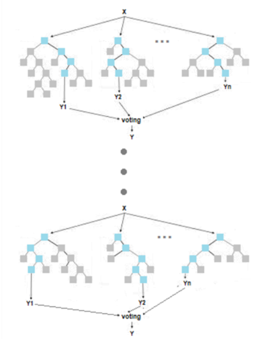
KU LEUVEN

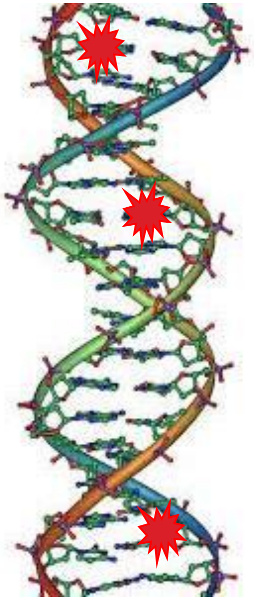
Dušan Popović

A computational framework for prioritization of disease-causing mutations



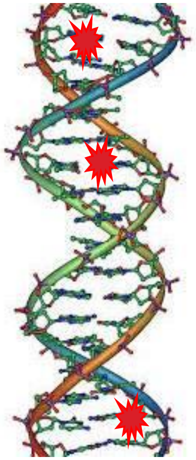
PREVALENCE OF GENETIC DISORDERS

PREVELANCE OF GENETIC DISORDERS



CAUSED BY MUTATIONS

PREVELANCE OF GENETIC DISORDERS



CAUSED BY MUTATIONS

NEXT-GENERATION SEQUENCING

NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

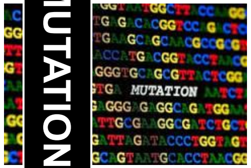
DISCOVERY OF MUTATION CAUSING THE DISEASE

NOVEL GENETIC DISORDER

SAMPLE

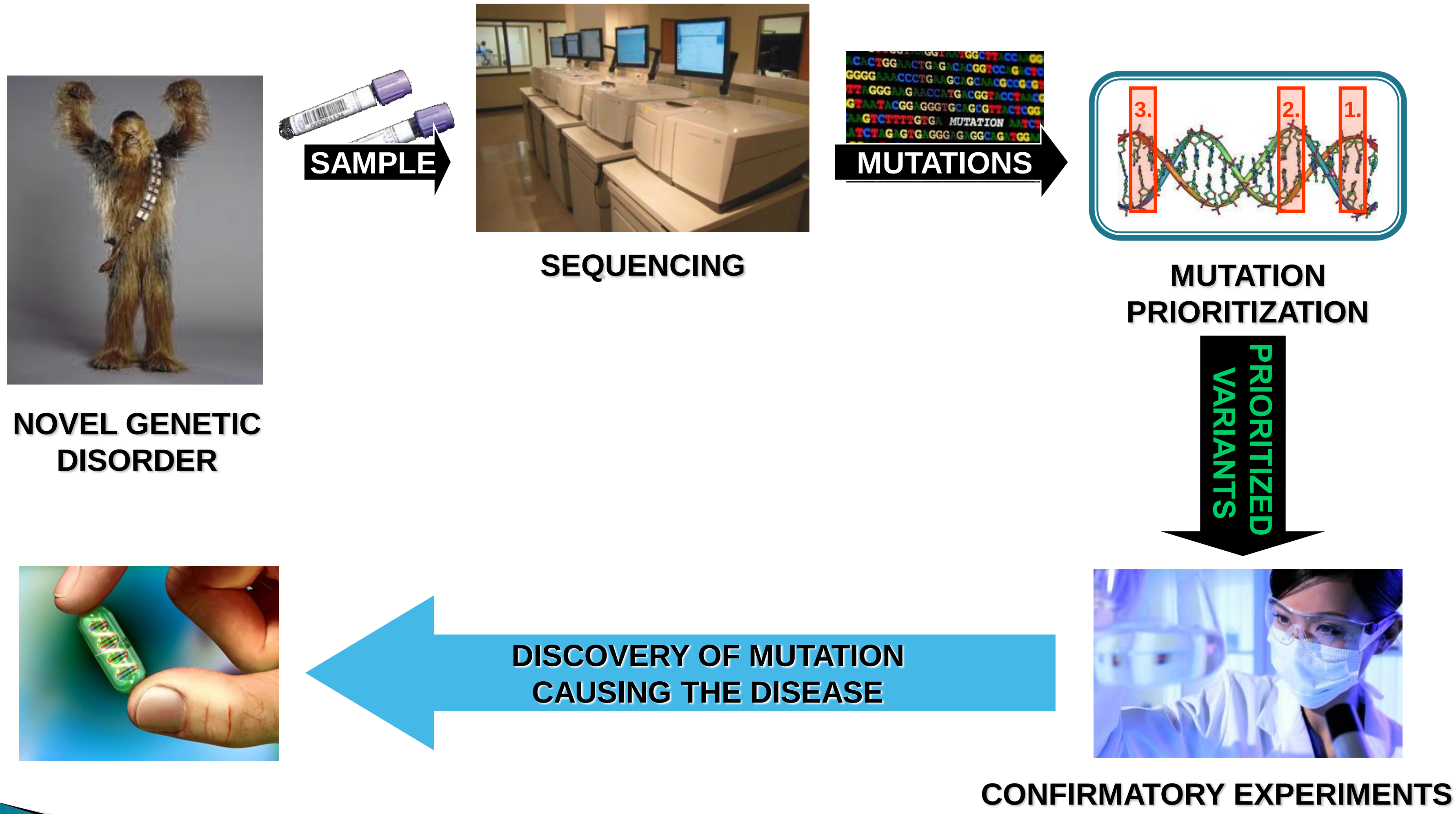
SEQUENCING

MUTATIONS



CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

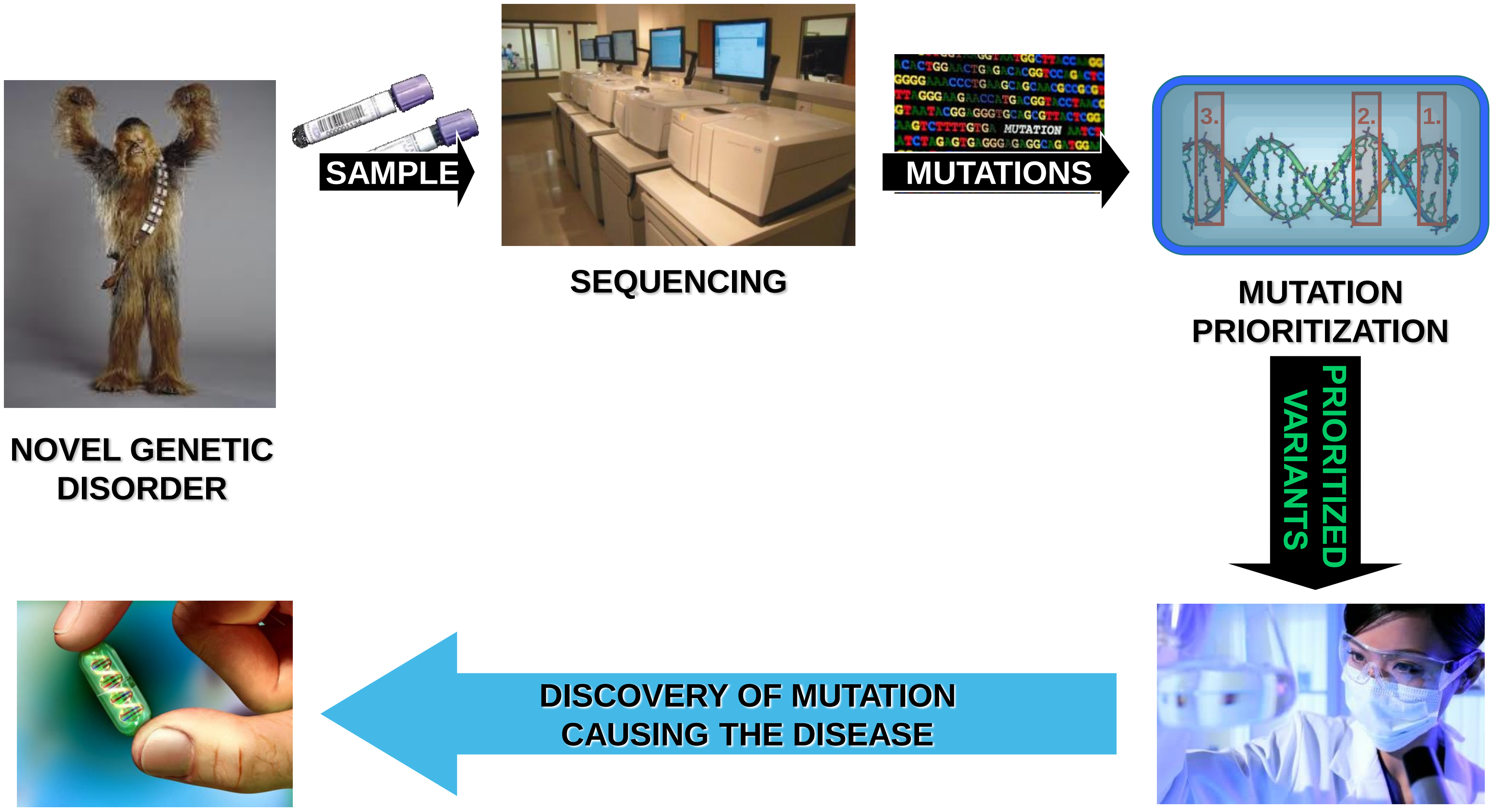
MUTATIONS

MUTATION PRIORITIZATION

PRIORITIZED VARIANTS

CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

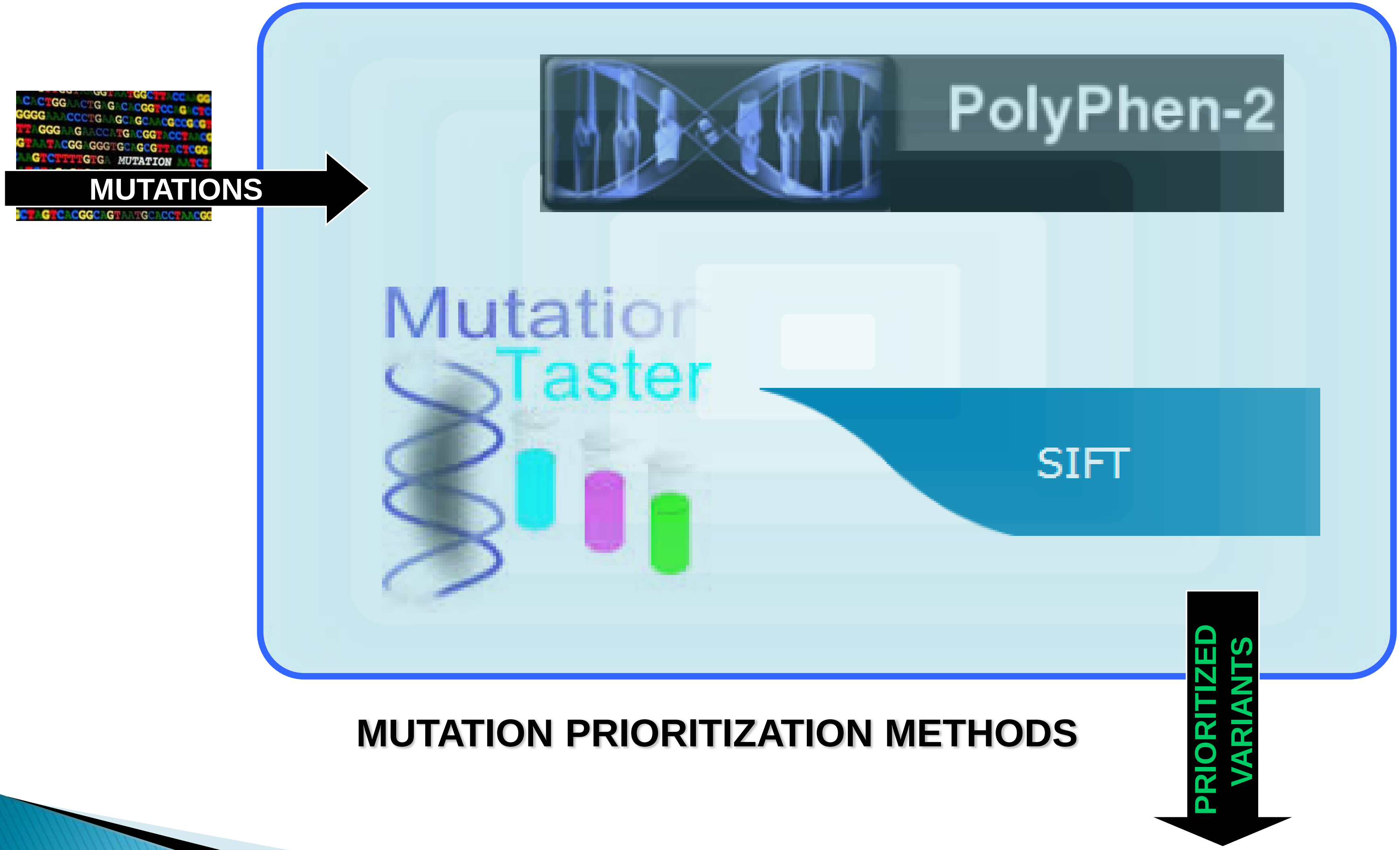
MUTATIONS

MUTATION PRIORITIZATION

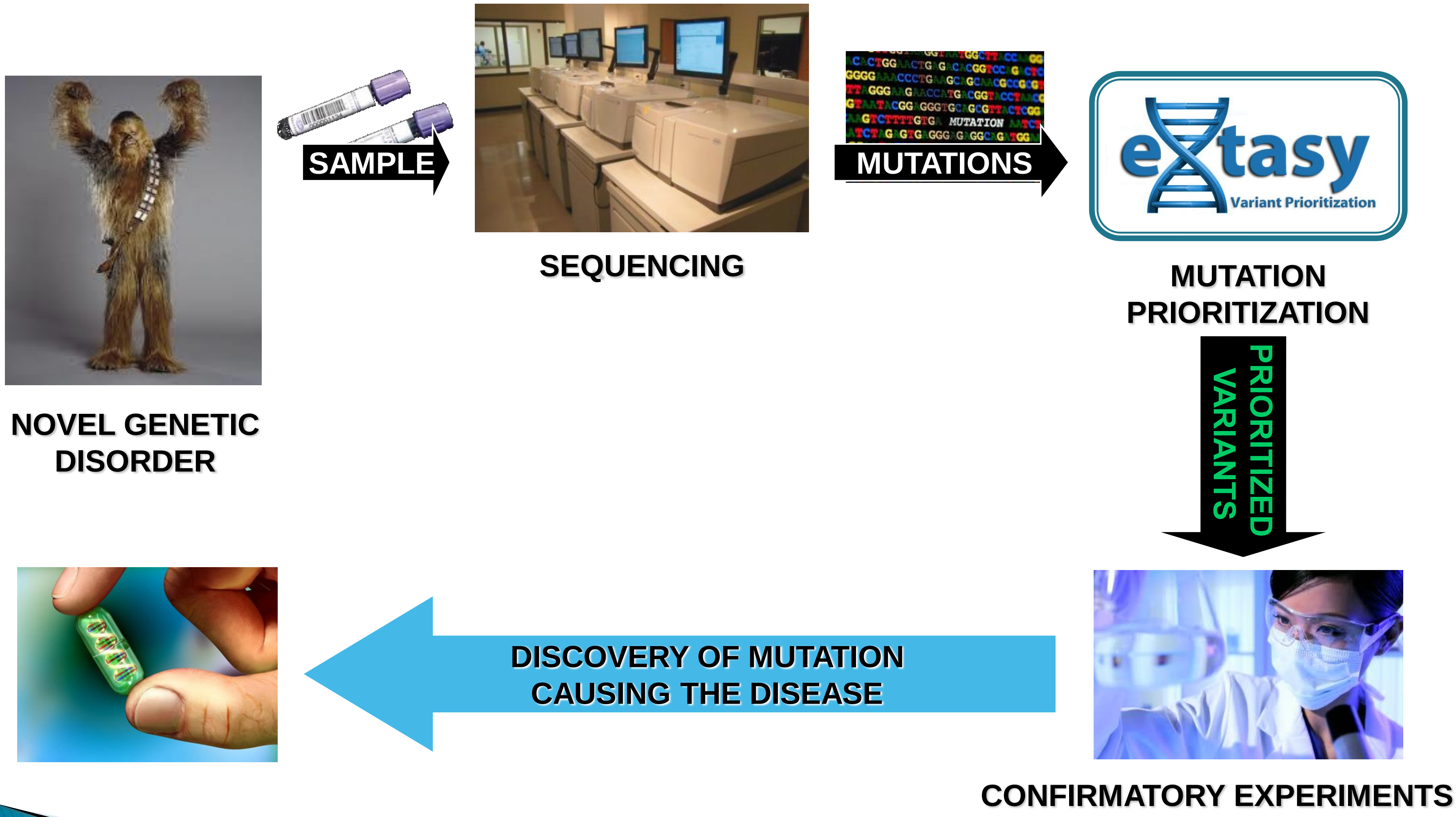
PRIORITIZED VARIANTS

CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



MUTATION PRIORITIZATION METHODS



NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

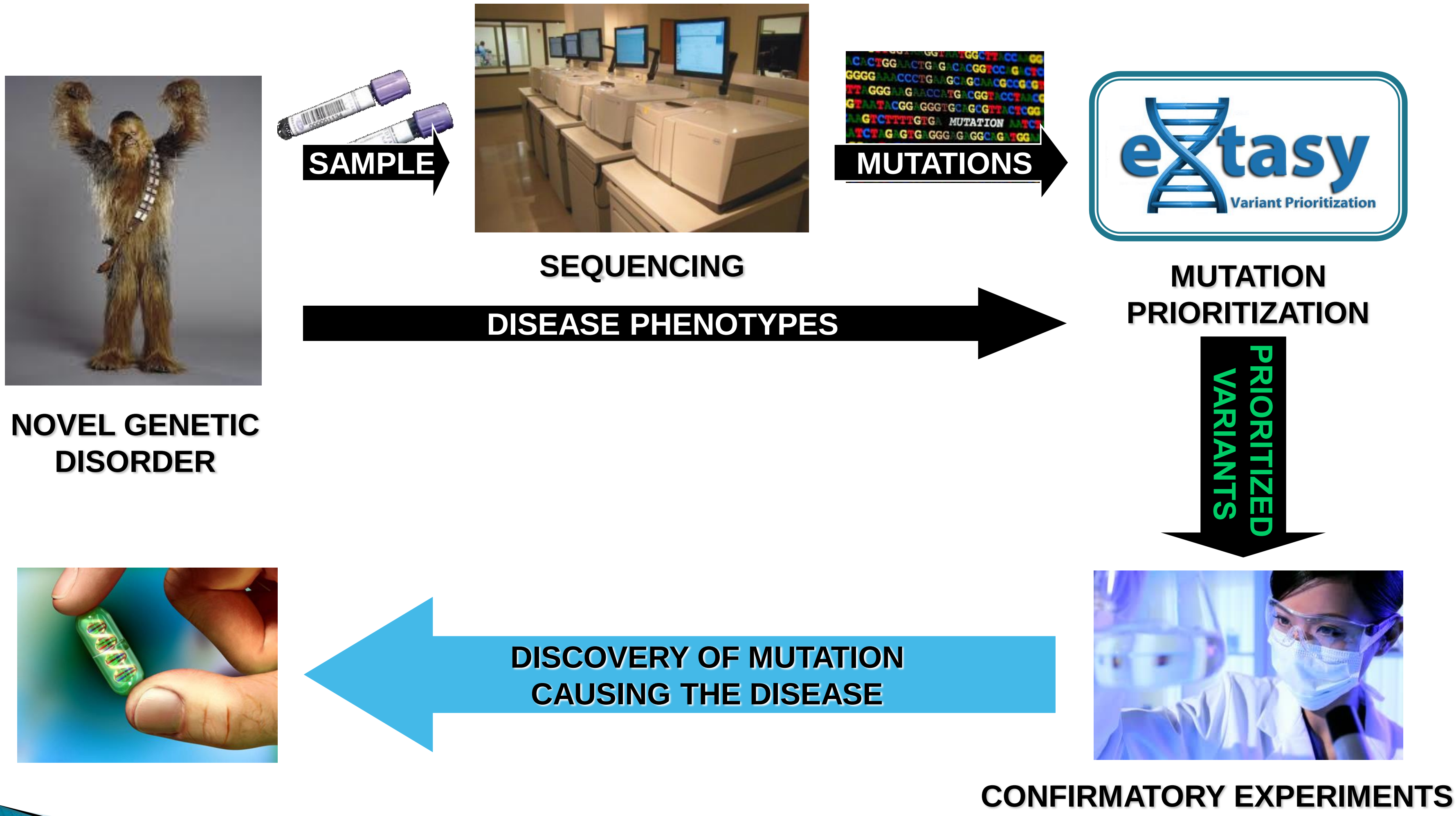
MUTATIONS

MUTATION PRIORITIZATION

PRIORITIZED VARIANTS

CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

MUTATIONS

extasy Variant Prioritization

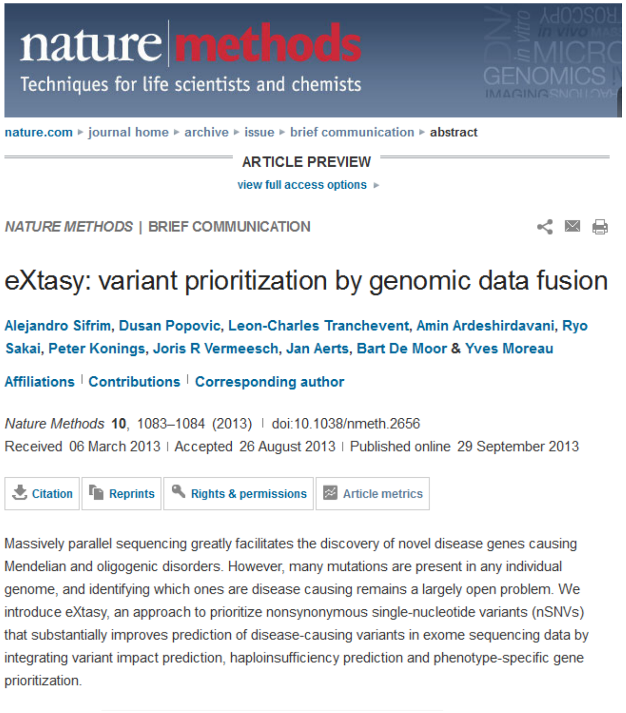
MUTATION PRIORITIZATION

DISEASE PHENOTYPES

PRIORITIZED VARIANTS

CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



nature methods
Techniques for life scientists and chemists

nature.com ▸ journal home ▸ archive ▸ issue ▸ brief communication ▸ abstract

ARTICLE PREVIEW
view full access options ▸

NATURE METHODS | BRIEF COMMUNICATION

eXtasy: variant prioritization by genomic data fusion

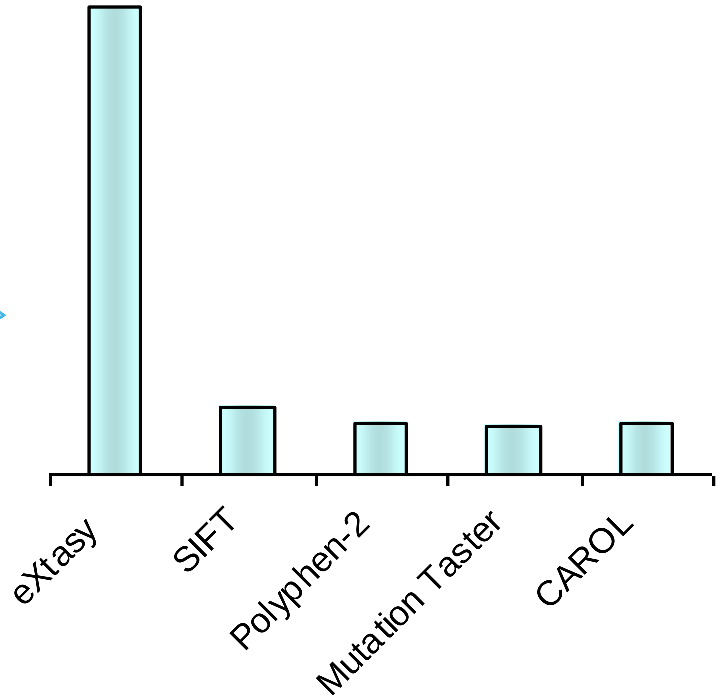
Alejandro Sifrim, Dusan Popovic, Leon-Charles Tranchevent, Amin Ardeshirdavani, Ryo Sakai, Peter Konings, Joris R Vermeesch, Jan Aerts, Bart De Moor & Yves Moreau

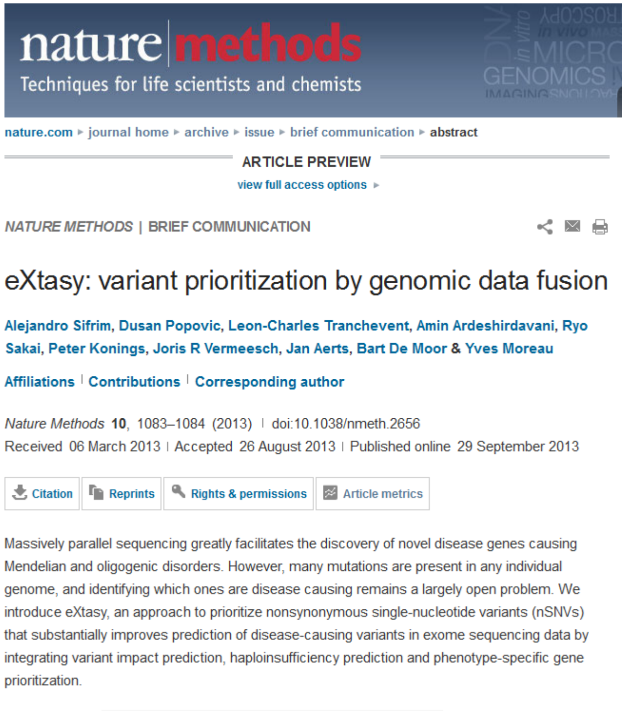
Affiliations | Contributions | Corresponding author

Nature Methods **10**, 1083–1084 (2013) | doi:10.1038/nmeth.2656
Received 06 March 2013 | Accepted 26 August 2013 | Published online 29 September 2013

Citation | Reprints | Rights & permissions | Article metrics

Massively parallel sequencing greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce eXtasy, an approach to prioritize nonsynonymous single-nucleotide variants (nSNVs) that substantially improves prediction of disease-causing variants in exome sequencing data by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization.





