Variants prioritization: annotation and filtration steps

2nd Variant Effect Prediction Training
Prague, Czech Republic, 2017

Prof. Christophe Béroud,
Aix-Marseille University
INSERM UMR_S910 Medical Genetics and Functional Genomics

RD Connect

WP1: Coordination

Hanns Lochmüller Newcastle and TREAT-NMD

WP2: Patient registries

Domenica Taruscio
ISS and EPIRARE

WP3: Biobanks

Lucia Monaco
Fondaz. Telethon & EuroBioBank

WP4: Bioinformatics

Christophe Béroud
AMU & INSERM Marseille

WP5: Unified platform

Ivo Gut
CNAG Barcelona

WP6 Ethical/legal/social

Mats Hansson
Uppsala

WP7: Impact and innovation

Kate Bushby
Newcastle and EUCERD/ EJARD

The "genetics and bioinformatics" team

Marseille - France

Medical school of la Timone

Aix-Marseille Université and INSERM institute dedicated to Medical Genetics and functional Genomics

"Translational research in Rare Diseases"

Our team

Genetics and Bioinformatics team – Multidisciplinary group

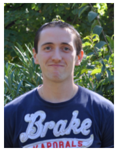
Team Leader

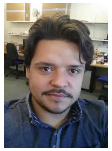
Christophe Béroud
Professor

Bioinformatics

David Salgado
PhD

Céline Guien
Engineer


Jean-Pierre Desvignes
Engineer


Marc Garibal
Engineer

Coralie Grattepanche
Master *Student*

Mélanie Corcuff
Master *Student*

Genetics of Dystonia

Gwenaëlle Collod-Béroud
Researcher

Arnaud Blanchard
PhD

Kahina Medjber
PhD

Constance Renault
Master Student

Our main projects

Locus Specific Mutation databases - UMD (Universal Mutation databases)

Gather >208,000 manually expert curated mutations in more than 60 databases:

- Genes involved in cancers (*APC, BRCA1, BRCA2, TP53, RB1, MEN1, SUR1, VHL, WT1*…)
- Genes involved in genetic disorders (*FBN1, LDLR, DMD, VLCAD, MCAD, LMNA, EMD, FKRP, SGCG, SGCA, ATP7B, TREAT-NMD_DMD, TREAT-NMD_SMA*…)

Patients registries – Population databases

Global TREAT-NMD DMD & SMA
International Dysferlinopathy Registry
French Database for Marfan and related Syndromes
National databases CNVs (BANCCO) + SNVs (RDVD Rare disease variant database)

Pathogenicity prediction systems

UMD-Predictor (http://umd-predictor.eu)
Human Splicing Finder (http://umd.be/HSF3/)

Next generation sequencing

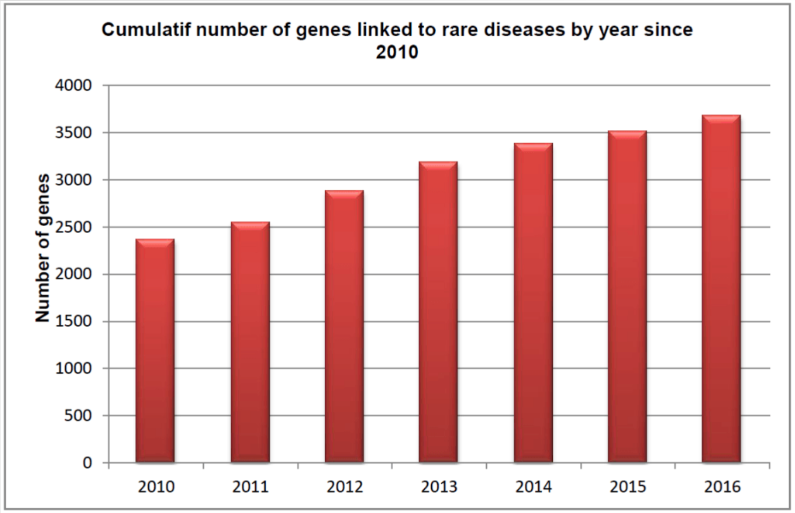
NGS Data analysis
Variant Annotation and Filtration tool (VarAFT)

Clinical tools

Skip-E for Antisense oligonucleotides to induce exon skipping
NR-Analyzer for nonsense mutations eligible for non-sense read-through
Crawfish for trans-splicing

Context

- Next generation Sequencing has facilitated the discovery of new genes and genetic variants in a multitude of human disorders
- 1st Whole-Exome Sequencing (WES) done by Ng et al 2009 in Miller Syndrome
- In 2013, >150 Mendelian disorders were studied by WES (Rabbani *et al*, J Hum Genet, 2014)
- IRDiRC and Orphanet ➛ 3,700 genes involved in RD, >1,300 identified between 2010 and 2016 (IRDiRC website)



IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

…But…

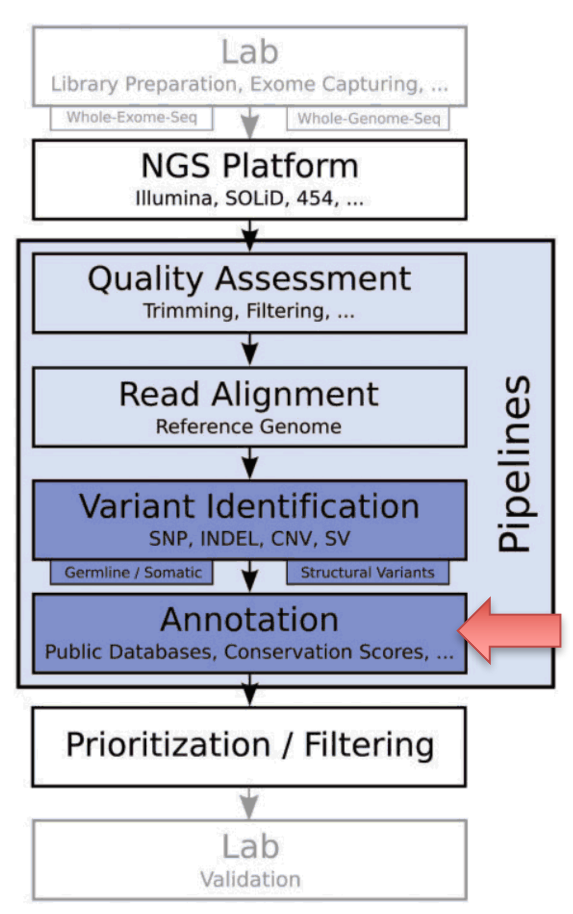
- Despite all these encouraging figures
  - Only 23-26% of WES are successful (higher rate if several individuals from the same family are sequenced 34-37% for a trio) [Farwell et al. 2015, Sawyer et al. 2016]
    - Technical factors (homopolymers, GC reach regions, poor quality at read ends …)
    - Type of disease causing mutations (not captured, triplet repeat expansions, CNVs, pseudogenes …)
    - Bioinformatics pipeline to generate VCF (same sequencing technology, not same VCF)
    - Wrong annotations/filtrations

…But…

- Despite all these encouraging figures
  - Only 23-26% of WES are successful (higher rate if several individuals from the same family are sequenced 34-37% for a trio) [Farwell et al. 2015, Sawyer et al. 2016]
    - Technical factors (homopolymers, GC reach regions, poor quality at read ends …)
    - Type of disease causing mutations (not captured, triplet repeat expansions, CNVs, pseudogenes …)
    - Bioinformatics pipeline to generate VCF (same sequencing technology, not same VCF)
    - **Wrong annotations/filtrations**

Variant annotation

- Part of the data analysis process
  - Mandatory for prioritization and filtration of variants
- Two objectives
  - Help to refine our estimate of how likely a variant is to be true, genotype, quality …
  - Provide functional annotations to determine the links between a genetic variation and a disease

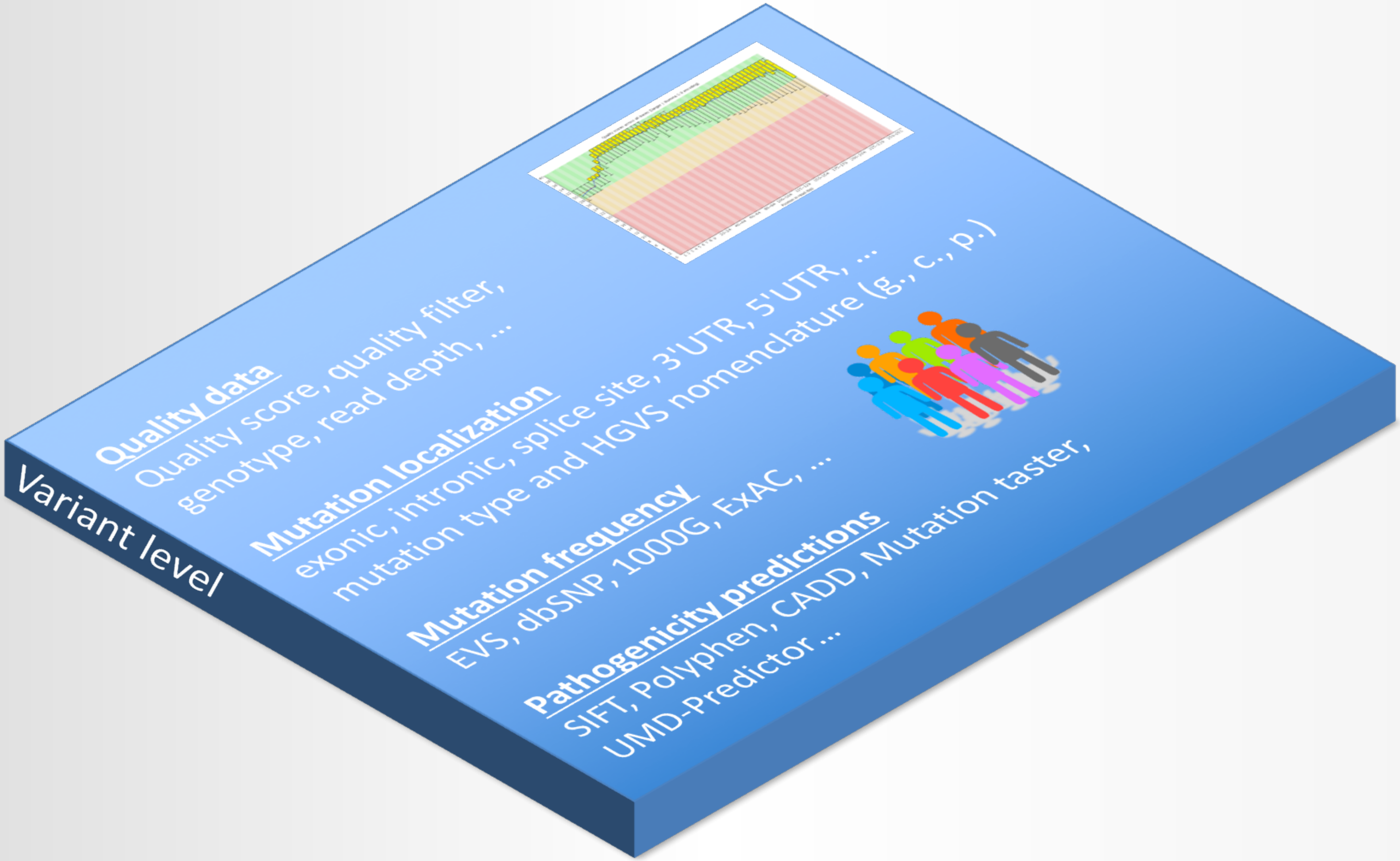


Adapted from Pabinger et al. Briefings in Bioinformatics 2014

Variant annotation

- It is performed at various levels

Variant annotation



Quality data
Quality score, quality filter,
genotype, read depth, …

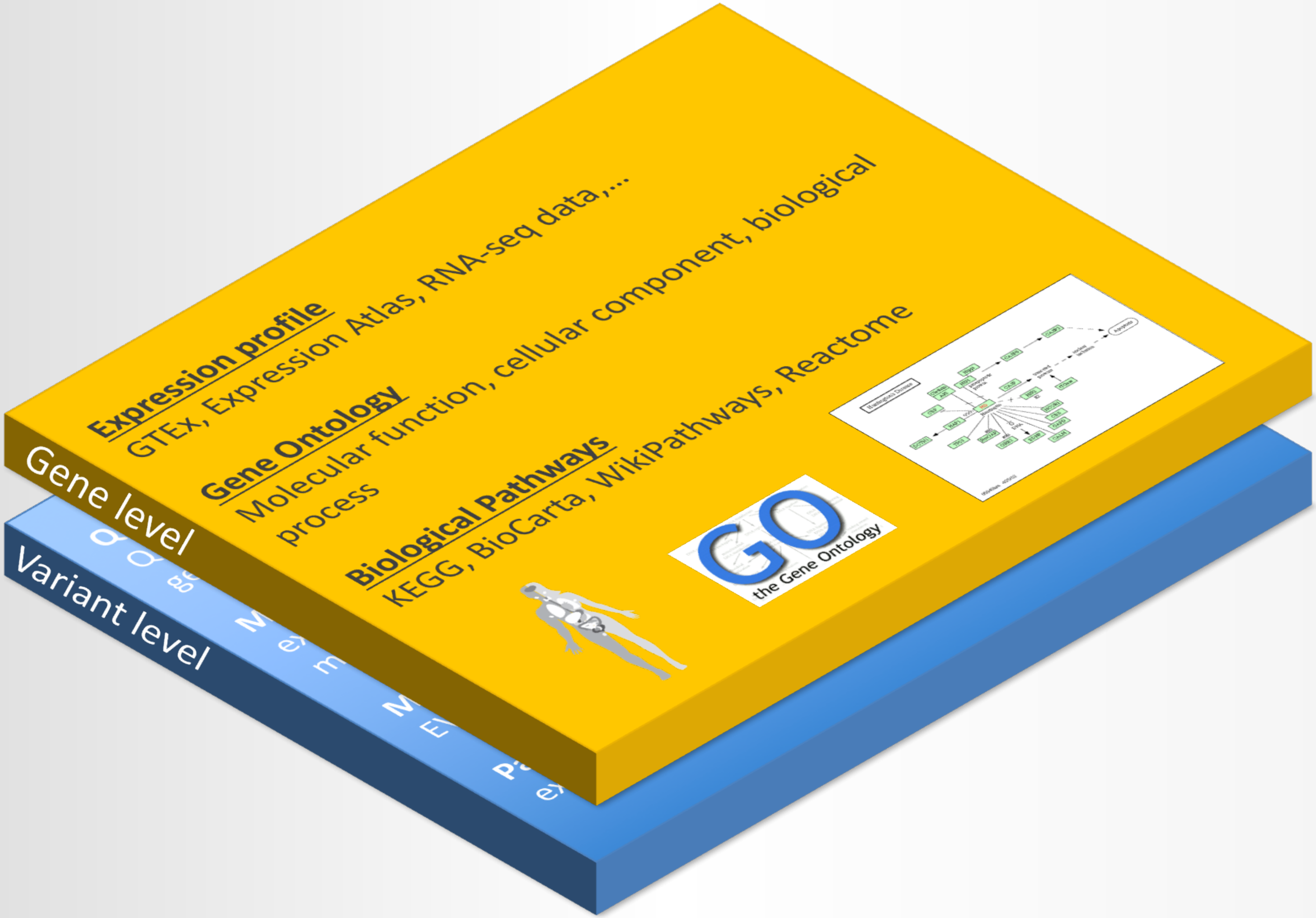
Mutation localization
exonic, intronic, splice site, 3'UTR, 5'UTR, …
mutation type and HGVS nomenclature (g., c., p.)

