University of Helsinki

Biometry and Bioinformatics II
Fall 2013

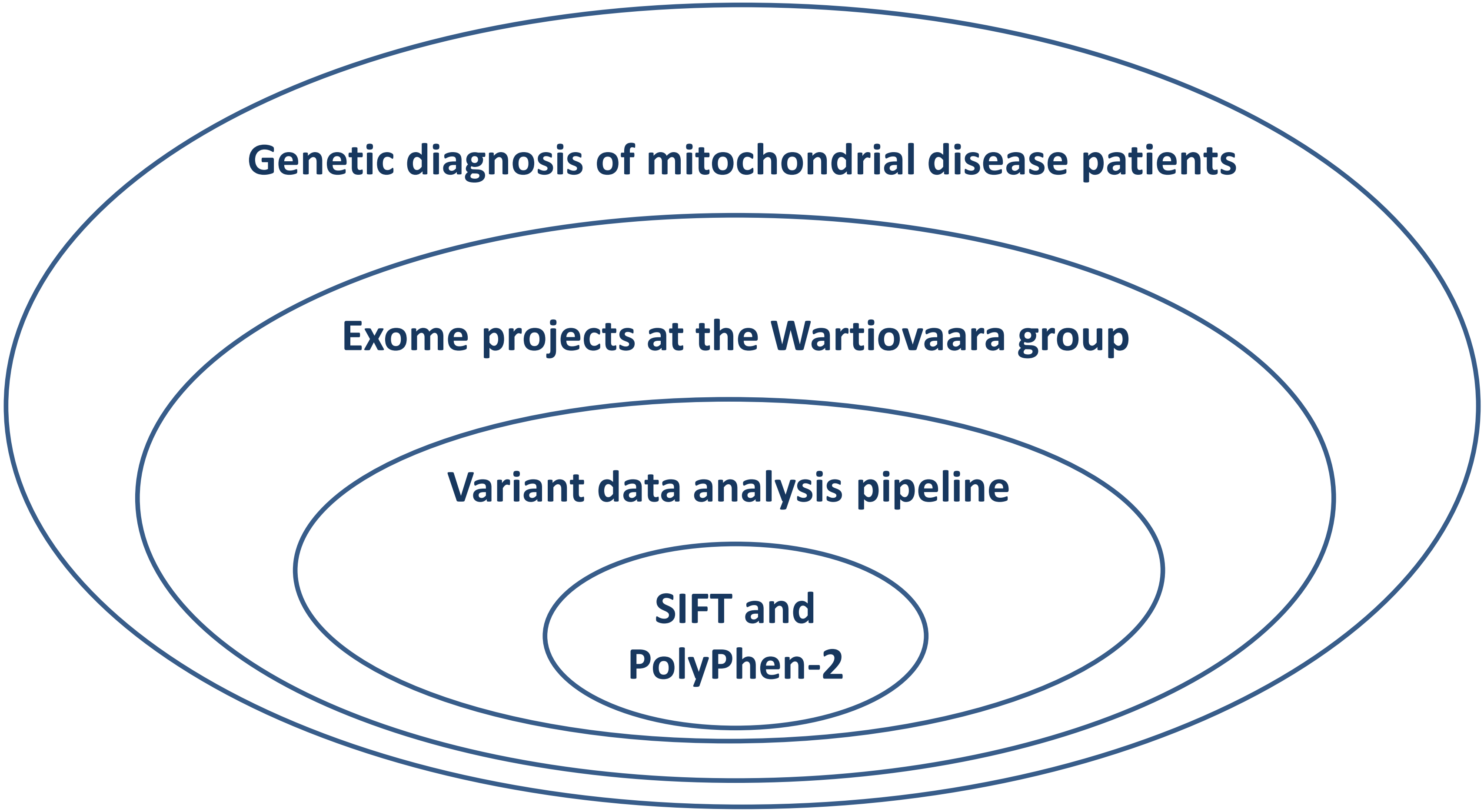
In silico prediction of protein-damaging single nucleotide variants

Virginia Brilhante

Helsinki, 9.10.2013

Outline (1/2)

- Background



Genetic diagnosis of mitochondrial disease patients

Exome projects at the Wartiovaara group

Variant data analysis pipeline

SIFT and PolyPhen-2

Outline (2/2)

- Software tools for prediction of protein-damaging SNVs
  - What are they (for)?
  - Some indicators of value
  - Evolutionary conservation premise

SIFT

- Algorithm overview
- Score, prediction and confidence

PolyPhen-2

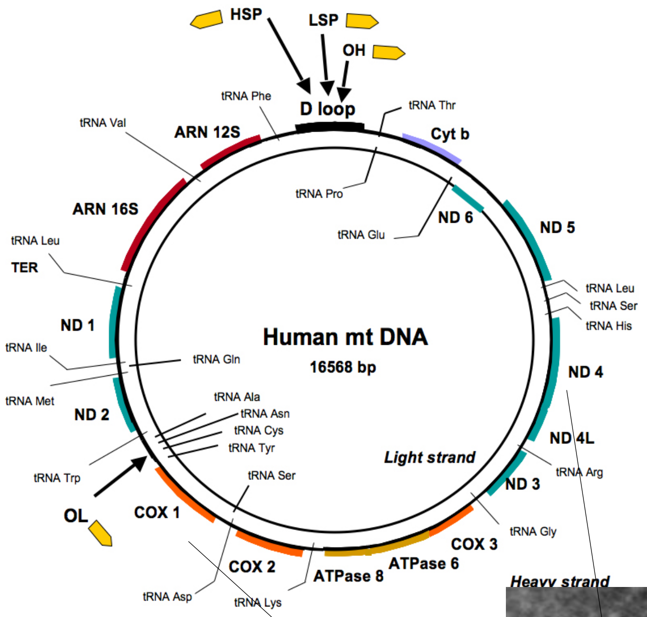
- Algorithm overview
- Score, prediction and additional estimates

- Distinct tools, distinct predictions
- Considerations on accuracy and use in diagnostics
- Summary

Application: genetic diagnosis of mito disease patients

- Genetic diagnosis of suspected mitochondrial disease patients; better understanding of mitochondrial disorders
  - mitochondria are the organelles where cellular energy is generated
  - mitochondrial dysfunction
    - defective mitochondria-located proteins
  - bigenomic

