Matricola n.000772140

ALMA MATER STUDIORUM
UNIVERSITA' DI BOLOGNA

SCUOLADI SCIENZE
CORSO DI LAUREA MAGISTRALE IN
BIOINFORMATICS

Detecting cancer causing genes and Variants in Colon Adenocarcinoma

Tesi di laurea in BIOINFORMATICS

Relatore
Dr. Emidio Capriotti

Presentata da
Luigi Chiricosta

Sessione 2
Anno Accademico 2016/2017

Outline

- Introduction
- Databases
- Method
- Result
- Variant interpretation
- Conclusions and future perspectives

Cancer definition

- Cancer is the name given to a collection of related diseases.

- Cancer can start in almost any tissue of the human body.

- When cancer develops, cells change morphology, survive longer and divide without stopping forming the tumor.

(*National Institute of Health, https://www.cancer.gov/*)

Cancer origins

Environmental factors*:*

Electromagnetical fields

Chemical agents

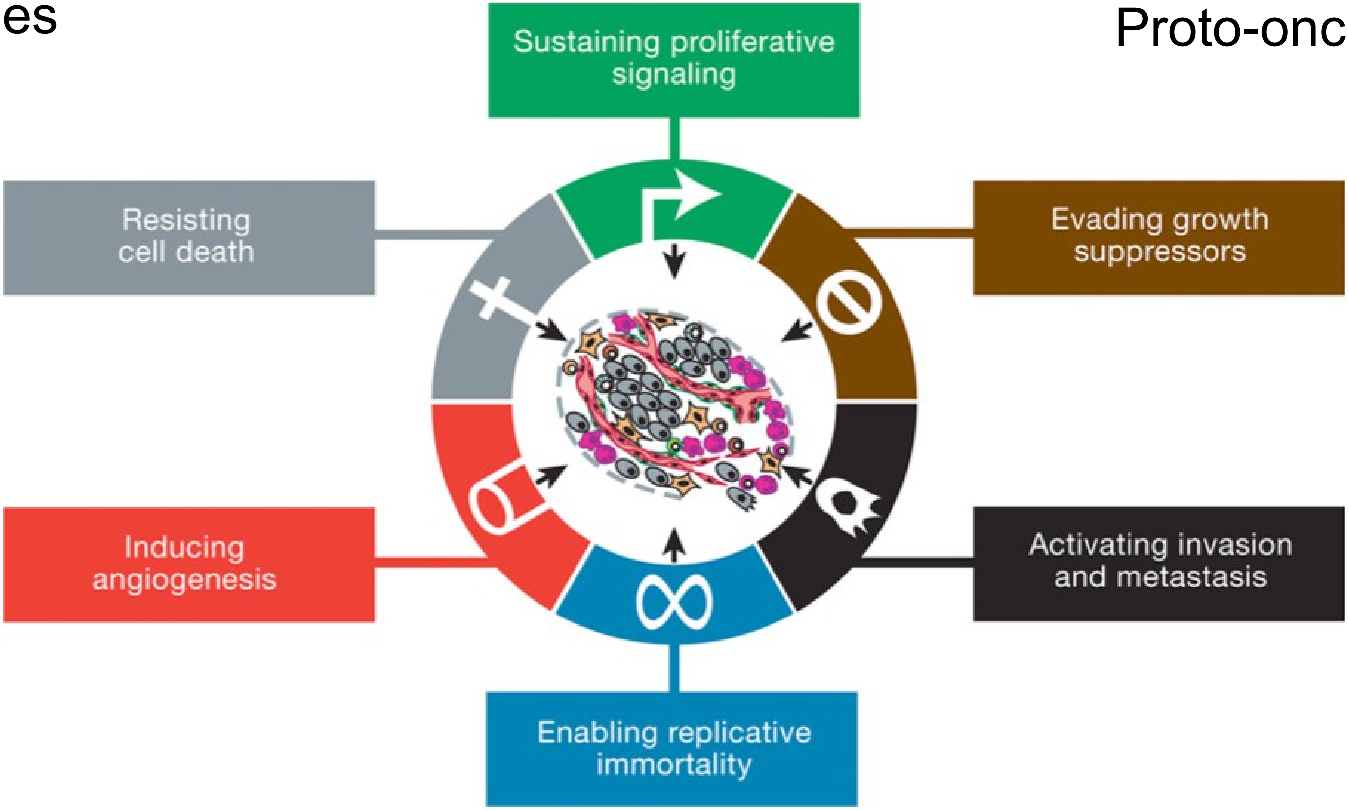
Oncoviruses

Genetic factors:

Tumor-suppressor genes (TSGs)

DNA repair genes

Proto-oncogenes


Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

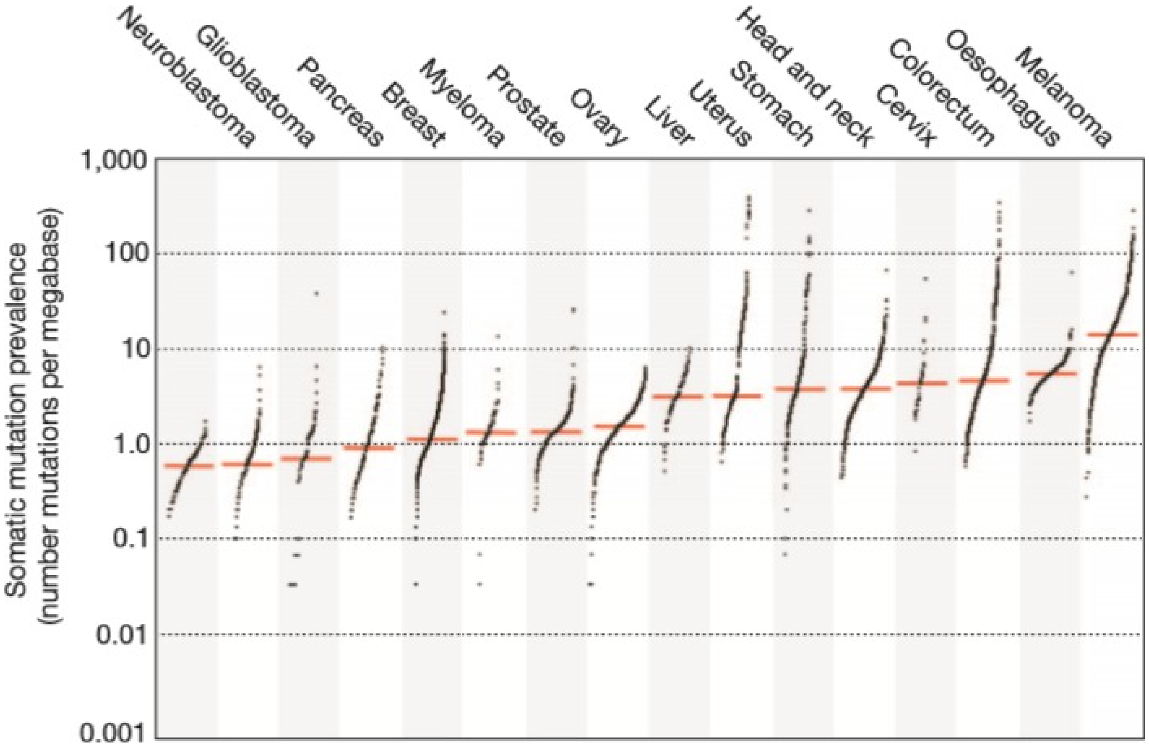
(Hanahan and Weinbeg, Cell, 2011)

Cancer heterogeneity

For the same cancer:

Inter-heterogeneity

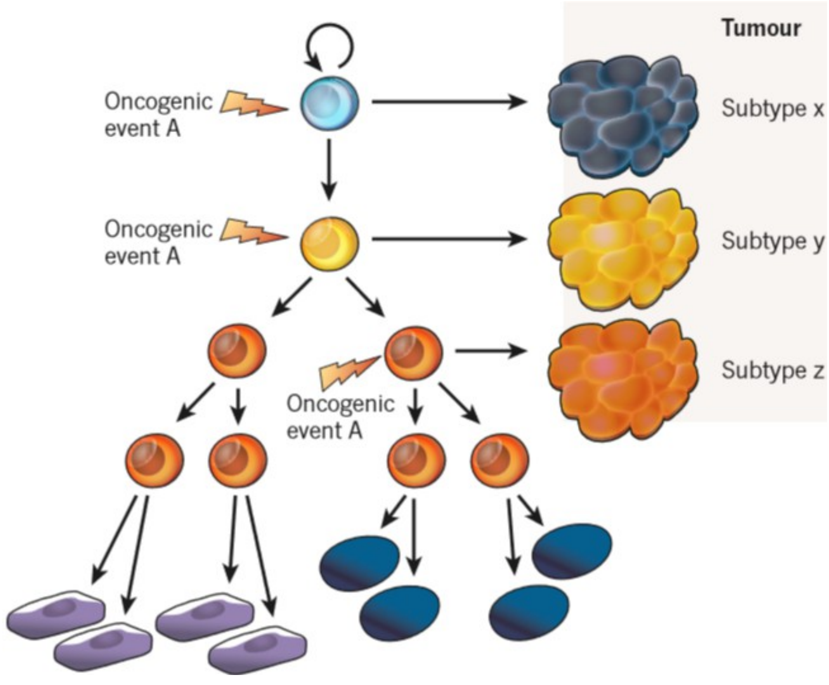
(among the individuals)



(*Alexandrov et. al, Nature, 2013*)

Intra-heterogeneity

(in the tissue; breast cancer has *luminal A*, *luminal B*, *triple negative/basal-like*, *HER2 type*)



(*Visvader, Nature, 2011*)

Experimental analysis

- DNA methylation

(histone modifications and CpG island methylations can cause the silencing or the activation of genes respectivelly by hypermethylation or hypomethylation)

- Transcription profile

(different splicing alterations like exon skip and intron retention not typically recognized in human transcriptome are found by RNA-sequencing)

- Structural Variants

(insertions, deletions, duplications, inversions, translocations and copy-number variants)

- Single Nucleotide Variants (SNVs)

(synonymous, nonsynonymous, frameshift insertion, frameshift deletion, stopgain, stoploss)

Germline vs somatic

Germline: variant present both in tumor and normal tissue.

Somatic: variant present only in tumor tissue.

- Rare germline variants used to estimante the background mutation rate of a gene.
- Compare the frequencies of somatic vs germline to detect possible disease associated genes.

Outline

- Introduction
- Databases
- Method
- Result
- Variant interpretation
- Conclusions and future perspectives

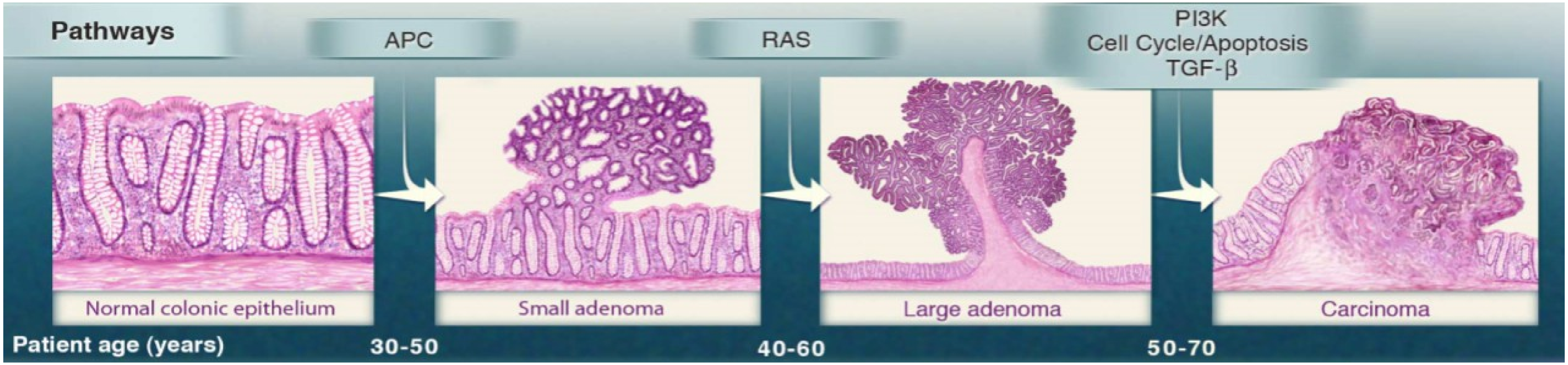
Colon Adenocarcinoma

COAD is one of the most common form of cancer.