Mutation frequency
EVS, dbSNP, 1000G, ExAC, …

Pathogenicity predictions
SIFT, Polyphen, CADD, Mutation taster,
UMD-Predictor …

Variant level

Variant annotation



Gene level

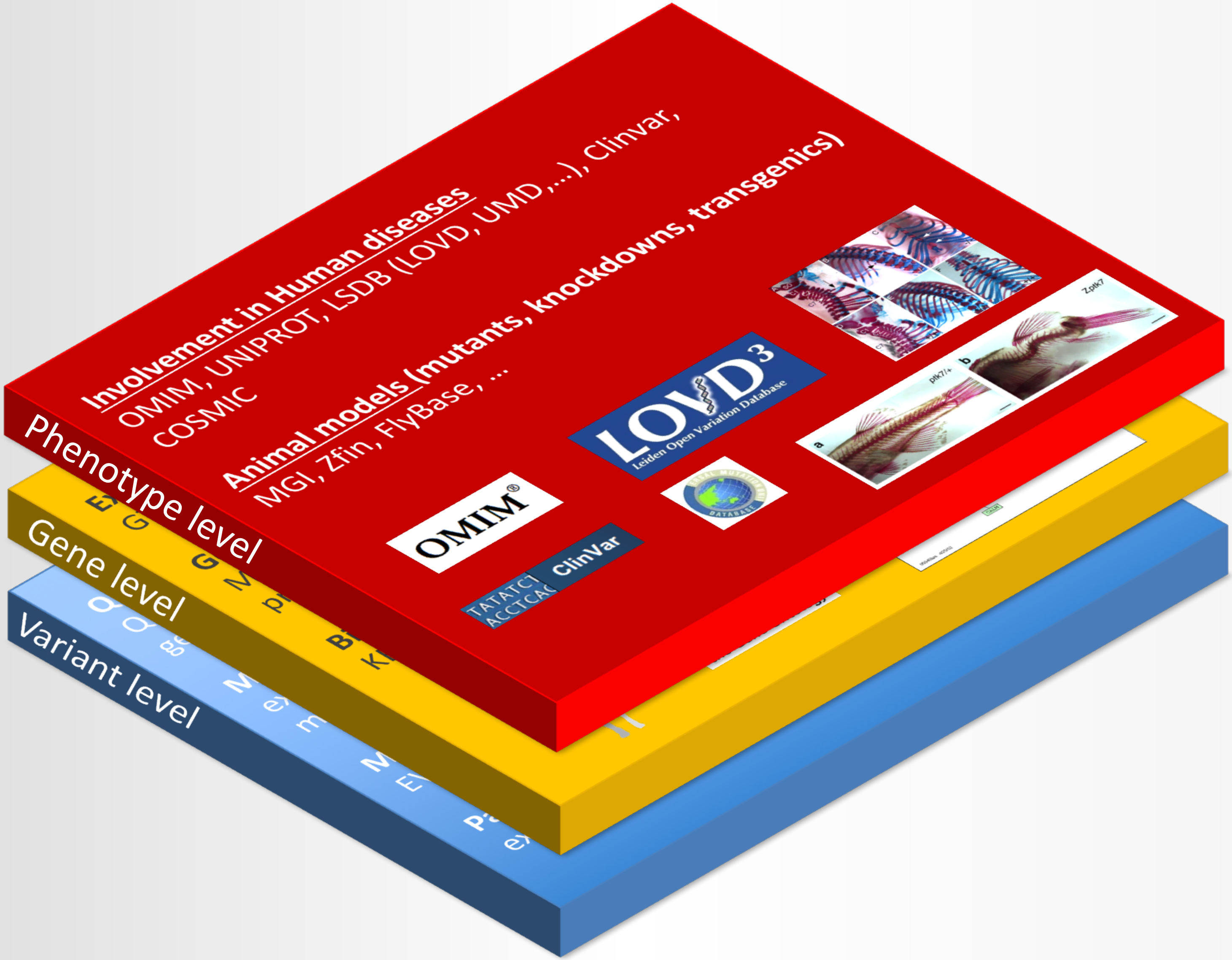
Expression profile
GTEx, Expression Atlas, RNA-seq data,…

Gene Ontology
Molecular function, cellular component, biological process

Biological Pathways
KEGG, BioCarta, WikiPathways, Reactome

Variant level

Variant annotation



Involvement in Human diseases

OMIM, UNIPROT, LSDB (LOVD, UMD,...), Clinvar, COSMIC

Animal models (mutants, knockdowns, transgenics)

MGI, Zfin, FlyBase, ...

Phenotype level

Gene level

Variant level

Variant annotation

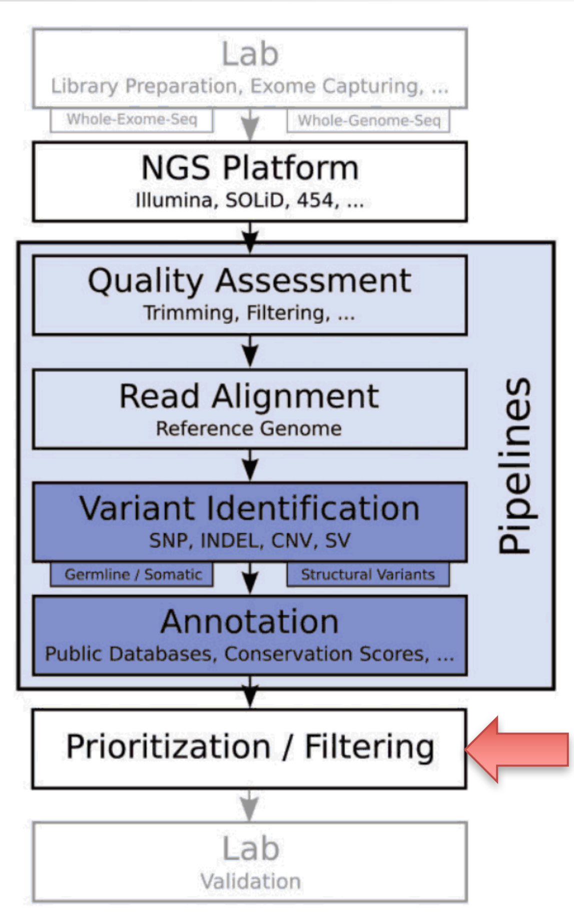
Variant annotation systems

	Annovar	SNPeff	Ensembl VEP	SeattleSeq	AnnTools	Oncotator	Vanno	Variant Annotation Tools
Availability	Command line	Command line	Command line Webservices Web	web	command line	Command line Web	Web	Command line
Variant quality	Yes	Yes	Yes	-	Yes	-	Yes	Yes
Variant localization	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gene/transcript annotation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Genotype	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes
Population frequency	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Impact at the RNA level	Yes	Yes	Yes	-	-	-	-	-
Impact at the protein level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Conservation	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes
Reported impact	-	-	Yes	Yes	-	Yes	Yes	Yes
Predicted pathogenicity	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes
Gene ontology	Yes	-	-	-	-	Yes	Yes	-
Pathways	-	-	-	Yes	-	-	Yes	Yes
Tissue expression	-	-	-	-	-	-	-	-

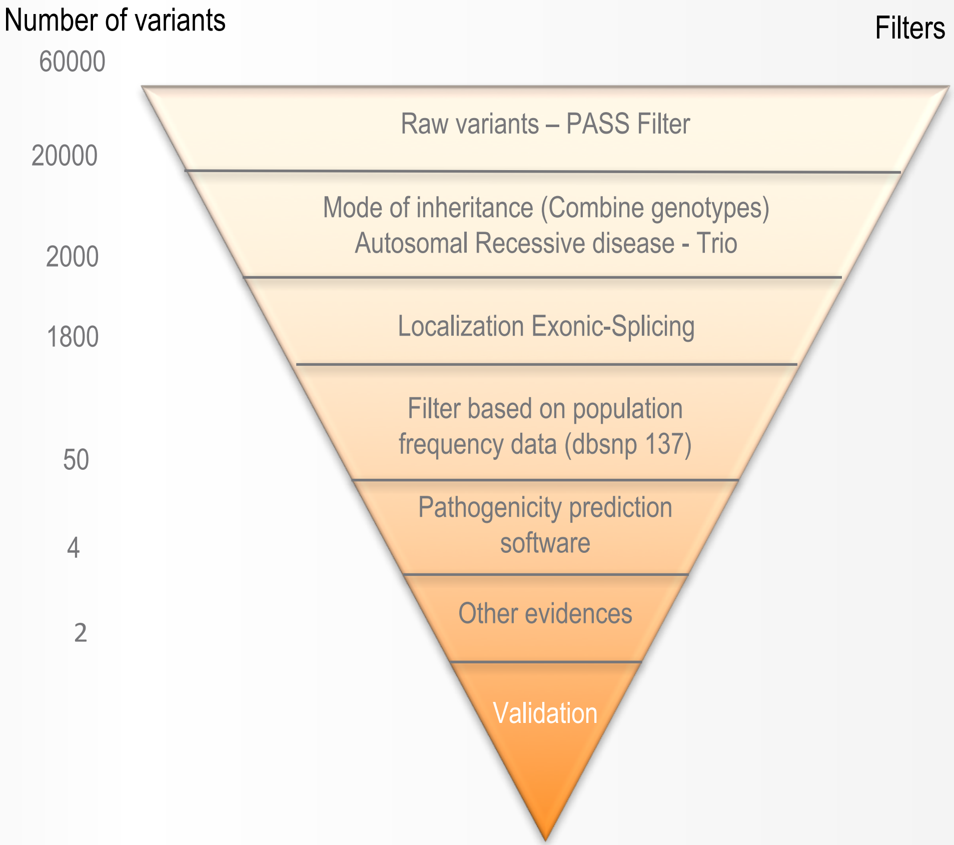
Various types of annotation software are available (command line/web)

No system is providing annotations at all levels ➛ need to be combined

Variant filtration



Adapted from Pabinger et al. Briefings in Bioinformatics 2014


Number of variants
Filters
60000
20000
2000
1800
50
4
2
Raw variants – PASS Filter
Mode of inheritance (Combine genotypes)
Autosomal Recessive disease - Trio
Localization Exonic-Splicing
Filter based on population
frequency data (dbsnp 137)
Pathogenicity prediction
software
Other evidences
Validation

Identify disease causing/private mutations

Filtration tools

- Automatic systems
  - Disease causing genes based on pedigree and phenotypic data

Software name	Availability	Mode of inheritance	Custom analysis	Mutation localization	Mutation type	Mutation frequency	Pathogenicity predictions	Functional evidences	Clinical report	Prioritization score
ExomeWalker	Web App	Yes	-	-	-	Yes	No but provided	No but provided	Yes	Yes
Exomiser	Command line	Yes	-	-	-	Yes	Yes	-	Yes	Yes
eXtasy	Command line Web App	-	-	-	-	-	No but provided	No but provided	-	Yes
MirTRIOS	Web App	Yes	-	Yes	Yes	Yes	Yes	-	No - But provided	Yes
OMIM Explorer	Web App	Yes*	-	-	-	-	-	-	Yes	Yes
OVA	Web App	Yes	-	Yes	Yes	Yes (2)	-	Yes	Yes	Yes
wKGGSeq	Web App	Yes	-	Yes	Yes	Yes	Yes	Yes	Yes	Yes

- Fast and accurate method for known genes/diseases
- Automatically gather additional information
- Work only with known genes/diseases with annotations
- Limited flexibility

* Does not combine multiple samples

Filtration tools

- Semi automatic/manual systems
  - Users can select candidate mutations by applying various set of filters

Software name	Availability	Mode of inheritance	Custom analysis	Mutation localization	Mutation type	Mutation frequency	Pathogenicity predictions	Functional evidences	Clinical report
ANNOVAR	Command line	-	-	-	-	Yes	Yes	-	Yes
BIERapp	Web App	Yes	Yes	Yes	Yes	Yes	-	-	-
FILTUS	GUI	Yes	Yes	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided
FMFilter	GUI	Yes	-	Yes - if provided	Yes - if provided	Yes - if provided	-	-	-
Gemini	Command line Web App	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vanno	Web App	-	-	Yes	Yes	Yes	No but provided	Yes	Yes
VarAFT	GUI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No - But provided
VarSifter	Command line Web App	Yes	Yes	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided
VCF-MINER	Local Web App	Yes	Yes	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided

- Provide a good flexibility and traceability
- Can be tedious and difficult
- No filtration at all annotation levels

Challenges

Variant annotation and filtration **are not a simple** and **solved problem**

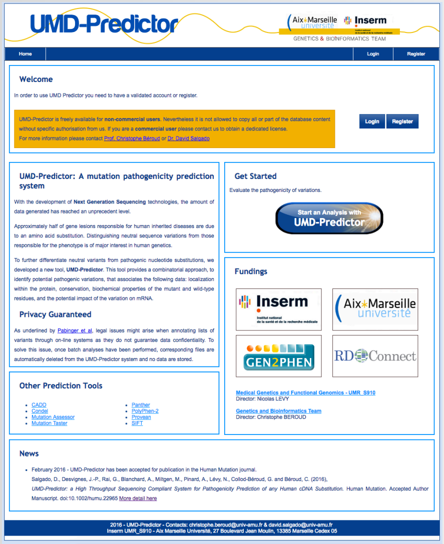
- **Variant Annotation**
  - Require the management of multiple and large data sources (local)
  - Incorrect or incomplete annotations can cause researchers to overlook and dilute interesting variants in a pool of false positives
  - From Davis McCarthy (Genome Medecine 2014)
    - Choice of transcript set and software can have a large effect on the variant annotation
      - Matching annotation for ANNOVAR on Ensembl or RefSeq genesets is only 83% for all exonic variants (lof, missenses, synonymous …)
      - Comparison between ANNOVAR and VEP on Ensembl transcripts 87% of all exonic variants were in agreement
      - Even more discrepancies with splicing variants
    - Where a specific tissue of interest is known, restrict annotation to transcripts known to be expressed in that tissue (GENCODE)

Challenges

- **Variant filtration/prioritization**
  - Good phenotypic description of patients
  - Mode of inheritance should be known (identify the right model, hypothesis on the penetrance …)
  - No Gold standard but frequently used filters
    - Frequency in the population
    - Genotype
    - Mutation type / Pathogenicity
    - Need to be done interactively (add/remove filters)
  - Discrepancies between pathogenicity predictors (may introduce false +/-)
  - Population frequency (not all databases gather healthy individuals, ethnicity) e.g. presence in dbSNP does not mean “polymorphism”
  - Privacy issue may arise when using “online” system

Our systems

To improve and facilitate disease causing mutation identification



Welcome

In order to use UMD Predictor you need to have a validated account or register.