nature methods
Techniques for life scientists and chemists

nature.com ▸ journal home ▸ archive ▸ issue ▸ brief communication ▸ abstract

ARTICLE PREVIEW
view full access options ▸

NATURE METHODS | BRIEF COMMUNICATION

eXtasy: variant prioritization by genomic data fusion

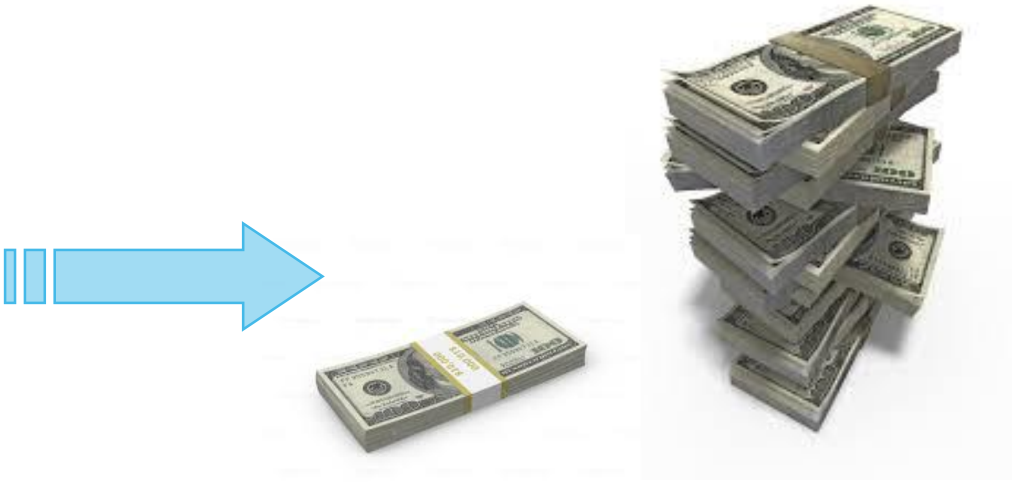
Alejandro Sifrim, Dusan Popovic, Leon-Charles Tranchevent, Amin Ardeshirdavani, Ryo Sakai, Peter Konings, Joris R Vermeesch, Jan Aerts, Bart De Moor & Yves Moreau

Affiliations | Contributions | Corresponding author

Nature Methods **10**, 1083–1084 (2013) | doi:10.1038/nmeth.2656
Received 06 March 2013 | Accepted 26 August 2013 | Published online 29 September 2013

Citation | Reprints | Rights & permissions | Article metrics

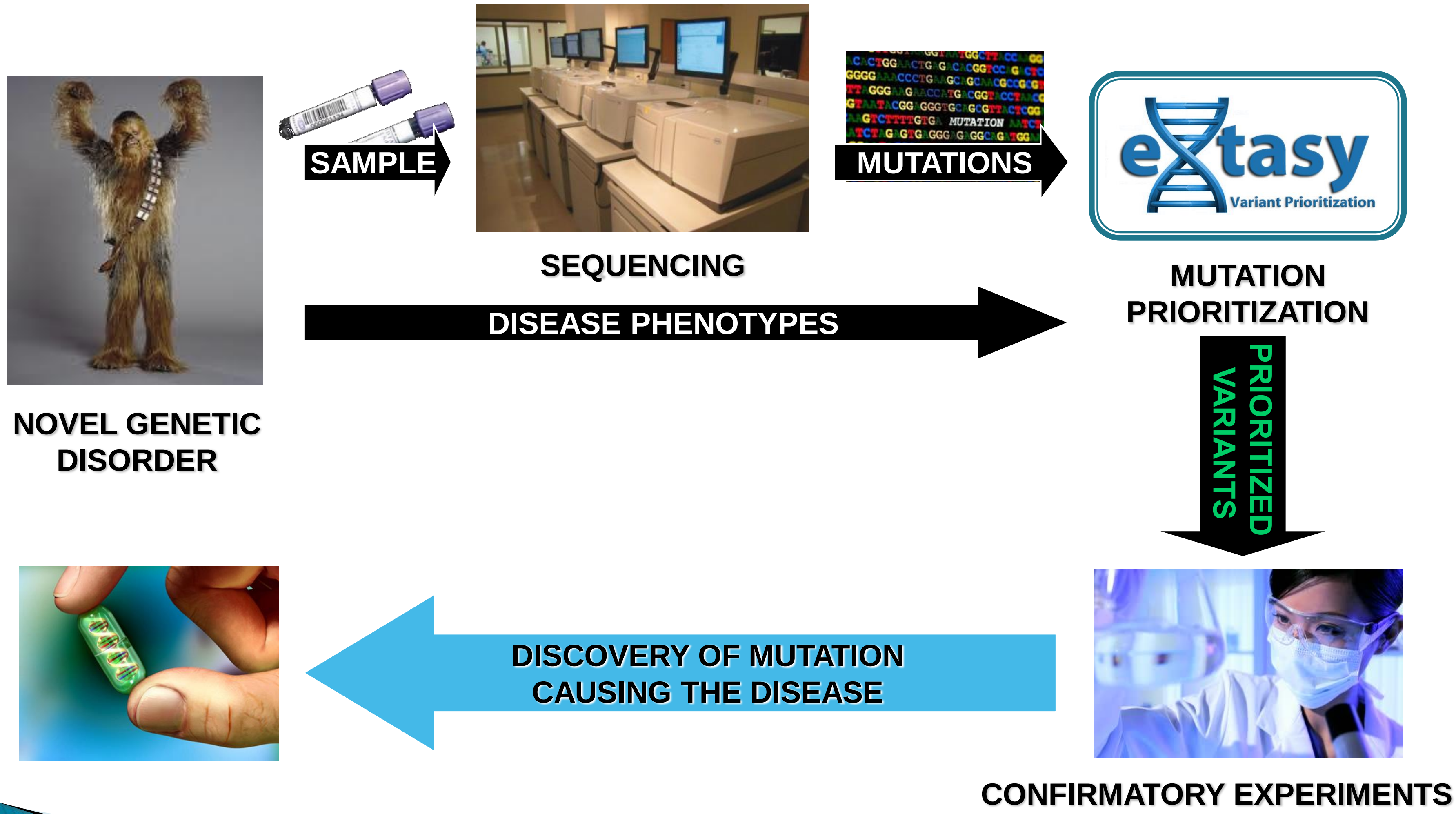
Massively parallel sequencing greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce eXtasy, an approach to prioritize nonsynonymous single-nucleotide variants (nSNVs) that substantially improves prediction of disease-causing variants in exome sequencing data by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization.



EXTASY

OTHER METHODS

HOW DOES EXTASY WORK?



NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

MUTATIONS

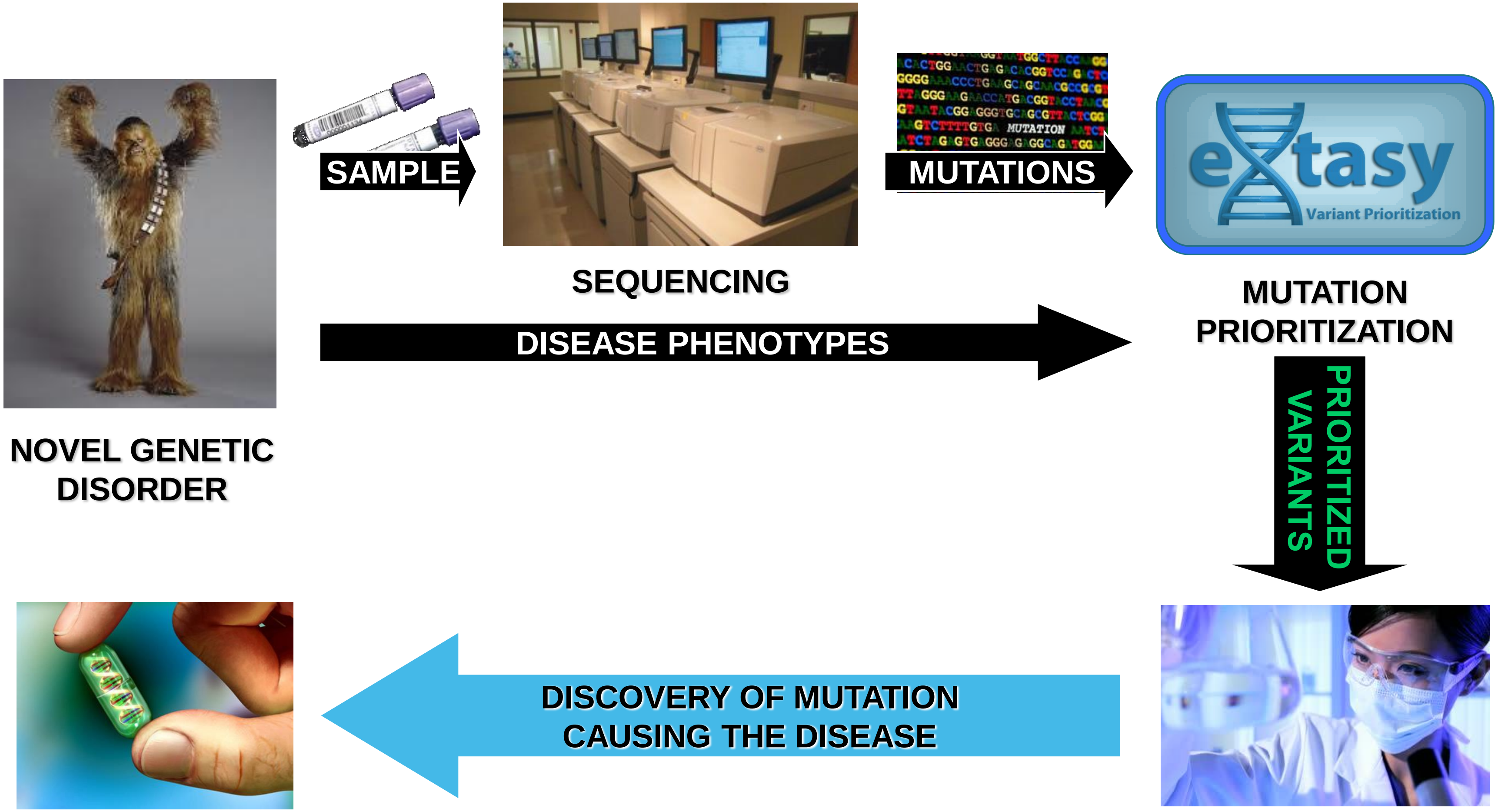
DISEASE PHENOTYPES

MUTATION PRIORITIZATION

PRIORITIZED VARIANTS

CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



NOVEL GENETIC DISORDER

SAMPLE

SEQUENCING

MUTATIONS

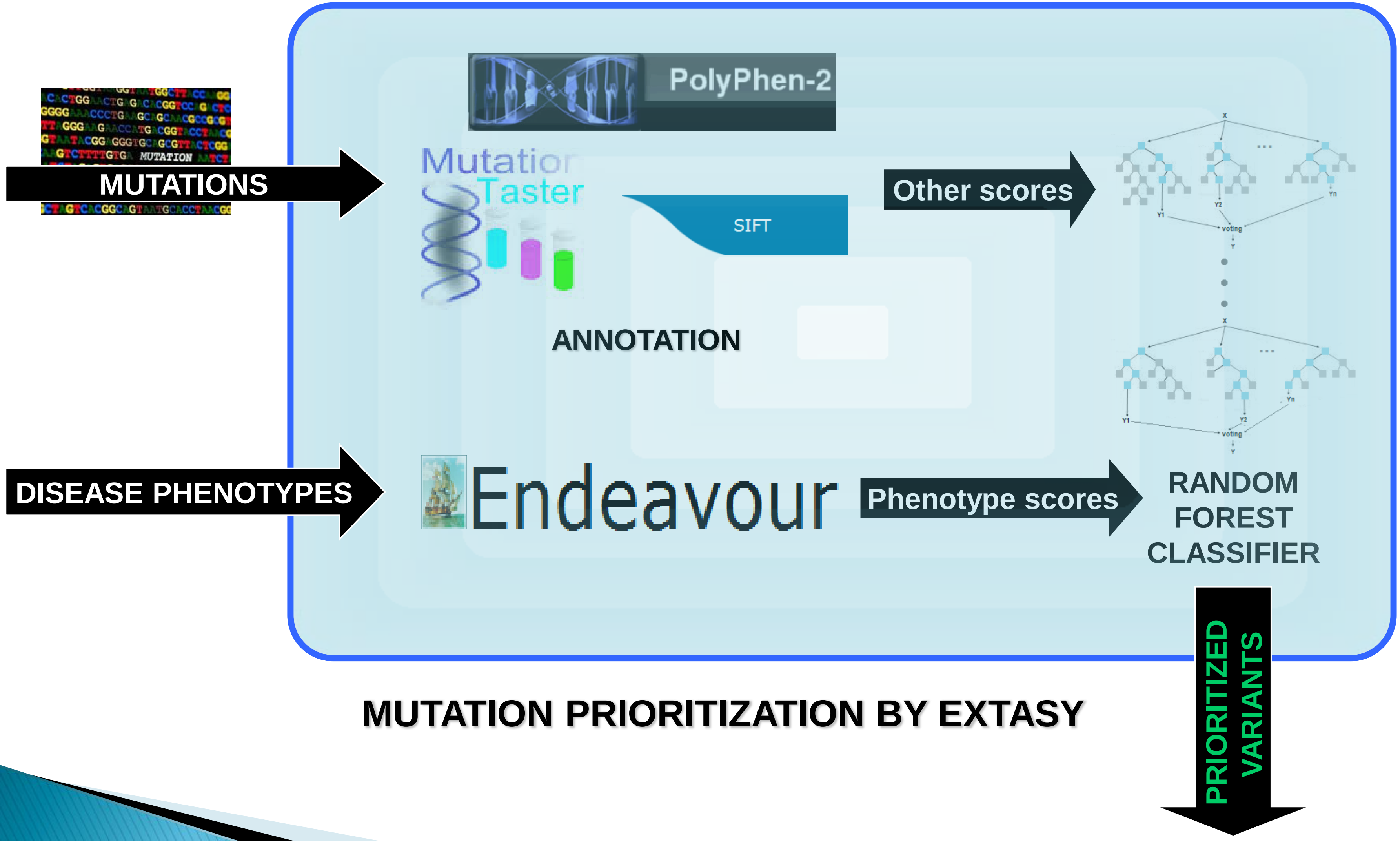
MUTATION PRIORITIZATION

DISEASE PHENOTYPES

PRIORITIZED VARIANTS

CONFIRMATORY EXPERIMENTS

DISCOVERY OF MUTATION CAUSING THE DISEASE



MUTATIONS

DISEASE PHENOTYPES

PolyPhen-2

MutationTaster

SIFT

ANNOTATION

Endeavour

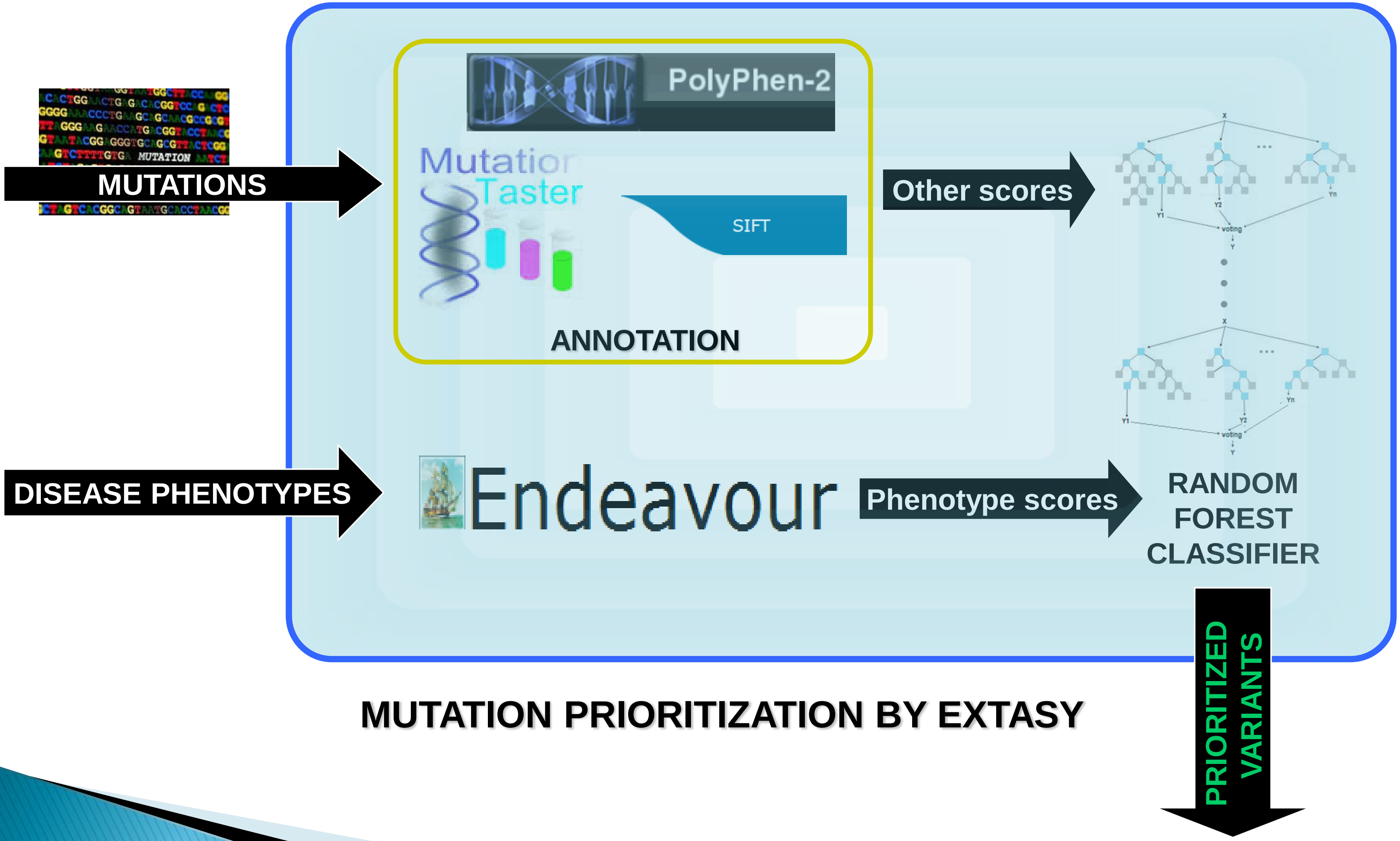
Other scores

Phenotype scores

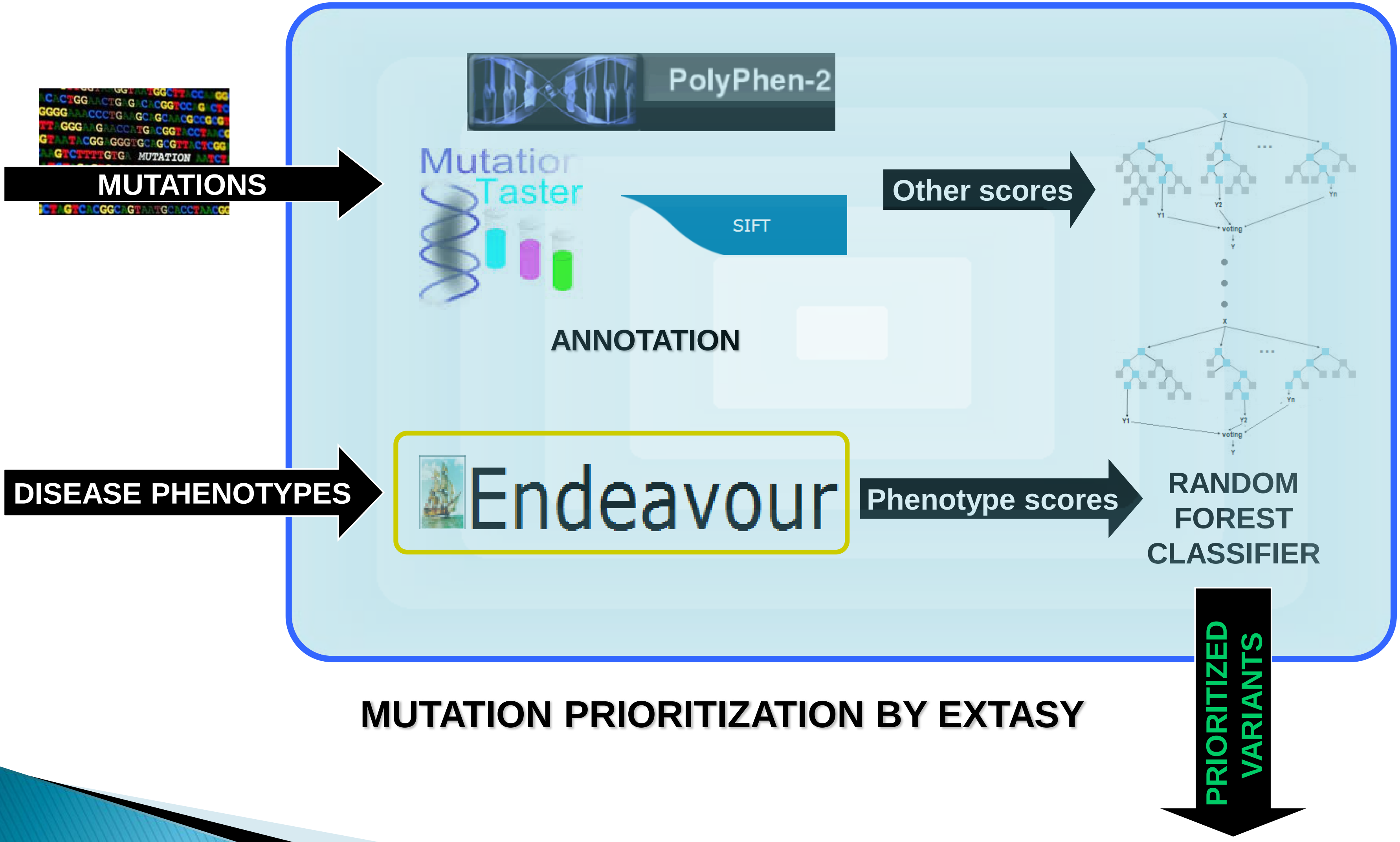
RANDOM FOREST CLASSIFIER

PRIORITIZED VARIANTS

MUTATION PRIORITIZATION BY EXTASY



MUTATION PRIORITIZATION BY EXTASY



MUTATIONS

DISEASE PHENOTYPES

PolyPhen-2

MutationTaster

SIFT

ANNOTATION

Other scores

Endeavour

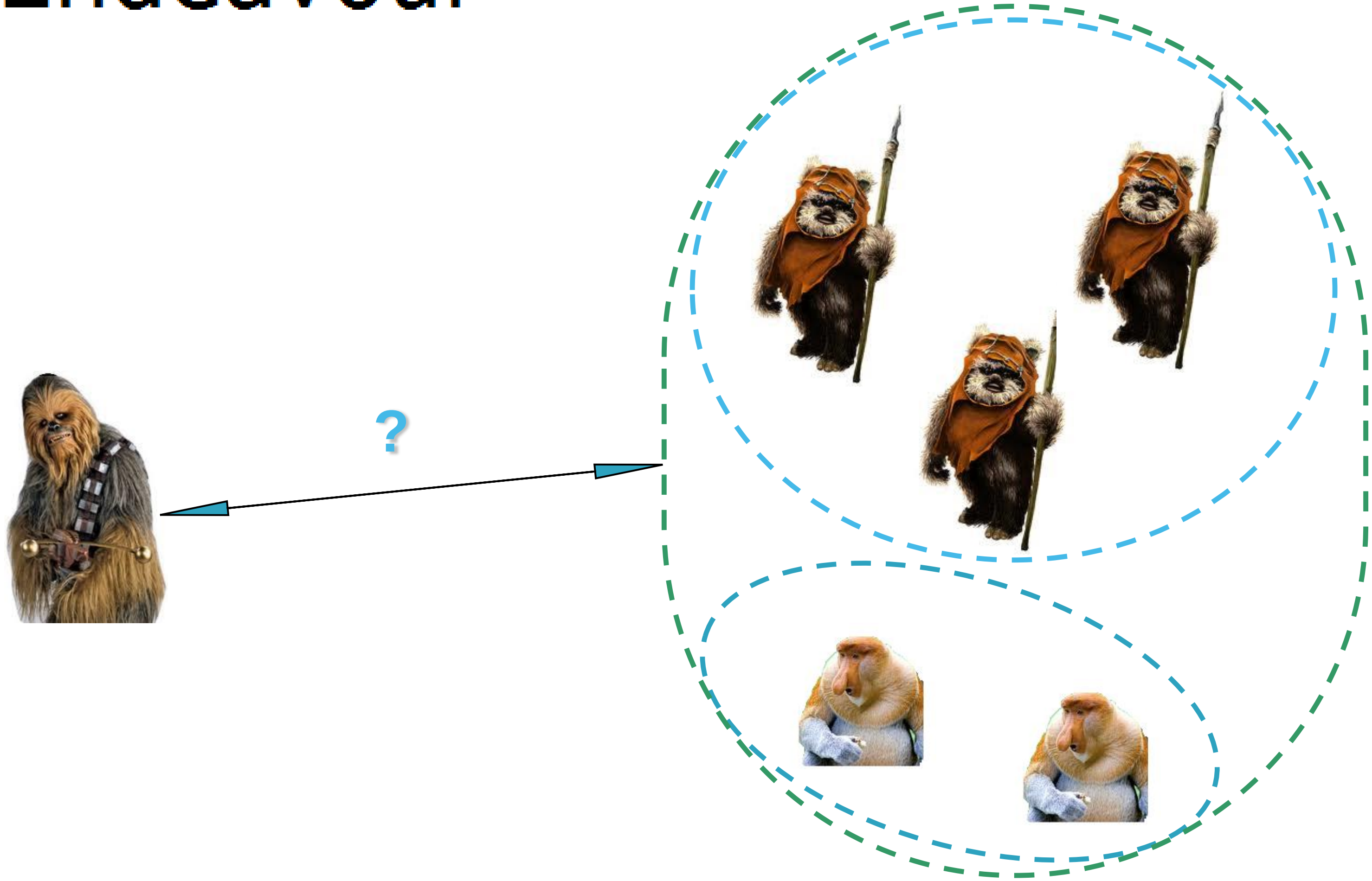
Phenotype scores

RANDOM FOREST CLASSIFIER

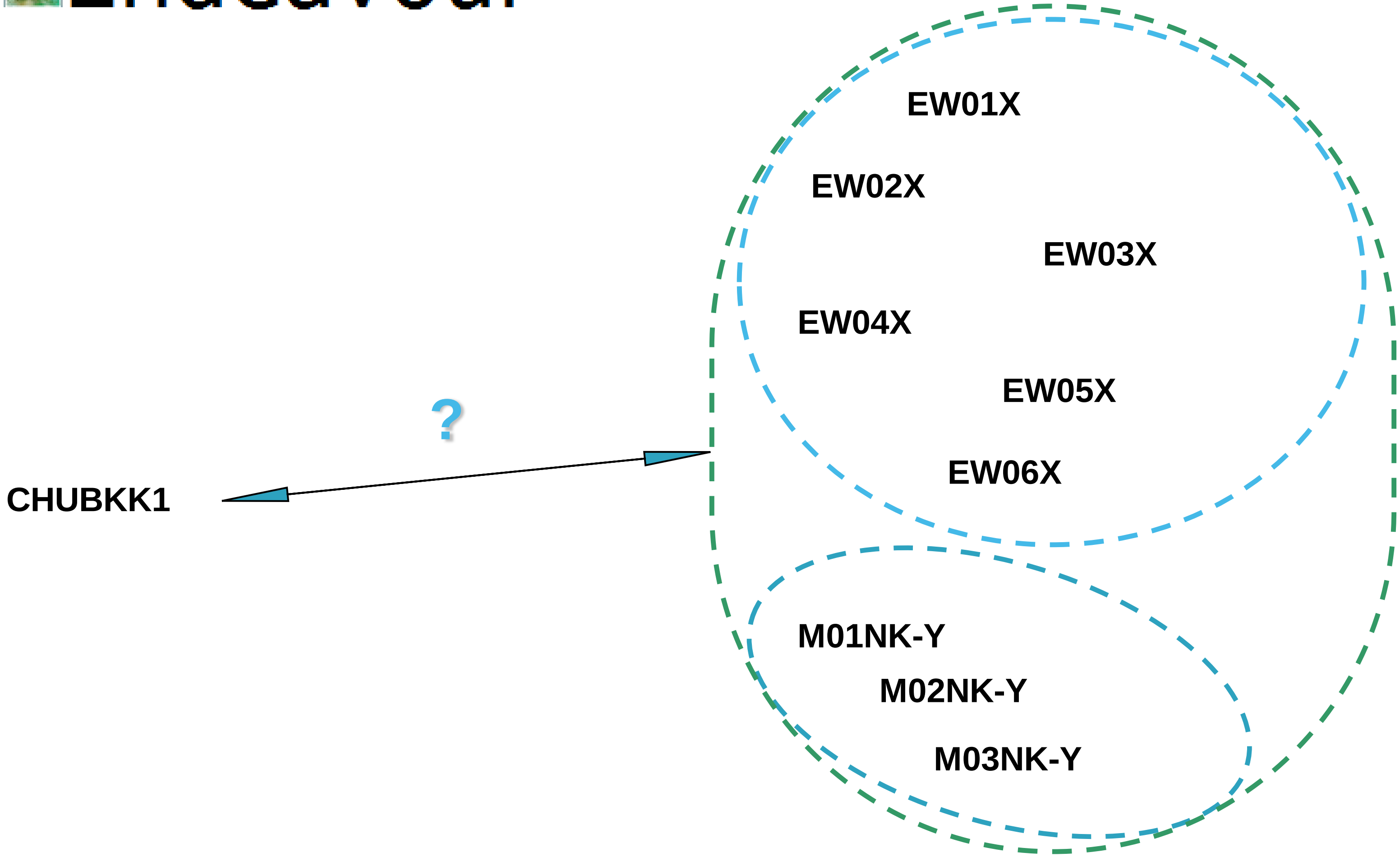
PRIORITIZED VARIANTS

MUTATION PRIORITIZATION BY EXTASY

Endeavour



Endeavour



?