Colon: organ in which a cell become carcinogenic.

Adenoma: benign tumor with glandular origins.

Carcinoma: malign tumor in the epithelial tissue.



(*Vogelstein et al., Science, 2013*)

Databases

- The Cancer Genome Atlas:

  (https://cancergenome.nih.gov/)

  - Baylor College of Medicine (BCoM)

    (220 patients)

  - Broad Institute

    (456 patients)

Cancer

- 1000 Genomes Project

  (http://www.internationalgenome.org/)

  (2504 individuals)

Healthy

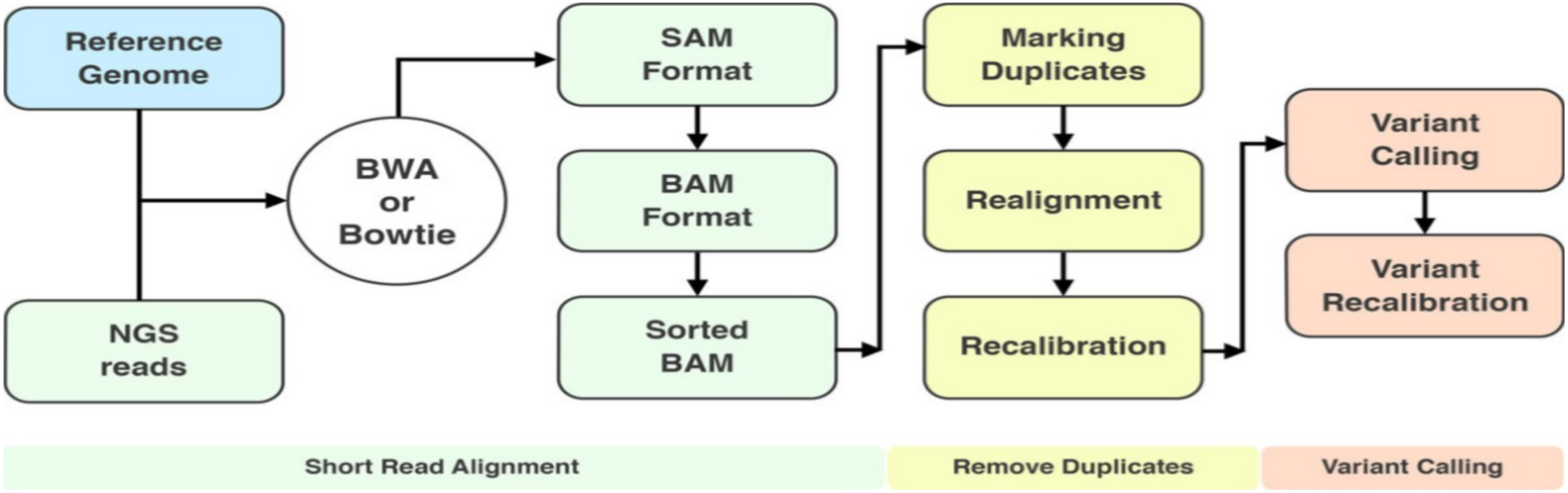
Outline

- Introduction
- Databases
- Method
- Result
- Variant interpretation
- Conclusions and future perspectives

Variant calling

DNA sequencing data are compared with the human reference genome (*hg19*) to extract small variants.

Variant calling parameters are used to filter out low quality variants (depth, base quality, allele fraction).



(Tian *et al., BMC Genomics, 2015*)

Variant annotation

ANNOVAR tool retrieves for each mutation:

- Type of mutational effect
- Mutated residue and its position
- Mutated gene
- Allele frequency

Dataset	Synonymous	Nonsynonymous	Frameshift insertion	Frameshift deletion	Stop gain	Stop loss	Total
BCoM	19.4%	**73.4%**	1.3%	3.9%	1.9%	0.1%	~360,000
Broad	17.8%	**74.4%**	1.9%	3.6%	2.3%	0.1%	~600,000
1000G	40.2%	**58.2%**	0.2%	0.4%	1.0%	0.1%	~2,000,000

Variant selection and classification

- Each mutation with allele frequency > 0.5% is removed.