UMD-Predictor is freely available for **non-commercial users**. Nevertheless it is not allowed to copy all or part of the database content without specific authorisation from us. If you are a **commercial user** please contact us to obtain a dedicated license.
For more information please contact Prof. Christophe Béroud or Dr. David Salgado

UMD-Predictor: A mutation pathogenicity prediction system

With the development of **Next Generation Sequencing** technologies, the amount of data generated has reached an unprecedent level.

Approximately half of gene lesions responsible for human inherited diseases are due to an amino acid substitution. Distinguishing neutral sequence variations from those responsible for the phenotype is of major interest in human genetics.

To further differentiate neutral variants from pathogenic nucleotide substitutions, we developed a new tool, **UMD-Predictor**. This tool provides a combinatorial approach, to identify potential pathogenic variations, that associates the following data: localization within the protein, conservation, biochemical properties of the mutant and wild-type residues, and the potential impact of the variation on mRNA.

Privacy Guaranteed

As underlined by Pabinger et al, legal issues might arise when annotating lists of variants through on-line systems as they do not guarantee data confidentiality. To solve this issue, once batch analyses have been performed, corresponding files are automatically deleted from the UMD-Predictor system and no data are stored.

Get Started

Evaluate the pathogenicity of variations.

Fundings

Inserm — Institut national de la santé et de la recherche médicale
Aix*Marseille université
GEN2PHEN
RD Connect

Medical Genetics and Functional Genomics - UMR_S910
Director: Nicolas LEVY

Genetics and Bioinformatics Team
Director: Christophe BEROUD

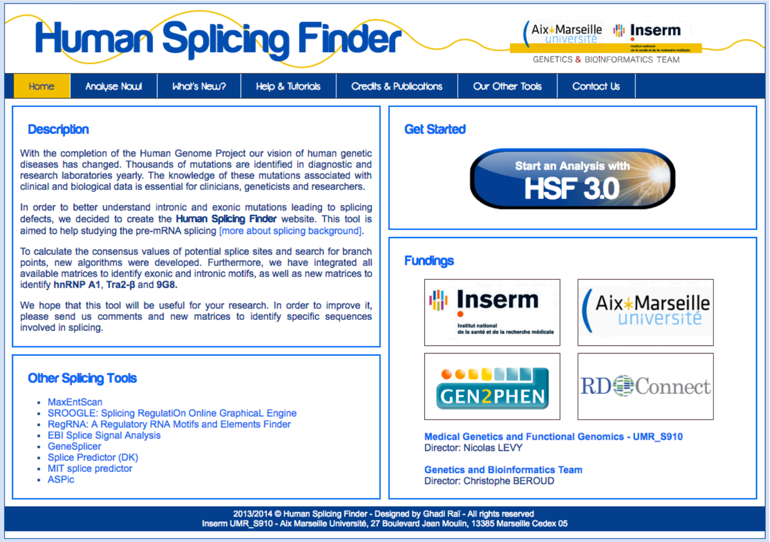
Other Prediction Tools

- CADD
- Condel
- Mutation Assessor
- Mutation Taster
- Panther
- PolyPhen-2
- Provean
- SIFT

News

- February 2016 - UMD-Predictor has been accepted for publication in the Human Mutation journal.
  Salgado, D., Desvignes, J.-P., Rai, G., Blanchard, A., Miltgen, M., Pinard, A., Lévy, N., Collod-Béroud, G. and Béroud, C. (2016), *UMD-Predictor: a High Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution.* Human Mutation. Accepted Author Manuscript. doi:10.1002/humu.22965 More detail here

2016 - UMD-Predictor - Contacts: christophe.beroud@univ-amu.fr & david.salgado@univ-amu.fr
Inserm UMR_S910 - Aix Marseille Université, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05



Description

With the completion of the Human Genome Project our vision of human genetic diseases has changed. Thousands of mutations are identified in diagnostic and research laboratories yearly. The knowledge of these mutations associated with clinical and biological data is essential for clinicians, geneticists and researchers.

In order to better understand intronic and exonic mutations leading to splicing defects, we decided to create the **Human Splicing Finder** website. This tool is aimed to help studying the pre-mRNA splicing [more about splicing background].

To calculate the consensus values of potential splice sites and search for branch points, new algorithms were developed. Furthermore, we have integrated all available matrices to identify exonic and intronic motifs, as well as new matrices to identify **hnRNP A1**, **Tra2-β** and **9G8.**

We hope that this tool will be useful for your research. In order to improve it, please send us comments and new matrices to identify specific sequences involved in splicing.

Get Started

Fundings

Inserm — Institut national de la santé et de la recherche médicale
Aix*Marseille université
GEN2PHEN
RD Connect

Medical Genetics and Functional Genomics - UMR_S910
Director: Nicolas LEVY

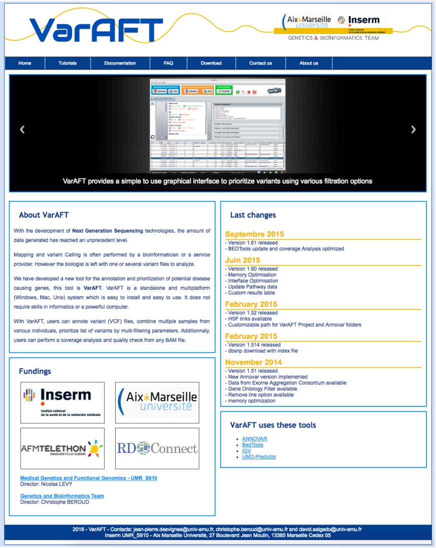
Genetics and Bioinformatics Team
Director: Christophe BEROUD

Other Splicing Tools

- MaxEntScan
- SROOGLE: Splicing RegulatiOn Online GraphicaL Engine
- RegRNA: A Regulatory RNA Motifs and Elements Finder
- EBI Splice Signal Analysis
- GeneSplicer
- Splice Predictor (DK)
- MIT splice predictor
- ASPic

2013/2014 © Human Splicing Finder - Designed by Ghadi Raï - All rights reserved
Inserm UMR_S910 - Aix Marseille Université, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05

Pathogenicity prediction systems
variant annotation



VarAFT provides a simple to use graphical interface to prioritize variants using various filtration options

About VarAFT

With the development of **Next Generation Sequencing** technologies, the amount of data generated has reached an unprecedent level.

Mapping and variant Calling is often performed by a bioinformatician or a service provider. However the biologist is left with one or several variant files to analyze.

We have developed a new tool for the annotation and prioritization of potential disease causing genes, this tool is **VarAFT**. VarAFT is a standalone and multiplatform (Windows, Mac, Unix) system which is easy to install and easy to use. It does not require skills in informatics or a powerful computer.

With VarAFT, users can annote variant (VCF) files, combine multiple samples from various individuals, prioritize list of variants by multi-filtering parameters. Additionnaly, users can perform a coverage analysis and quality check from any BAM file.

Last changes

Septembre 2015
- Version 1.61 released
- BEDTools update and coverage Analysis optimized

Juin 2015
- Version 1.60 released
- Memory Optimisation
- Interface Optimisation
- Update Pathway data
- Custom results table

February 2015
- Version 1.52 released
- HSF links available
- Customizable path for VarAFT Project and Annovar folders

February 2015
- Version 1.514 released
- dbsnp download with index file

November 2014
- Version 1.51 released
- New Annovar version implemented
- Data from Exome Aggregation Consortium available
- Gene Ontology Filter available
- Remove line option available
- memory optimization

Fundings

Inserm — Institut national de la santé et de la recherche médicale
Aix*Marseille université
AFMTELETHON — INNOVER POUR GUERIR
RD Connect

Medical Genetics and Functional Genomics - UMR_S910
Director: Nicolas LEVY

Genetics and Bioinformatics Team
Director: Christophe BEROUD

VarAFT uses these tools

- ANNOVAR
- BedTools
- IGV
- UMD-Predictor

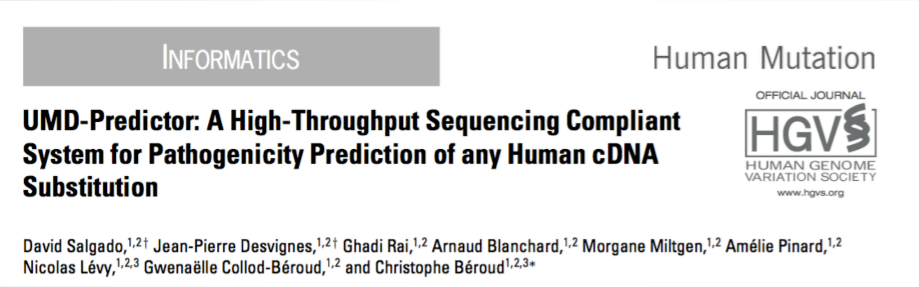
2016 - VarAFT - Contacts: jean-pierre.desvignes@univ-amu.fr, christophe.beroud@univ-amu.fr and david.salgado@univ-amu.fr
Inserm UMR_S910 - Aix Marseille Université, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05

Variant annotation and
filtration

UMD-Predictor

http://umd-predictor.eu

- Pathogenicity prediction system for any Human cDNA substitutions
  - Precomputed all possible substitutions for all nucleotides of any human transcripts (280,315,899 substitutions)
  - Combined multiple features in a unique score (0-100)
    - AA change – substitution and biochemical matrices (BLOSUM/Yu)
    - Exonic splicing signal (HSF - Acceptors and Donors splice sites)
    - Protein key residues (UNIPROT HCD)
    - Conserved and functional domains -100 species protein alignments (Phastcons) + Grantham
    - Allele frequency
- Available through a web application and webservices



INFORMATICS

Human Mutation

OFFICIAL JOURNAL

HGVS

HUMAN GENOME VARIATION SOCIETY

www.hgvs.org

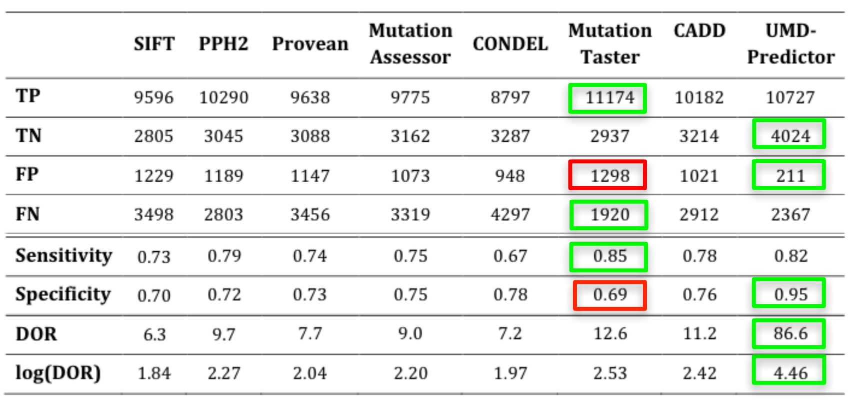
UMD-Predictor: A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution

David Salgado,[1,2†] Jean-Pierre Desvignes,[1,2†] Ghadi Rai,[1,2] Arnaud Blanchard,[1,2] Morgane Miltgen,[1,2] Amélie Pinard,[1,2] Nicolas Lévy,[1,2,3] Gwenaëlle Collod-Béroud,[1,2] and Christophe Béroud[1,2,3*]

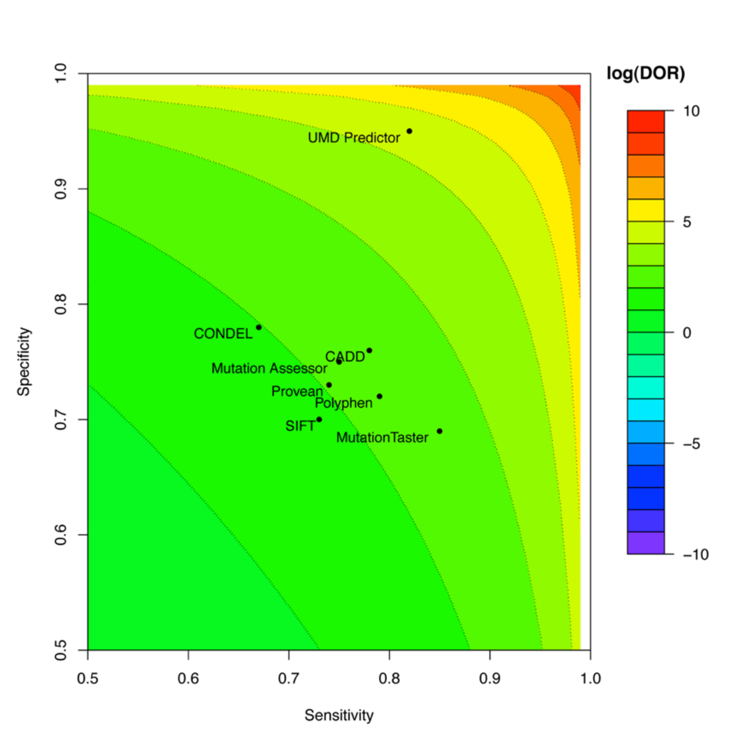
UMD-Predictor
Evaluation

- 4 datasets (more than 140,000 mutations) - Varibench + dbSNP, Uniprot, ClinVar and PredictSNP
- 7 references pathogenicity predictors

Varibench + dbSNP = 17,329 variations



	SIFT	PPH2	Provean	Mutation Assessor	CONDEL	Mutation Taster	CADD	UMD-Predictor
TP	9596	10290	9638	9775	8797	11174	10182	10727
TN	2805	3045	3088	3162	3287	2937	3214	4024
FP	1229	1189	1147	1073	948	1298	1021	211
FN	3498	2803	3456	3319	4297	1920	2912	2367
Sensitivity	0.73	0.79	0.74	0.75	0.67	0.85	0.78	0.82
Specificity	0.70	0.72	0.73	0.75	0.78	0.69	0.76	0.95
DOR	6.3	9.7	7.7	9.0	7.2	12.6	11.2	86.6
log(DOR)	1.84	2.27	2.04	2.20	1.97	2.53	2.42	4.46