CHUBKK1

EW01X

EW02X

EW03X

EW04X

EW05X

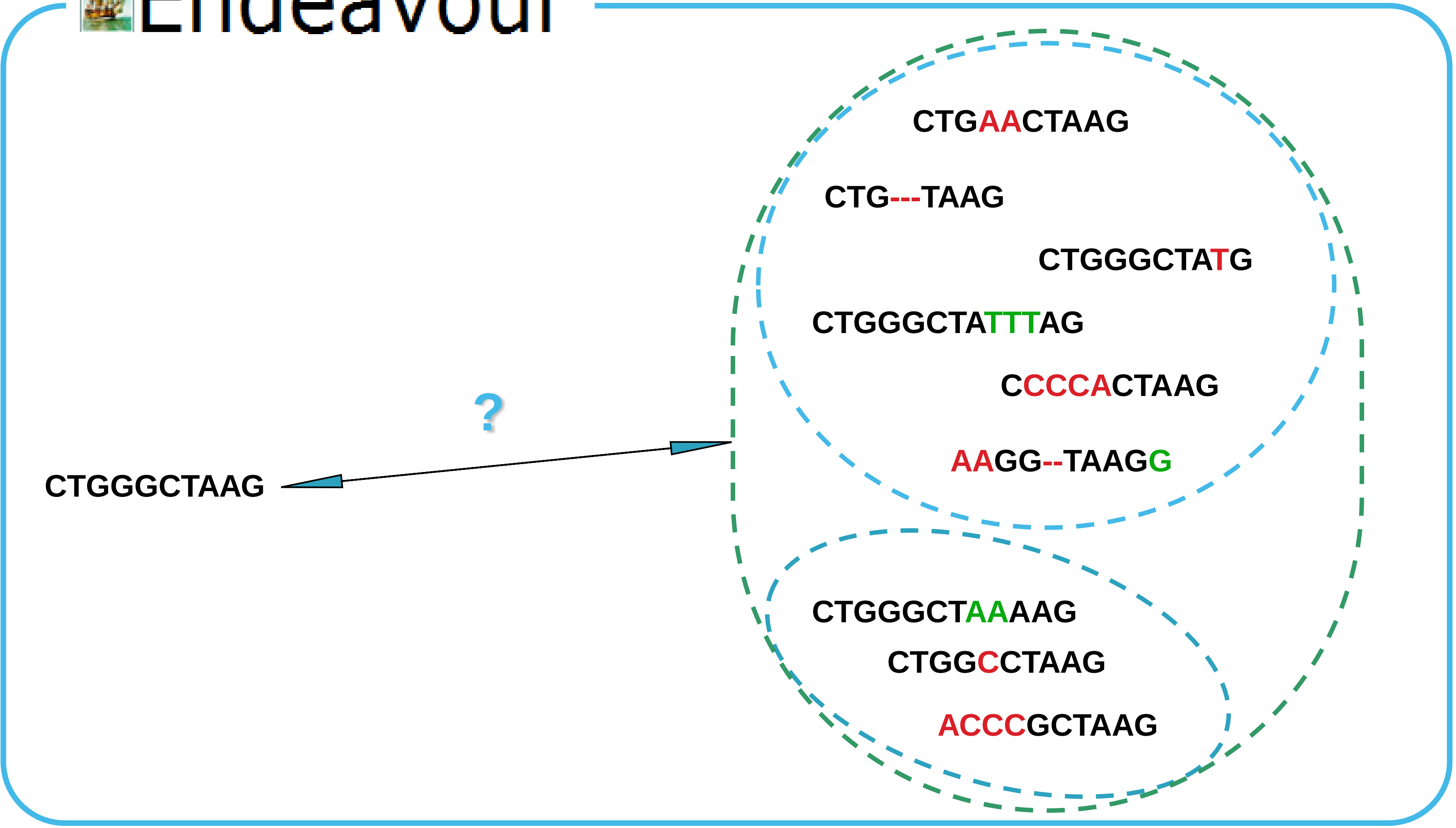
EW06X

M01NK-Y

M02NK-Y

M03NK-Y

Endeavour



CTGGGCTAAG

?

CTGAACTAAG

CTG---TAAG

CTGGGCTATG

CTGGGCTATTTAG

CCCCACTAAG

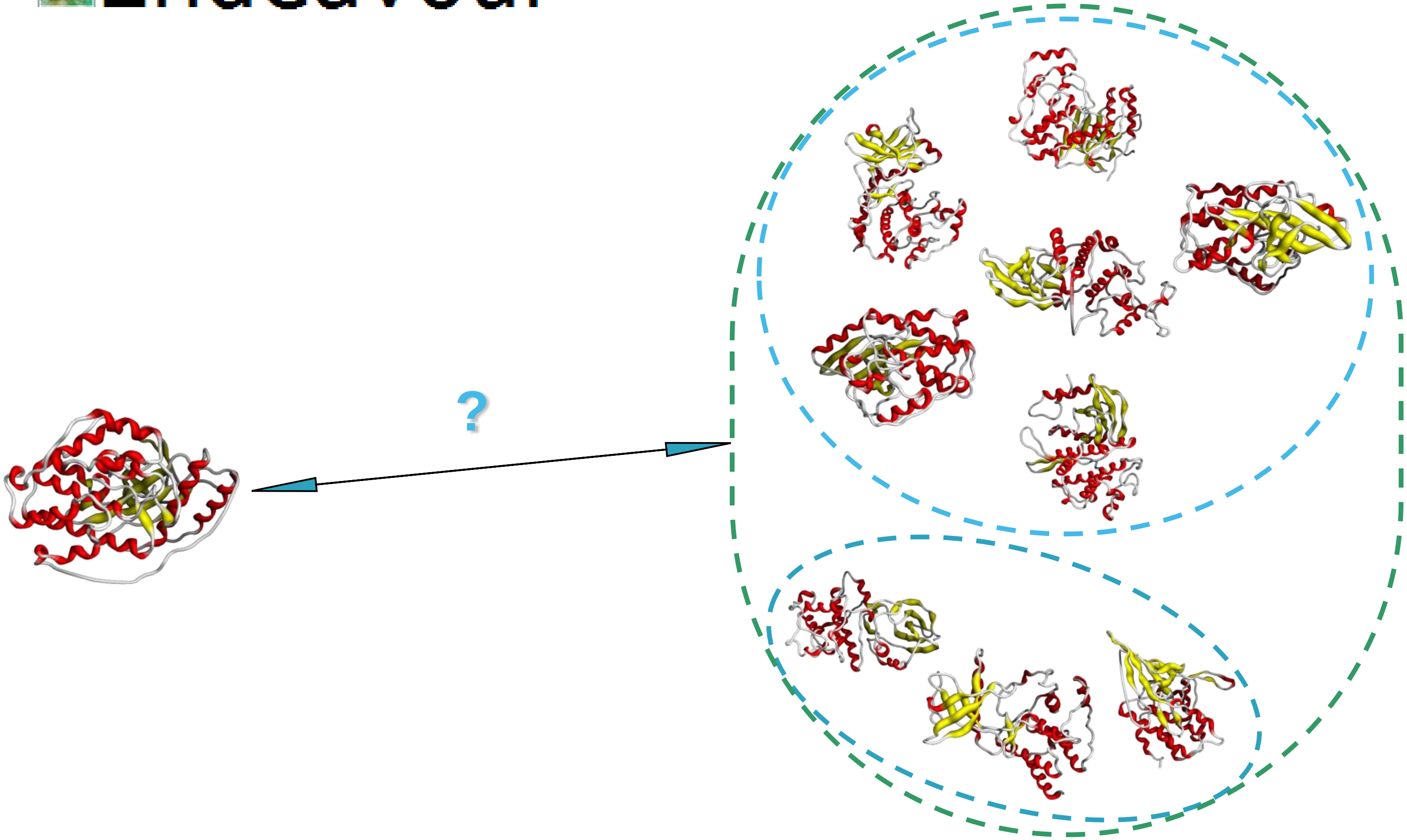
AAGG--TAAGG

CTGGGCTAAAAG

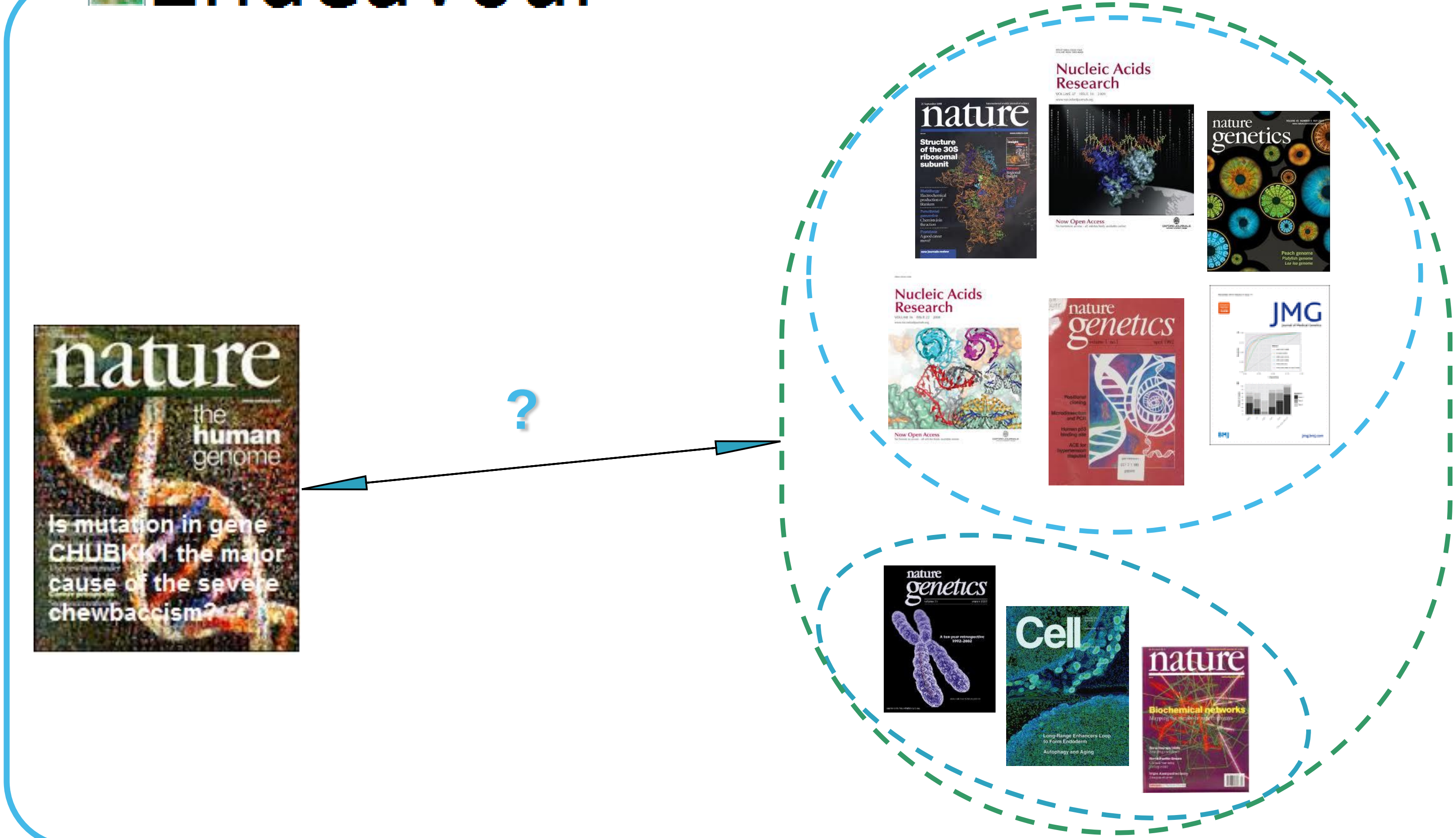
CTGGCCTAAG

ACCCGCTAAG

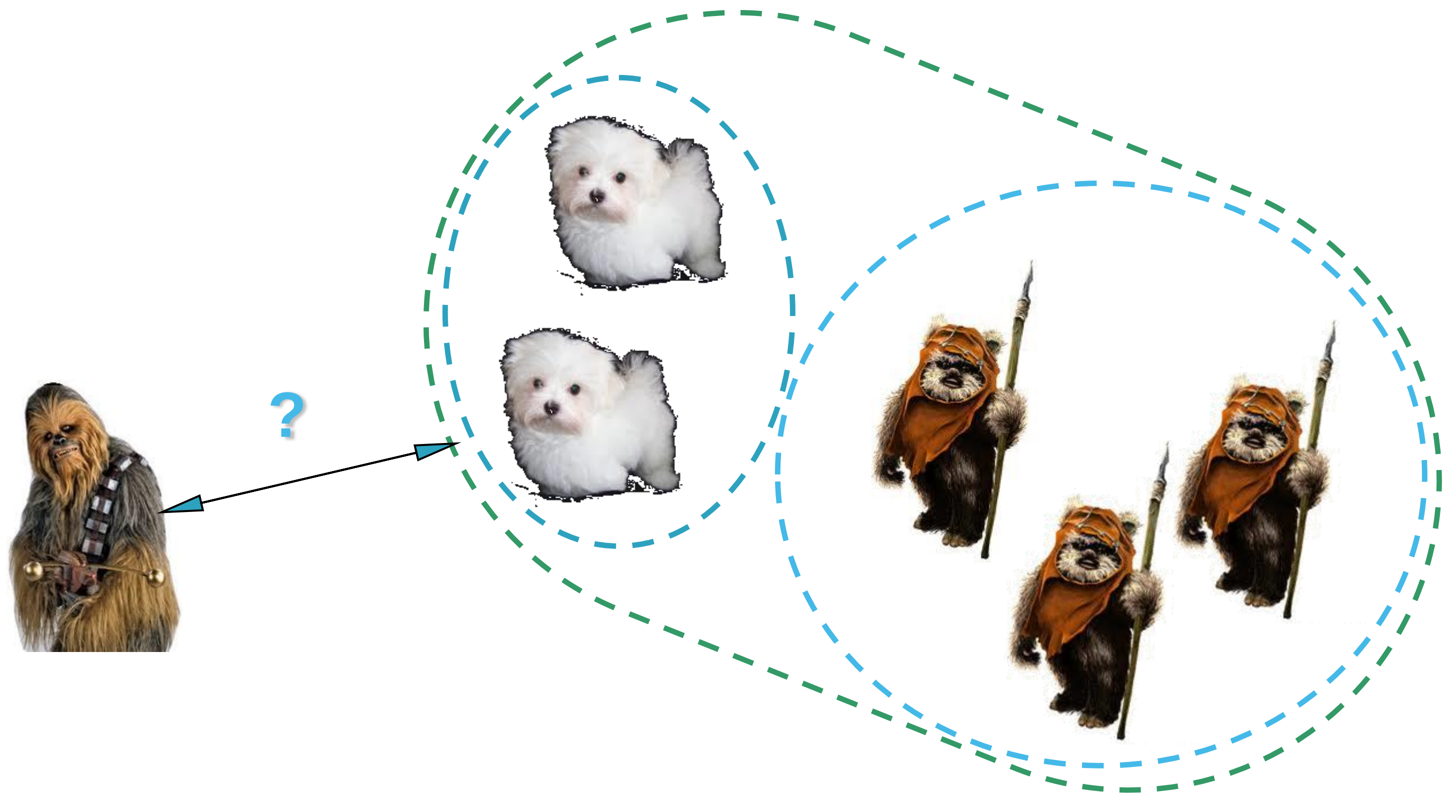
Endeavour

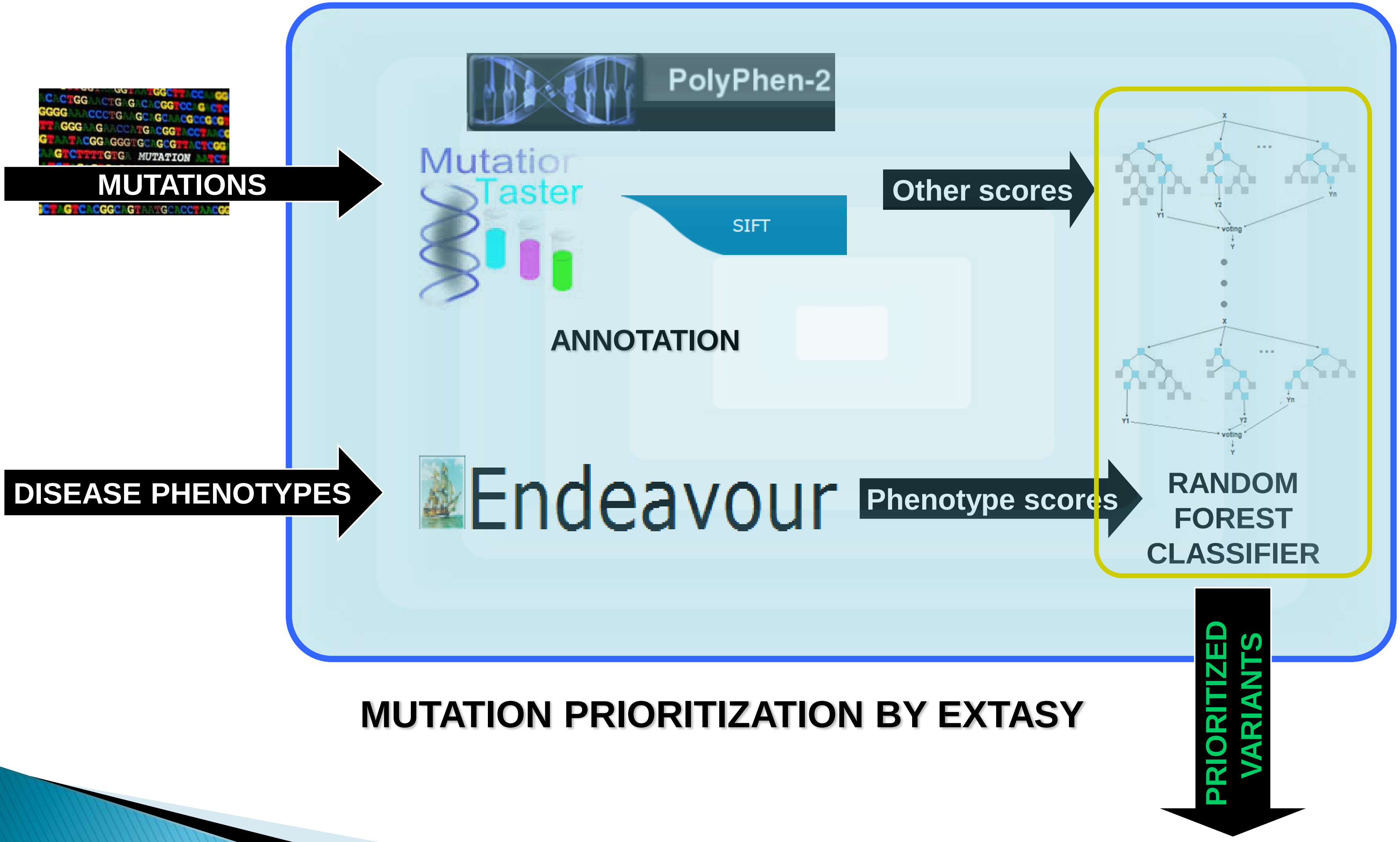


Endeavour


nature
the human genome
Is mutation in gene CHUBKK1 the major cause of the severe chewbaccism?
?
nature
Nucleic Acids Research
nature genetics
Nucleic Acids Research
nature genetics
JMG
nature genetics
Cell
nature