An organelle that needs two genomes

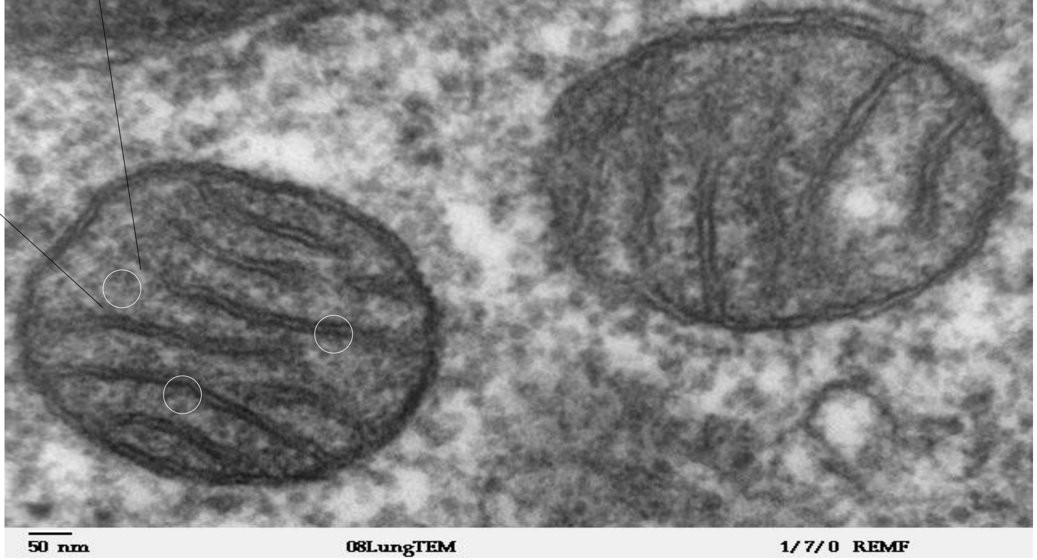


Source:
[Bellance 2009]

mtDNA
encodes
13 proteins

Nuclear DNA
encodes
~1500 proteins
imported into
mitochondria

Source: Wikipedia



Source: Wikipedia

Mitochondria

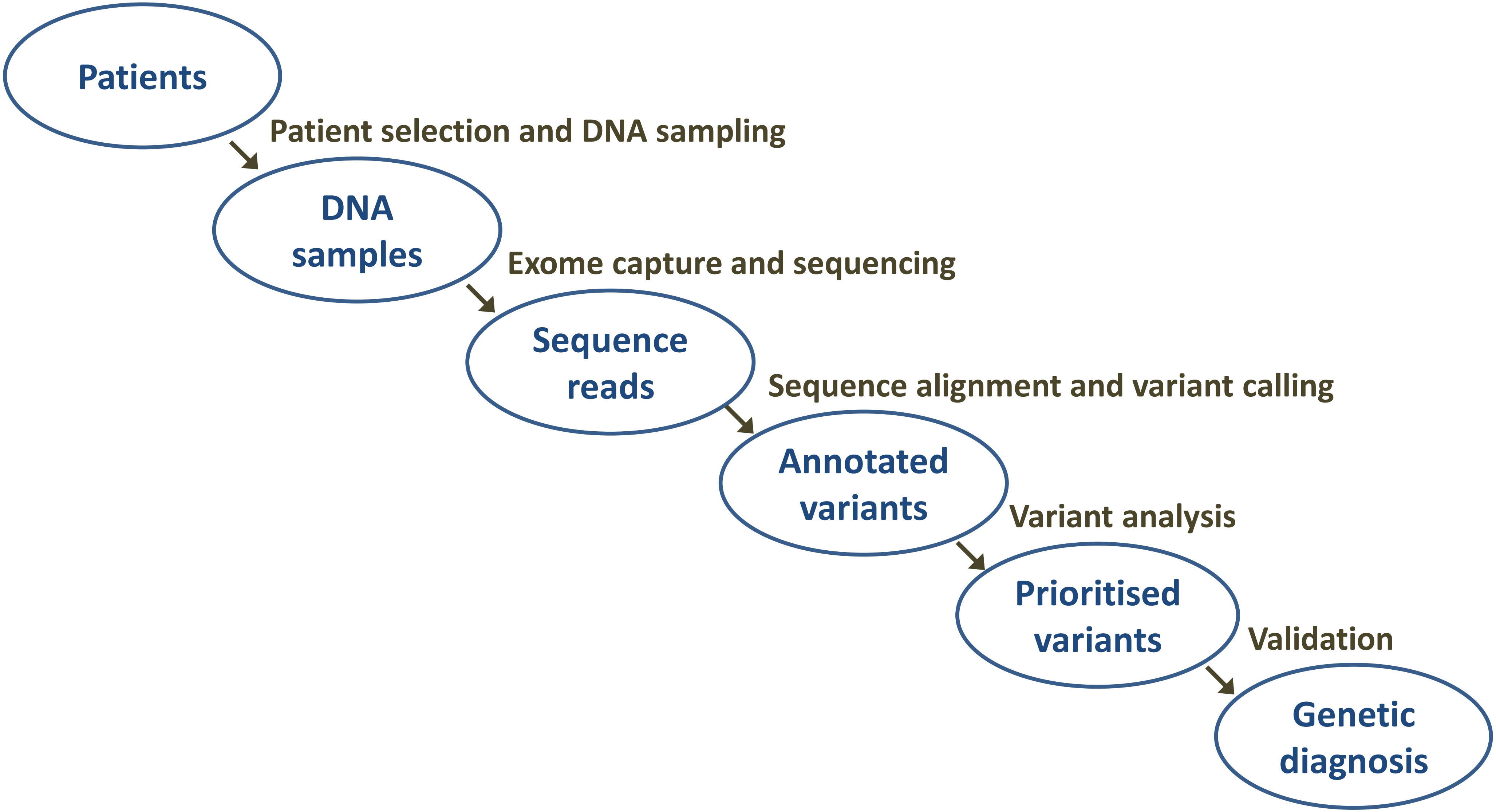
Application: genetic diagnosis of mito disease patients

- Causative mutations of mito disease
  - inherited
    - maternal (mtDNA), X-linked, autosomal dominant
    - autosomal recessive
      - supported by:
        - population structure in Finland
          - increased likelihood of some degree of parental consanguinity
        - suspected disorders in patient cohort
      - homozygous and compound heterozygous variants
  - *de novo* (sporadic)