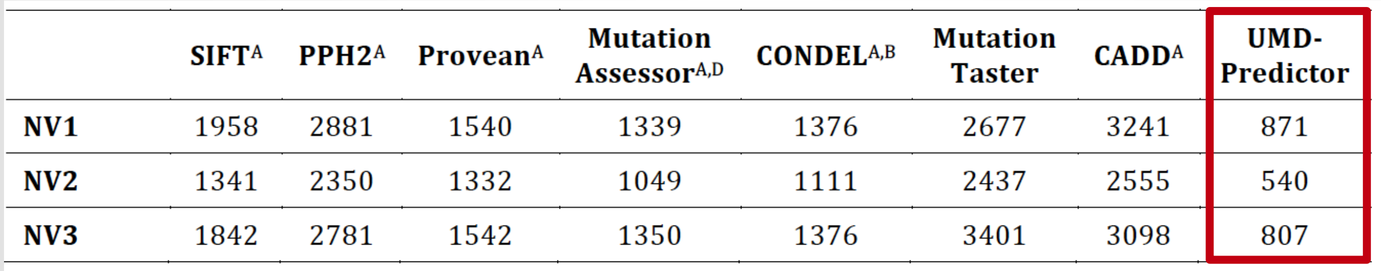


Similar results with other datasets (Cf. Salgado, Desvignes, et al. Human Mutation 2016)

DOR : Measure the effectiveness of a diagnostic test
Trade-off between sensitivity and specificity

UMD-Predictor
Real data evaluation

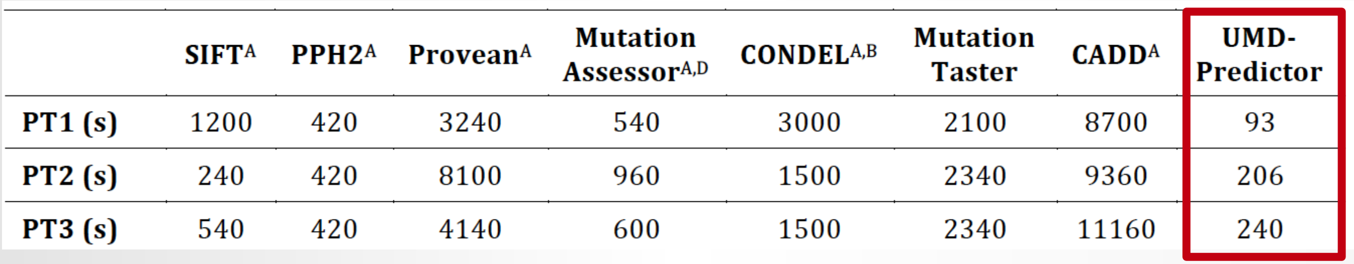
Comparison using 3 WES performed in a clinical diagnostic context



	SIFT^A	PPH2^A	Provean^A	Mutation Assessor^A,D	CONDEL^A,B	Mutation Taster	CADD^A	UMD-Predictor
NV1	1958	2881	1540	1339	1376	2677	3241	871
NV2	1341	2350	1332	1049	1111	2437	2555	540
NV3	1842	2781	1542	1350	1376	3401	3098	807

Shortest list of potential pathogenic mutations

Time required to process VCF files



	SIFT^A	PPH2^A	Provean^A	Mutation Assessor^A,D	CONDEL^A,B	Mutation Taster	CADD^A	UMD-Predictor
PT1 (s)	1200	420	3240	540	3000	2100	8700	93
PT2 (s)	240	420	8100	960	1500	2340	9360	206
PT3 (s)	540	420	4140	600	1500	2340	11160	240

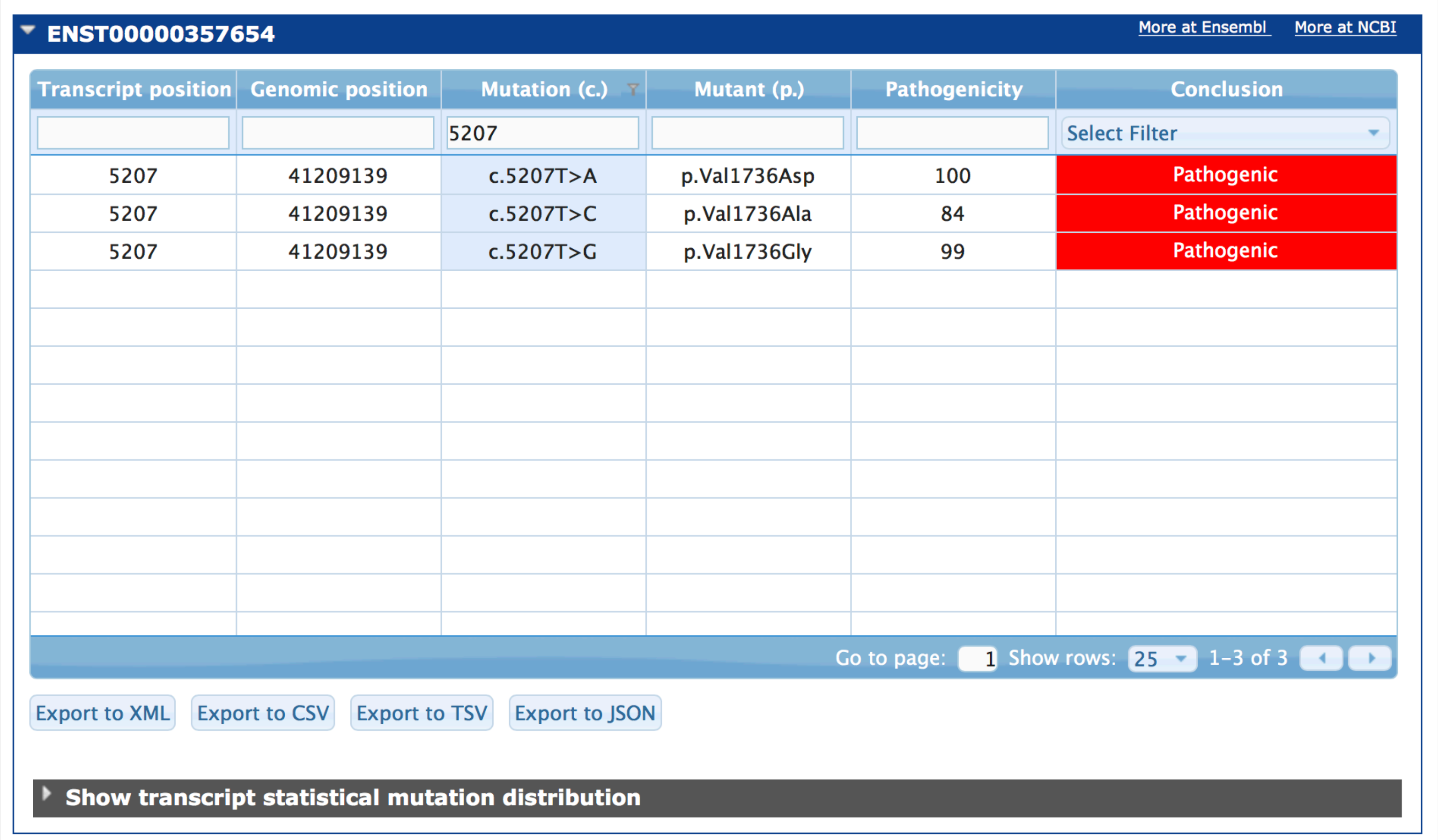
Even faster by using webservices

Fastest system to process variations from VCF files

Salgado D, et al. Human Mutation 2016

UMD-Predictor (http://umd-predictor.eu)
Real data evaluation

Data from yesterday presentation: *BRCA1* c.5207T>C



Transcript position	Genomic position	Mutation (c.)	Mutant (p.)	Pathogenicity	Conclusion
		5207			Select Filter
5207	41209139	c.5207T>A	p.Val1736Asp	100	Pathogenic
5207	41209139	c.5207T>C	p.Val1736Ala	84	Pathogenic
5207	41209139	c.5207T>G	p.Val1736Gly	99	Pathogenic

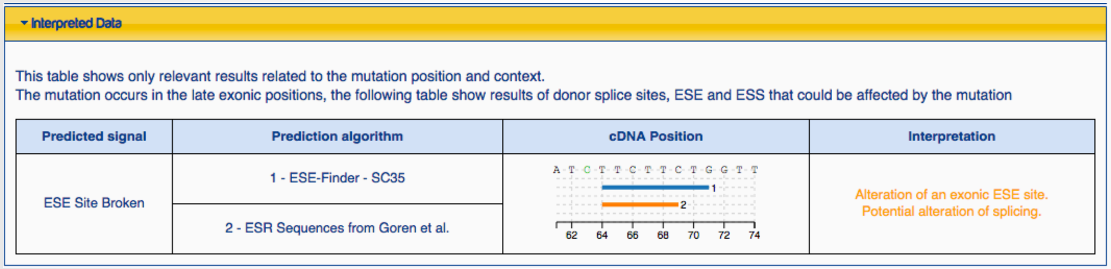
Go to page: 1 Show rows: 25 1–3 of 3
Export to XML
Export to CSV
Export to TSV
Export to JSON
Show transcript statistical mutation distribution

Salgado D, et al. Human Mutation 2016

Human Splicing Finder

http://umd.be/HSF3/

- Pathogenicity prediction system for any mutations on splicing signals
- Reference system (828 citations, Web of Science; 1,125, Google Scholar)
- A “One stop-Shop” system
  - Splicing signals
  - Branch points
  - Auxiliary signals (ESE, ESS, ISE, ISS)
- Combine various predictive systems, matrices and specific algorithms
- Expert system for data interpretation (establishment of rules to provide a conclusion) e.g. *MSH2* c.274_276del



▾ Interpreted Data

This table shows only relevant results related to the mutation position and context.
The mutation occurs in the late exonic positions, the following table show results of donor splice sites, ESE and ESS that could be affected by the mutation

Predicted signal	Prediction algorithm	cDNA Position	Interpretation
ESE Site Broken	1 - ESE-Finder - SC35		Alteration of an exonic ESE site. Potential alteration of splicing.
	2 - ESR Sequences from Goren et al.		

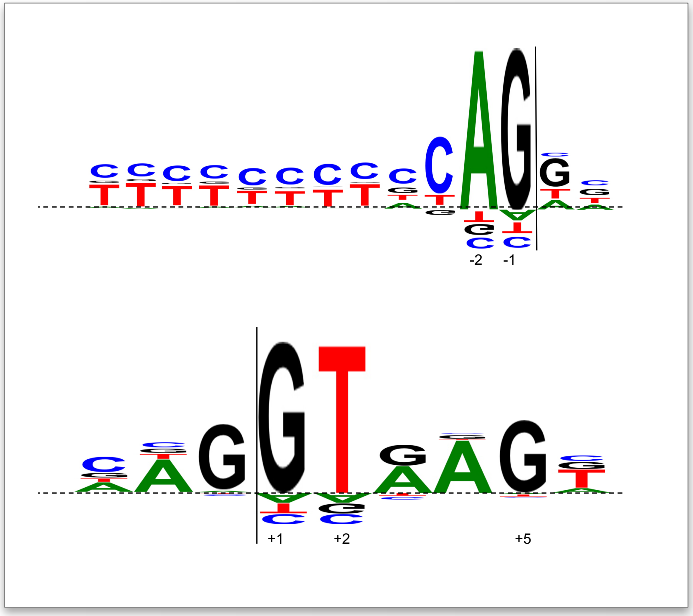
- Compliant with NGS technologies: webservices (available in the HSF3-Pro version)

Human Splicing Finder

http://umd.be/HSF3/

- Use a *BRCA1/2* mutations dataset from ClinVar
  - 5' ss (3 last exonic nt + 6 first intronic nt)
  - 3' ss (12 last intronic nt + 2 first exonic nt)

  - *BRCA1*
    - 135 pathogenic mutations
    - 16 non-pathogenic mutations
  - *BRCA2*
    - 88 pathogenic mutations
    - 15 non-pathogenic mutations



Human Splicing Finder

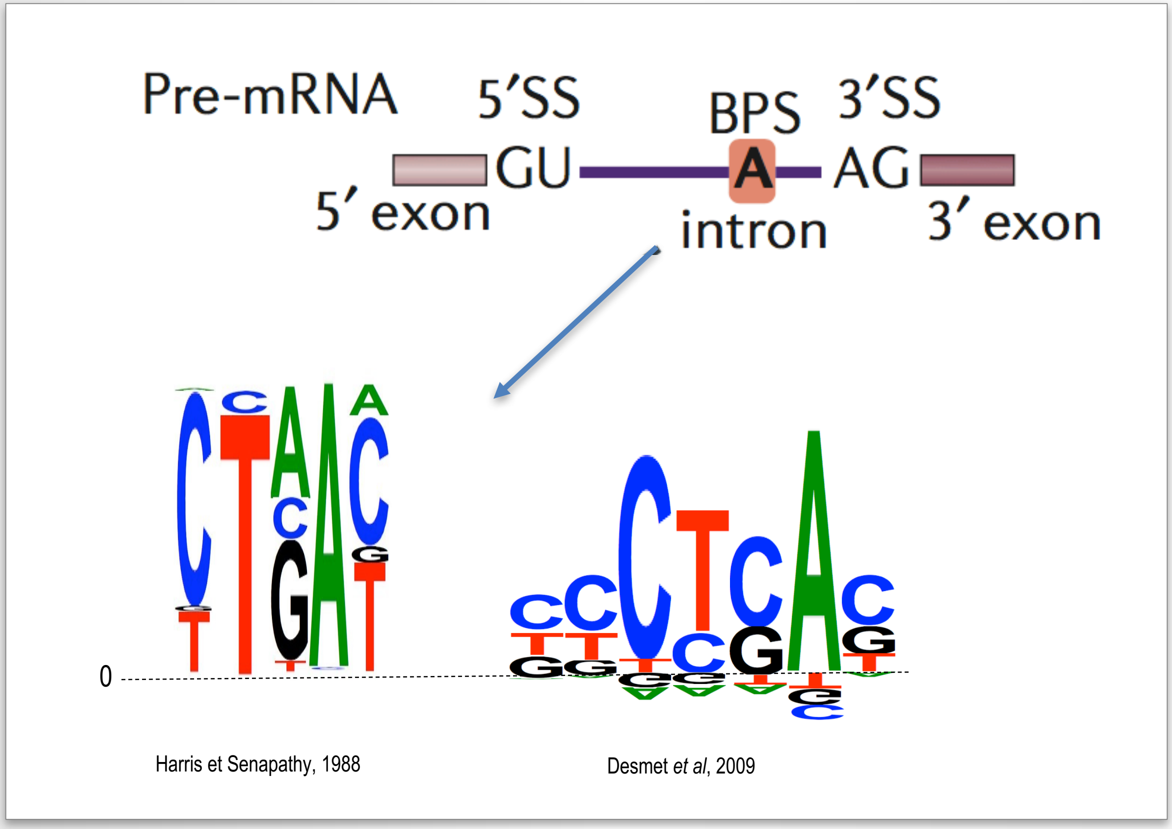
http://umd.be/HSF3/

***BRCA1* & *BRCA2* dataset**	
True Positives	223
True Negatives	28
False Positives	3
False Negatives	0
Positive Predictive Value	0.986
Negative Predictive Value	1.000
Sensitivity	1.000
Specificity	0.903
Accuracy	0.988
Matthews Correlation Coefficient	0.944