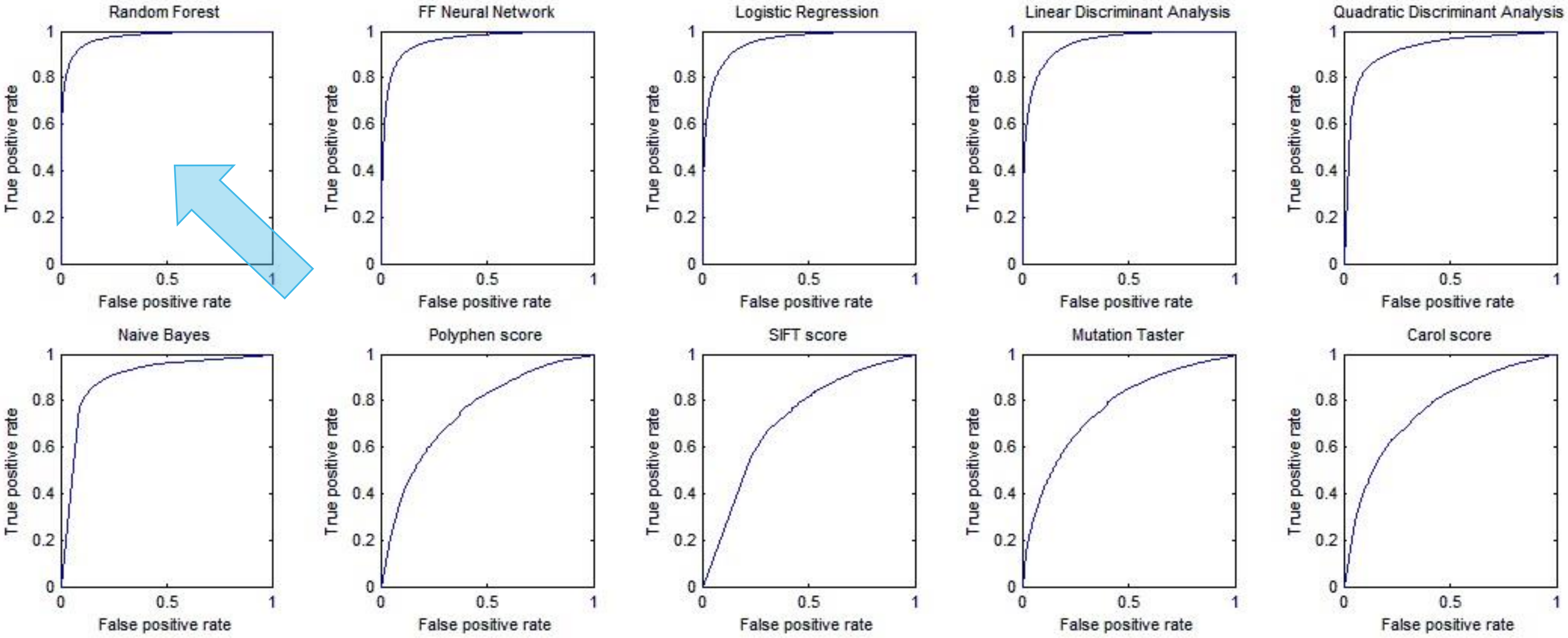
Endeavour

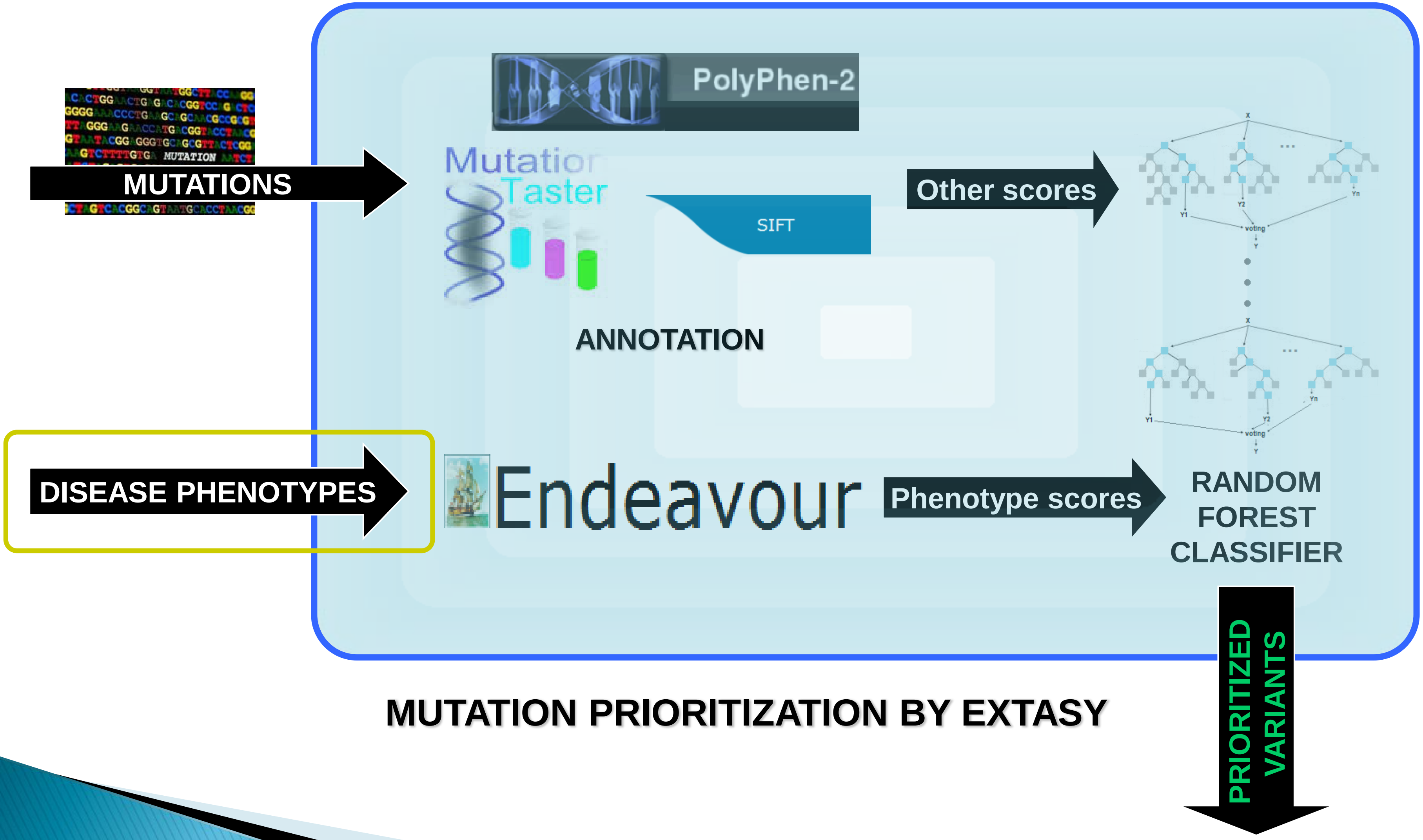




MUTATION PRIORITIZATION BY EXTASY

RANDOM FOREST CLASSIFIER





MUTATIONS

DISEASE PHENOTYPES

PolyPhen-2

MutationTaster

SIFT

ANNOTATION

Endeavour

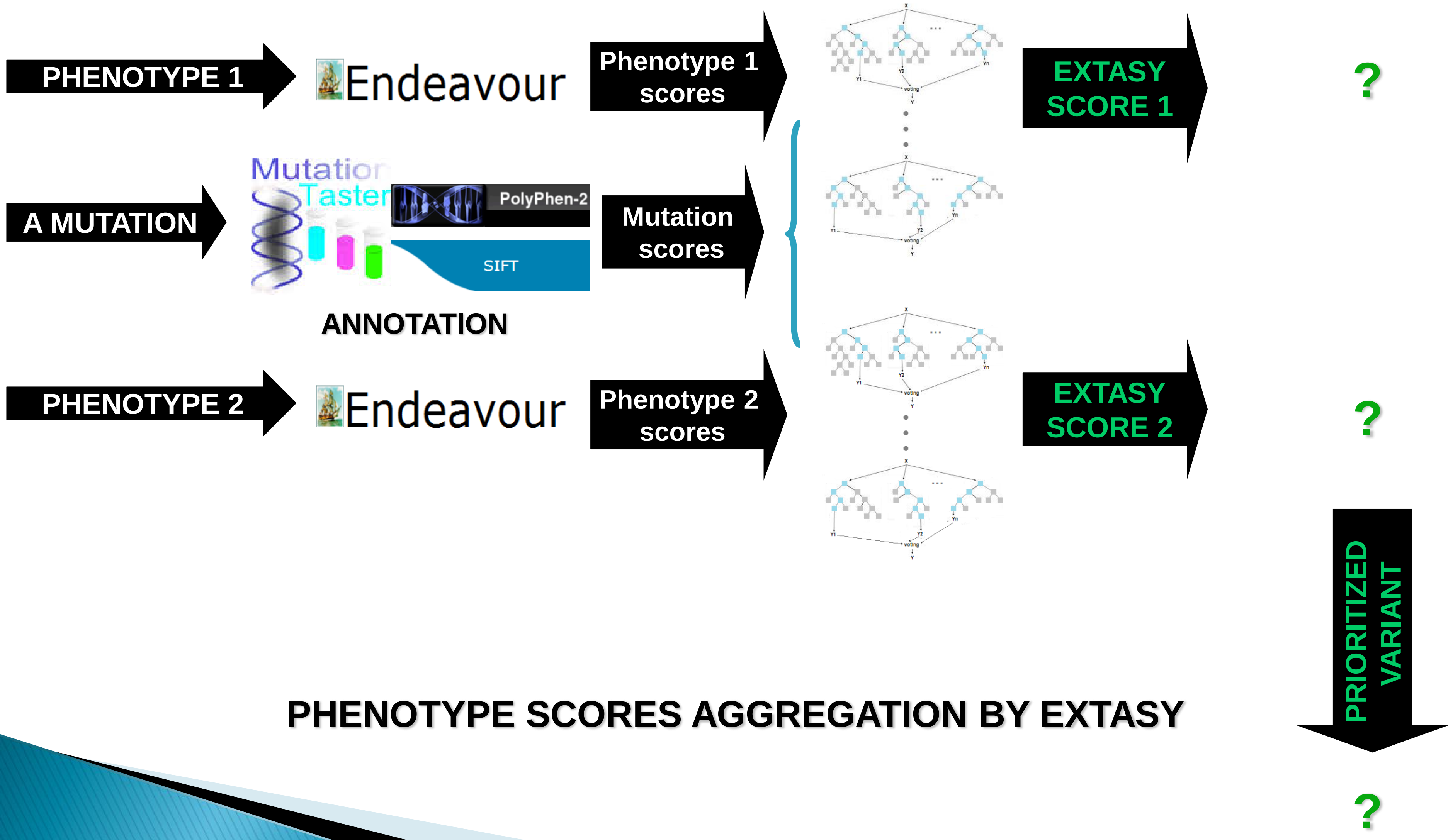
Other scores

Phenotype scores

RANDOM FOREST CLASSIFIER

PRIORITIZED VARIANTS

MUTATION PRIORITIZATION BY EXTASY



PHENOTYPE 1

Endeavour

Phenotype 1 scores

EXTASY SCORE 1

?

A MUTATION

MutationTaster

PolyPhen-2

SIFT

ANNOTATION

Mutation scores

PHENOTYPE 2

Endeavour

Phenotype 2 scores

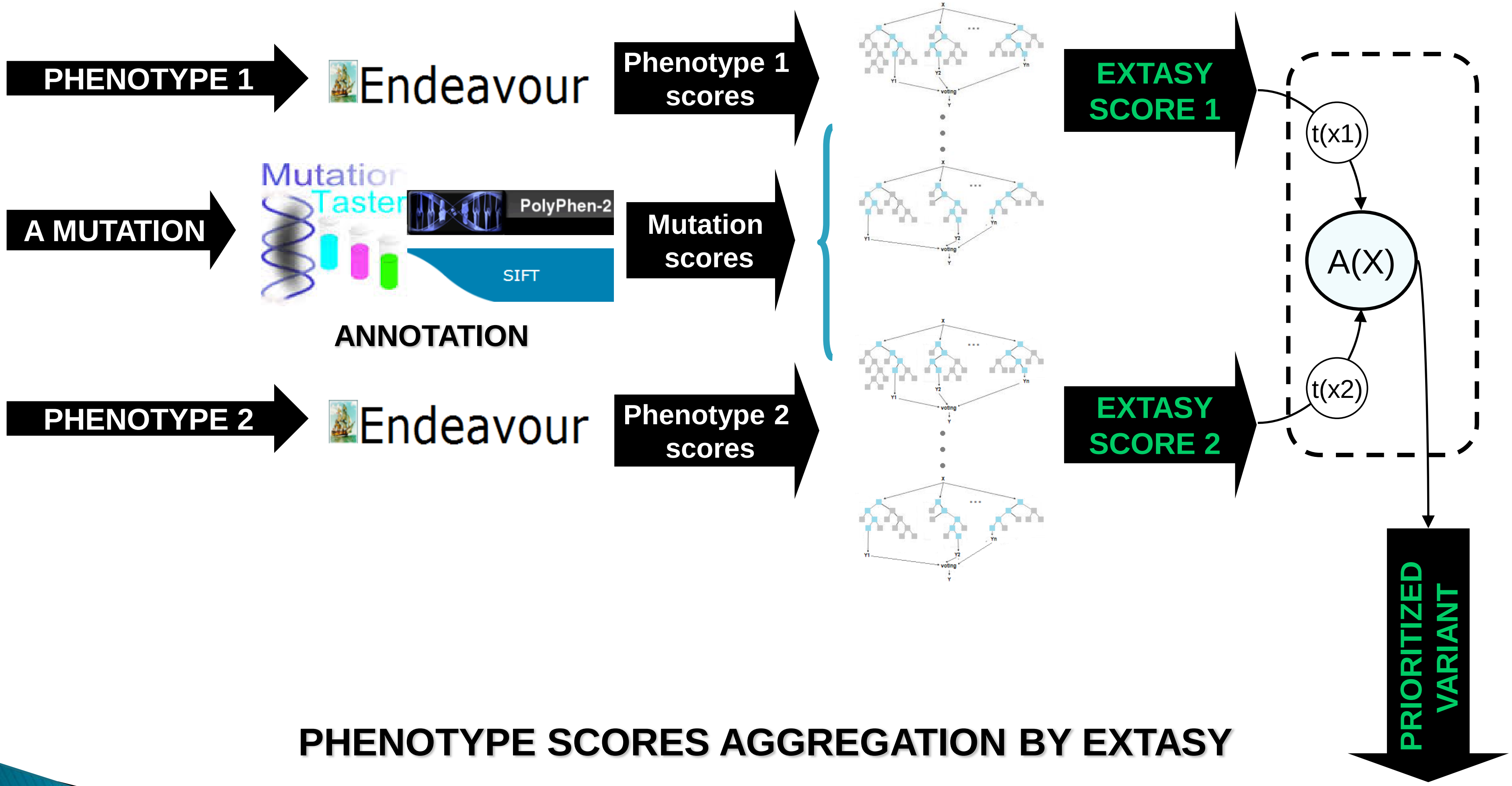
EXTASY SCORE 2

?

PRIORITIZED VARIANT

?

PHENOTYPE SCORES AGGREGATION BY EXTASY



PHENOTYPE 1

Endeavour

Phenotype 1 scores

A MUTATION

Mutation Taster

PolyPhen-2

SIFT

ANNOTATION

Mutation scores

PHENOTYPE 2

Endeavour

Phenotype 2 scores

EXTASY SCORE 1

EXTASY SCORE 2

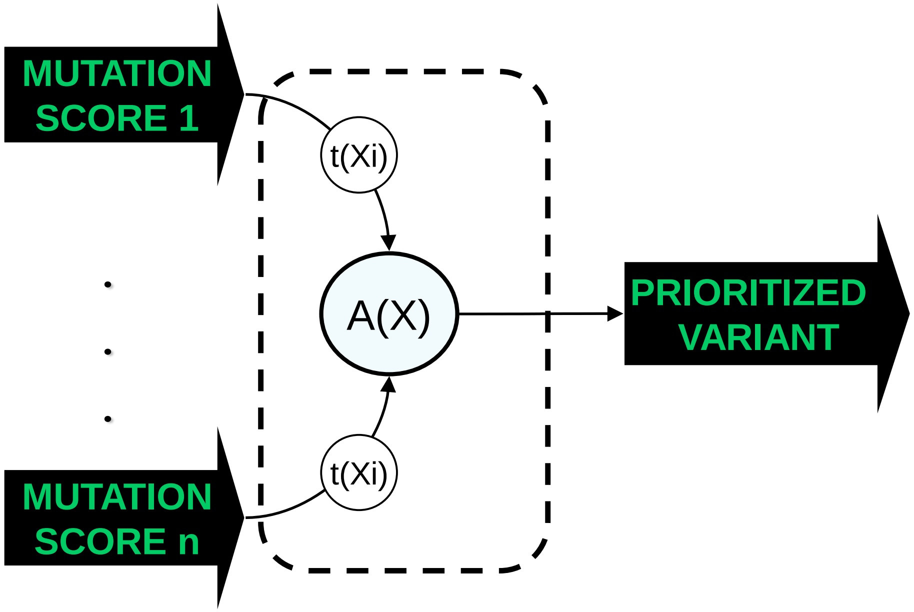
t(x1)

A(X)

t(x2)

PRIORITIZED VARIANT

PHENOTYPE SCORES AGGREGATION BY EXTASY



1. MAXIMUM :

$t(Xi) = Xi$

$A(X) = max(X)$

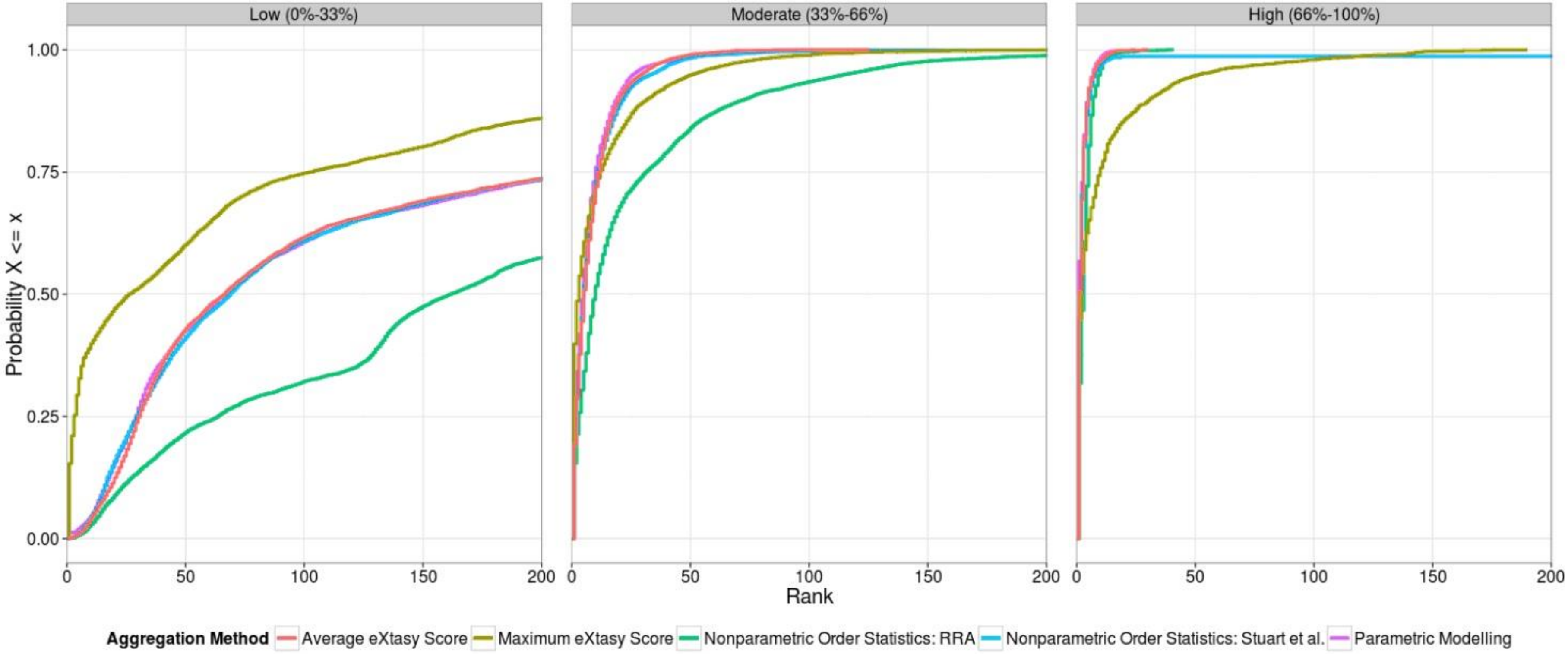
2. PARAMETRIC MODELING :

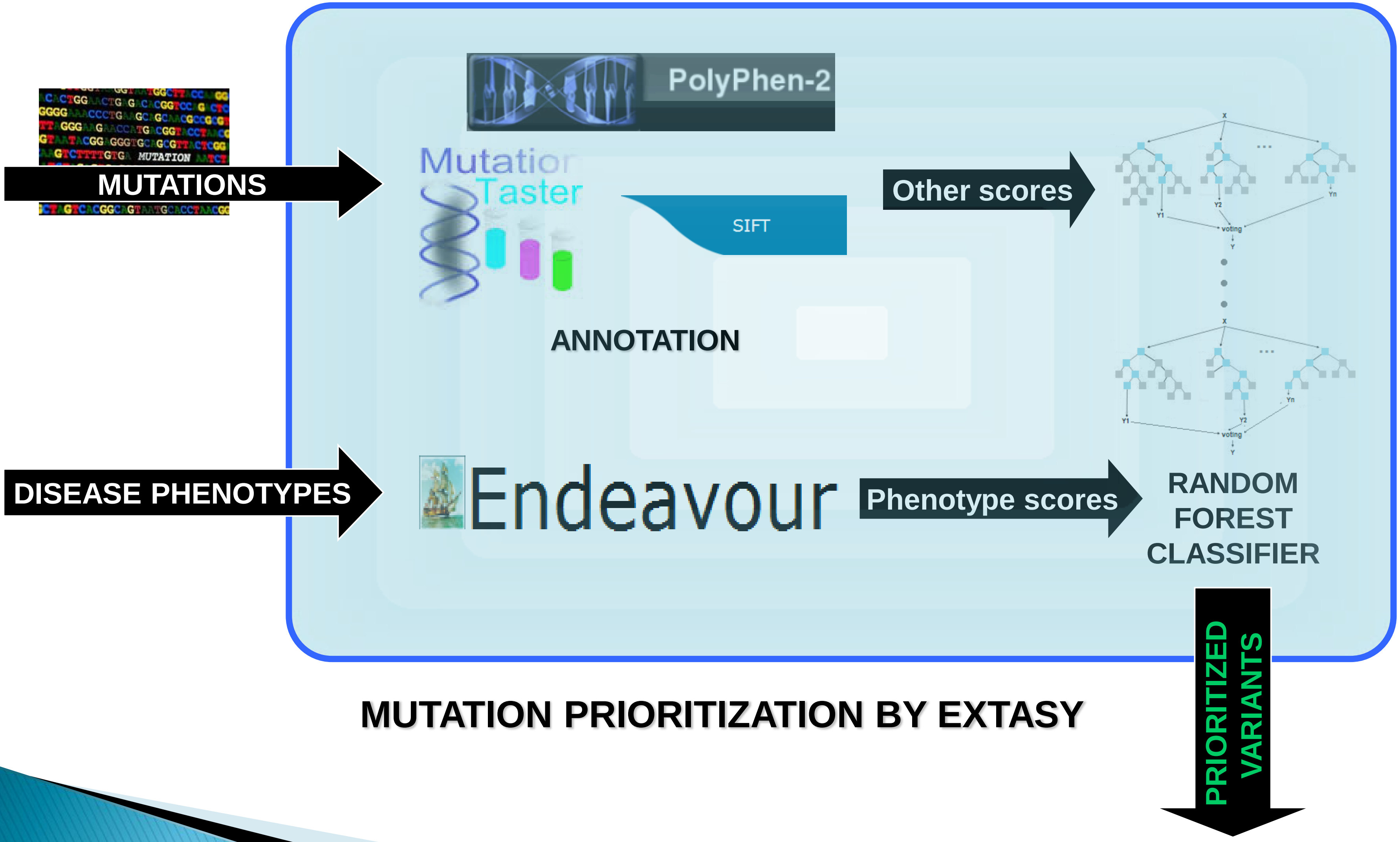
$t(Xi) = P(Xi)$

$A(X) = F(X)$

P – p-value of Gamma distribution

F – Fisher’s omnibus statistics


Low (0%-33%)
Moderate (33%-66%)
High (66%-100%)
Probability X <= x
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Rank
Aggregation Method
Average eXtasy Score
Maximum eXtasy Score
Nonparametric Order Statistics: RRA
Nonparametric Order Statistics: Stuart et al.
Parametric Modelling



MUTATIONS

DISEASE PHENOTYPES

ANNOTATION

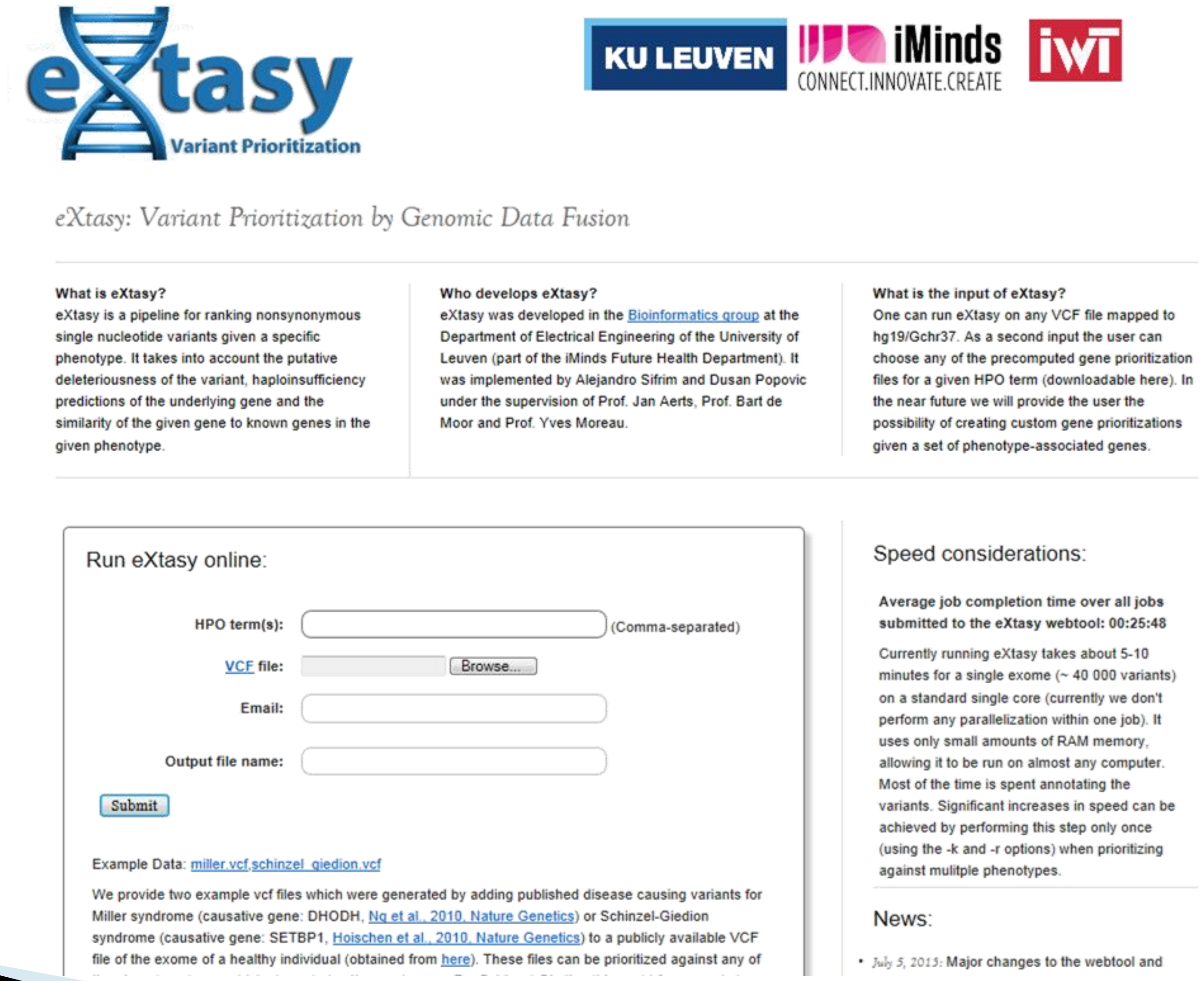
Other scores

Phenotype scores

RANDOM FOREST CLASSIFIER

PRIORITIZED VARIANTS

MUTATION PRIORITIZATION BY EXTASY



eXtasy: Variant Prioritization by Genomic Data Fusion

What is eXtasy?
eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXtasy?
eXtasy was developed in the Bioinformatics group at the Department of Electrical Engineering of the University of Leuven (part of the iMinds Future Health Department). It was implemented by Alejandro Sifrim and Dusan Popovic under the supervision of Prof. Jan Aerts, Prof. Bart de Moor and Prof. Yves Moreau.

What is the input of eXtasy?
One can run eXtasy on any VCF file mapped to hg19/Gchr37. As a second input the user can choose any of the precomputed gene prioritization files for a given HPO term (downloadable here). In the near future we will provide the user the possibility of creating custom gene prioritizations given a set of phenotype-associated genes.

Run eXtasy online:

HPO term(s): (Comma-separated)

VCF file:

Email:

Output file name:

Example Data: miller.vcf,schinzel_giedion.vcf

We provide two example vcf files which were generated by adding published disease causing variants for Miller syndrome (causative gene: DHODH, Ng et al., 2010, Nature Genetics) or Schinzel-Giedion syndrome (causative gene: SETBP1, Hoischen et al., 2010, Nature Genetics) to a publicly available VCF file of the exome of a healthy individual (obtained from here). These files can be prioritized against any of

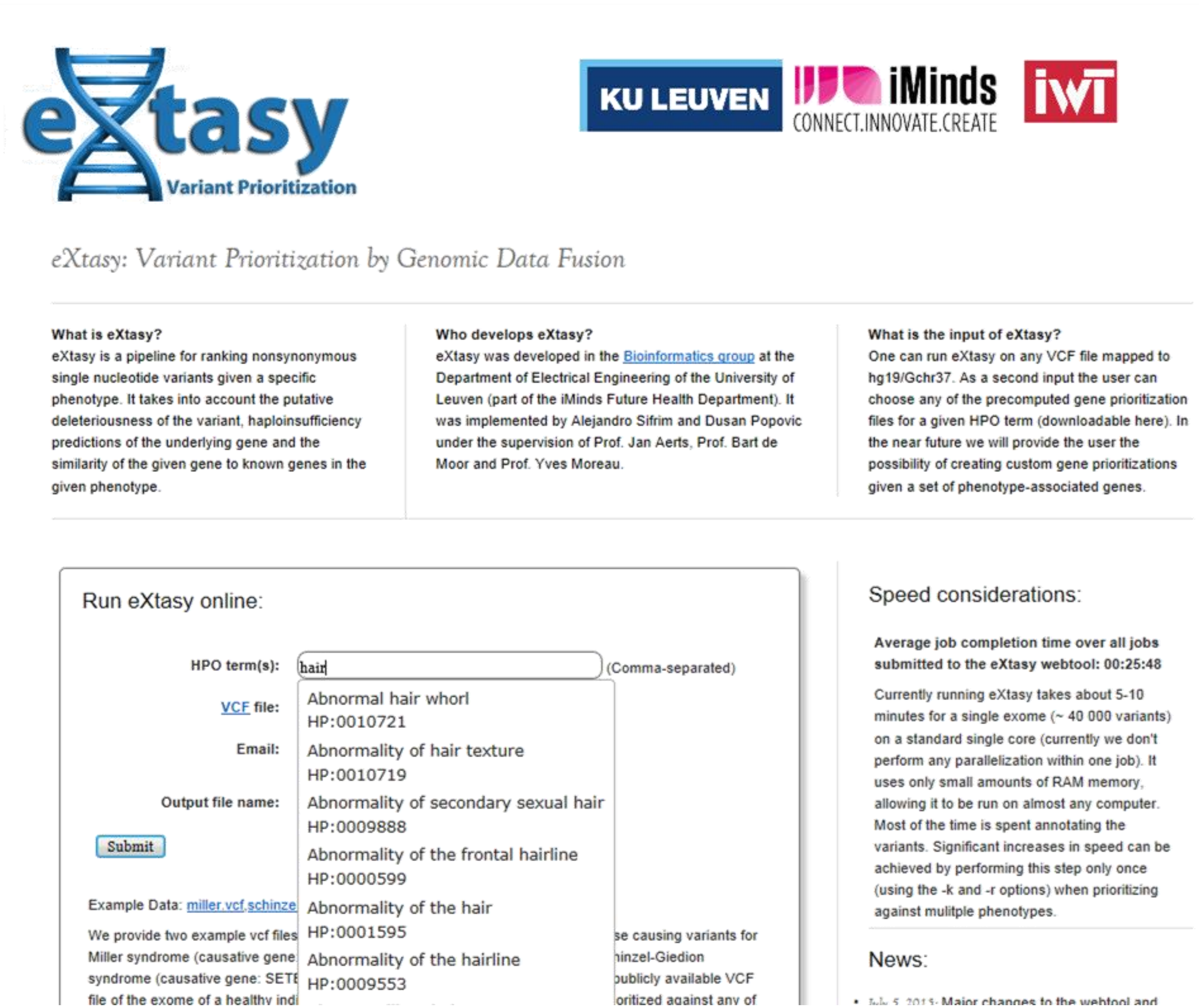
Speed considerations:

Average job completion time over all jobs submitted to the eXtasy webtool: 00:25:48

Currently running eXtasy takes about 5-10 minutes for a single exome (~ 40 000 variants) on a standard single core (currently we don't perform any parallelization within one job). It uses only small amounts of RAM memory, allowing it to be run on almost any computer. Most of the time is spent annotating the variants. Significant increases in speed can be achieved by performing this step only once (using the -k and -r options) when prioritizing against mulitple phenotypes.