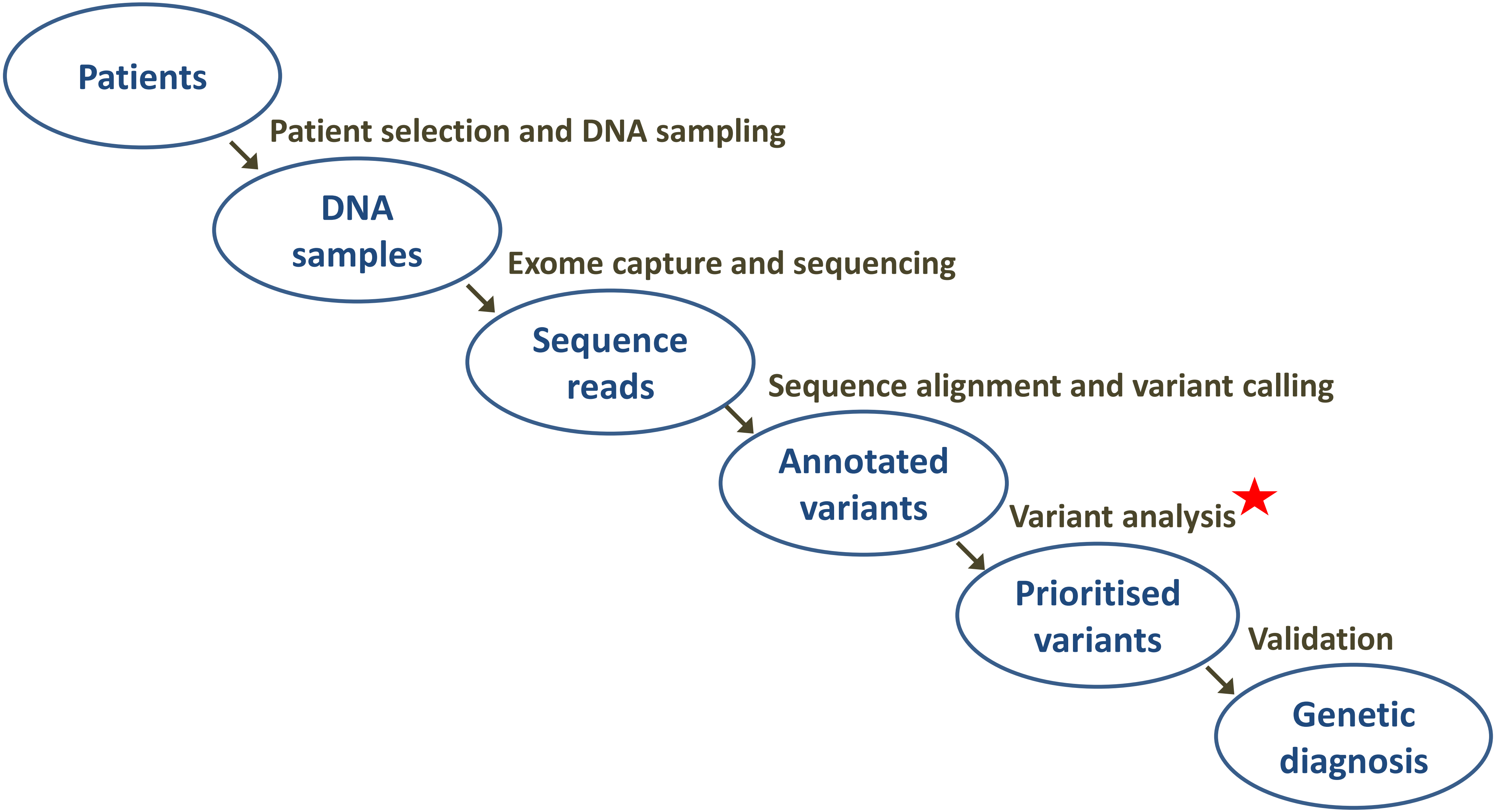
Prediction of protein-damaging SNVs: one step within an exome variant data analysis pipeline

- Exome projects at the Wartiovaara group
  - partnership with the Institute for Molecular Medicine Finland (FIMM)
    - exome sequencing and variant calling
  - ~ 100 patients sequenced so far
- Exome
  - all exons of all genes in a genome – protein coding regions
  - ~1% of the human genome
  - holds majority of mutations currently known to associate with hereditary diseases

Prediction of protein-damaging SNVs:
one step within an exome variant data analysis pipeline



Prediction of protein-damaging SNVs:
one step within an exome variant data analysis pipeline