Human Splicing Finder

http://umd.be/HSF3/

- Impact of mutations on branch points



Harris et Senapathy, 1988

Desmet *et al*, 2009

Human Splicing Finder

http://umd.be/HSF3/

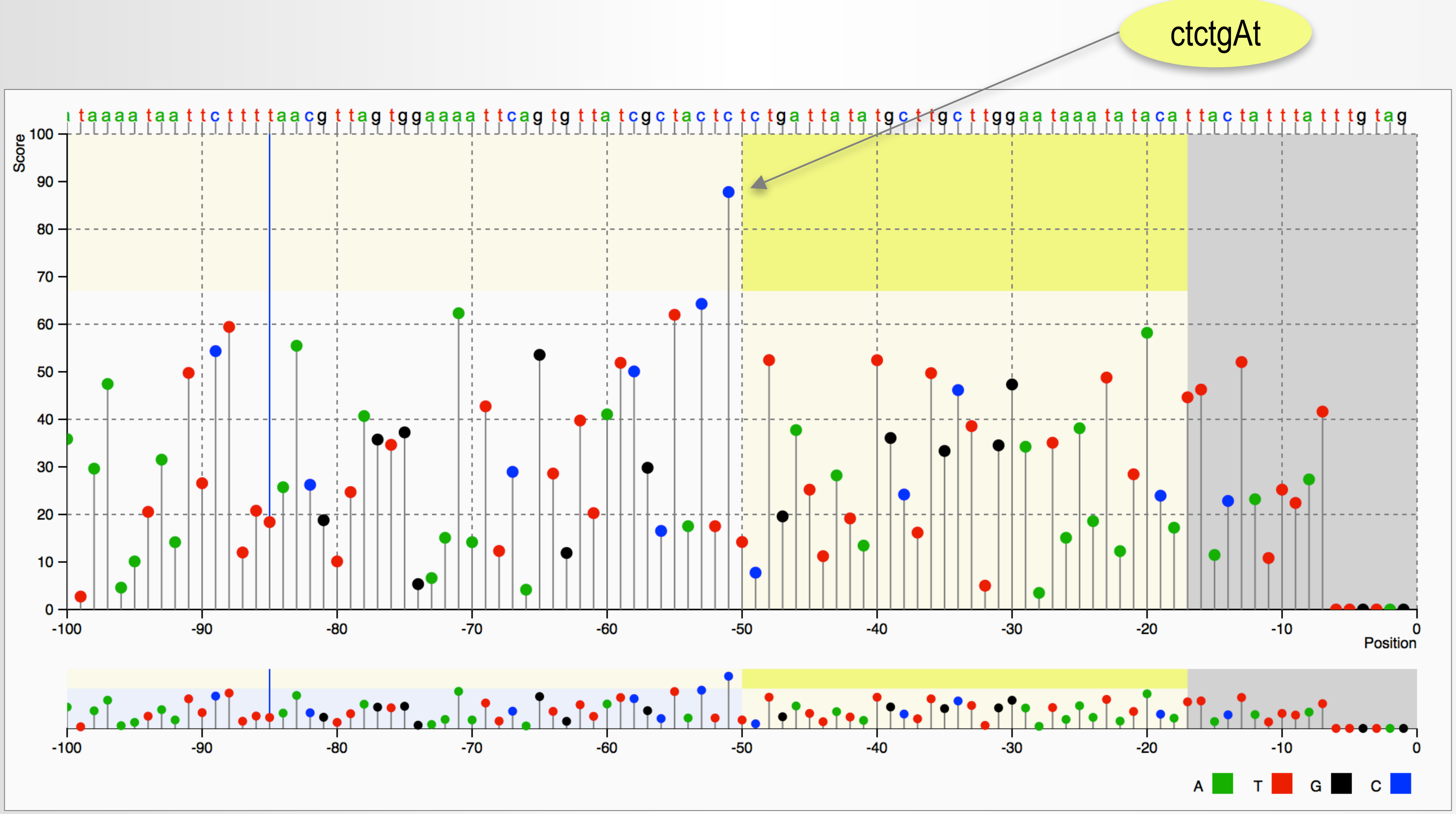
- Impact of mutations on branch points

- Few BPS are described in the literature
  - BRAF c.1141-51 C>G (Pupo et al. 2017), BPS identification
  - IKBKG/NEMO c.519-23A>T (Jorgensen et al. 2016), alteration of the BPS
  - C21orf2 c.643-23A>T (Wang et al. 2016), alteration of the BPS
  - ITGB4 c.1762-25T>A (Masunaga et al. 2015), alteration of the BPS
  - PC c.1369-29A>G (Ostergaardet al. 2012), alteration of the BPS
  - KCNH2 c.IVS9-28A>G (Crotti et al. 2009), alteration of the BPS
  - …

Human Splicing Finder

http://umd.be/HSF3/

- Few BPS are described in the literature
  - BRAF c.1141-51 C>G (Pupo et al. 2017), BPS identification



ctctgAt

Aix*Marseille université

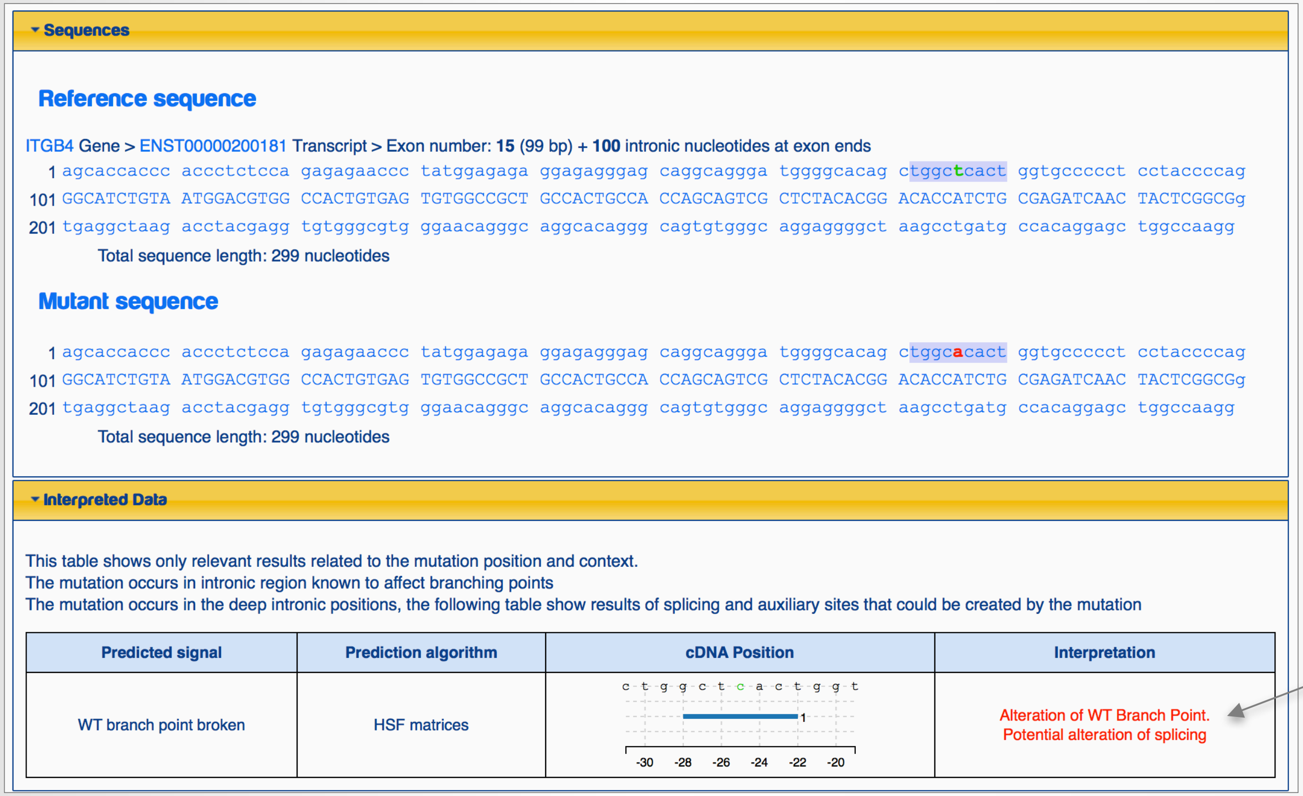
Inserm
Institut national de la santé et de la recherche médicale

GENETICS & BIOINFORMATICS TEAM

Human Splicing Finder

http://umd.be/HSF3/

- Few BPS are described in the literature
  - ITGB4 c.1762-25T>A (Masunaga et al. 2015), alteration of the BPS



▾ Sequences

Reference sequence

ITGB4 Gene > ENST00000200181 Transcript > Exon number: **15** (99 bp) + **100** intronic nucleotides at exon ends

```
  1 agcaccaccc accctctcca gagagaaccc tatggagaga ggagagggag caggcaggga tggggcacag ctggctcact ggtgccccct cctaccccag
101 GGCATCTGTA ATGGACGTGG CCACTGTGAG TGTGGCCGCT GCCACTGCCA CCAGCAGTCG CTCTACACGG ACACCATCTG CGAGATCAAC TACTCGGCGg
201 tgaggctaag acctacgagg tgtgggcgtg ggaacagggc aggcacaggg cagtgtgggc aggaggggct aagcctgatg ccacaggagc tggccaagg
```

Total sequence length: 299 nucleotides

Mutant sequence

```
  1 agcaccaccc accctctcca gagagaaccc tatggagaga ggagagggag caggcaggga tggggcacag ctggcacact ggtgccccct cctaccccag
101 GGCATCTGTA ATGGACGTGG CCACTGTGAG TGTGGCCGCT GCCACTGCCA CCAGCAGTCG CTCTACACGG ACACCATCTG CGAGATCAAC TACTCGGCGg
201 tgaggctaag acctacgagg tgtgggcgtg ggaacagggc aggcacaggg cagtgtgggc aggaggggct aagcctgatg ccacaggagc tggccaagg
```

Total sequence length: 299 nucleotides

▾ Interpreted Data

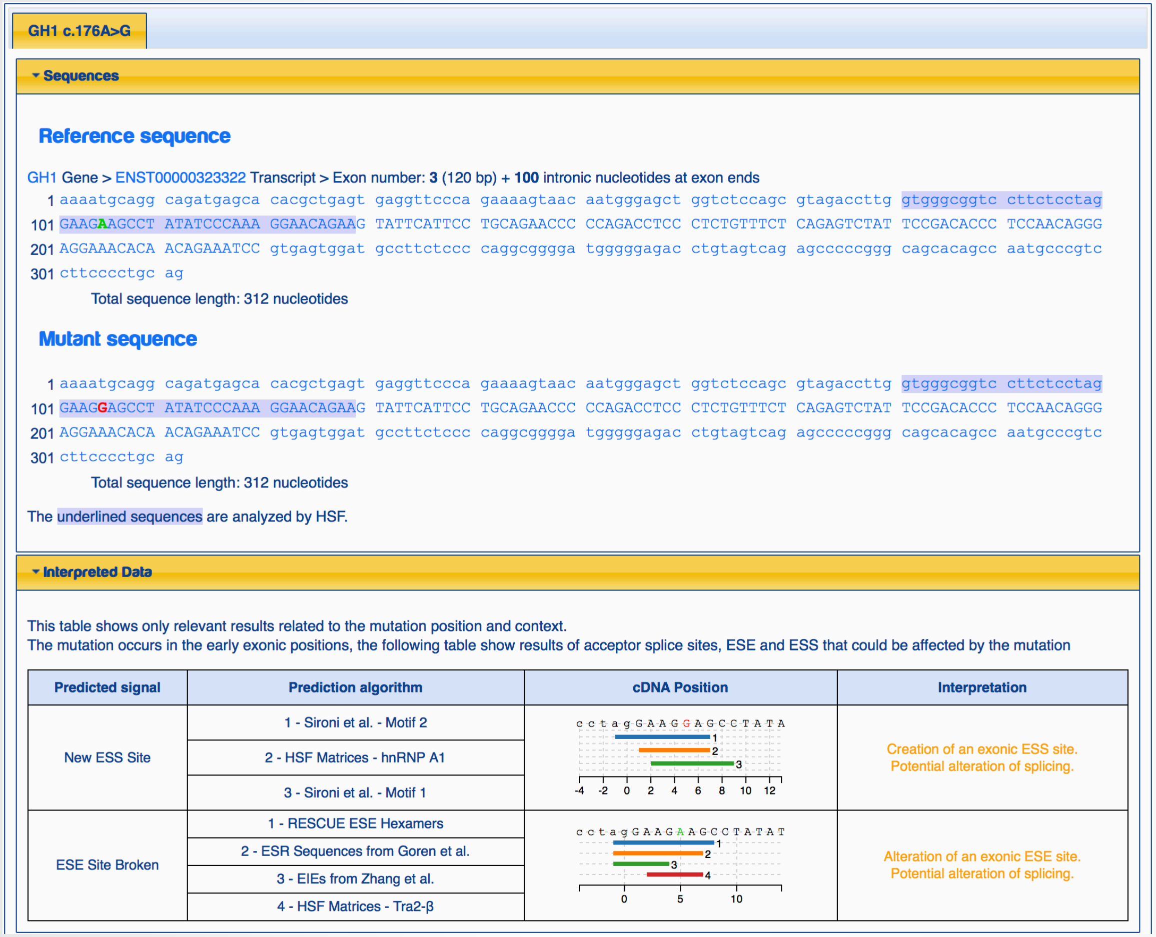
This table shows only relevant results related to the mutation position and context.
The mutation occurs in intronic region known to affect branching points
The mutation occurs in the deep intronic positions, the following table show results of splicing and auxiliary sites that could be created by the mutation

Predicted signal	Prediction algorithm	cDNA Position	Interpretation
WT branch point broken	HSF matrices	c t g g c t c a c t g g t 1 -30 -28 -26 -24 -22 -20	Alteration of WT Branch Point. Potential alteration of splicing

Human Splicing Finder

http://umd.be/HSF3/

- Examples from Andreas Laner yesterday (impact on ESE/ESS)



GH1 c.176A>G

▾ Sequences

Reference sequence

GH1 Gene > ENST00000323322 Transcript > Exon number: **3** (120 bp) + **100** intronic nucleotides at exon ends

```
  1 aaaatgcagg cagatgagca cacgctgagt gaggttccca gaaaagtaac aatgggagct ggtctccagc gtagaccttg gtgggcggtc cttctcctag
101 GAAGAAGCCT ATATCCCAAA GGAACAGAAG TATTCATTCC TGCAGAACCC CCAGACCTCC CTCTGTTTCT CAGAGTCTAT TCCGACACCC TCCAACAGGG
201 AGGAAACACA ACAGAAATCC gtgagtggat gccttctccc caggcgggga tgggggagac ctgtagtcag agcccccggg cagcacagcc aatgcccgtc
301 cttcccctgc ag
```

Total sequence length: 312 nucleotides

Mutant sequence

```
  1 aaaatgcagg cagatgagca cacgctgagt gaggttccca gaaaagtaac aatgggagct ggtctccagc gtagaccttg gtgggcggtc cttctcctag
101 GAAGGAGCCT ATATCCCAAA GGAACAGAAG TATTCATTCC TGCAGAACCC CCAGACCTCC CTCTGTTTCT CAGAGTCTAT TCCGACACCC TCCAACAGGG
201 AGGAAACACA ACAGAAATCC gtgagtggat gccttctccc caggcgggga tgggggagac ctgtagtcag agcccccggg cagcacagcc aatgcccgtc
301 cttcccctgc ag
```

Total sequence length: 312 nucleotides

The underlined sequences are analyzed by HSF.

