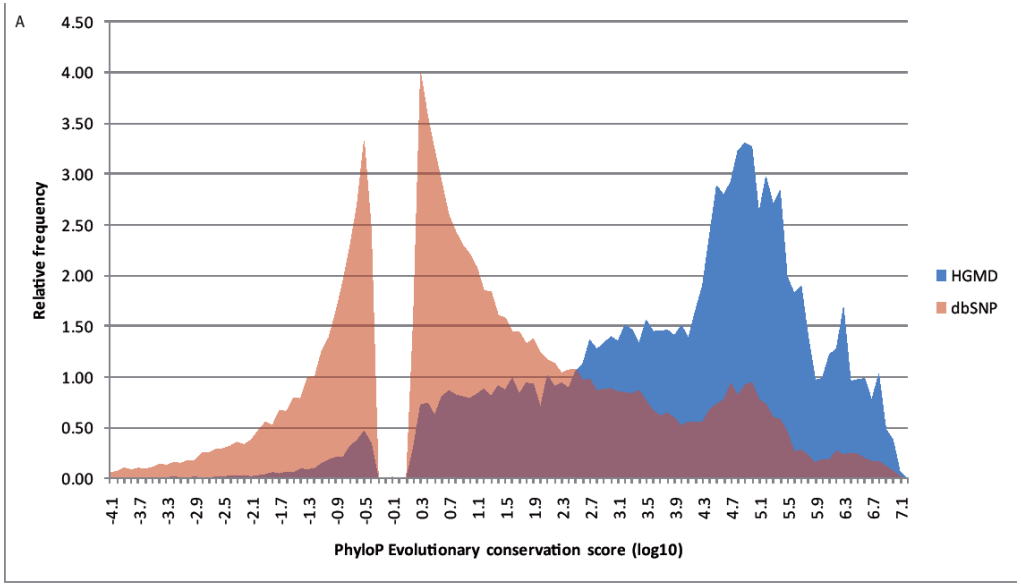
Candidate strategy

- Selection of variants based on **variant** and **gene** interpretation
- Traditional gene prioritization techniques[12]
- Variant interpretation: Polyphen, SIFT, Mutpred, *etc*.
- Evolutionary conservation



[12]Erlich *et al*. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 2011

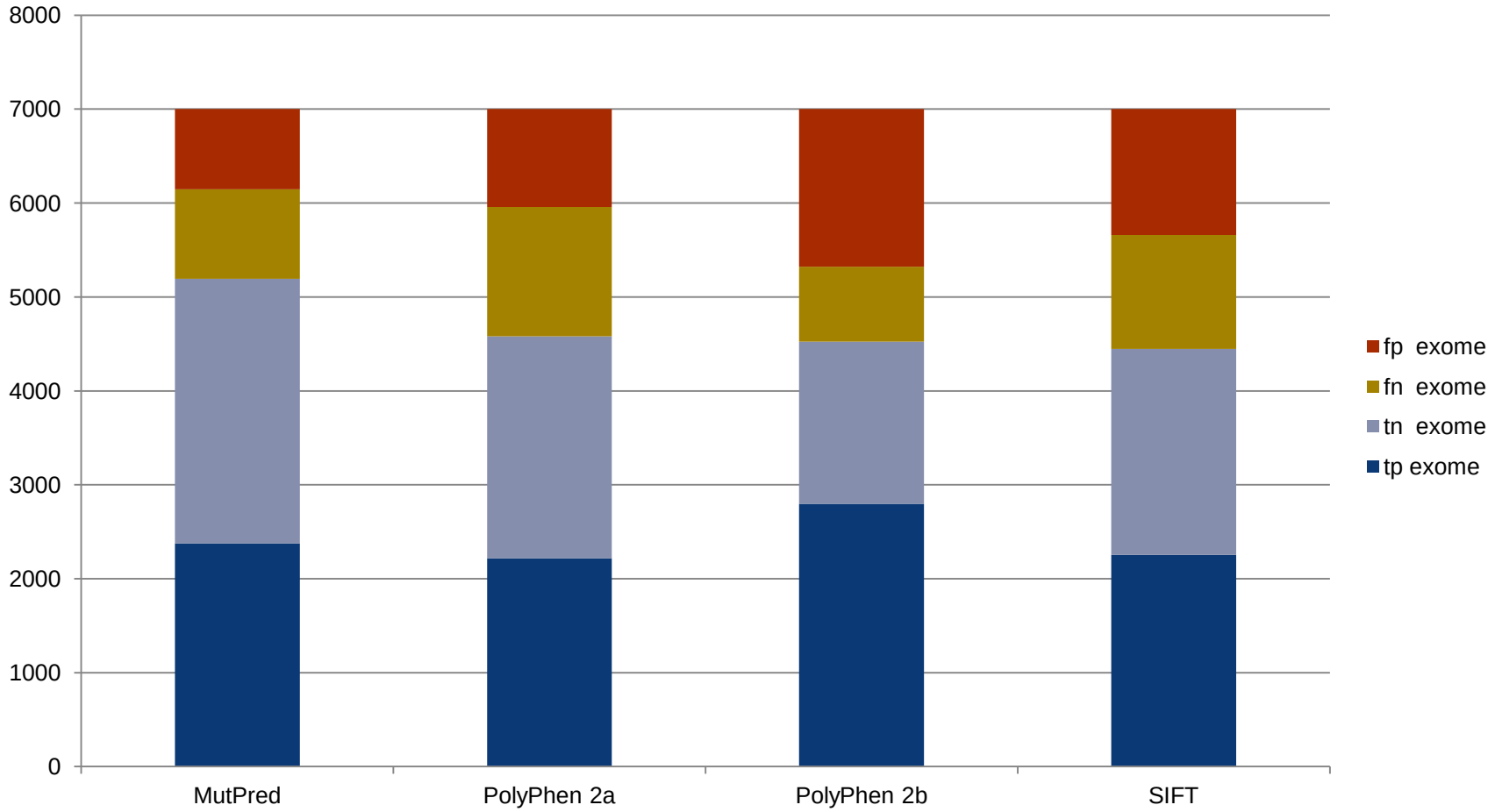
Evolutionary conservation for variant prioritization