News:

- *July 5, 2013:* Major changes to the webtool and



eXtasy: Variant Prioritization by Genomic Data Fusion

What is eXtasy?
eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXtasy?
eXtasy was developed in the Bioinformatics group at the Department of Electrical Engineering of the University of Leuven (part of the iMinds Future Health Department). It was implemented by Alejandro Sifrim and Dusan Popovic under the supervision of Prof. Jan Aerts, Prof. Bart de Moor and Prof. Yves Moreau.

What is the input of eXtasy?
One can run eXtasy on any VCF file mapped to hg19/Gchr37. As a second input the user can choose any of the precomputed gene prioritization files for a given HPO term (downloadable here). In the near future we will provide the user the possibility of creating custom gene prioritizations given a set of phenotype-associated genes.

Run eXtasy online:

HPO term(s): hair (Comma-separated)

Abnormal hair whorl
HP:0010721
Abnormality of hair texture
HP:0010719
Abnormality of secondary sexual hair
HP:0009888
Abnormality of the frontal hairline
HP:0000599
Abnormality of the hair
HP:0001595
Abnormality of the hairline
HP:0009553

VCF file:

Email:

Output file name:

Submit

Example Data: miller.vcf,schinze

We provide two example vcf files ... se causing variants for Miller syndrome (causative gene ... hinzel-Giedion syndrome (causative gene: SETE ... publicly available VCF file of the exome of a healthy indi ... oritized against any of

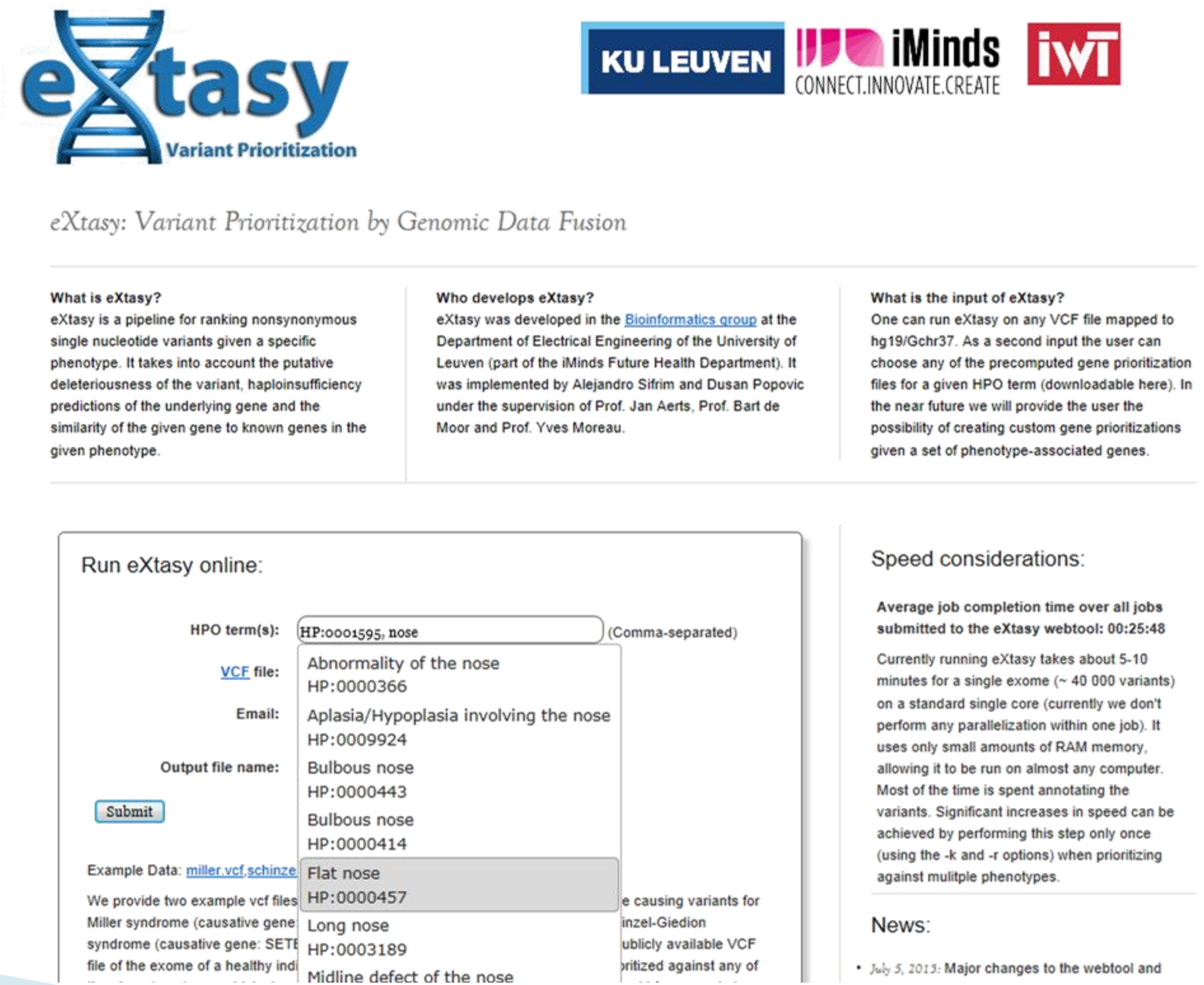
Speed considerations:

Average job completion time over all jobs submitted to the eXtasy webtool: 00:25:48

Currently running eXtasy takes about 5-10 minutes for a single exome (~ 40 000 variants) on a standard single core (currently we don't perform any parallelization within one job). It uses only small amounts of RAM memory, allowing it to be run on almost any computer. Most of the time is spent annotating the variants. Significant increases in speed can be achieved by performing this step only once (using the -k and -r options) when prioritizing against mulitple phenotypes.

News:

- July 5, 2013: Major changes to the webtool and



eXtasy: Variant Prioritization by Genomic Data Fusion

What is eXtasy?

eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXtasy?

eXtasy was developed in the Bioinformatics group at the Department of Electrical Engineering of the University of Leuven (part of the iMinds Future Health Department). It was implemented by Alejandro Sifrim and Dusan Popovic under the supervision of Prof. Jan Aerts, Prof. Bart de Moor and Prof. Yves Moreau.

What is the input of eXtasy?

One can run eXtasy on any VCF file mapped to hg19/Gchr37. As a second input the user can choose any of the precomputed gene prioritization files for a given HPO term (downloadable here). In the near future we will provide the user the possibility of creating custom gene prioritizations given a set of phenotype-associated genes.

Run eXtasy online:

HPO term(s): HP:0001595, nose (Comma-separated)

VCF file:

Email:

Output file name:

Submit

Abnormality of the nose
HP:0000366
Aplasia/Hypoplasia involving the nose
HP:0009924
Bulbous nose
HP:0000443
Bulbous nose
HP:0000414
Flat nose
HP:0000457
Long nose
HP:0003189
Midline defect of the nose

Example Data: miller.vcf,schinze

We provide two example vcf files … e causing variants for Miller syndrome (causative gene … inzel-Giedion syndrome (causative gene: SETE … ublicly available VCF file of the exome of a healthy indi … oritized against any of

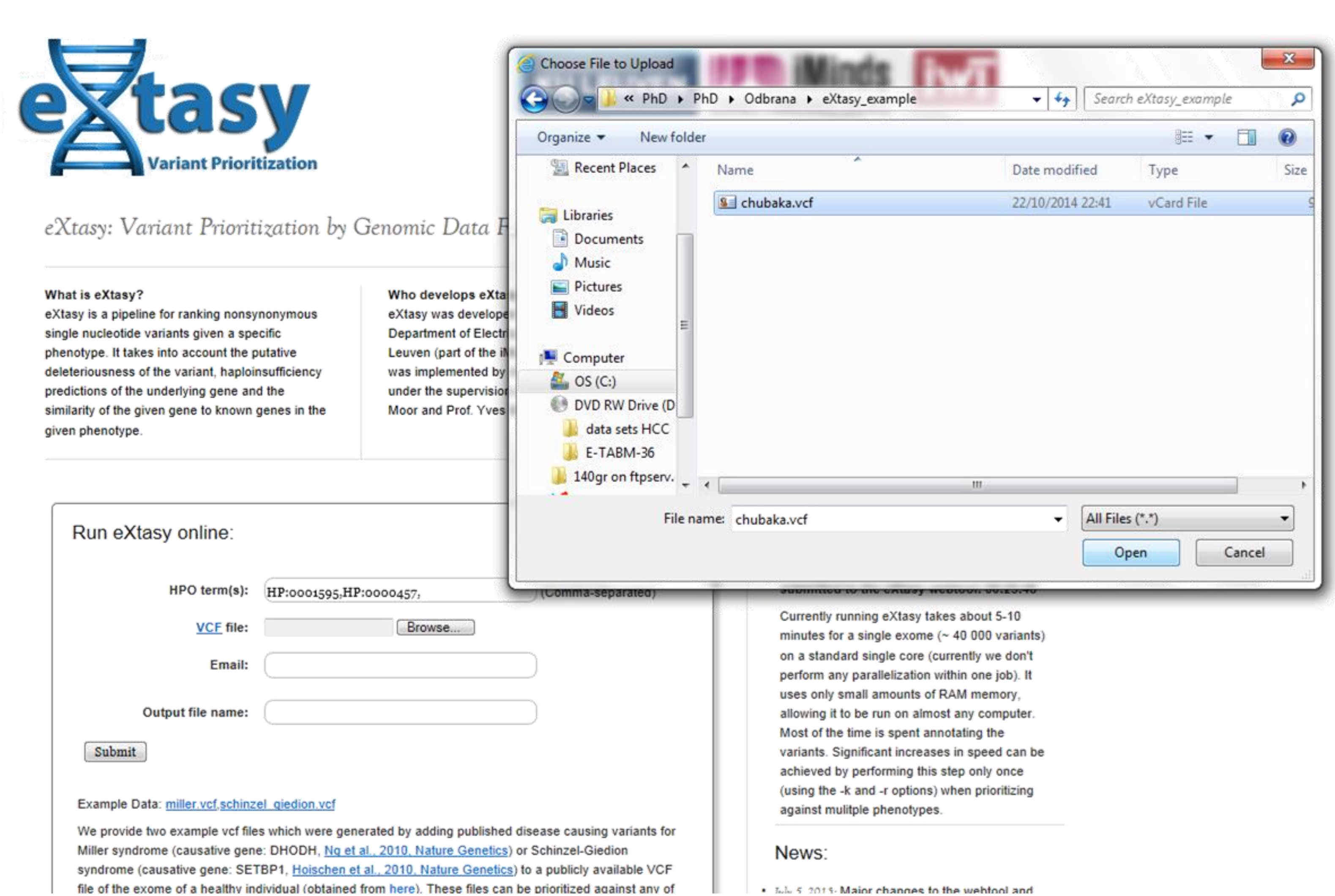
Speed considerations:

Average job completion time over all jobs submitted to the eXtasy webtool: 00:25:48

Currently running eXtasy takes about 5-10 minutes for a single exome (~ 40 000 variants) on a standard single core (currently we don't perform any parallelization within one job). It uses only small amounts of RAM memory, allowing it to be run on almost any computer. Most of the time is spent annotating the variants. Significant increases in speed can be achieved by performing this step only once (using the -k and -r options) when prioritizing against mulitple phenotypes.

News:

- *July 5, 2013:* Major changes to the webtool and



eXtasy Variant Prioritization

eXtasy: Variant Prioritization by Genomic Data F

What is eXtasy?
eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXta
eXtasy was develope
Department of Electr
Leuven (part of the i
was implemented by
under the supervisio
Moor and Prof. Yves

Run eXtasy online:

HPO term(s):	HP:0001595,HP:0000457, (Comma-separated)
VCF file:	Browse...
Email:	
Output file name:	

Submit

Example Data: miller.vcf,schinzel_giedion.vcf

We provide two example vcf files which were generated by adding published disease causing variants for Miller syndrome (causative gene: DHODH, Ng et al., 2010, Nature Genetics) or Schinzel-Giedion syndrome (causative gene: SETBP1, Hoischen et al., 2010, Nature Genetics) to a publicly available VCF file of the exome of a healthy individual (obtained from here). These files can be prioritized against any of

Currently running eXtasy takes about 5-10 minutes for a single exome (~ 40 000 variants) on a standard single core (currently we don't perform any parallelization within one job). It uses only small amounts of RAM memory, allowing it to be run on almost any computer. Most of the time is spent annotating the variants. Significant increases in speed can be achieved by performing this step only once (using the -k and -r options) when prioritizing against mulitple phenotypes.

News:

- July 5, 2013: Major changes to the webtool and

Choose File to Upload

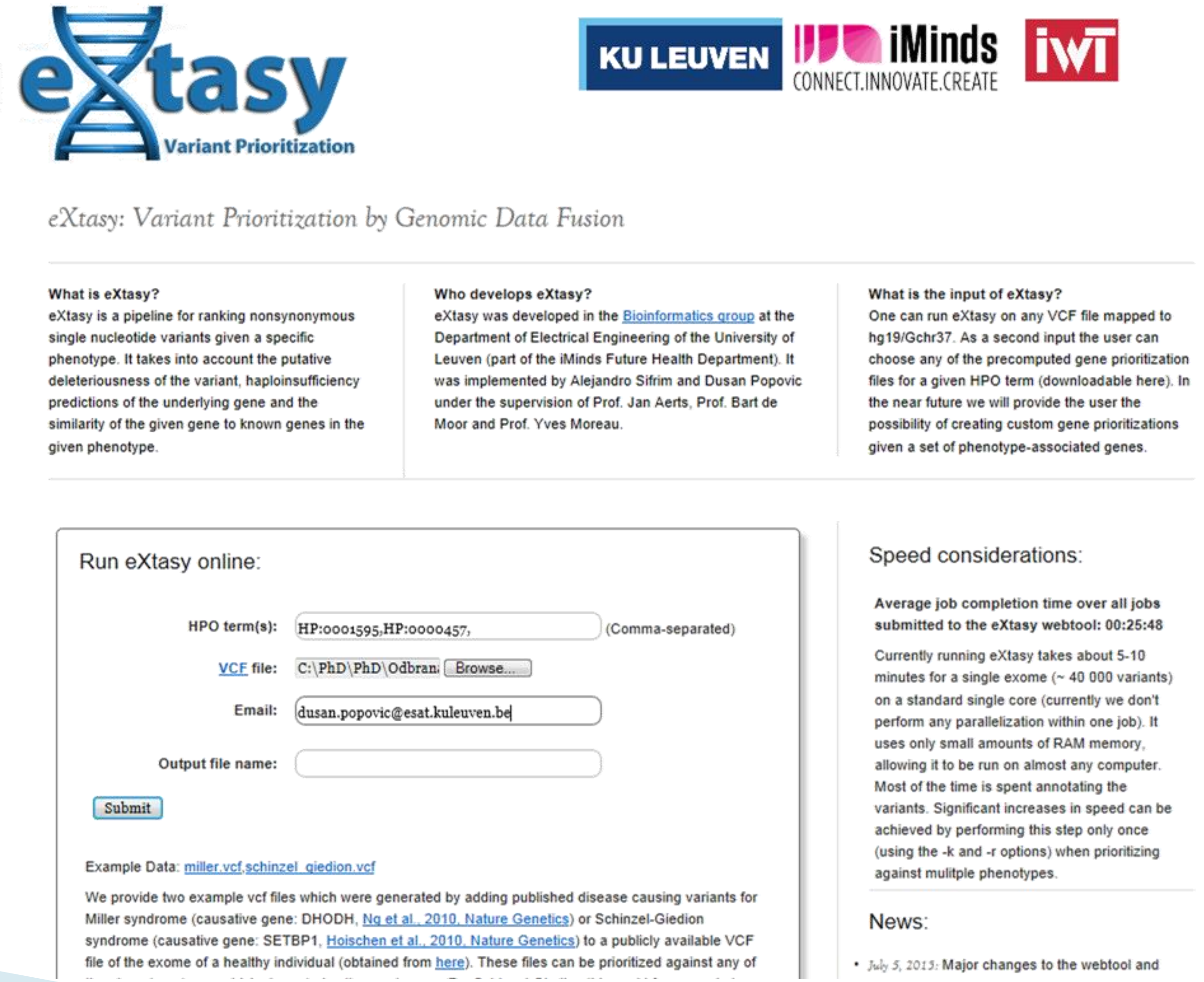
« PhD ▸ PhD ▸ Odbrana ▸ eXtasy_example

Search eXtasy_example

Name	Date modified	Type	Size
chubaka.vcf	22/10/2014 22:41	vCard File	

File name: chubaka.vcf

All Files (*.*)



eXtasy: Variant Prioritization by Genomic Data Fusion

What is eXtasy?
eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXtasy?
eXtasy was developed in the Bioinformatics group at the Department of Electrical Engineering of the University of Leuven (part of the iMinds Future Health Department). It was implemented by Alejandro Sifrim and Dusan Popovic under the supervision of Prof. Jan Aerts, Prof. Bart de Moor and Prof. Yves Moreau.

What is the input of eXtasy?
One can run eXtasy on any VCF file mapped to hg19/Gchr37. As a second input the user can choose any of the precomputed gene prioritization files for a given HPO term (downloadable here). In the near future we will provide the user the possibility of creating custom gene prioritizations given a set of phenotype-associated genes.

Run eXtasy online:

HPO term(s): HP:0001595,HP:0000457, (Comma-separated)

VCF file: C:\PhD\PhD\Odbran: Browse...

Email: dusan.popovic@esat.kuleuven.be

Output file name:

Submit

Example Data: miller.vcf,schinzel_giedion.vcf

We provide two example vcf files which were generated by adding published disease causing variants for Miller syndrome (causative gene: DHODH, Ng et al., 2010, Nature Genetics) or Schinzel-Giedion syndrome (causative gene: SETBP1, Hoischen et al., 2010, Nature Genetics) to a publicly available VCF file of the exome of a healthy individual (obtained from here). These files can be prioritized against any of

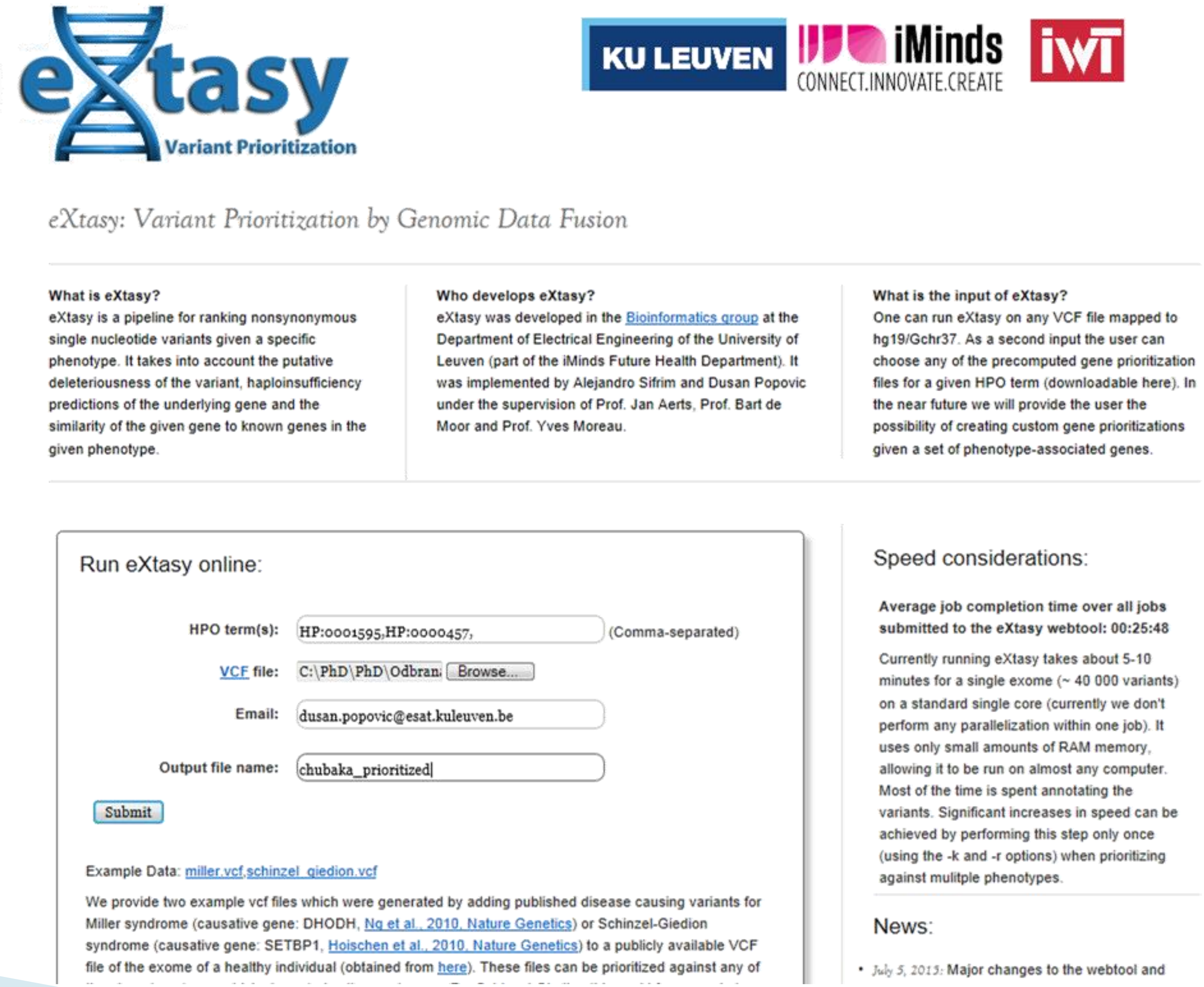
Speed considerations:

Average job completion time over all jobs submitted to the eXtasy webtool: 00:25:48

Currently running eXtasy takes about 5-10 minutes for a single exome (~ 40 000 variants) on a standard single core (currently we don't perform any parallelization within one job). It uses only small amounts of RAM memory, allowing it to be run on almost any computer. Most of the time is spent annotating the variants. Significant increases in speed can be achieved by performing this step only once (using the -k and -r options) when prioritizing against mulitple phenotypes.

News:

- *July 5, 2015:* Major changes to the webtool and



eXtasy: Variant Prioritization by Genomic Data Fusion

What is eXtasy?
eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXtasy?
eXtasy was developed in the Bioinformatics group at the Department of Electrical Engineering of the University of Leuven (part of the iMinds Future Health Department). It was implemented by Alejandro Sifrim and Dusan Popovic under the supervision of Prof. Jan Aerts, Prof. Bart de Moor and Prof. Yves Moreau.

What is the input of eXtasy?
One can run eXtasy on any VCF file mapped to hg19/Gchr37. As a second input the user can choose any of the precomputed gene prioritization files for a given HPO term (downloadable here). In the near future we will provide the user the possibility of creating custom gene prioritizations given a set of phenotype-associated genes.

Run eXtasy online:

HPO term(s):	HP:0001595,HP:0000457,	(Comma-separated)
VCF file:	C:\PhD\PhD\Odbran: Browse...	
Email:	dusan.popovic@esat.kuleuven.be	
Output file name:	chubaka_prioritized	

Submit

Example Data: miller.vcf,schinzel_giedion.vcf

We provide two example vcf files which were generated by adding published disease causing variants for Miller syndrome (causative gene: DHODH, Ng et al., 2010, Nature Genetics) or Schinzel-Giedion syndrome (causative gene: SETBP1, Hoischen et al., 2010, Nature Genetics) to a publicly available VCF file of the exome of a healthy individual (obtained from here). These files can be prioritized against any of

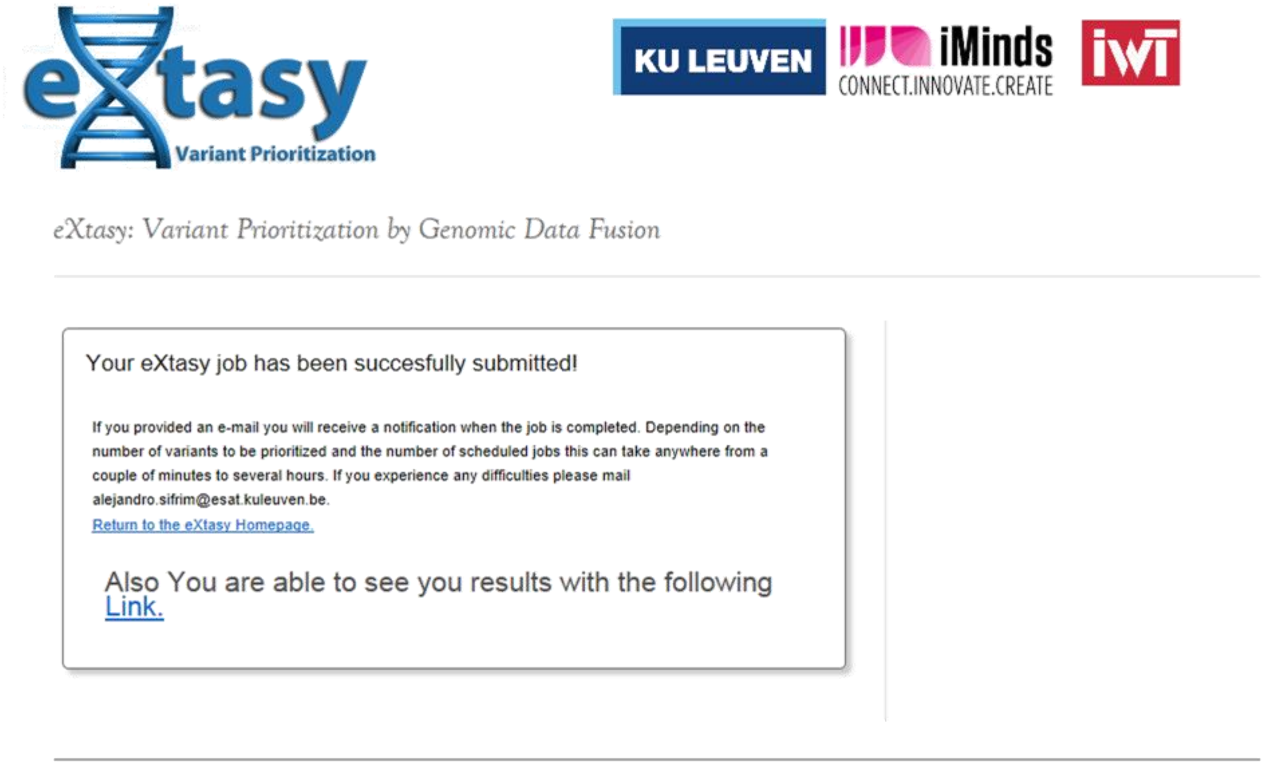
Speed considerations:

Average job completion time over all jobs submitted to the eXtasy webtool: 00:25:48

Currently running eXtasy takes about 5-10 minutes for a single exome (~ 40 000 variants) on a standard single core (currently we don't perform any parallelization within one job). It uses only small amounts of RAM memory, allowing it to be run on almost any computer. Most of the time is spent annotating the variants. Significant increases in speed can be achieved by performing this step only once (using the -k and -r options) when prioritizing against mulitple phenotypes.