Patients

Patient selection and DNA sampling

DNA samples

Exome capture and sequencing

Sequence reads

Sequence alignment and variant calling

Annotated variants

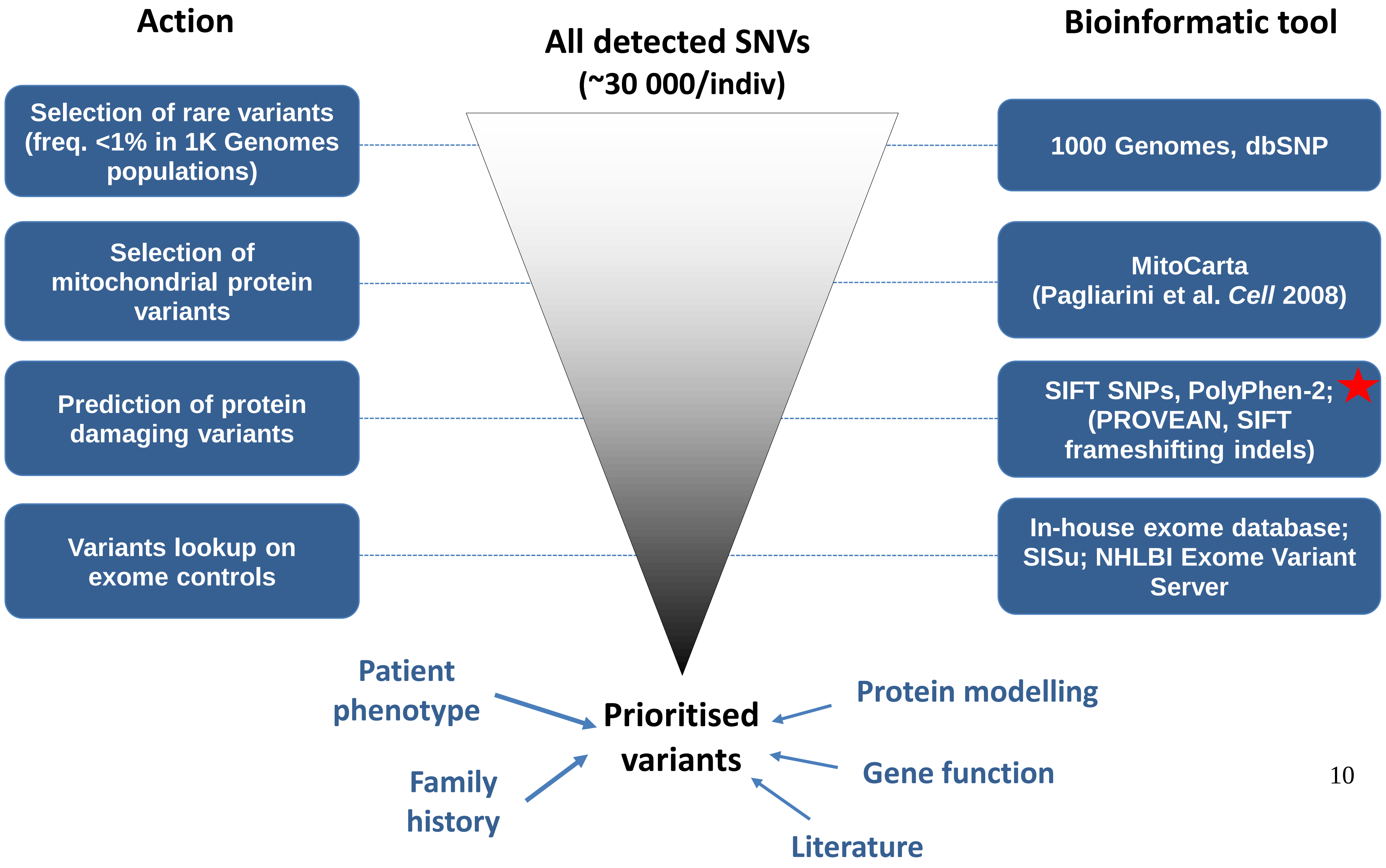
Variant analysis ★

Prioritised variants

Validation

Genetic diagnosis

Exome variant data analysis



All detected SNVs (~30 000/indiv)

Action	Bioinformatic tool
Selection of rare variants (freq. <1% in 1K Genomes populations)	1000 Genomes, dbSNP
Selection of mitochondrial protein variants	MitoCarta (Pagliarini et al. *Cell* 2008)
Prediction of protein damaging variants	SIFT SNPs, PolyPhen-2; (PROVEAN, SIFT frameshifting indels)
Variants lookup on exome controls	In-house exome database; SISu; NHLBI Exome Variant Server

Prioritised variants

Patient phenotype

Family history

Protein modelling

Gene function

Literature

Group Wartiovaara, Research Program for Molecular Neurology
Biomedicum Helsinki, University of Helsinki

Mission: To understand the molecular background of mitochondrial disorders, and use that knowledge to develop diagnosis and therapy.

Software tools for prediction of protein-damaging SNVs

- What are they for?
  - prediction of the propensity of individual amino acid changes to damage protein function
  - restricted to aa substitutions caused by non-synonymous single nucleotide variants (nsSNVs) in DNA
    - make up more than 50% of human genetic variation known to be involved in inherited diseases
      - missense deleterious (or pathogenic) mutations

- In Craig Venter's genome:
  - 3 213 401 SNVs
  - 3 882 nsSNVs

Some indicators of value (1/2)

- SIFT and PolyPhen-2 are widely used
  - publicly available, Web-based tools

- Many other tools exist: Condel, Mutation taster, Panther, MAPP, etc.

SIFT

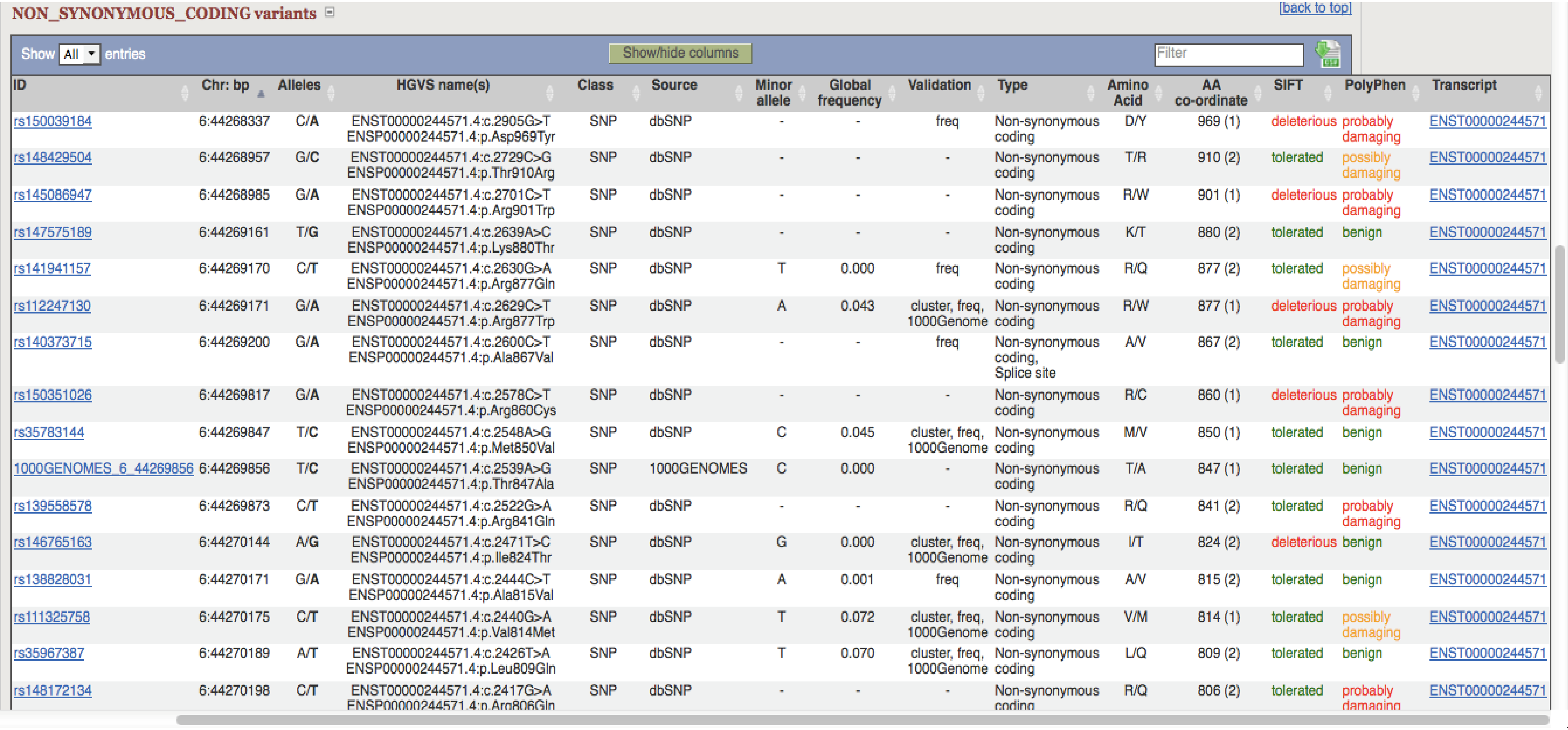
- developed at the Fred Hutchinson Cancer Research Center
- first published in 2001
- published in **nature protocols** in 2009
- server in J. Craig Venter Institute for about 6 years
- open source

PolyPhen-2

- main authors affiliated to Harvard Medical School and Max Planck Institute
- successor of PolyPhen published in 2002
- published in **nature methods** in 2010

Some indicators of value (2/2)