Part III – Computational Predictions

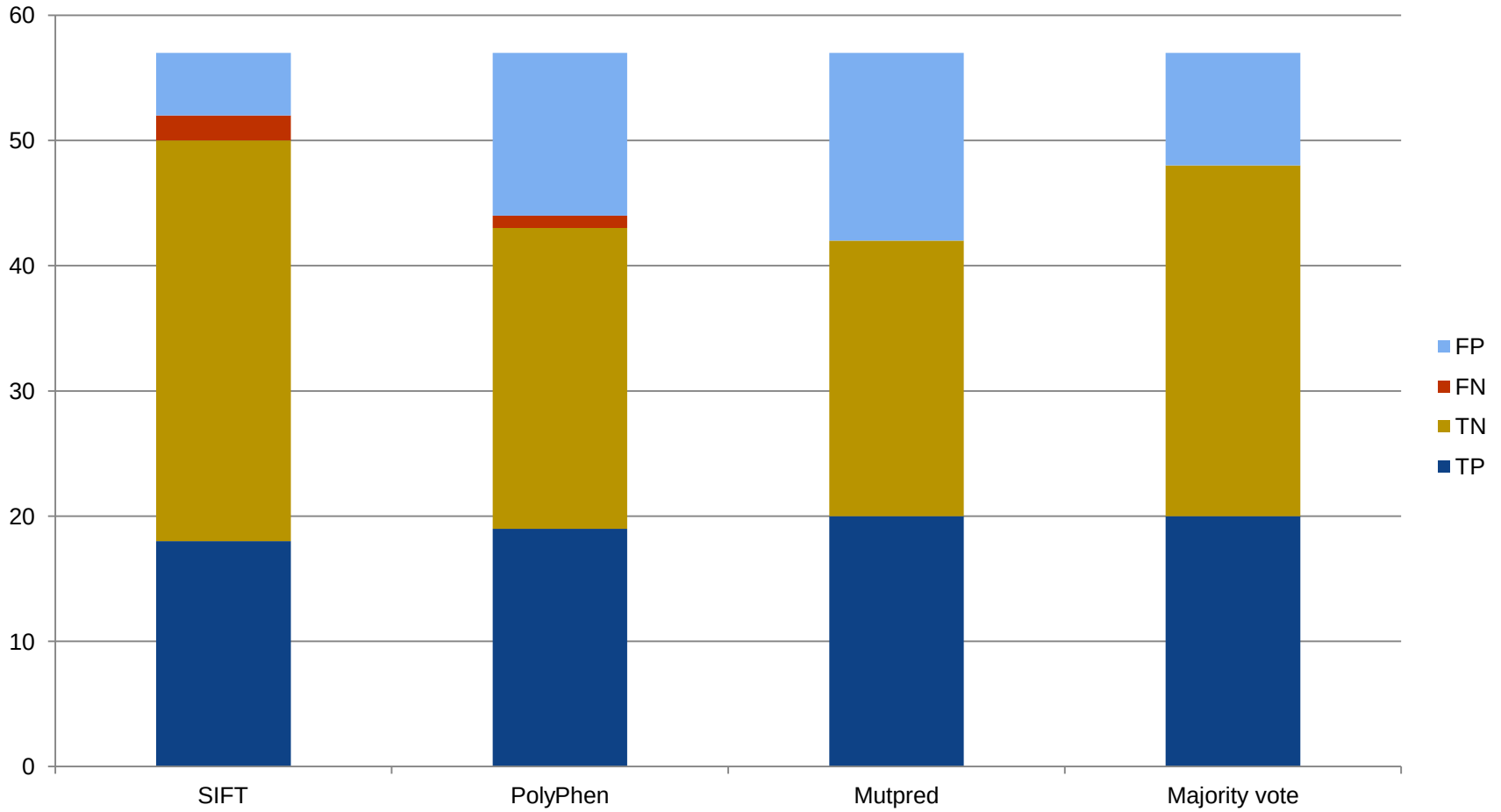
- **Polyphen2:** Bayesian classification based on sequence/structure attributes and MSA (http://genetics.bwh.harvard.edu/pph2/)
- **Mutpred**: Random forest classification on protein structure attributes and evolutionary attributes. (http://mutpred.mutdb.org/)
- **SIFT:** probability of substitution tolerance based on MSA (http://sift.jcvi.org/)
- **Mutation taster**: Naïve bayes classifier, sequence distribution and protein domains (http://www.mutationtaster.org/)

Performance comparison of prediction programs



Thusberg *et. al* Hum mut. 2011

Prediction on 57 blindness variants



Neveling *et al. Hum mut. 2012*

Conclusions

- Open source annotation tools available for variant annotation
- Think about your method of prioritization before starting any experiments. Most successful studies:
  - Clear Mendelian disorders
  - Good control dataset
  - Family members available for follow up
  - Cohort available for finding recurrence
- Pathogenicity prediction can help but should be used with care.


Joris Veltman


Alex Hoischen


Lisenka Vissers


Bregje van Bon


Han Brunner













All families
& clinicians
involved!