- The mutational impact characterizes 3 lists:
  - *Synonymoys*
  - *Nonsynonymous*
  - *Functional*

- For each individual we generated one file for germline and one for somatic variants.

Gene prioritization

- The fraction of rare variants across samples is used for the prioritization of cancer associated genes.

- The p-value one-tailed is calculated by Fisher's Exact test.

- The contingency table is:

	Mutated	**No mutated**
Tumor	Amounts of mutated samples in tumor	Amount of no mutated samples in tumor
Control	Amount of mutated samples in control	Amount of no mutated samples in control

$$Score = -\log_{10}(p - value_{1\,\text{tail}})$$

Outline

- Introduction
- Databases
- Method
- Result
- Variant interpretation
- Conclusions and future perspectives

ContrastRank vs Thesis

Comparison of the nonsynonymous-based and ContrastRank prioritization lists, which use a different ranking score.

	ContrastRank	Thesis
# of genes	18,537	17,005
1st	KRAS: 72.6	TP53: 35.2
2nd	TP53: 63.7	KRAS: 31.4
3rd	PIK3CA: 39.4	PIK3CA: 20.0
4th	BRAF: 29.9	BRAF: 10.2
5th	RYR2: 12.9	RYR2: 9.4
Spearman	0.71	
K-T	0.55	

Comparing TCGA datasets

(normal cell as control)

Comparison of the nonsynonymous and functional ranking lists from the TCGA datasets obtained with different variant calling procedures.

	BCoM	Broad	BCoM	Broad
	Nonsynonmous	Nonsynonmous	Functional	Functional
# of genes	17,005	18,012	17,405	18,191
1st	TP53: 35.2	KRAS: 67.8	APC: 52.5	APC: 116.3
2nd	KRAS: 31.4	TP53: 63.1	TP53: 45.5	TP53: 90.9
3rd	PIK3CA: 20.0	PIK3CA: 46.9	KRAS: 31.4	KRAS: 67.8
4th	BRAF: 10.1	TTN: 33.2	PIK3CA: 20.4	PIK3CA: 47.3
5th	RYR2: 9.4	RYR2: 24.3	BRAF: 11.2	TTN: 39.1
Spearman	0.73		0.75	
K-T	0.45		0.55	

Comparing Broad dataset

(normal cell as control)

Comparison of the synonymous, nonsynonymous and functional ranking lists from the Broad Institute dataset.

	Synonmous	Nonsynonmous	Functional
# of genes	16,703	18,012	18,191
1st	TTN: 20.1	KRAS: 67.8	APC: 116.3
2nd	MUC16: 12.2	TP53: 63.1	TP53: 90.9
3rd	FAT3: 11.6	PIK3CA: 46.9	KRAS: 67.8
4th	PCDH1: 9.7	TTN: 33.2	PIK3CA: 47.3
5th	OBSCN: 8.3	RYR2: 24.3	TTN: 39.1

TCGA vs 1000 Genomes Project

The detection of rare variants requires the analysis of a large set of samples.

For the detection of a variant with allele frequency below 1% more than 100 samples are needed.

The final score of each gene is calculated using a bootstrapping procedure.

The prioritization score of each gene is obtained comparing the mutation rate in tumor with TCGA normal and 1000 Genomes samples. The minimum of the two scores is selected.

Combined prioritization score

Comparison of the nonsynonymous and functional ranking lists of the TCGA datasets after the bootstrapping procedure.

	BCoM	Broad	BCoM	Broad
	Nonsynonmous	Nonsynonmous	Functional	Functional
# of genes	17,005	9,723	17,405	11,075
1st	TP53: 31.5	KRAS: 32.4	APC: 52.5	APC: 45.4
2nd	KRAS: 31.4	TP53: 27.7	TP53: 45.5	TP53: 40.8
3rd	PIK3CA: 20.0	PIK3CA: 21.3	KRAS: 31.4	KRAS: 32.8
4th	BRAF: 9.5	BRAF: 8.8	PIK3CA: 20.4	PIK3CA: 21.3
5th	RYR2: 7.1	RYR2: 6.8	BRAF: 11.2	BRAF: 8.8
Spearman	0.66		0.67	
K-T	0.48		0.48	

Outline

- Introduction
- Databases
- Method
- Result
- Variant interpretation
- Conclusions and future perspectives