News:

- *July 5, 2015:* Major changes to the webtool and



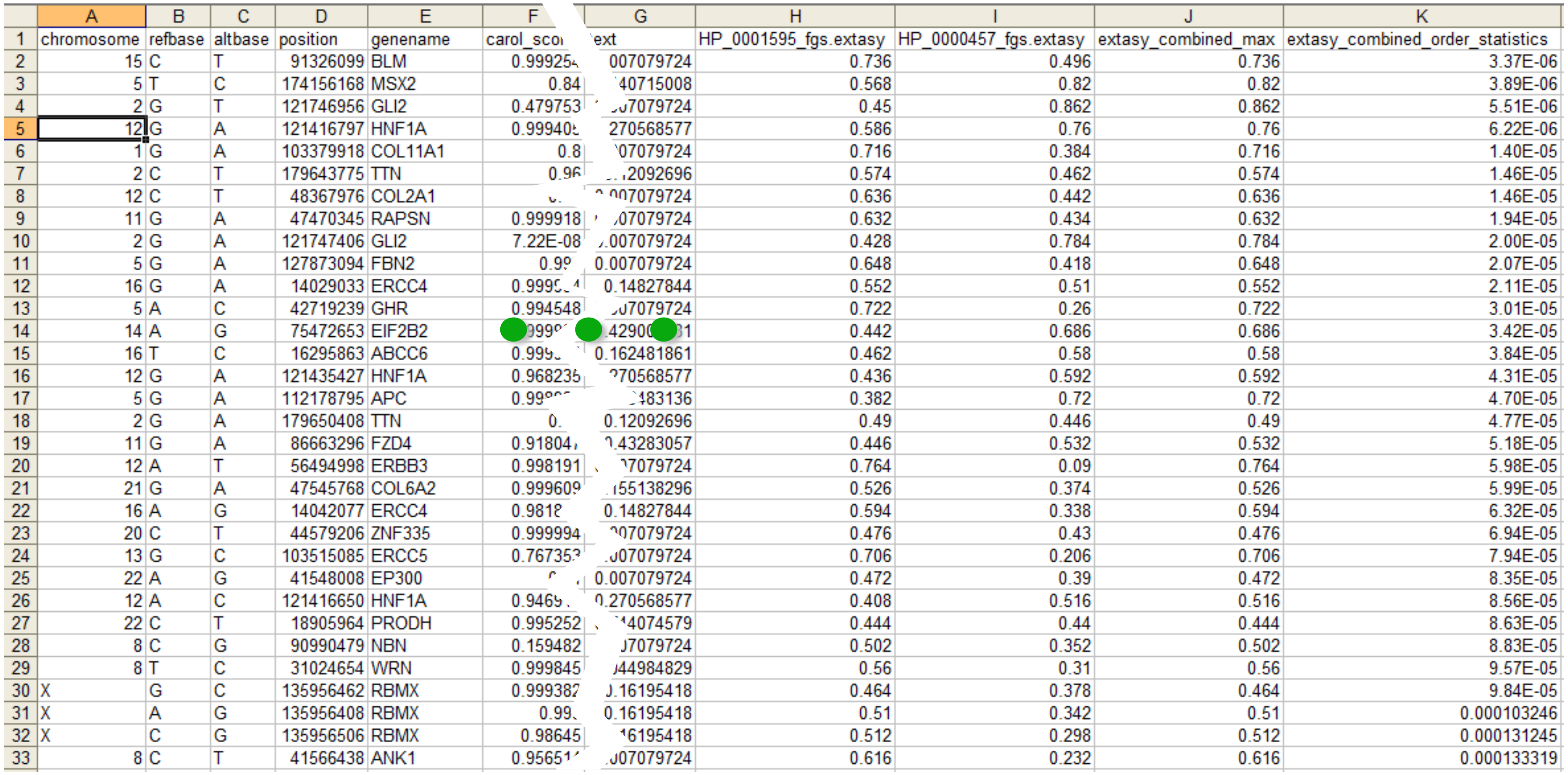
eXtasy: Variant Prioritization by Genomic Data Fusion

Your eXtasy job has been succesfully submitted!

If you provided an e-mail you will receive a notification when the job is completed. Depending on the number of variants to be prioritized and the number of scheduled jobs this can take anywhere from a couple of minutes to several hours. If you experience any difficulties please mail alejandro.sifrim@esat.kuleuven.be.

Return to the eXtasy Homepage.

Also You are able to see you results with the following Link.

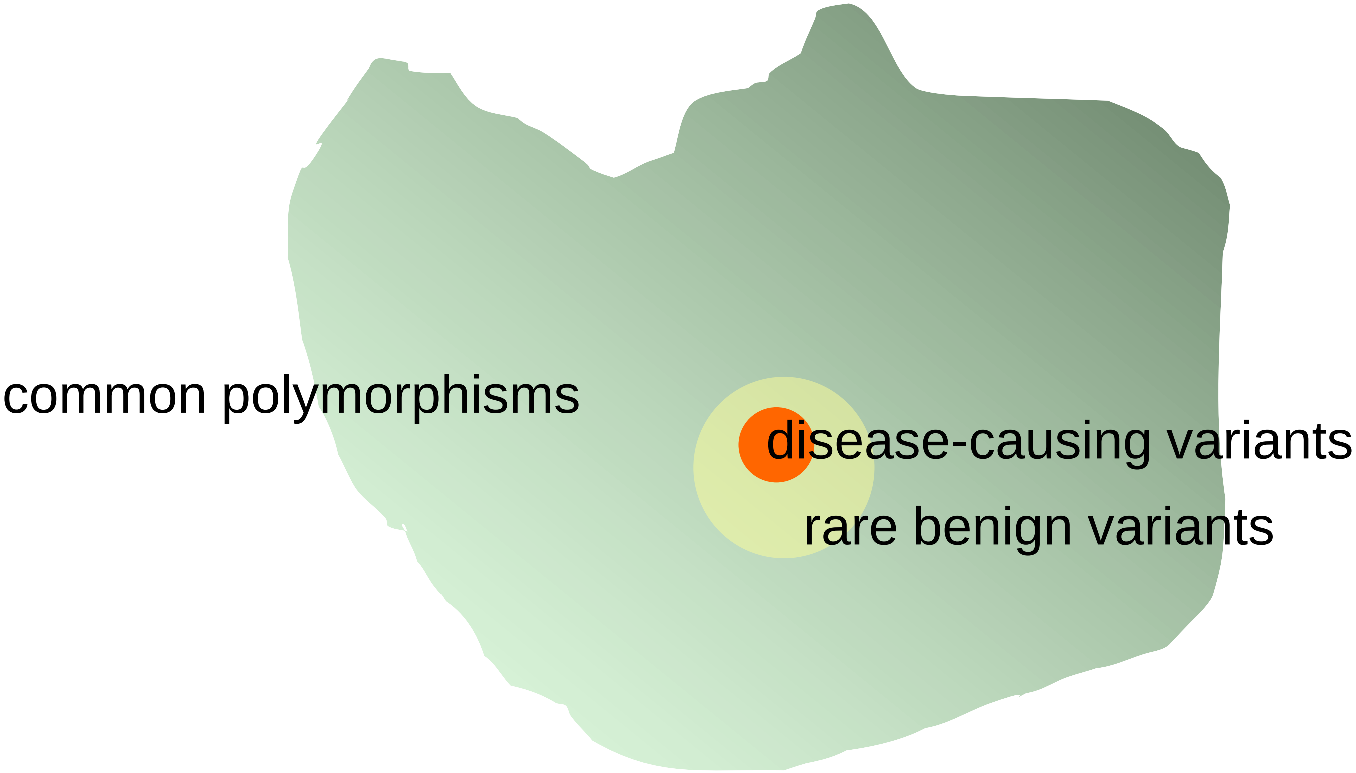


	A	B	C	D	E	F	G	H	I	J	K
1	chromosome	refbase	altbase	position	genename	carol_scor	ext	HP_0001595_fgs.extasy	HP_0000457_fgs.extasy	extasy_combined_max	extasy_combined_order_statistics
2	15	C	T	91326099	BLM	0.99925	007079724	0.736	0.496	0.736	3.37E-06
3	5	T	C	174156168	MSX2	0.84	40715008	0.568	0.82	0.82	3.89E-06
4	2	G	T	121746956	GLI2	0.479753	07079724	0.45	0.862	0.862	5.51E-06
5	12	G	A	121416797	HNF1A	0.99940	270568577	0.586	0.76	0.76	6.22E-06
6	1	G	A	103379918	COL11A1	0.8	07079724	0.716	0.384	0.716	1.40E-05
7	2	C	T	179643775	TTN	0.96	12092696	0.574	0.462	0.574	1.46E-05
8	12	C	T	48367976	COL2A1	0.	0.007079724	0.636	0.442	0.636	1.46E-05
9	11	G	A	47470345	RAPSN	0.999918	07079724	0.632	0.434	0.632	1.94E-05
10	2	G	A	121747406	GLI2	7.22E-08	.007079724	0.428	0.784	0.784	2.00E-05
11	5	G	A	127873094	FBN2	0.99	0.007079724	0.648	0.418	0.648	2.07E-05
12	16	G	A	14029033	ERCC4	0.999	0.14827844	0.552	0.51	0.552	2.11E-05
13	5	A	C	42719239	GHR	0.994548	07079724	0.722	0.26	0.722	3.01E-05
14	14	A	G	75472653	EIF2B2	999	.42900 31	0.442	0.686	0.686	3.42E-05
15	16	T	C	16295863	ABCC6	0.999	0.162481861	0.462	0.58	0.58	3.84E-05
16	12	G	A	121435427	HNF1A	0.968235	270568577	0.436	0.592	0.592	4.31E-05
17	5	G	A	112178795	APC	0.99	483136	0.382	0.72	0.72	4.70E-05
18	2	G	A	179650408	TTN	0.	0.12092696	0.49	0.446	0.49	4.77E-05
19	11	G	A	86663296	FZD4	0.91804	0.43283057	0.446	0.532	0.532	5.18E-05
20	12	A	T	56494998	ERBB3	0.998191	07079724	0.764	0.09	0.764	5.98E-05
21	21	G	A	47545768	COL6A2	0.999609	155138296	0.526	0.374	0.526	5.99E-05
22	16	A	G	14042077	ERCC4	0.9818	0.14827844	0.594	0.338	0.594	6.32E-05
23	20	C	T	44579206	ZNF335	0.999994	007079724	0.476	0.43	0.476	6.94E-05
24	13	G	C	103515085	ERCC5	0.767353	.007079724	0.706	0.206	0.706	7.94E-05
25	22	A	G	41548008	EP300		0.007079724	0.472	0.39	0.472	8.35E-05
26	12	A	C	121416650	HNF1A	0.9469	0.270568577	0.408	0.516	0.516	8.56E-05
27	22	C	T	18905964	PRODH	0.995252	14074579	0.444	0.44	0.444	8.63E-05
28	8	C	G	90990479	NBN	0.159482	07079724	0.502	0.352	0.502	8.83E-05
29	8	T	C	31024654	WRN	0.999845	044984829	0.56	0.31	0.56	9.57E-05
30	X	G	C	135956462	RBMX	0.999382	0.16195418	0.464	0.378	0.464	9.84E-05
31	X	A	G	135956408	RBMX	0.99	0.16195418	0.51	0.342	0.51	0.000103246
32	X	C	G	135956506	RBMX	0.98645	16195418	0.512	0.298	0.512	0.000131245
33	8	C	T	41566438	ANK1	0.95651	007079724	0.616	0.232	0.616	0.000133319

WHY DOES EXTASY WORK SO WELL?

1. PROBLEM-TAILORED CHOICE OF THE TRAINING SET

1. PROBLEM-TAILORED CHOICE OF THE TRAINING SET

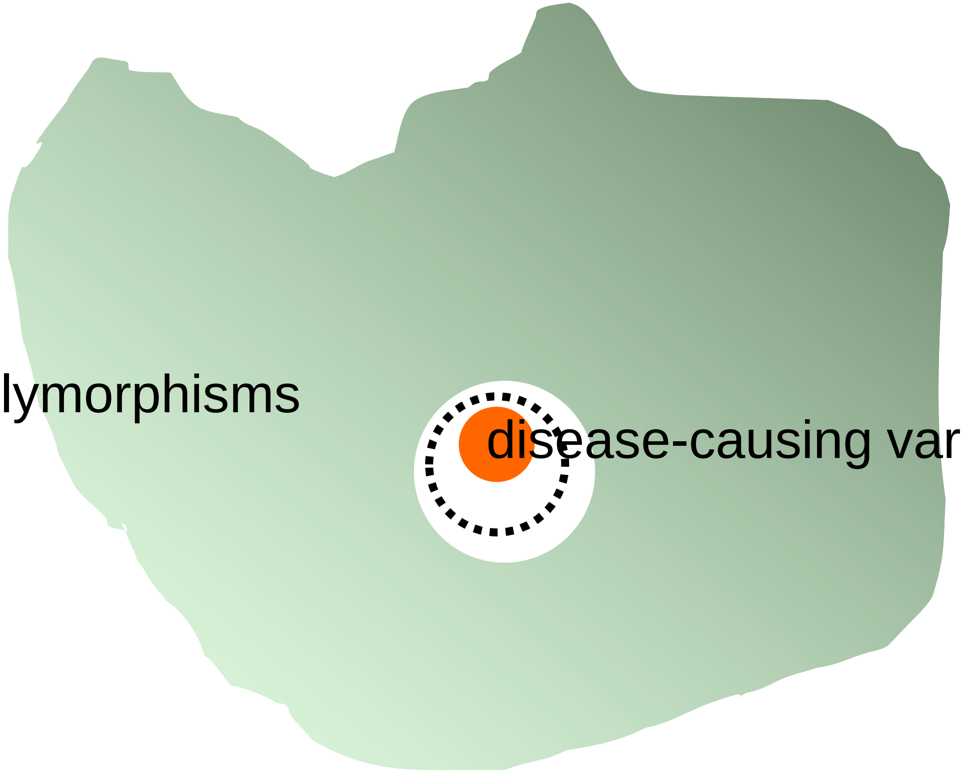


common polymorphisms

disease-causing variants

rare benign variants

1. PROBLEM-TAILORED CHOICE OF THE TRAINING SET



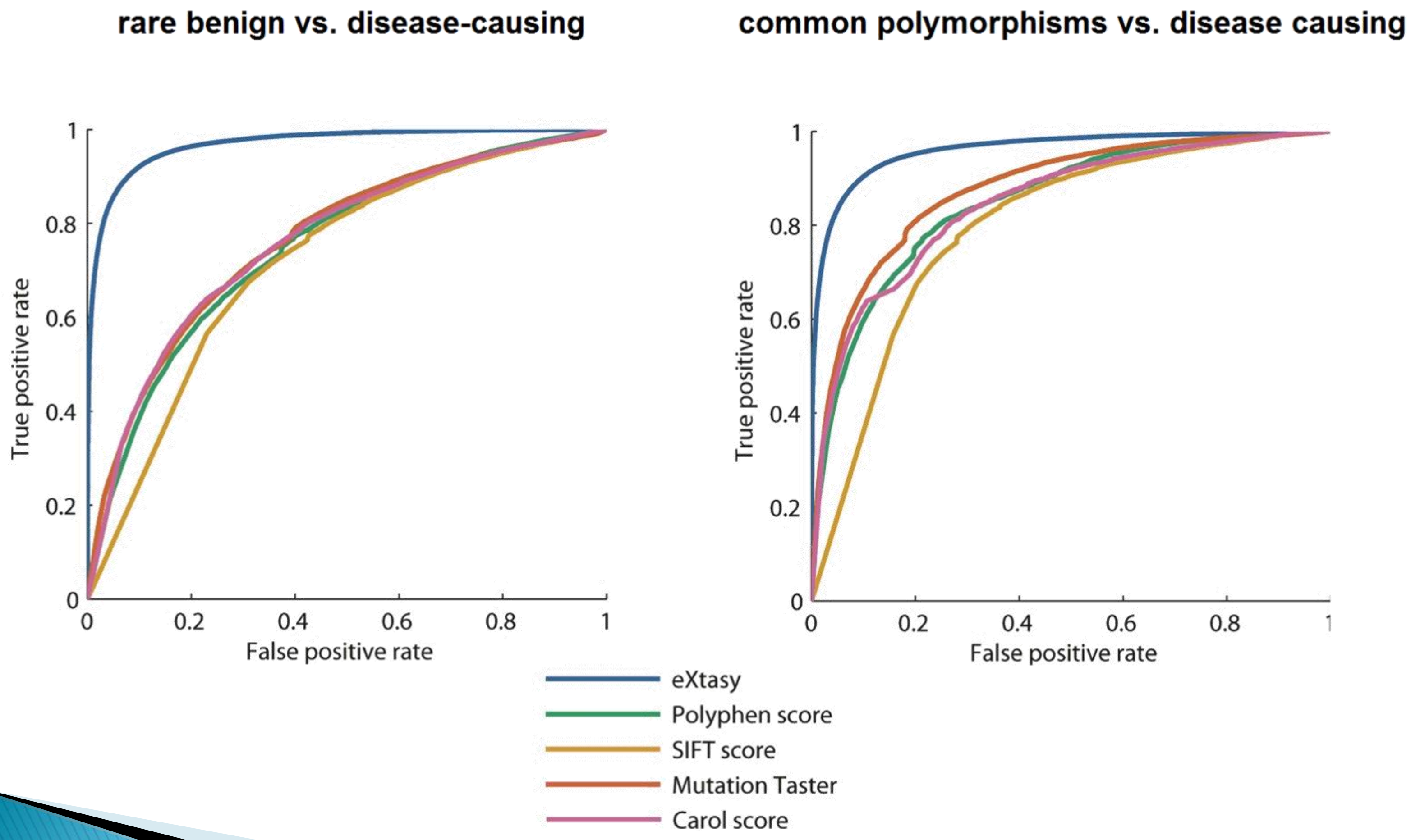
common polymorphisms

disease-causing variants

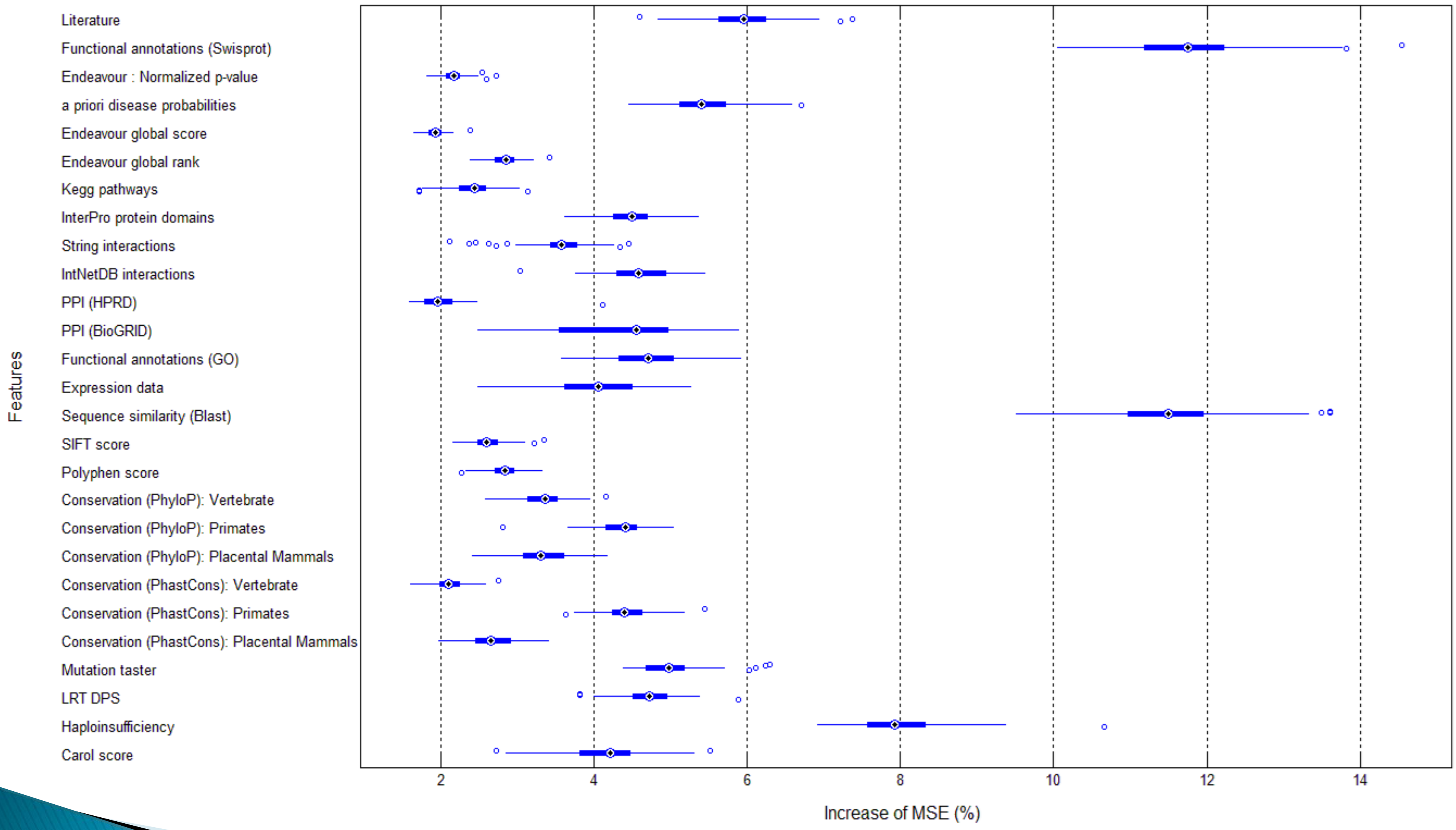
1. PROBLEM-TAILORED CHOICE OF THE TRAINING SET

disease-causing variants
rare benign variants

1. PROBLEM-TAILORED CHOICE OF THE TRAINING SET



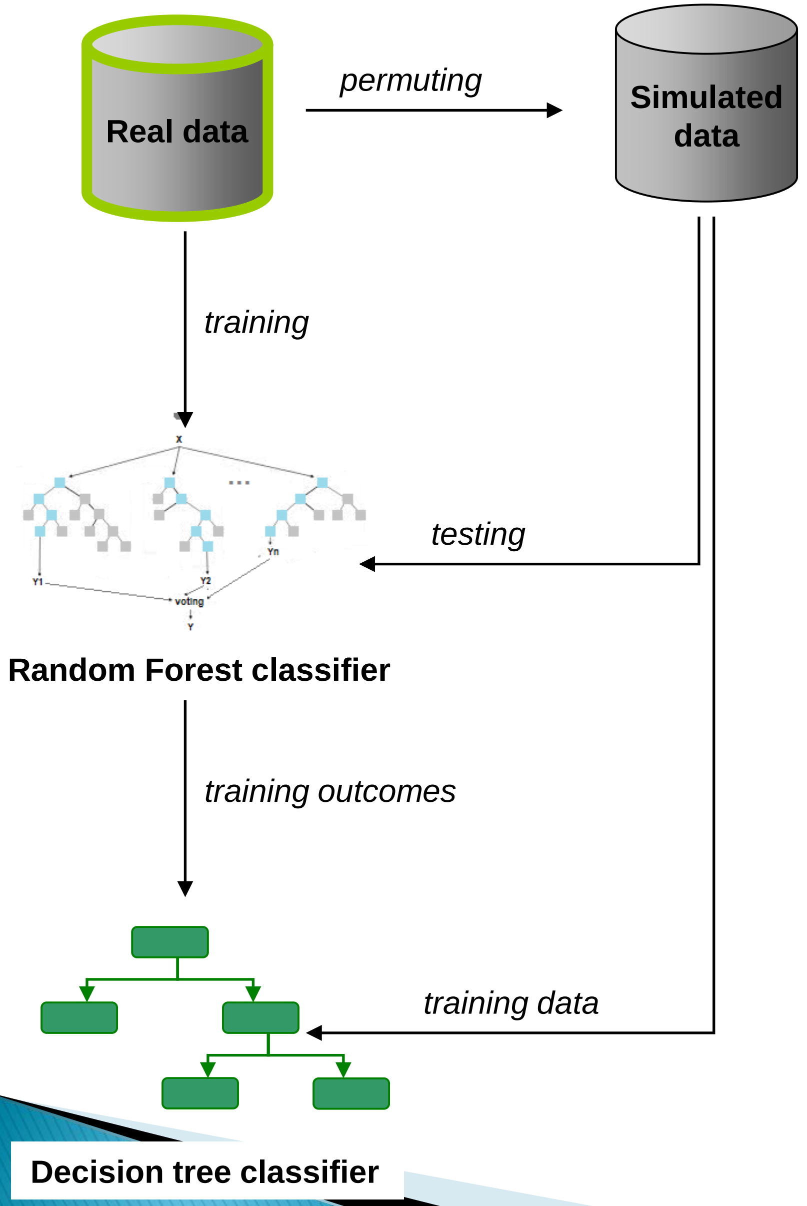
1. **PROBLEM-TAILORED CHOICE OF THE TRAINING SET**
2. **HETEROGENOUS DATA FUSION**


Literature
Functional annotations (Swisprot)
Endeavour : Normalized p-value
a priori disease probabilities
Endeavour global score
Endeavour global rank
Kegg pathways
InterPro protein domains
String interactions
IntNetDB interactions
PPI (HPRD)
PPI (BioGRID)
Functional annotations (GO)
Expression data
Sequence similarity (Blast)
SIFT score
Polyphen score
Conservation (PhyloP): Vertebrate
Conservation (PhyloP): Primates
Conservation (PhyloP): Placental Mammals
Conservation (PhastCons): Vertebrate
Conservation (PhastCons): Primates
Conservation (PhastCons): Placental Mammals
Mutation taster
LRT DPS
Haploinsufficiency
Carol score
Features
2
4
6
8
10
12
14
Increase of MSE (%)