- 1000 Genomes



ID	Chr: bp	Alleles	HGVS name(s)	Class	Source	Minor allele	Global frequency	Validation	Type	Amino Acid	AA co-ordinate	SIFT	PolyPhen	Transcript
rs150039184	6:44268337	C/A	ENST00000244571.4:c.2905G>T ENSP00000244571.4:p.Asp969Tyr	SNP	dbSNP	-	-	freq	Non-synonymous coding	D/Y	969 (1)	deleterious	probably damaging	ENST00000244571
rs148429504	6:44268957	G/C	ENST00000244571.4:c.2729C>G ENSP00000244571.4:p.Thr910Arg	SNP	dbSNP	-	-	-	Non-synonymous coding	T/R	910 (2)	tolerated	possibly damaging	ENST00000244571
rs145086947	6:44268985	G/A	ENST00000244571.4:c.2701C>T ENSP00000244571.4:p.Arg901Trp	SNP	dbSNP	-	-	-	Non-synonymous coding	R/W	901 (1)	deleterious	probably damaging	ENST00000244571
rs147575189	6:44269161	T/G	ENST00000244571.4:c.2639A>C ENSP00000244571.4:p.Lys880Thr	SNP	dbSNP	-	-	-	Non-synonymous coding	K/T	880 (2)	tolerated	benign	ENST00000244571
rs141941157	6:44269170	C/T	ENST00000244571.4:c.2630G>A ENSP00000244571.4:p.Arg877Gln	SNP	dbSNP	T	0.000	freq	Non-synonymous coding	R/Q	877 (2)	tolerated	possibly damaging	ENST00000244571
rs112247130	6:44269171	G/A	ENST00000244571.4:c.2629C>T ENSP00000244571.4:p.Arg877Trp	SNP	dbSNP	A	0.043	cluster, freq, 1000Genome	Non-synonymous coding	R/W	877 (1)	deleterious	probably damaging	ENST00000244571
rs140373715	6:44269200	G/A	ENST00000244571.4:c.2600C>T ENSP00000244571.4:p.Ala867Val	SNP	dbSNP	-	-	freq	Non-synonymous coding, Splice site	A/V	867 (2)	tolerated	benign	ENST00000244571
rs150351026	6:44269817	G/A	ENST00000244571.4:c.2578C>T ENSP00000244571.4:p.Arg860Cys	SNP	dbSNP	-	-	-	Non-synonymous coding	R/C	860 (1)	deleterious	probably damaging	ENST00000244571
rs35783144	6:44269847	T/C	ENST00000244571.4:c.2548A>G ENSP00000244571.4:p.Met850Val	SNP	dbSNP	C	0.045	cluster, freq, 1000Genome	Non-synonymous coding	M/V	850 (1)	tolerated	benign	ENST00000244571
1000GENOMES_6_44269856	6:44269856	T/C	ENST00000244571.4:c.2539A>G ENSP00000244571.4:p.Thr847Ala	SNP	1000GENOMES	C	0.000	-	Non-synonymous coding	T/A	847 (1)	tolerated	benign	ENST00000244571
rs139558578	6:44269873	C/T	ENST00000244571.4:c.2522G>A ENSP00000244571.4:p.Arg841Gln	SNP	dbSNP	-	-	-	Non-synonymous coding	R/Q	841 (2)	tolerated	probably damaging	ENST00000244571
rs146765163	6:44270144	A/G	ENST00000244571.4:c.2471T>C ENSP00000244571.4:p.Ile824Thr	SNP	dbSNP	G	0.000	cluster, freq, 1000Genome	Non-synonymous coding	I/T	824 (2)	deleterious	benign	ENST00000244571
rs138828031	6:44270171	G/A	ENST00000244571.4:c.2444C>T ENSP00000244571.4:p.Ala815Val	SNP	dbSNP	A	0.001	freq	Non-synonymous coding	A/V	815 (2)	tolerated	benign	ENST00000244571
rs111325758	6:44270175	C/T	ENST00000244571.4:c.2440G>A ENSP00000244571.4:p.Val814Met	SNP	dbSNP	T	0.072	cluster, freq, 1000Genome	Non-synonymous coding	V/M	814 (1)	tolerated	possibly damaging	ENST00000244571
rs35967387	6:44270189	A/T	ENST00000244571.4:c.2426T>A ENSP00000244571.4:p.Leu809Gln	SNP	dbSNP	T	0.070	cluster, freq, 1000Genome	Non-synonymous coding	L/Q	809 (2)	tolerated	benign	ENST00000244571
rs148172134	6:44270198	C/T	ENST00000244571.4:c.2417G>A ENSP00000244571.4:p.Arg806Gln	SNP	dbSNP	-	-	-	Non-synonymous coding	R/Q	806 (2)	tolerated	probably damaging	ENST00000244571

- Proprietary software for analysis of NGS data

Evolutionary conservation premise

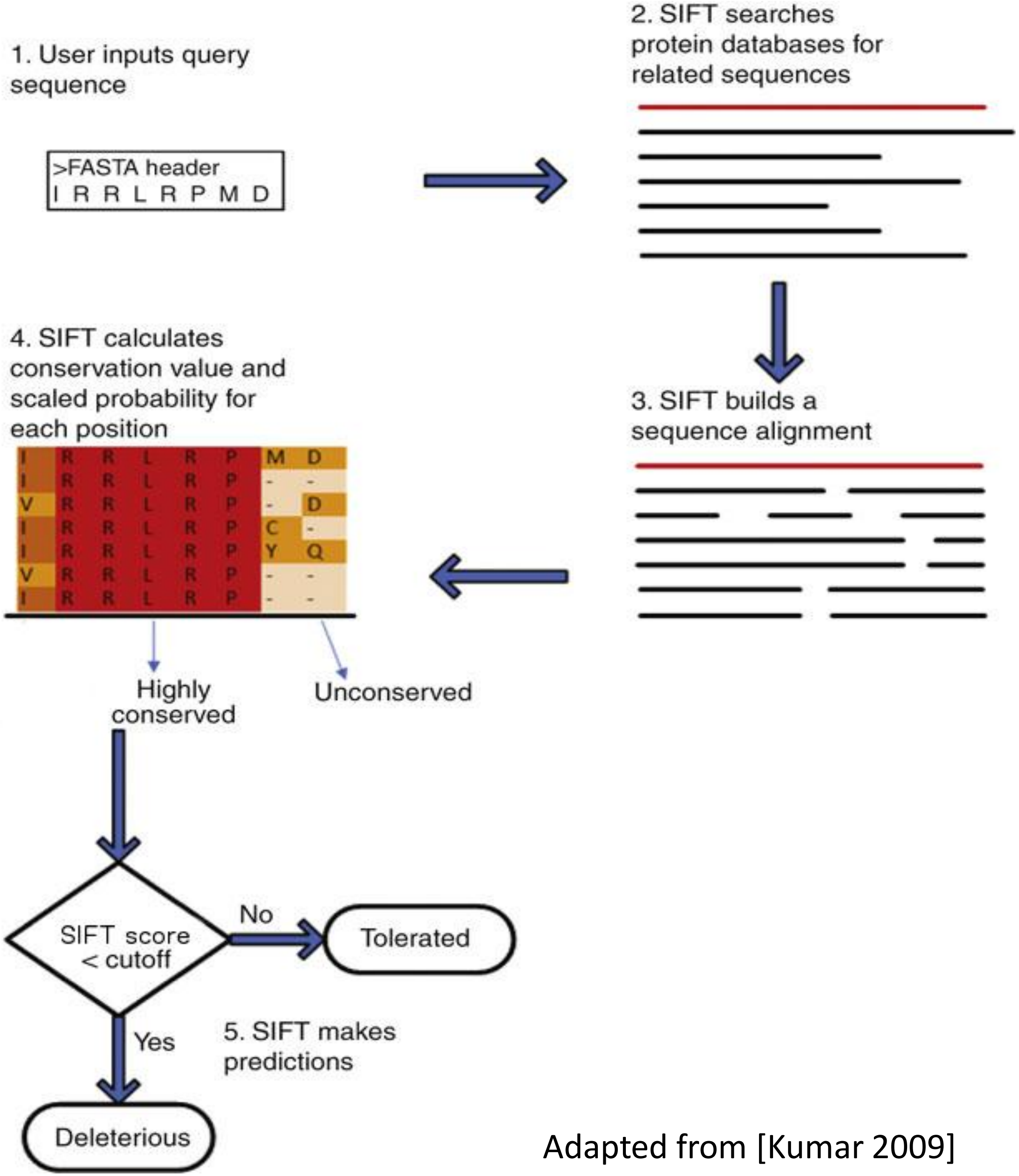
Important amino acids in a protein sequence are conserved