▾ Interpreted Data

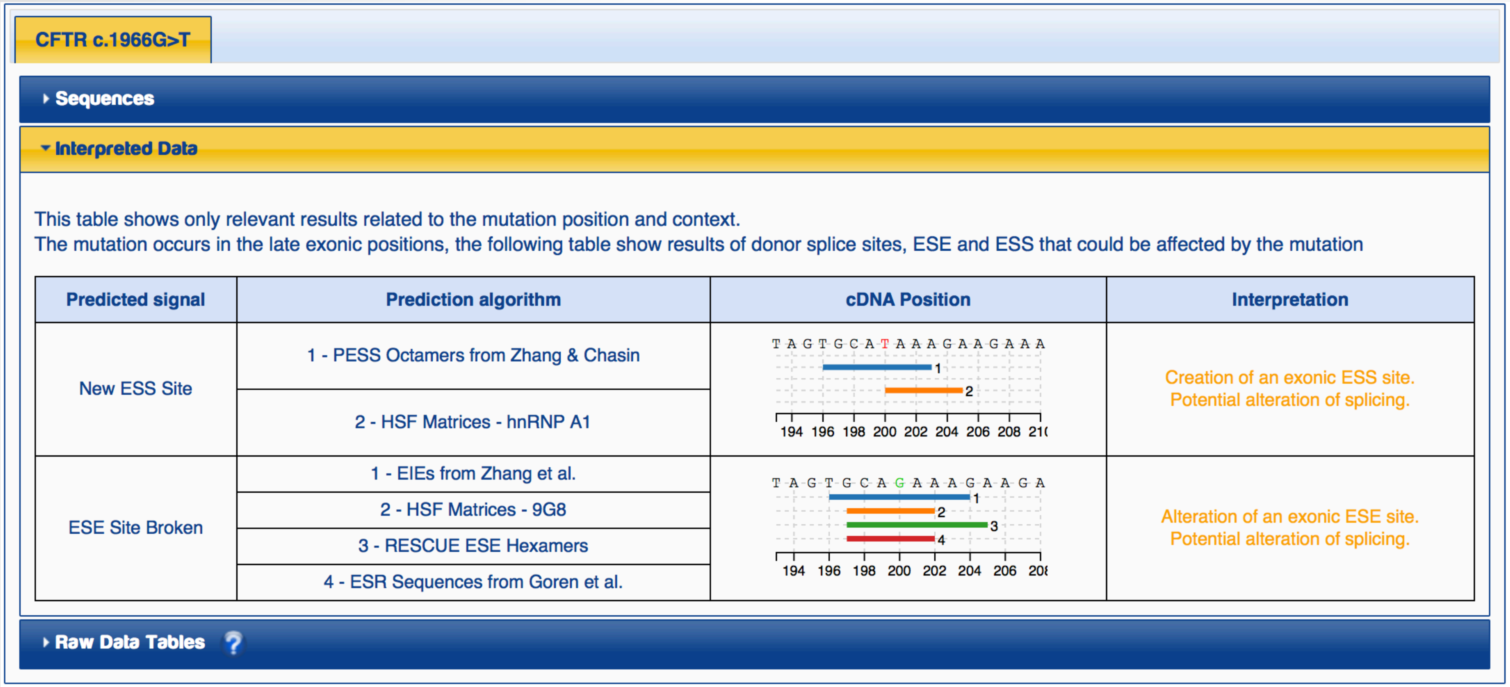
This table shows only relevant results related to the mutation position and context.
The mutation occurs in the early exonic positions, the following table show results of acceptor splice sites, ESE and ESS that could be affected by the mutation

Predicted signal	Prediction algorithm	cDNA Position	Interpretation
New ESS Site	1 - Sironi et al. - Motif 2	c c t a g G A A G G A G C C T A T A -4 -2 0 2 4 6 8 10 12	Creation of an exonic ESS site. Potential alteration of splicing.
	2 - HSF Matrices - hnRNP A1		
	3 - Sironi et al. - Motif 1		
ESE Site Broken	1 - RESCUE ESE Hexamers	c c t a g G A A G A A G C C T A T A T 0 5 10	Alteration of an exonic ESE site. Potential alteration of splicing.
	2 - ESR Sequences from Goren et al.		
	3 - EIEs from Zhang et al.		
	4 - HSF Matrices - Tra2-β		

Human Splicing Finder

http://umd.be/HSF3/

- Examples from Andreas Laner yesterday (impact on ESE/ESS)



This table shows only relevant results related to the mutation position and context.
The mutation occurs in the late exonic positions, the following table show results of donor splice sites, ESE and ESS that could be affected by the mutation

Predicted signal	Prediction algorithm	cDNA Position	Interpretation
New ESS Site	1 - PESS Octamers from Zhang & Chasin 2 - HSF Matrices - hnRNP A1	T A G T G C A T A A A G A A G A A A 1 2 194 196 198 200 202 204 206 208 21(	Creation of an exonic ESS site. Potential alteration of splicing.
ESE Site Broken	1 - EIEs from Zhang et al. 2 - HSF Matrices - 9G8 3 - RESCUE ESE Hexamers 4 - ESR Sequences from Goren et al.	T A G T G C A G A A A G A A G A 1 2 3 4 194 196 198 200 202 204 206 20ξ	Alteration of an exonic ESE site. Potential alteration of splicing.

Human Splicing Finder

http://umd.be/HSF3/

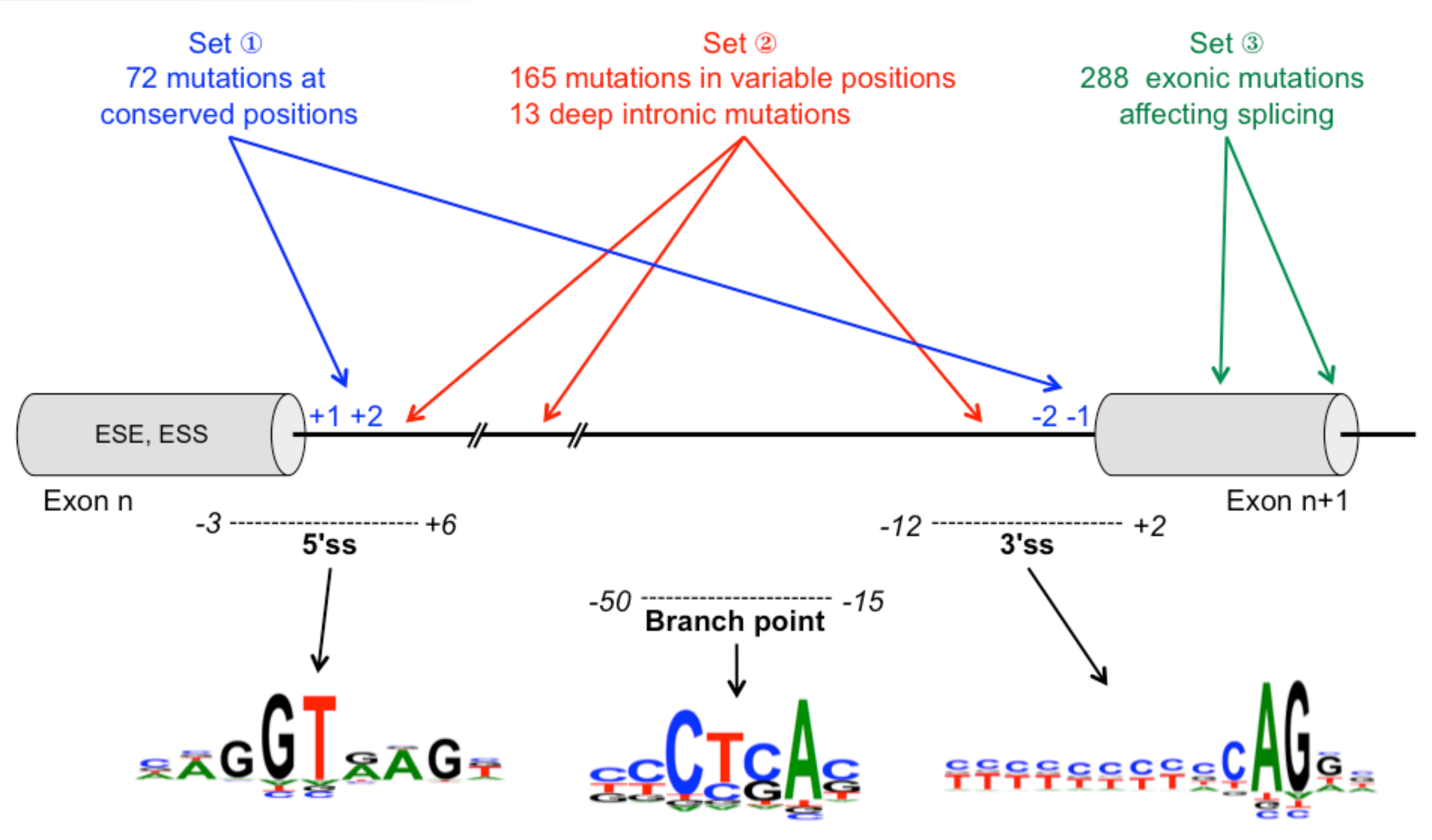
HSF system is highly accurate to predict the impact of mutations on 5' and 3' splice sites

Now contains specific matrices for non-canonical splice sites

HSF efficiently predicts the **BPS** and the impact of mutations on these sites

HSF predicts the impact of mutations on ESE/ESS

Splicing predictors comparison


Set ①
72 mutations at conserved positions
Set ②
165 mutations in variable positions
13 deep intronic mutations
Set ③
288 exonic mutations affecting splicing
ESE, ESS
+1 +2
-2 -1
Exon n
Exon n+1
-3 +6
5'ss
-12 +2
3'ss
-50 -15
Branch point

Splicing predictors comparison

Software	**Global**	**3’ss**	**5’ss**
GeneSplicer	63.89 %	82.61 %	55.1 %
GenScan	45.83 %	82.61 %	30.61 %
Human Splicing Finder	**100%**	**100%**	**100%**
MaxEntScan	**100%**	**100%**	**100%**
NNSplice	97.22 %	**100%**	95.92 %
SplicePort	87.5 %	82.61 %	89.8 %
SplicePredictor	87.5 %	95.65 %	83.67 %
SpliceView	**100%**	**100%**	**100%**
SROOGLE	**100%**	**100%**	**100%**
Average	***86.88 %***	***93.72 %***	***83.9 %***

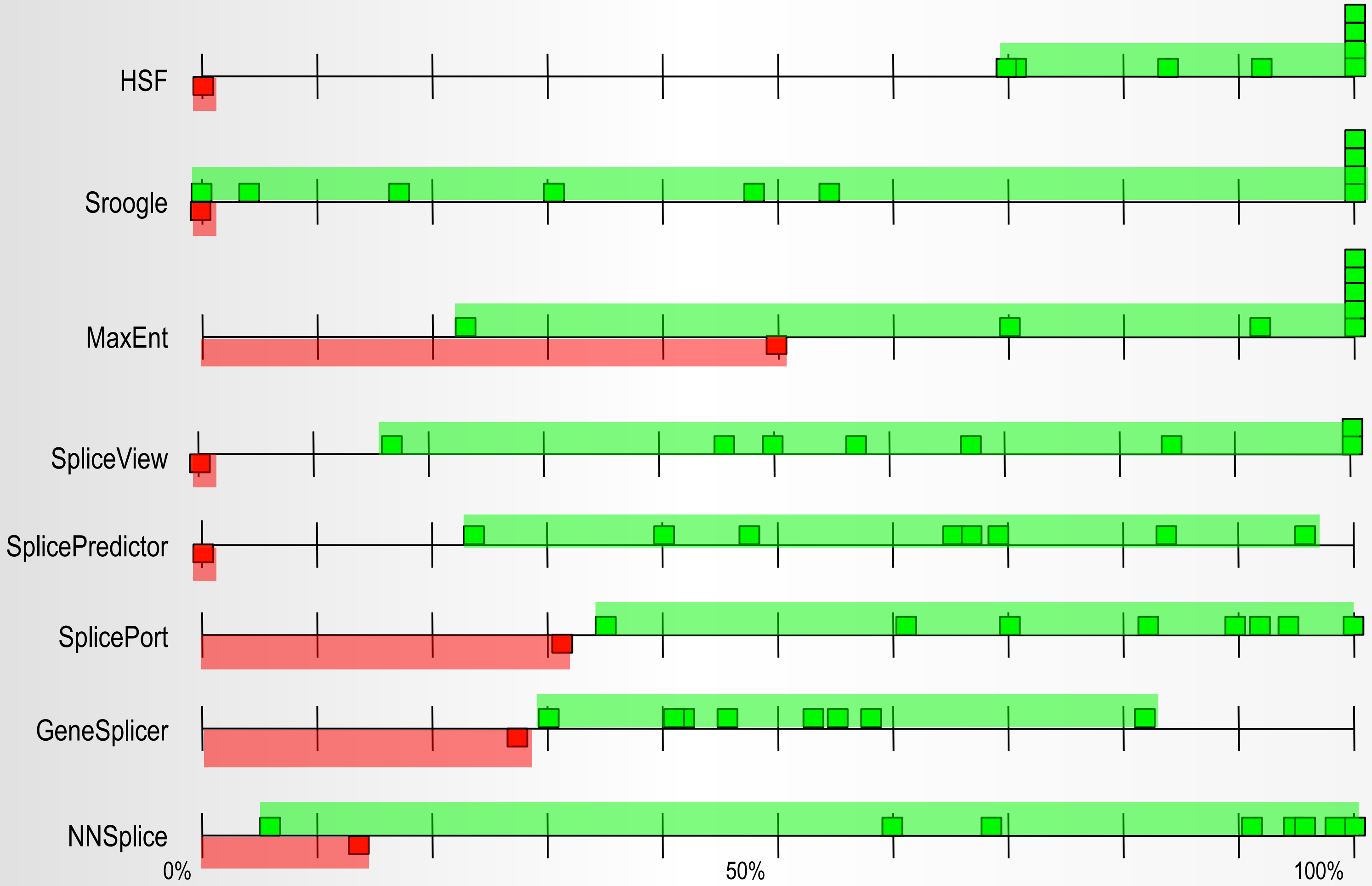
Splicing predictors comparison

Software	Intronic (<100 bp)	Intronic (> 100 bp)
GeneSplicer	41.18%	46.15%
GenScan	11.76%	0.00%
Human Splicing Finder	**70.59%**	**92.31%**
MaxEntScan	23.53%	**92.31%**
NNSplice	5.88%	69.23%
SplicePort	35.29%	61.54%
SplicePredictor	23.53%	69.23%
SpliceView	17.65%	84.62%
Sroogle	17.65%	30.77%
Average	***27.45%***	***60.68%***