1. PROBLEM-TAILORED CHOICE OF THE TRAINING SET

2. HETEROGENOUS DATA FUSION

3. PHENOTYPIC INFORMATION



Real data

permuting

Simulated data

training

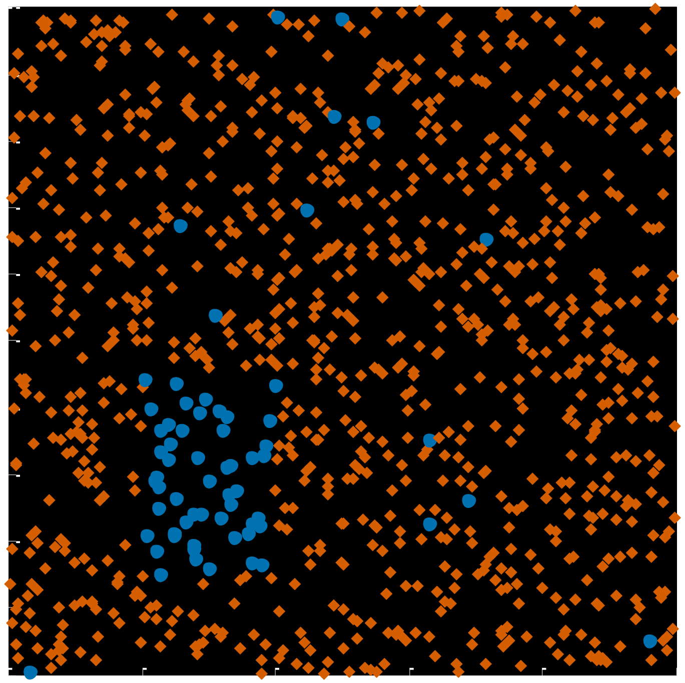
testing

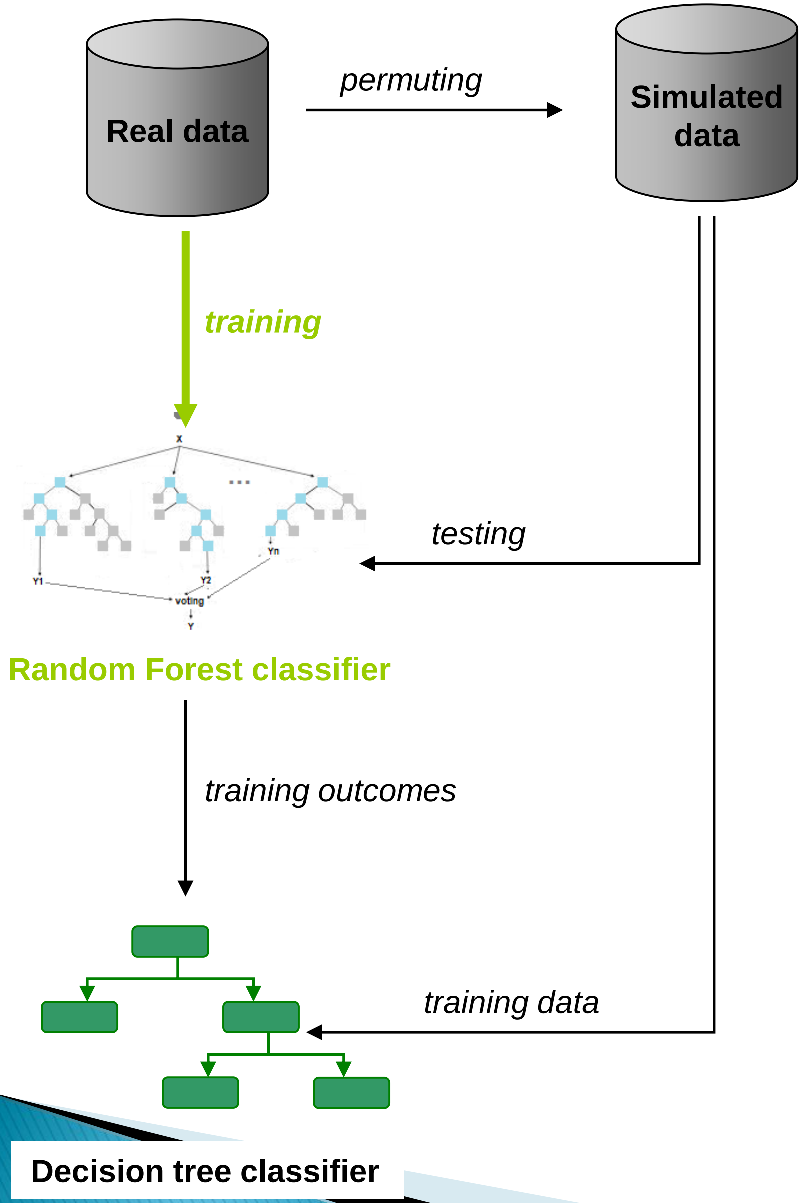
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

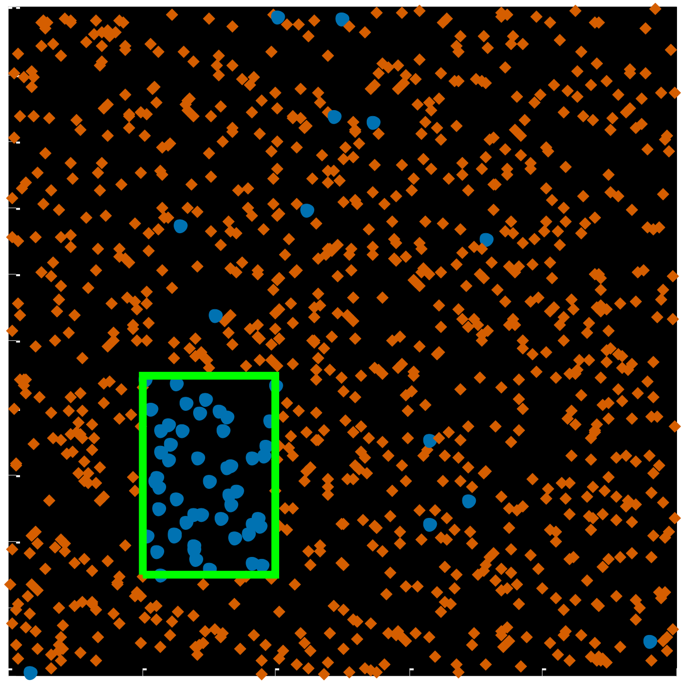
testing

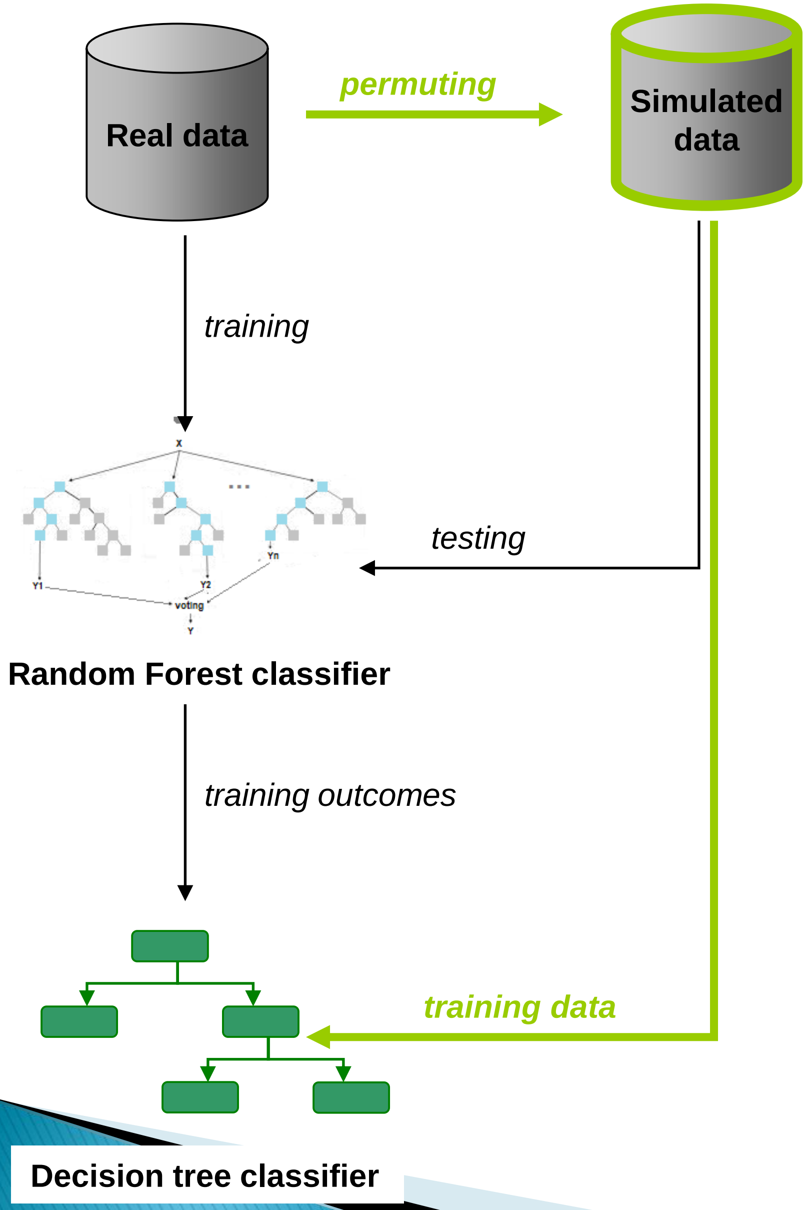
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

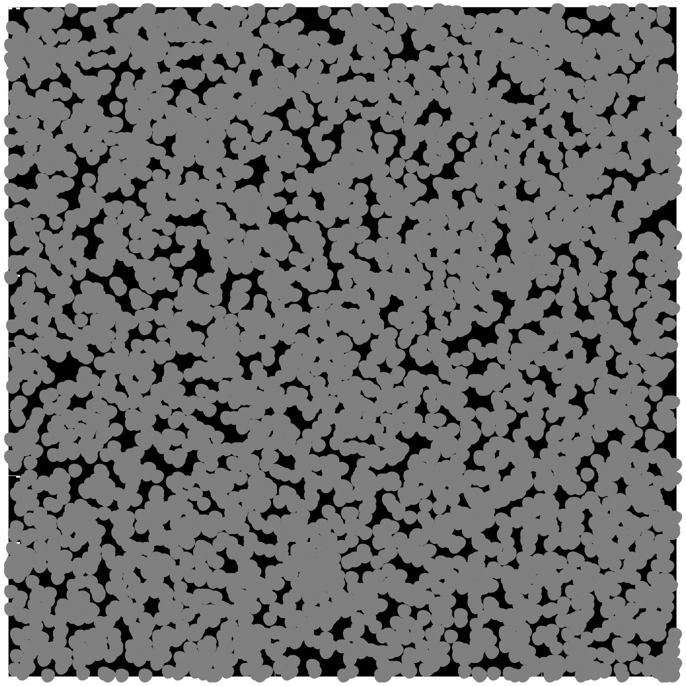
testing

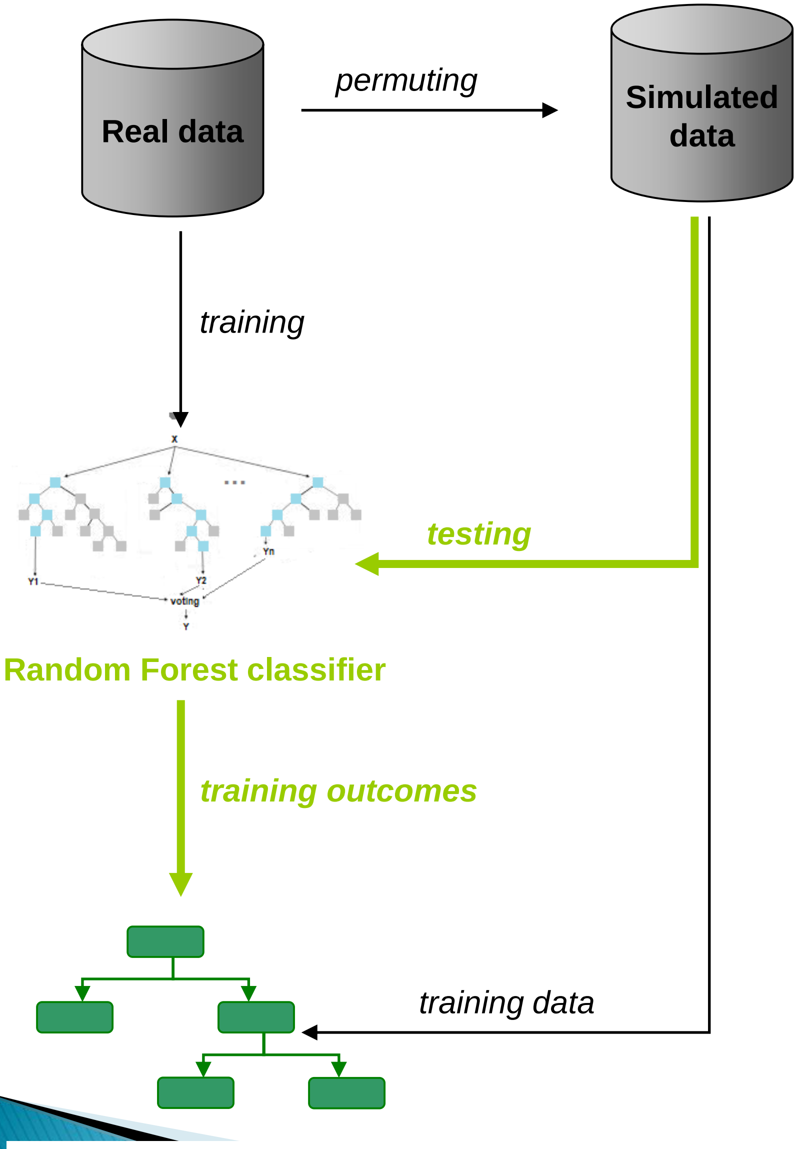
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

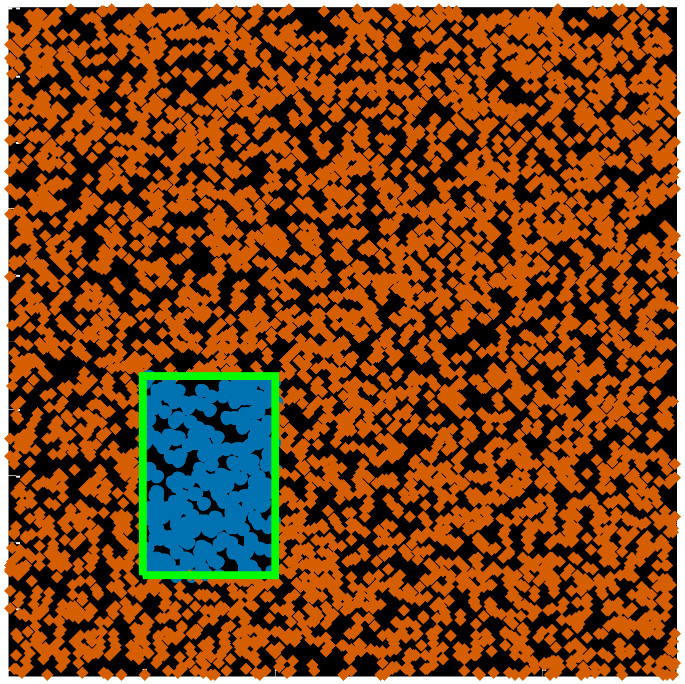
testing

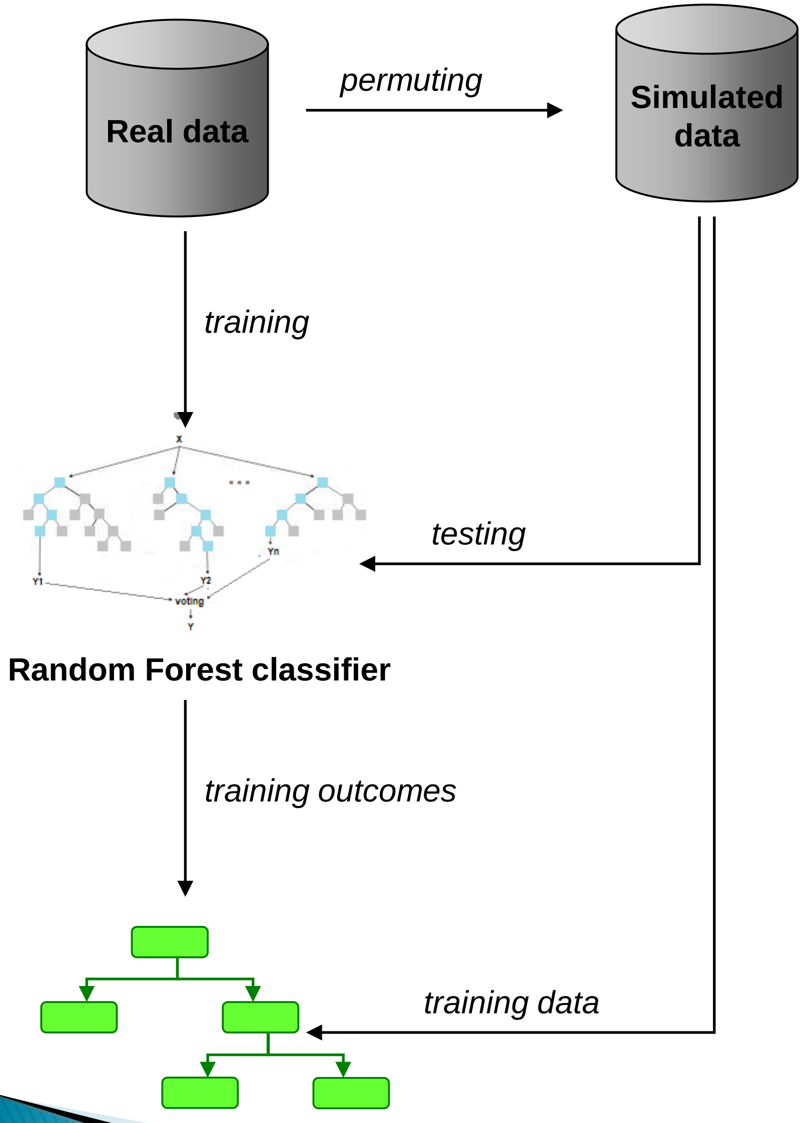
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

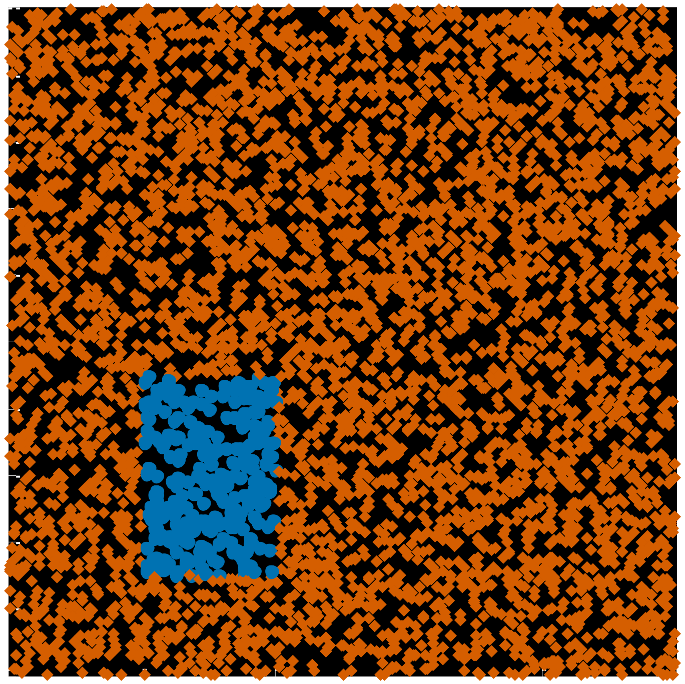
testing

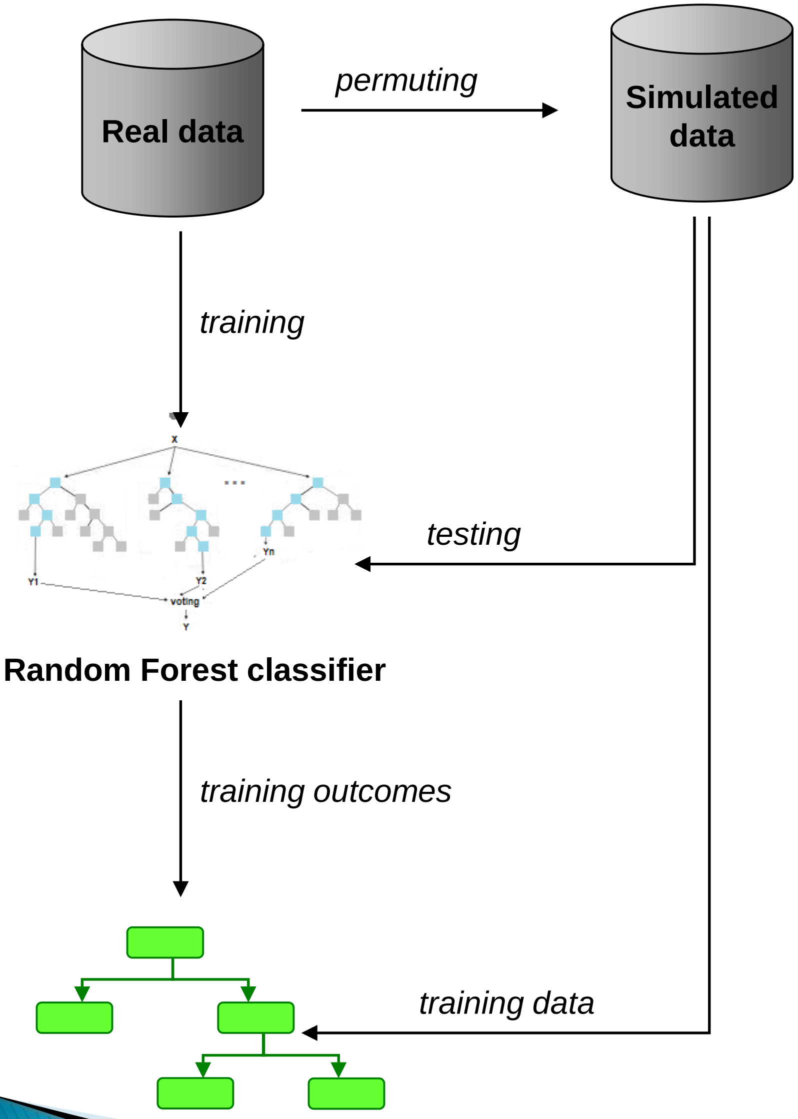
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

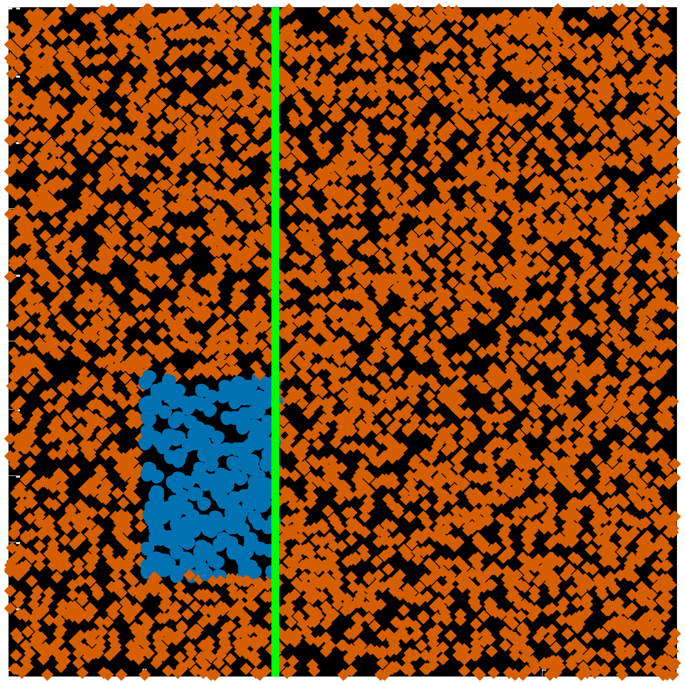
testing

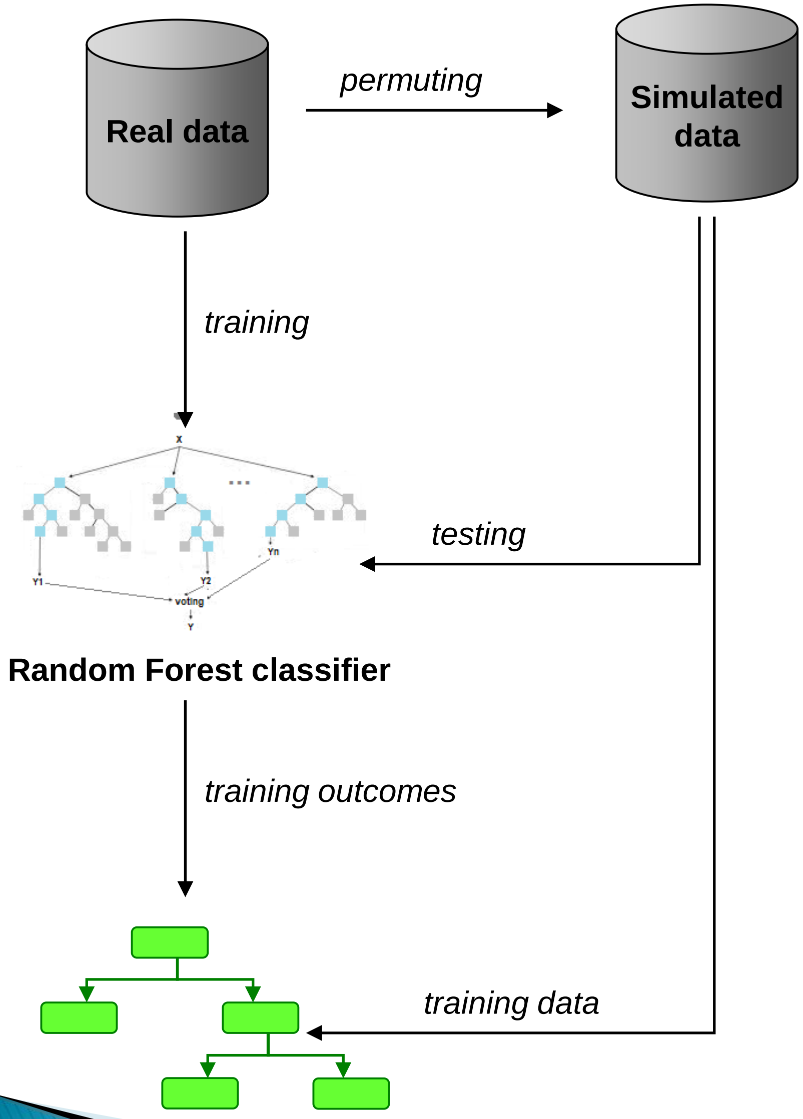
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

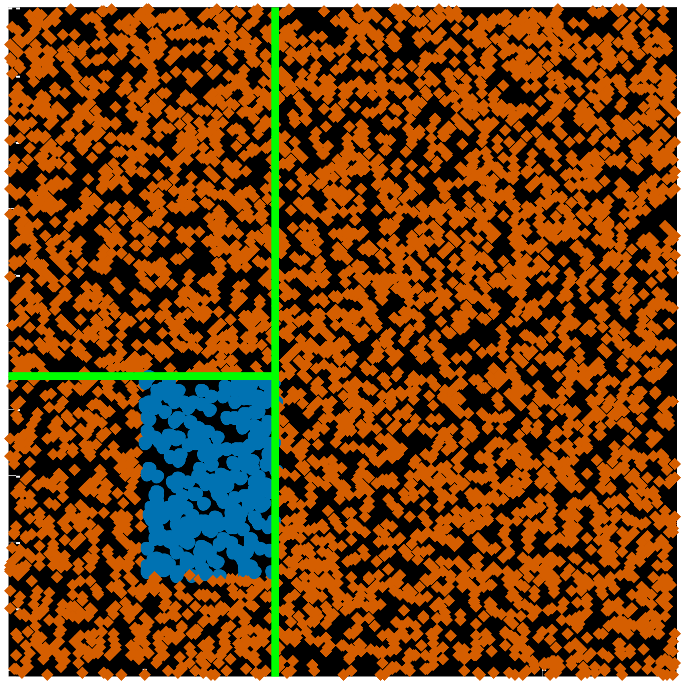
testing

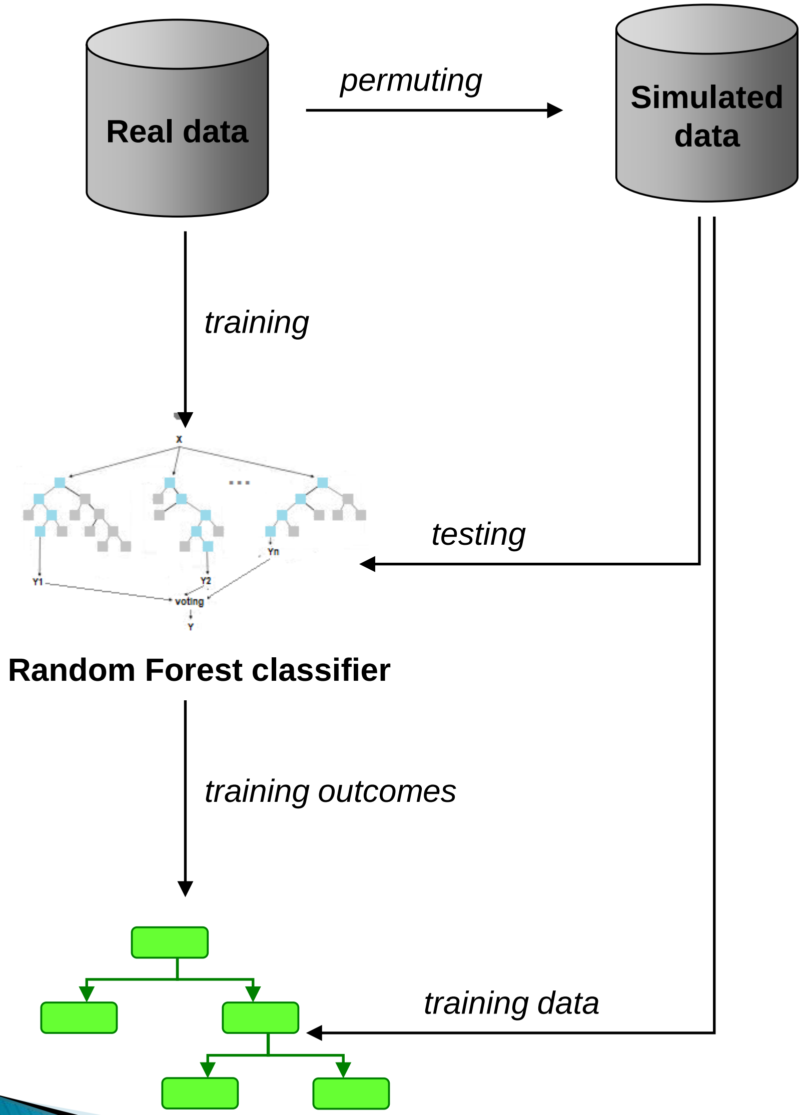
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