- Highly conserved amino acid positions in a protein sequence tend to be intolerant to substitution, whereas those with a low degree of conservation tolerate most substitutions
- Implicit assumption of change as deleterious
  - functional conservation
- Better applicability of the tools to monogenic diseases
  - similar conservation patterns between known complex disease nsSNVs and polymorphisms in the general population

SIFT

- Sorting Tolerant From Intolerant
- Predictions based only on conservation information obtained from a multiple alignment of homologous protein sequences

SIFT algorithm overview



1. User inputs query sequence

>FASTA header
I R R L R P M D

2. SIFT searches protein databases for related sequences

3. SIFT builds a sequence alignment

4. SIFT calculates conservation value and scaled probability for each position

I	R	R	L	R	P	M	D
I	R	R	L	R	P	-	-
V	R	R	L	R	P	-	D
I	R	R	L	R	P	C	-
I	R	R	L	R	P	Y	Q
V	R	R	L	R	P	-	-
I	R	R	L	R	P	-	-

Highly conserved

Unconserved

SIFT score < cutoff

No

Tolerated

Yes

Deleterious

5. SIFT makes predictions

Adapted from [Kumar 2009]

- Figure for protein input

2. BLAST algorithm;
UniProt and NCBI protein databases

3. Alignment of the query sequence with homologous sequences (MSA) found in step 2

4. Probabilities for all possible aa substitutions at each position used to estimate the SIFT score

- aa freqs. in MSA
- BLOSUM62 subst. scores

4. Conservation value is a measure of sequence diversity

SIFT score and prediction

- Score in the range [0, 1]
  - probability of an amino acid substitution caused by a nsSNV being tolerated

  - score ≥ 0.05: ‘TOLERATED’ prediction
    - functionally neutral substitution

  - score < 0.05: ‘DAMAGING’ prediction
    - substitution affects protein function

Sequence diversity and confidence in predictions (1/2)

- Apart from highly conserved protein families, too little diversity (or, too much conservation) between the homologous sequences is not desirable for prediction
  - e.g.
    - multiple sequences of the same organism/protein in the BLAST-searched databases
    - conservation by chance in elapsed evolutionary time
  - ideally, functionally conserved orthologous sequences

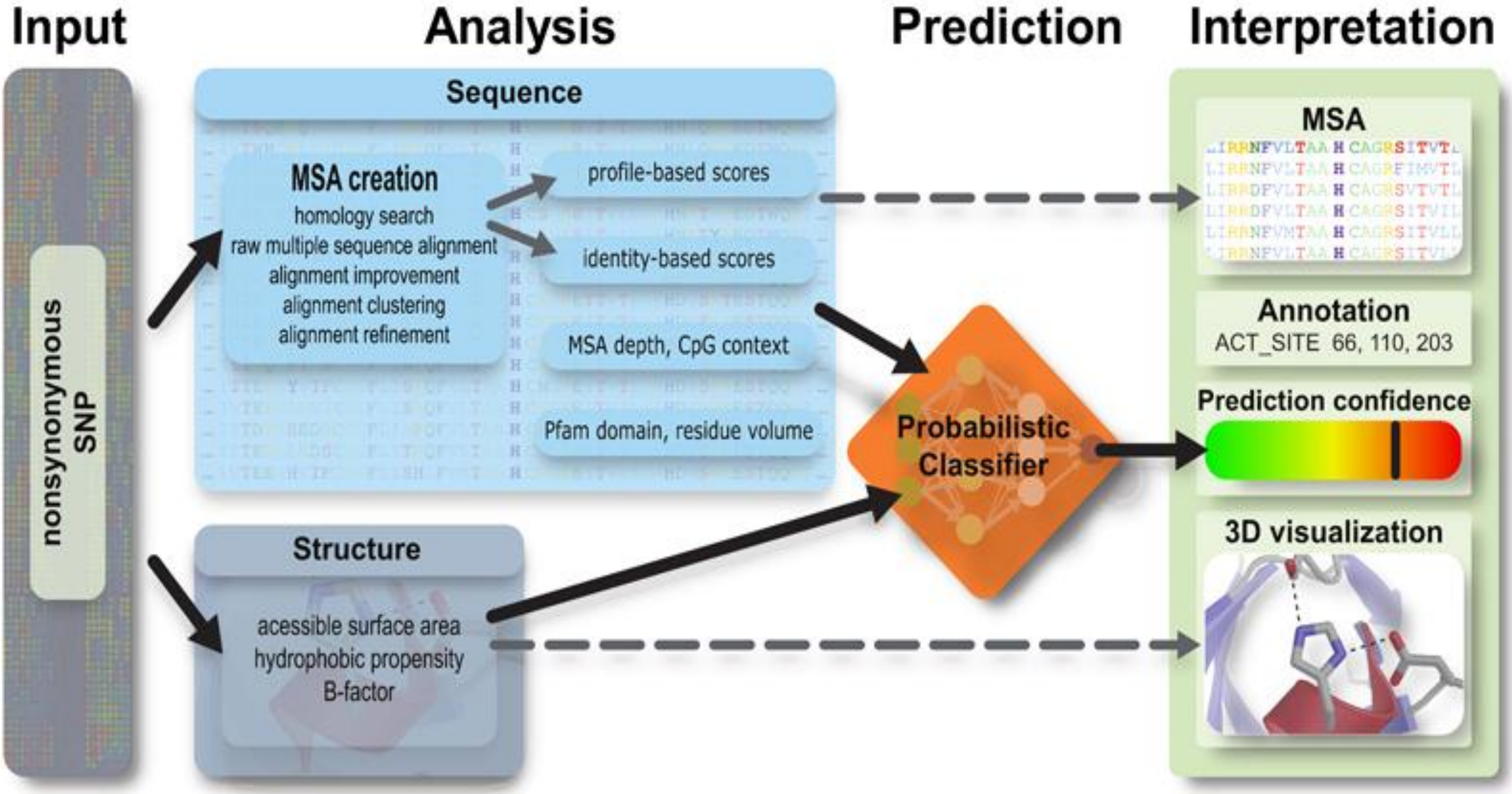
Sequence diversity and confidence in predictions (2/2)