Splicing predictors comparison

Software	**Polymorphisms**
GeneSplicer	50%
GenScan	**0%**
Human Splicing Finder	**0%**
MaxEntScan	50%
NNSplice	25%
SplicePort	100%
SplicePredictor	**0%**
SpliceView	**0%**
Sroogle	**0%**
Average	***10%***

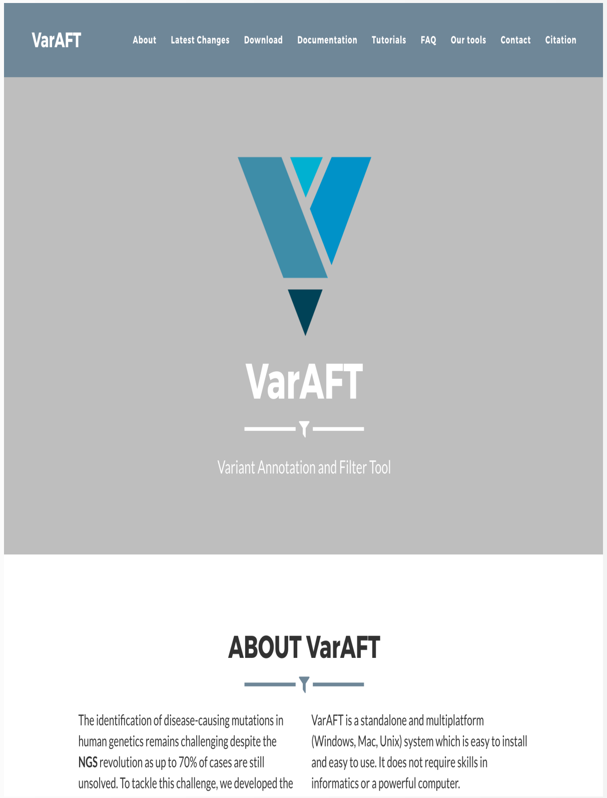
Splicing predictors comparison



VARiant Annotation and Filtration Tool

http://varaft.eu

- An user-friendly variant annotation and filtration tool
- Standalone multiplatform application
- Evaluation of the data coverage for WGS/WES/panel
- One click annotation (based on ANNOVAR) and other sources
- Provides UMD-Predictor and HSF annotations
- Interactive filtration at many levels
  - Combine multiple samples
  - Automatic selection of variants (mode of inheritance)
  - Genetic population studies
  - Cancers
- Standardize your variant analysis processes
  - save/reuse/share applied filters

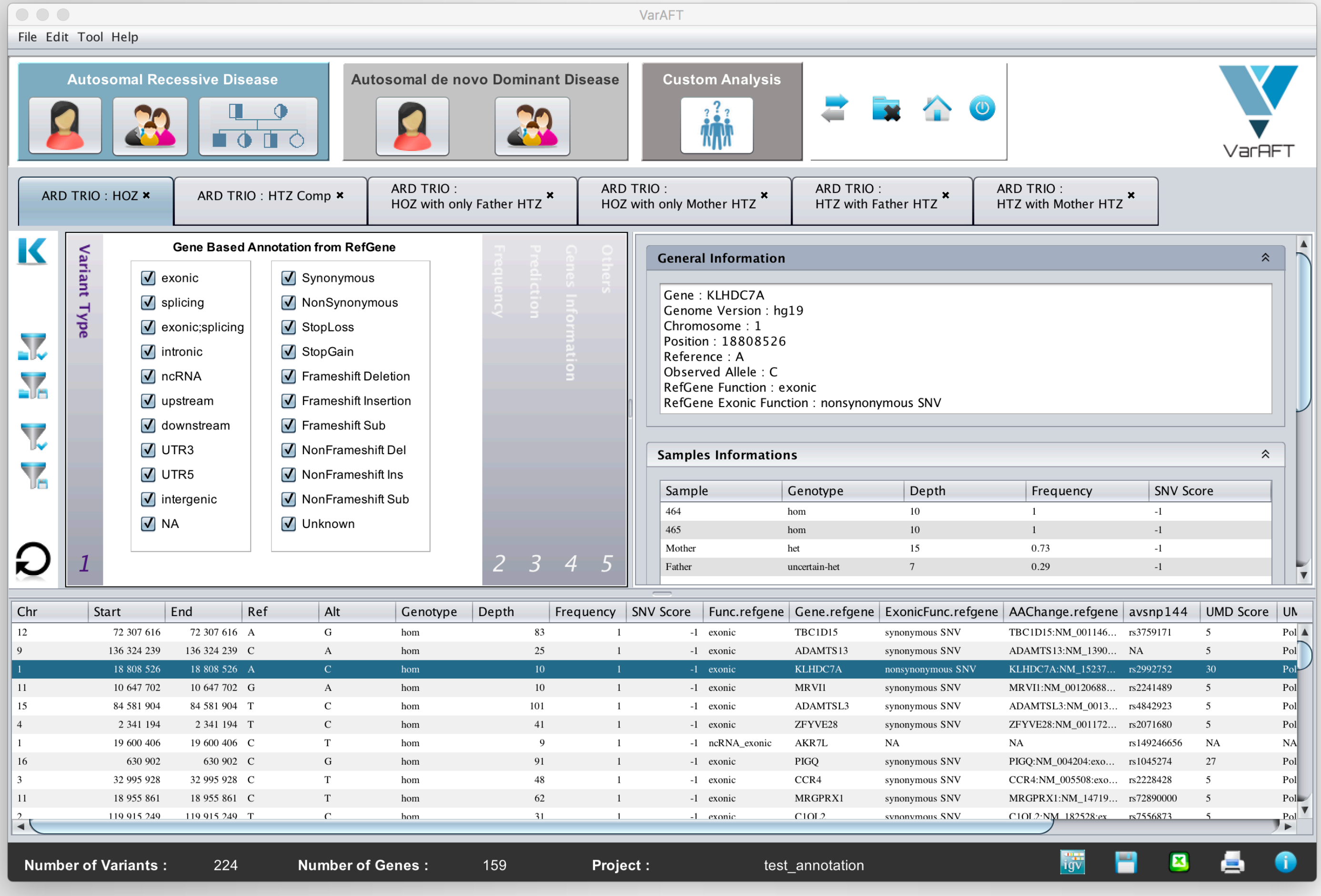


ABOUT VarAFT

The identification of disease-causing mutations in human genetics remains challenging despite the NGS revolution as up to 70% of cases are still unsolved. To tackle this challenge, we developed the

VarAFT is a standalone and multiplatform (Windows, Mac, Unix) system which is easy to install and easy to use. It does not require skills in informatics or a powerful computer.

Variant Annotation and Filtration Tool


VarAFT
File Edit Tool Help
Autosomal Recessive Disease
Autosomal de novo Dominant Disease
Custom Analysis
VarAFT
ARD TRIO : HOZ ×
ARD TRIO : HTZ Comp ×
ARD TRIO : HOZ with only Father HTZ ×
ARD TRIO : HOZ with only Mother HTZ ×
ARD TRIO : HTZ with Father HTZ ×
ARD TRIO : HTZ with Mother HTZ ×
Variant Type
1
Frequency
Prediction
Genes Information
Others
2
3
4
5
General Information
Samples Informations
igv

Gene Based Annotation from RefGene

☑ exonic
☑ splicing
☑ exonic;splicing
☑ intronic
☑ ncRNA
☑ upstream
☑ downstream
☑ UTR3
☑ UTR5
☑ intergenic
☑ NA

☑ Synonymous
☑ NonSynonymous
☑ StopLoss
☑ StopGain
☑ Frameshift Deletion
☑ Frameshift Insertion
☑ Frameshift Sub
☑ NonFrameshift Del
☑ NonFrameshift Ins
☑ NonFrameshift Sub
☑ Unknown

General Information

Gene : KLHDC7A
Genome Version : hg19
Chromosome : 1
Position : 18808526
Reference : A
Observed Allele : C
RefGene Function : exonic
RefGene Exonic Function : nonsynonymous SNV

Samples Informations

Sample	Genotype	Depth	Frequency	SNV Score
464	hom	10	1	-1
465	hom	10	1	-1
Mother	het	15	0.73	-1
Father	uncertain-het	7	0.29	-1

Chr	Start	End	Ref	Alt	Genotype	Depth	Frequency	SNV Score	Func.refgene	Gene.refgene	ExonicFunc.refgene	AAChange.refgene	avsnp144	UMD Score	UM
12	72 307 616	72 307 616	A	G	hom	83	1	-1	exonic	TBC1D15	synonymous SNV	TBC1D15:NM_001146...	rs3759171	5	Pol
9	136 324 239	136 324 239	C	A	hom	25	1	-1	exonic	ADAMTS13	synonymous SNV	ADAMTS13:NM_1390...	NA	5	Pol
1	18 808 526	18 808 526	A	C	hom	10	1	-1	exonic	KLHDC7A	nonsynonymous SNV	KLHDC7A:NM_15237...	rs2992752	30	Pol
11	10 647 702	10 647 702	G	A	hom	10	1	-1	exonic	MRVI1	synonymous SNV	MRVI1:NM_00120688...	rs2241489	5	Pol
15	84 581 904	84 581 904	T	C	hom	101	1	-1	exonic	ADAMTSL3	synonymous SNV	ADAMTSL3:NM_0013...	rs4842923	5	Pol
4	2 341 194	2 341 194	T	C	hom	41	1	-1	exonic	ZFYVE28	synonymous SNV	ZFYVE28:NM_001172...	rs2071680	5	Pol
1	19 600 406	19 600 406	C	T	hom	9	1	-1	ncRNA_exonic	AKR7L	NA	NA	rs149246656	NA	NA
16	630 902	630 902	C	G	hom	91	1	-1	exonic	PIGQ	synonymous SNV	PIGQ:NM_004204:exo...	rs1045274	27	Pol
3	32 995 928	32 995 928	C	T	hom	48	1	-1	exonic	CCR4	synonymous SNV	CCR4:NM_005508:exo...	rs2228428	5	Pol
11	18 955 861	18 955 861	C	T	hom	62	1	-1	exonic	MRGPRX1	synonymous SNV	MRGPRX1:NM_14719...	rs72890000	5	Pol
2	119 915 249	119 915 249	T	C	hom	31	1	-1	exonic	C1QL2	synonymous SNV	C1QL2:NM_182528:ex...	rs7556873	5	Pol

Number of Variants : 224 **Number of Genes :** 159 **Project :** test_annotation

Conclusions

- No ideal annotation/filtration systems
  - Many of them are built to be used by bioinformaticians (command line systems)
- To maximize chances of identify causing mutations
  - be aware of challenges posed by each steps of the data analysis pipeline
  - use family members (when possible)
  - collect exact and complete phenotypic information
- Many challenges remain to be solved for both annotation and filtration systems - benchmarking initiatives – CAGI challenges
- Most of the current available systems are created for WES and need to be adapted to WGS
- Many more issues arise with WGS
  - Annotation for non-coding regions but also for non-protein coding genes
  - Need to develop and improve pathogenicity prediction system for non-exonic mutations
- Facilitated with data-sharing initiatives (shared variants and conclusion)

A simple use case

- A trio with an affected daughter and two healthy parents
- Mabry syndrome: intellectual disability, distinctive facial features, hyperphosphatasia, and other signs and symptoms.