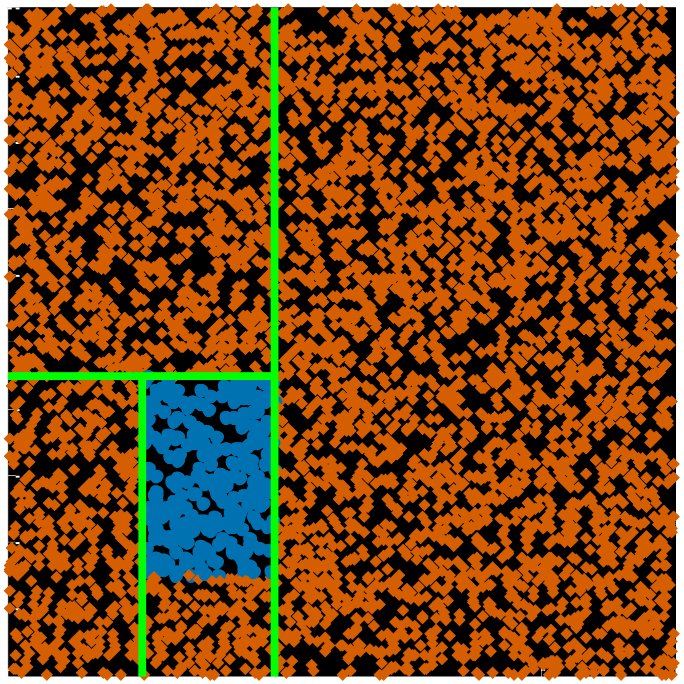
testing

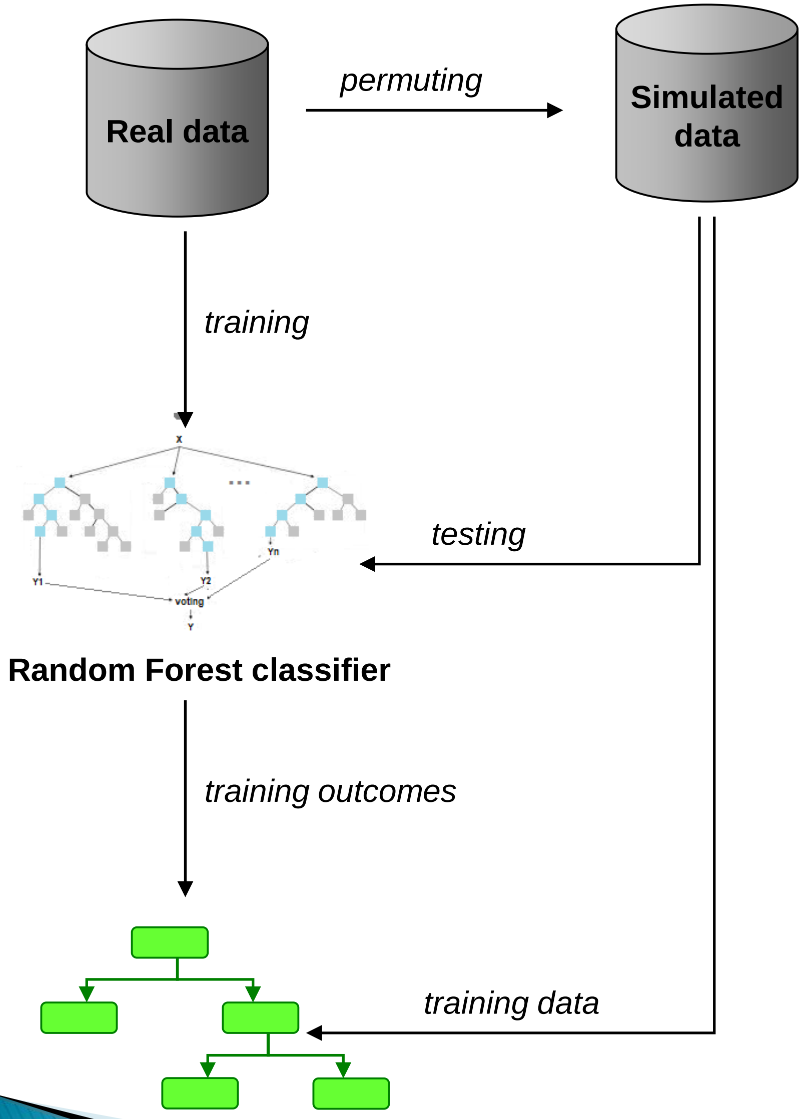
Random Forest classifier

training outcomes

training data

Decision tree classifier





Real data

permuting

Simulated data

training

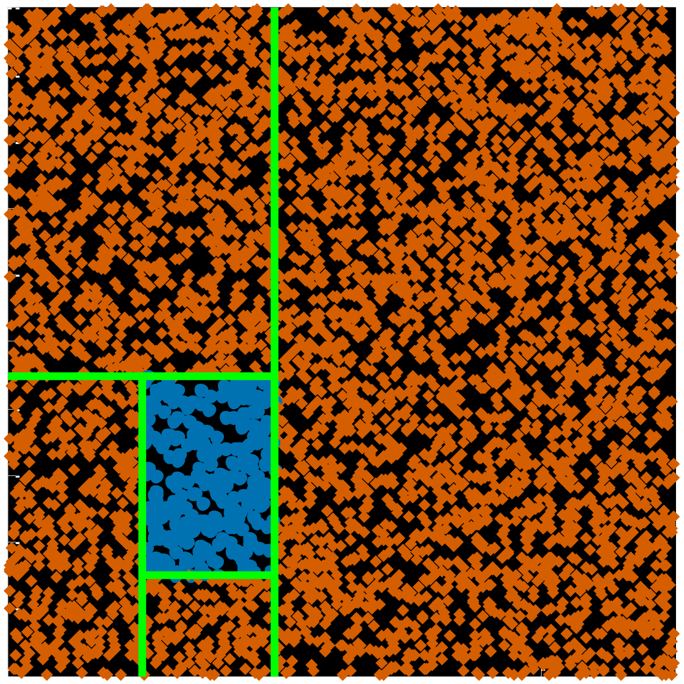
testing

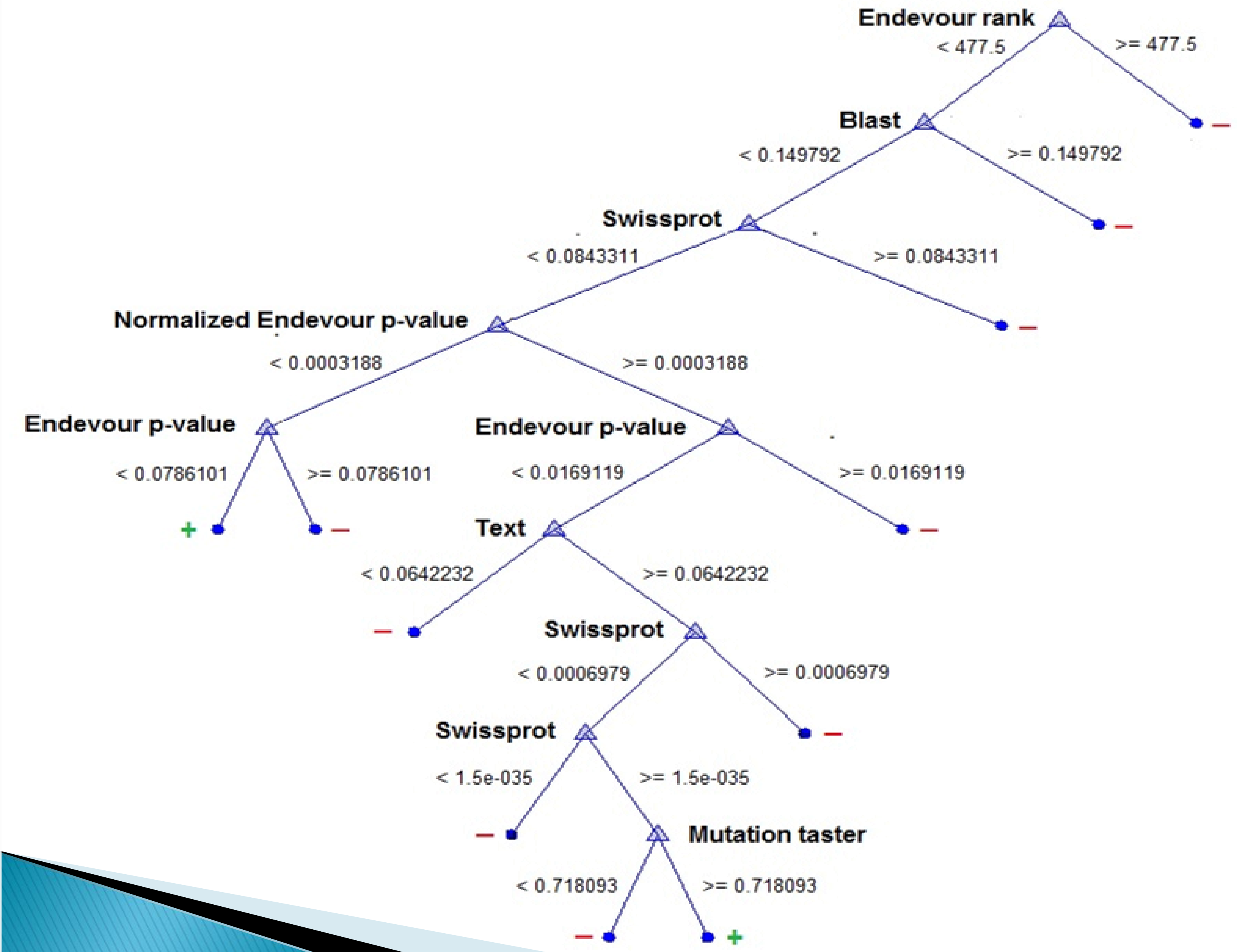
Random Forest classifier

training outcomes

training data

Decision tree classifier




Endevour rank
< 477.5
>= 477.5
–
Blast
< 0.149792
>= 0.149792
–
Swissprot
< 0.0843311
>= 0.0843311
–
Normalized Endevour p-value
< 0.0003188
>= 0.0003188
Endevour p-value
< 0.0786101
>= 0.0786101
+
–
Endevour p-value
< 0.0169119
>= 0.0169119
–
Text
< 0.0642232
>= 0.0642232
–
Swissprot
< 0.0006979
>= 0.0006979
–
Swissprot
< 1.5e-035
>= 1.5e-035
–
Mutation taster
< 0.718093
>= 0.718093
–
+

1. **PROBLEM-TAILORED CHOICE OF THE TRAINING SET**

2. **HETEROGENOUS DATA FUSION**

3. **PHENOTYPIC INFORMATION**

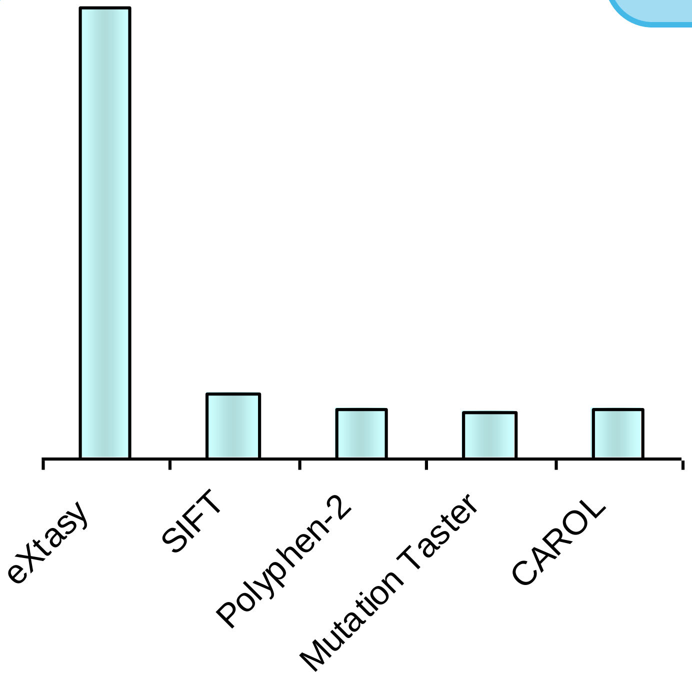
WHAT COMES NEXT?

Data :

- Appropriate control set
- Heterogenous sources
- Phenotypic information

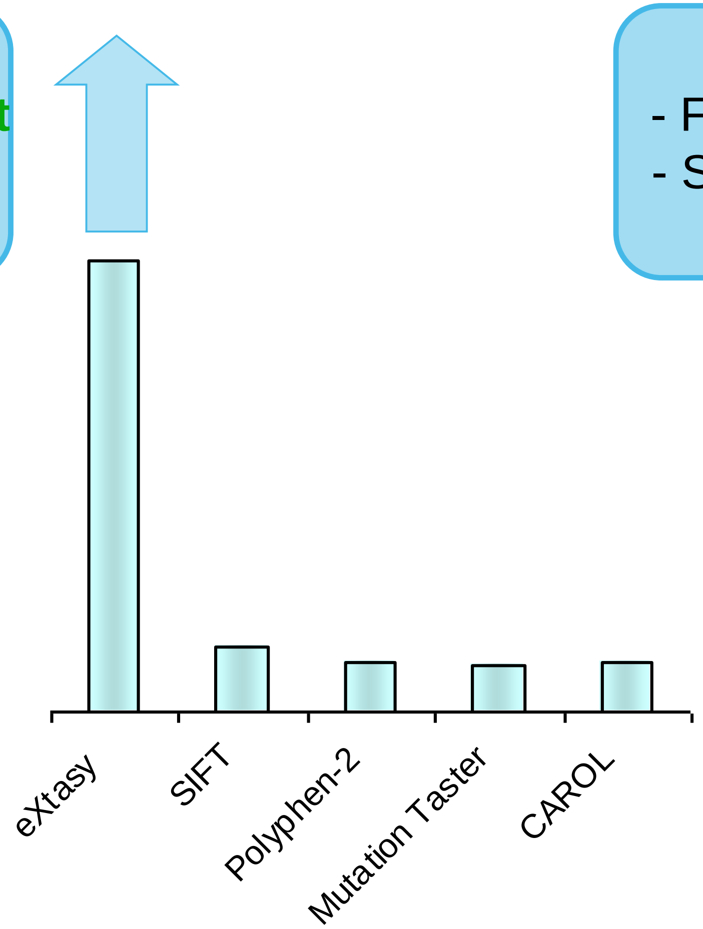
Algorithms :

- Flexible learner
- Score aggregation



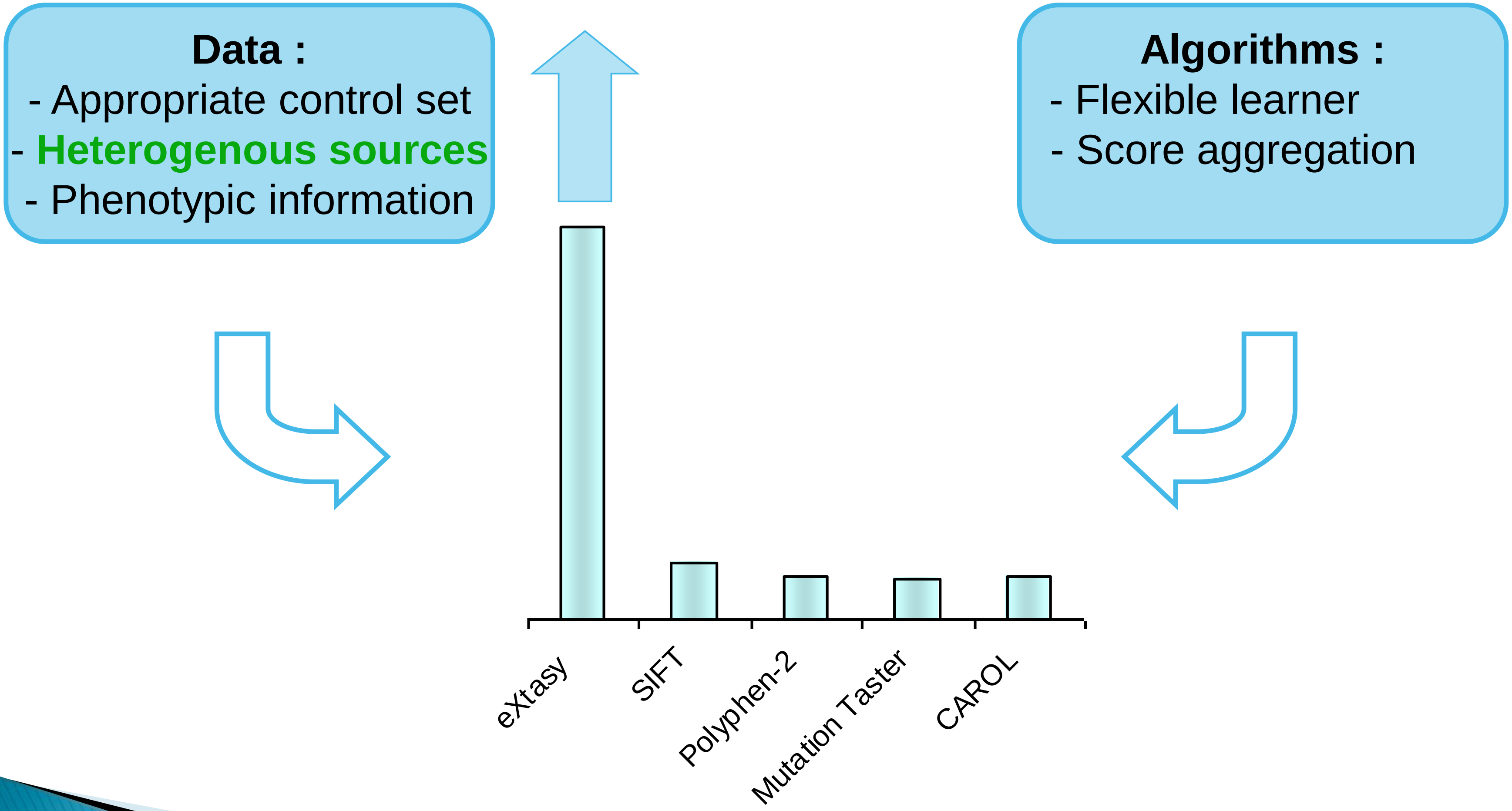
Data :

- **Appropriate control set**
- Heterogenous sources
- Phenotypic information



Algorithms :

- Flexible learner
- Score aggregation



Data :

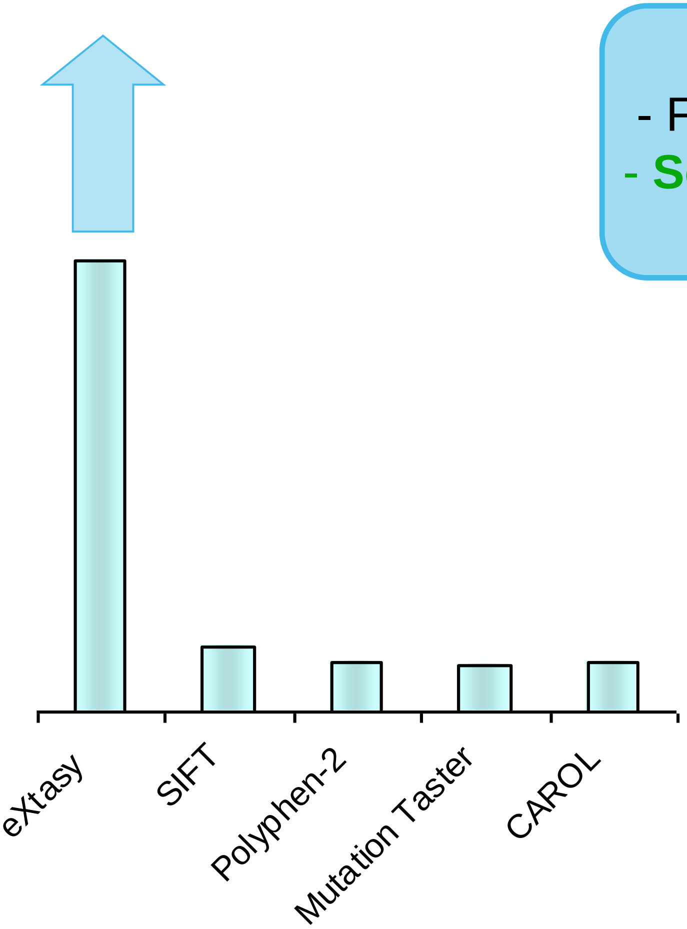
- Appropriate control set
- **Heterogenous sources**
- Phenotypic information

Algorithms :

- Flexible learner
- Score aggregation

Data :

- Appropriate control set
- Heterogenous sources
- Phenotypic information

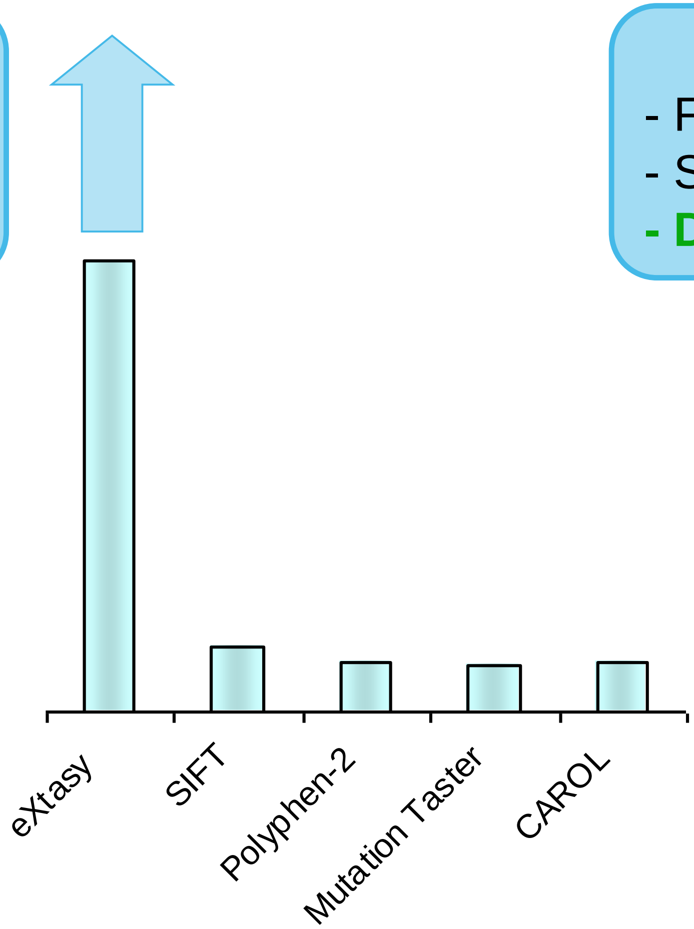


Algorithms :

- Flexible learner
- **Score aggregation**

Data :

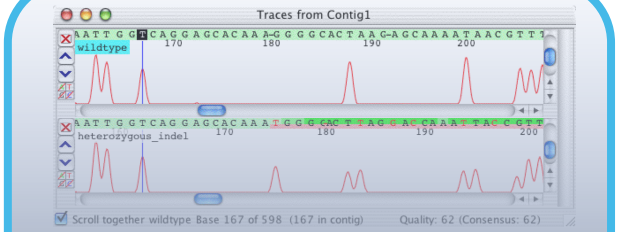
- Appropriate control set
- Heterogenous sources
- Phenotypic information



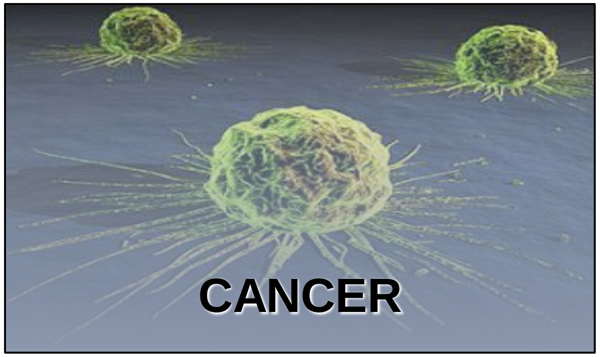
Algorithms :

- Flexible learner
- Score aggregation
- **De-biased model**

eXtasy
Variant Prioritization

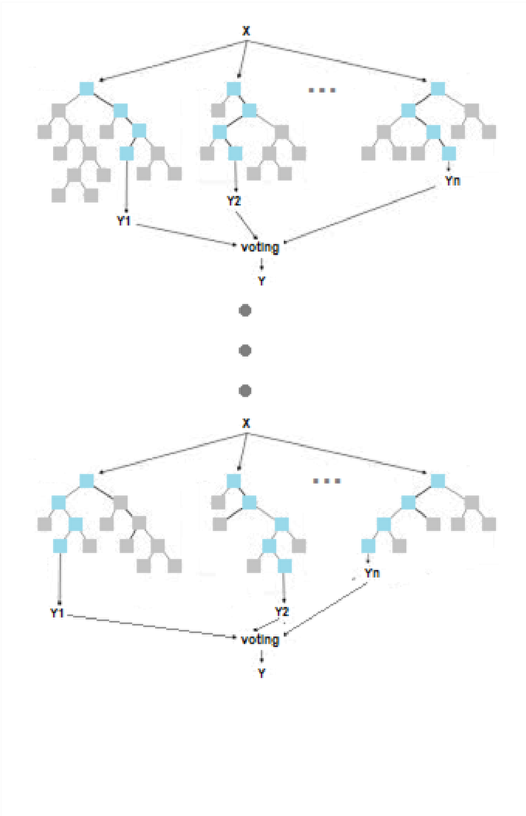

Traces from Contig1
AATTGGTCAGGAGCACAAA-GGGGCACTAAG-AGCAAAATAACGTTT
wildtype
170
180
190
200
AATTGGTCAGGAGCACAAATGGGCACTTAGGACCAAATTACCGTT
heterozygous_indel
170
180
190
200
☑ Scroll together wildtype Base 167 of 598 (167 in contig)
Quality: 62 (Consensus: 62)

OTHER TYPES OF MUTATIONS :
nonsense, splice sites, indels…



CANCER

THANK YOU!



Q & A