- SIFT uses a conservation value (Median Information Content) for each position in the sequence alignment
    - range [0, $\log_2 20$ (=4.32)] for protein sequences
        - 0: “min” conservation – all 20 amino acids are observed
        - 4.32: “max” conservation – only one amino acid is observed
        - ~3: target median conservation value of final set of SIFT-aligned sequences
            - aiming at optimum diversity within selected sequences

PolyPhen-2

- Employs a combination of features for prediction of pathogenicity of missense mutations:
  - sequence homology (SIFT uses just this)
  - protein structure information
  - physicochemical properties of amino acids

PolyPhen-2 algorithm overview (1/3)



From [Adzhubei 2009]

PolyPhen-2 algorithm overview (2/3)

- Sequence-based and structure-based predictive features
  - latter limited to proteins with known 3D structures
- Homology search using the BLAST algorithm over the UniProt database
- Multi-step alignment algorithm:

1. initial alignment (MAFFT -- Multiple Alignment using Fast Fourier Transform)
2. refinement of poorly aligned segments (Leon)
3. phylogenetic clustering (ClusPack); cluster containing query seq. is selected
4. alignment of selected cluster (MAFFT again)

PolyPhen-2 algorithm overview (3/3)

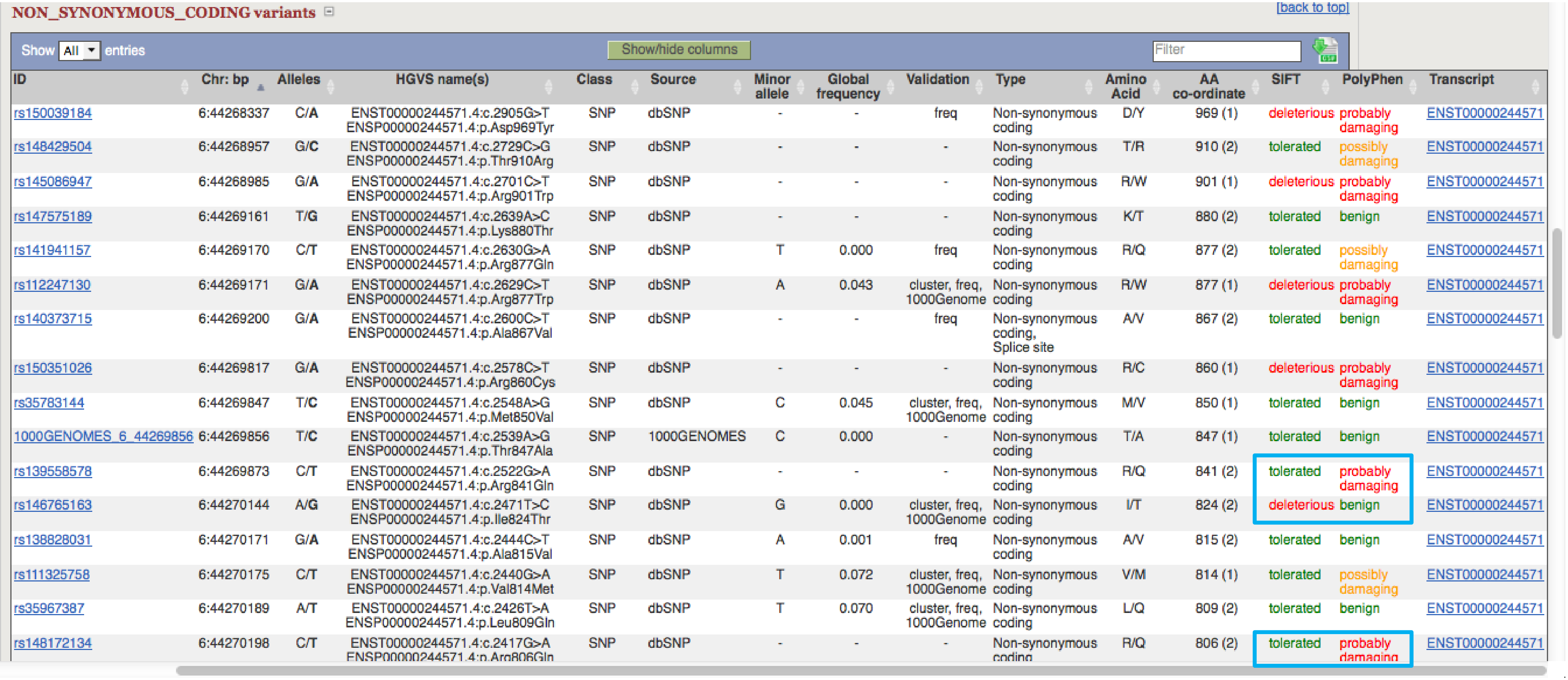
- Profile-based and identity-based scores
  - distinct MSA scopes
  - scores of conservation of an amino acid position using BLOSUM62 and considering, respectively:
    - the relatedness of the homologous sequences and the pattern of substitutions in the MSA as a whole
    - sequence identity between the query sequence and its closest homologues
- Probability (score) that a nsSNV is damaging (affects protein function) by a naïve Bayes classifier
  - assumptions of independence between the predictive features

PolyPhen-2 score and prediction

- nsSNV classes
  - $0.00 \leq \text{score} \leq 0.15$: BENIGN
  - $0.15 < \text{score} \leq 0.85$: POSSIBILY DAMAGING
  - $0.85 < \text{score} \leq 1.00$: PROBABLY DAMAGING

- Additional estimates
  - true positive rate (sensitivity)
  - true negative rate (specificity)