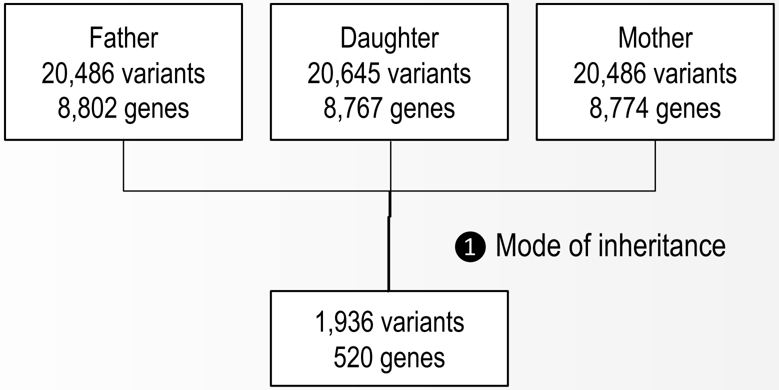
Father
20,486 variants
8,802 genes

Daughter
20,645 variants
8,767 genes

Mother
20,486 variants
8,774 genes

A simple use case

- A trio with an affected daughter and two healthy parents
- Mabry syndrome: intellectual disability, distinctive facial features, hyperphosphatasia, and other signs and symptoms.
- 4 steps filtration workflow:
  - Mode of inheritance (Autosomal recessive)
    - ➛ Homozygous or compound heterozygous



Father
20,486 variants
8,802 genes

Daughter
20,645 variants
8,767 genes

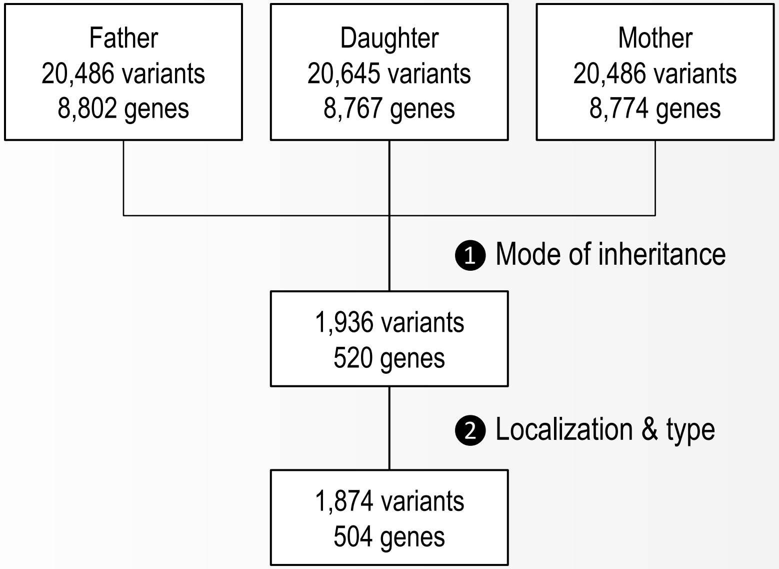
Mother
20,486 variants
8,774 genes

❶ Mode of inheritance

1,936 variants
520 genes

A simple use case

- A trio with an affected daughter and two healthy parents
- Mabry syndrome: intellectual disability, distinctive facial features, hyperphosphatasia, and other signs and symptoms.
- 4 steps filtration workflow:
  - Mode of inheritance (Autosomal recessive)
    - ➛ Homozygous or compound heterozygous
  - Mutation localization and type


Father
20,486 variants
8,802 genes

Daughter
20,645 variants
8,767 genes

Mother
20,486 variants
8,774 genes

❶ Mode of inheritance

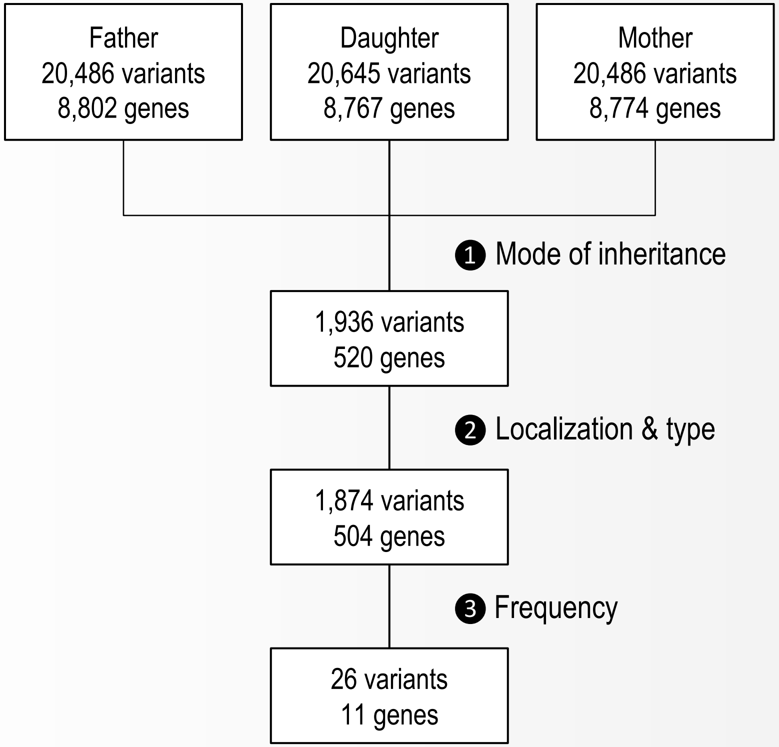
1,936 variants
520 genes

❷ Localization & type

1,874 variants
504 genes

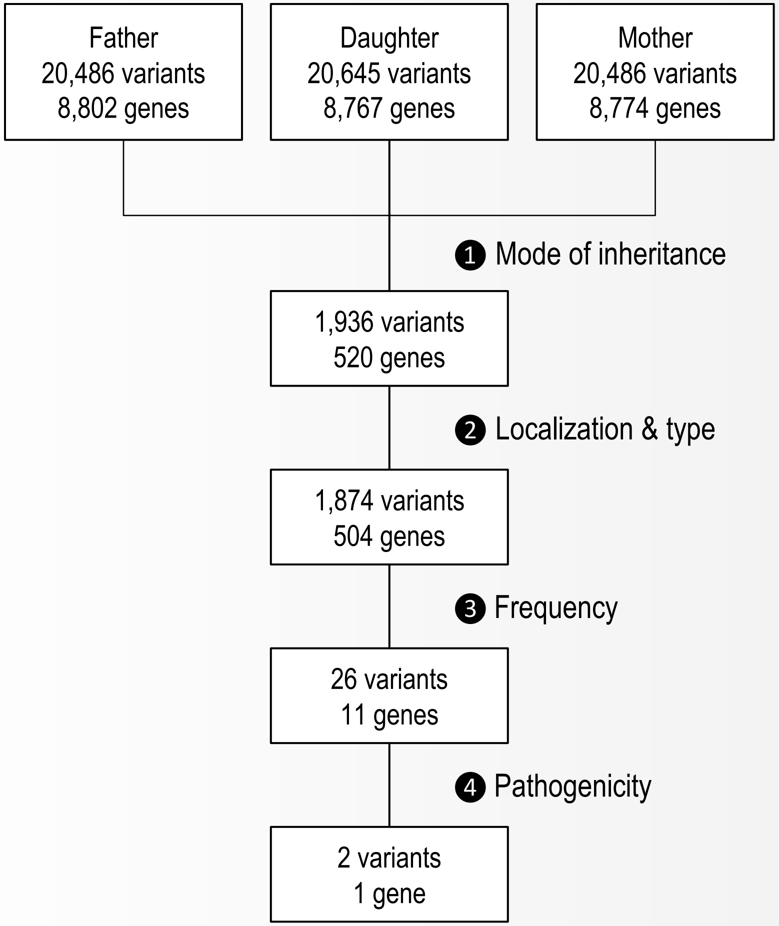
A simple use case

- A trio with an affected daughter and two healthy parents
- Mabry syndrome: intellectual disability, distinctive facial features, hyperphosphatasia, and other signs and symptoms.
- 4 steps filtration workflow:
  - Mode of inheritance (Autosomal recessive)
    - ➛ Homozygous or compound heterozygous
  - Mutation localization and type
  - Allele frequency



A simple use case

- A trio with an affected daughter and two healthy parents
- Mabry syndrome: intellectual disability, distinctive facial features, hyperphosphatasia, and other signs and symptoms.
- 4 steps filtration workflow:
  - Mode of inheritance (Autosomal recessive)
    - ➛ Homozygous or compound heterozygous
  - Mutation localization and type
  - Allele frequency
  - Pathogenicity predictions



Father
20,486 variants
8,802 genes

Daughter
20,645 variants
8,767 genes

Mother
20,486 variants
8,774 genes

1 Mode of inheritance

1,936 variants
520 genes

2 Localization & type

1,874 variants
504 genes

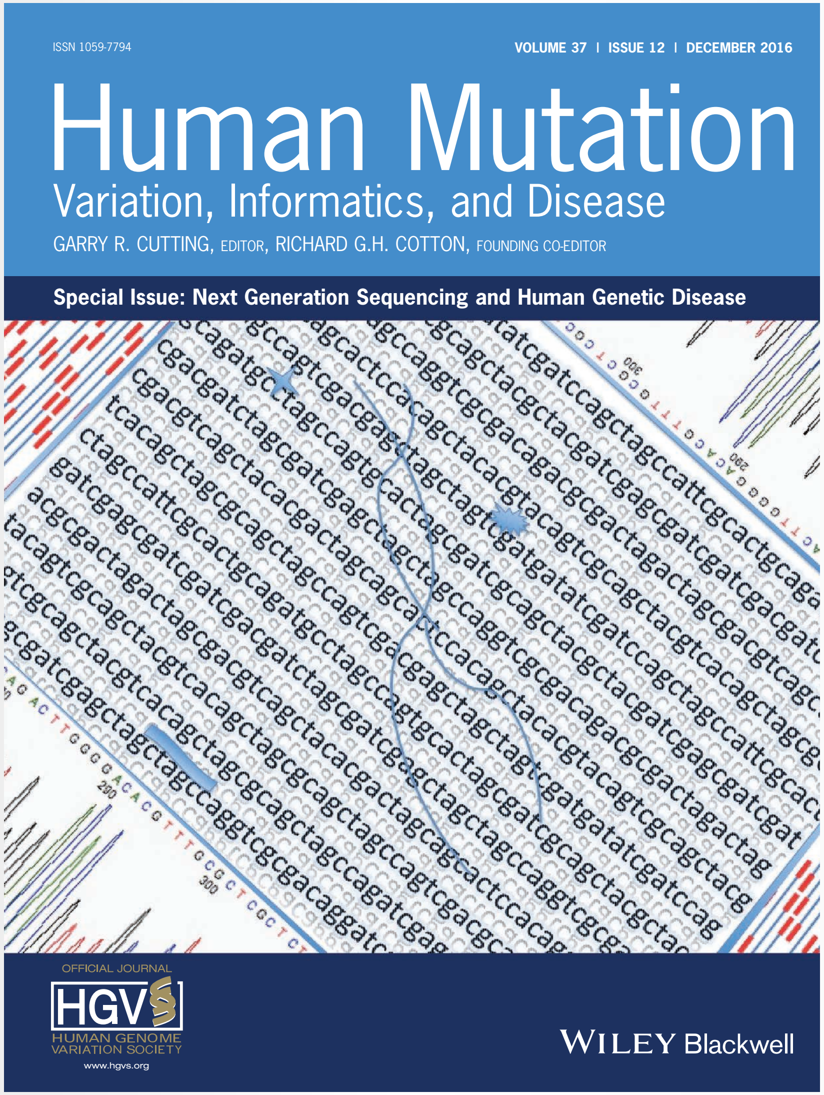
3 Frequency

26 variants
11 genes

4 Pathogenicity

2 variants
1 gene

More details



ISSN 1059-7794

VOLUME 37 I ISSUE 12 I DECEMBER 2016

Human Mutation

Variation, Informatics, and Disease

GARRY R. CUTTING, EDITOR, RICHARD G.H. COTTON, FOUNDING CO-EDITOR

Special Issue: Next Generation Sequencing and Human Genetic Disease

OFFICIAL JOURNAL

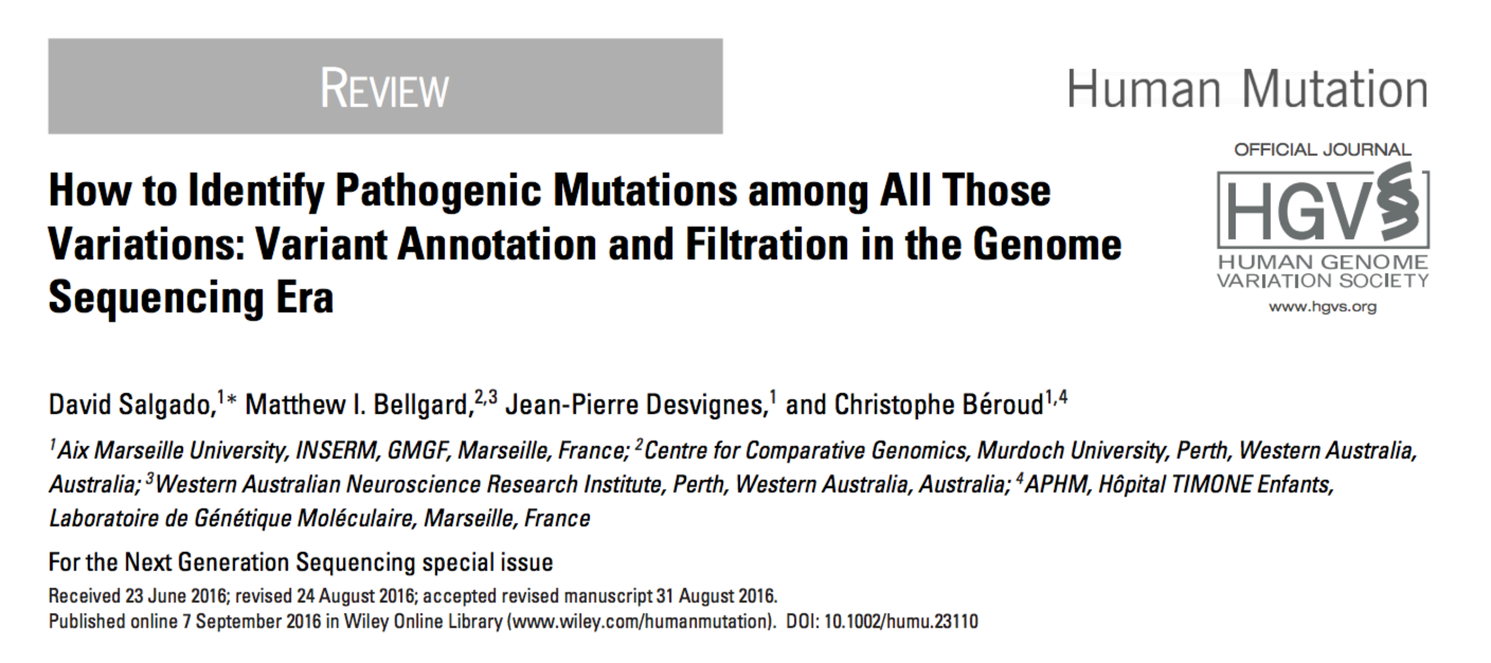
HGVS

HUMAN GENOME VARIATION SOCIETY

www.hgvs.org

WILEY Blackwell

More details



REVIEW

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era

David Salgado,[1]* Matthew I. Bellgard,[2,3] Jean-Pierre Desvignes,[1] and Christophe Béroud[1,4]

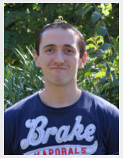
[1]Aix Marseille University, INSERM, GMGF, Marseille, France; [2]Centre for Comparative Genomics, Murdoch University, Perth, Western Australia, Australia; [3]Western Australian Neuroscience Research Institute, Perth, Western Australia, Australia; [4]APHM, Hôpital TIMONE Enfants, Laboratoire de Génétique Moléculaire, Marseille, France

For the Next Generation Sequencing special issue

Received 23 June 2016; revised 24 August 2016; accepted revised manuscript 31 August 2016.
Published online 7 September 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.23110

Acknowledgements

David Salgado



Jean-Pierre Desvignes

all members of the
Genetics & Bioinformatics Team

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

RD Connect

AFMTÉLÉTHON
INNOVER POUR GUÉRIR

NeurOmics

GenOmnis company (https://genomnis.com)

UMD-Predictor and HSF professional version for commercial users

GENOMNIS