Distinct tools may give distinct predictions



NON_SYNONYMOUS_CODING variants

ID	Chr: bp	Alleles	HGVS name(s)	Class	Source	Minor allele	Global frequency	Validation	Type	Amino Acid	AA co-ordinate	SIFT	PolyPhen	Transcript
rs150039184	6:44268337	C/A	ENST00000244571.4:c.2905G>T ENSP00000244571.4:p.Asp969Tyr	SNP	dbSNP	-	-	freq	Non-synonymous coding	D/Y	969 (1)	deleterious	probably damaging	ENST00000244571
rs148429504	6:44268957	G/C	ENST00000244571.4:c.2729C>G ENSP00000244571.4:p.Thr910Arg	SNP	dbSNP	-	-	-	Non-synonymous coding	T/R	910 (2)	tolerated	possibly damaging	ENST00000244571
rs145086947	6:44268985	G/A	ENST00000244571.4:c.2701C>T ENSP00000244571.4:p.Arg901Trp	SNP	dbSNP	-	-	-	Non-synonymous coding	R/W	901 (1)	deleterious	probably damaging	ENST00000244571
rs147575189	6:44269161	T/G	ENST00000244571.4:c.2639A>C ENSP00000244571.4:p.Lys880Thr	SNP	dbSNP	-	-	-	Non-synonymous coding	K/T	880 (2)	tolerated	benign	ENST00000244571
rs141941157	6:44269170	C/T	ENST00000244571.4:c.2630G>A ENSP00000244571.4:p.Arg877Gln	SNP	dbSNP	T	0.000	freq	Non-synonymous coding	R/Q	877 (2)	tolerated	possibly damaging	ENST00000244571
rs112247130	6:44269171	G/A	ENST00000244571.4:c.2629C>T ENSP00000244571.4:p.Arg877Trp	SNP	dbSNP	A	0.043	cluster, freq, 1000Genome	Non-synonymous coding	R/W	877 (1)	deleterious	probably damaging	ENST00000244571
rs140373715	6:44269200	G/A	ENST00000244571.4:c.2600C>T ENSP00000244571.4:p.Ala867Val	SNP	dbSNP	-	-	freq	Non-synonymous coding, Splice site	A/V	867 (2)	tolerated	benign	ENST00000244571
rs150351026	6:44269817	G/A	ENST00000244571.4:c.2578C>T ENSP00000244571.4:p.Arg860Cys	SNP	dbSNP	-	-	-	Non-synonymous coding	R/C	860 (1)	deleterious	probably damaging	ENST00000244571
rs35783144	6:44269847	T/C	ENST00000244571.4:c.2548A>G ENSP00000244571.4:p.Met850Val	SNP	dbSNP	C	0.045	cluster, freq, 1000Genome	Non-synonymous coding	M/V	850 (1)	tolerated	benign	ENST00000244571
1000GENOMES_6_44269856	6:44269856	T/C	ENST00000244571.4:c.2539A>G ENSP00000244571.4:p.Thr847Ala	SNP	1000GENOMES	C	0.000	-	Non-synonymous coding	T/A	847 (1)	tolerated	benign	ENST00000244571
rs139558578	6:44269873	C/T	ENST00000244571.4:c.2522G>A ENSP00000244571.4:p.Arg841Gln	SNP	dbSNP	-	-	-	Non-synonymous coding	R/Q	841 (2)	tolerated	probably damaging	ENST00000244571
rs146765163	6:44270144	A/G	ENST00000244571.4:c.2471T>C ENSP00000244571.4:p.Ile824Thr	SNP	dbSNP	G	0.000	cluster, freq, 1000Genome	Non-synonymous coding	I/T	824 (2)	deleterious	benign	ENST00000244571
rs138828031	6:44270171	G/A	ENST00000244571.4:c.2444C>T ENSP00000244571.4:p.Ala815Val	SNP	dbSNP	A	0.001	freq	Non-synonymous coding	A/V	815 (2)	tolerated	benign	ENST00000244571
rs111325758	6:44270175	C/T	ENST00000244571.4:c.2440G>A ENSP00000244571.4:p.Val814Met	SNP	dbSNP	T	0.072	cluster, freq, 1000Genome	Non-synonymous coding	V/M	814 (1)	tolerated	possibly damaging	ENST00000244571
rs35967387	6:44270189	A/T	ENST00000244571.4:c.2426T>A ENSP00000244571.4:p.Leu809Gln	SNP	dbSNP	T	0.070	cluster, freq, 1000Genome	Non-synonymous coding	L/Q	809 (2)	tolerated	benign	ENST00000244571
rs148172134	6:44270198	C/T	ENST00000244571.4:c.2417G>A ENSP00000244571.4:p.Arg806Gln	SNP	dbSNP	-	-	-	Non-synonymous coding	R/Q	806 (2)	tolerated	probably damaging	ENST00000244571

- Distinctions in composition of predictive features and algorithms

SIFT prediction accuracy

- SIFT [Kumar 2009]
  - when applied to a dataset of nsSNVs found in disease-affected individuals:
    - 69% of the disease-associated variants predicted to affect protein function (true positive rate)
  - when applied to a dataset of nsSNVs in healthy individuals:
    - 19% of the variants predicted to affected protein function (false positive rate)

PolyPhen-2 prediction accuracy

- PolyPhen-2 [Adzhubei 2010]
  - applied to two datasets compiled from UniProt with variants annotated as disease-causing and non-annotated variants (assumed benign)
    - variants associated with human Mendelian diseases
      - 92% true positive rate
      - 20% false positive rate
    - variants associated with human genetic disease, more generally
      - 73% true positive rate
      - 20% false positive rate

Prediction tools in diagnostics

- “SIFT is intended to guide future experiments and not intended for direct use in a clinical setting, because *in silico* predictions are not a substitute for laboratory experiments.” [Kumar 2009]

- Diagnostics of Mendelian diseases is mentioned as one of the applications of PolyPhen-2 in [Adzhubei 2010]

Summary

- SIFT and PolyPhen-2 are tools for predicting pathogenicity (damage to protein function) of missense mutations
  - great demand for computational prediction tools as sequencing technologies became more accessible
  - main underlying premise for prediction is evolutionary conservation
    - PolyPhen-2 uses amino acid chemistry and protein structure properties as added features
  - widely used in monogenic disease research settings with application in assisting genetic diagnosis
- SIFT and PolyPhen-2 often disagree and can be used as complementary tools

References

- Adzhubei I, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; 7(4):248–9.
- Bellance N, Lestienne P and Rossignol R. Mitochondria: from bioenergetics to the metabolic regulation of carcinogenesis. *Front Biosci* 2009; 14:4015–34.
- Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; 4(7):1073–81.
- Kumar S, Dudley JT, Filipski A, Liu L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Gene*t 2011; 27(9):377–86.