Variant interpretation

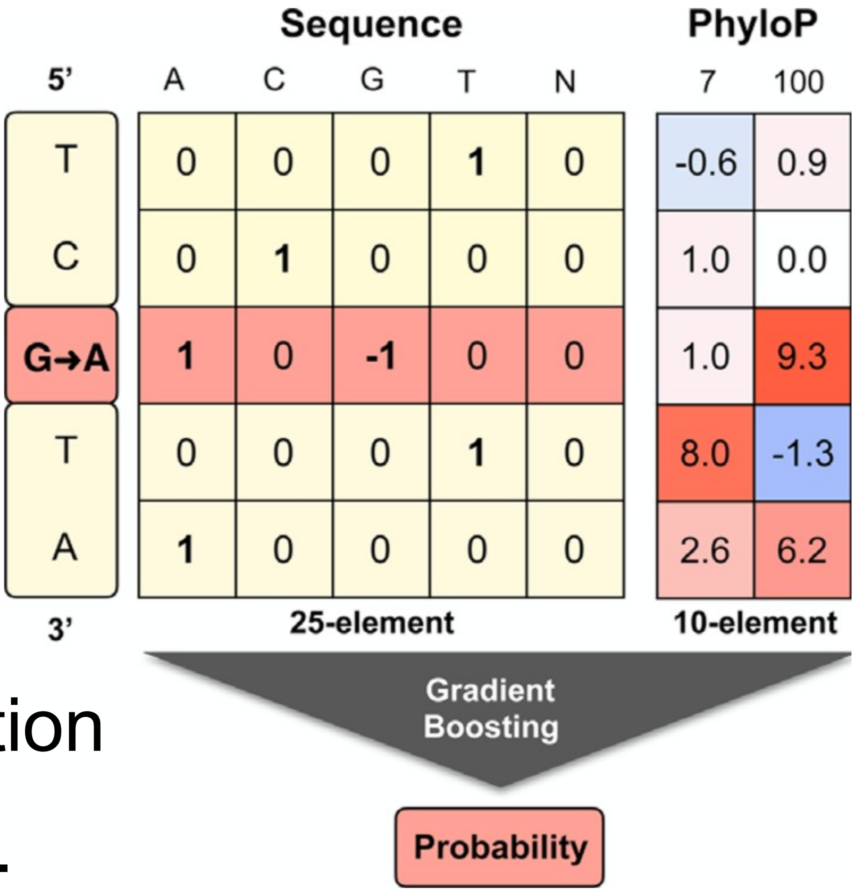
PhD-SNP^g is a binary classificator based on gradient boosting algorithm trained on ClinVar.

It classifies variants in:

BENIGN: *Score < 0.5*

PATHOGENIC: *Score ≥ 0.5*

The main output values are the prediction score and the conservation score across species (PhyloP100).



	Sequence					PhyloP	
5'	A	C	G	T	N	7	100
T	0	0	0	1	0	-0.6	0.9
C	0	1	0	0	0	1.0	0.0
G→A	1	0	-1	0	0	1.0	9.3
T	0	0	0	1	0	8.0	-1.3
A	1	0	0	0	0	2.6	6.2
3'	25-element					10-element	

Gradient Boosting

Probability

(Capriotti and Fariselli., *Nucleic Acid Research, 2017*)

Prioritization and causing variants

Comparison of the nonsynonymous and functional ranking lists of the TCGA datasets after removing benign variants predicted by PhD-SNP^g.

	BCoM	Broad	BCoM	Broad
	Nonsynonmous	Nonsynonmous	Functional	Functional
# of genes	15,070	18,012	16,023	18,191
1st	TP53: 33.8	KRAS: 66.5	APC: 58.2	APC: 123.7
2nd	KRAS: 31.0	TP53: 62.6	TP53: 44.0	TP53: 90.7
3rd	PIK3CA: 20.0	PIK3CA: 46.9	KRAS: 31.0	KRAS: 66.5
4th	BRAF: 10.2	TTN: 31.7	PIK3CA: 20.4	PIK3CA: 47.3
5th	RYR2: 8.5	FAT4: 22.2	BRAF: 11.2	TTN: 39.1
Spearman	0.72		0.74	
K-T	0.12		0.13	

Variant analysis

TCGA variants are analysed according to:

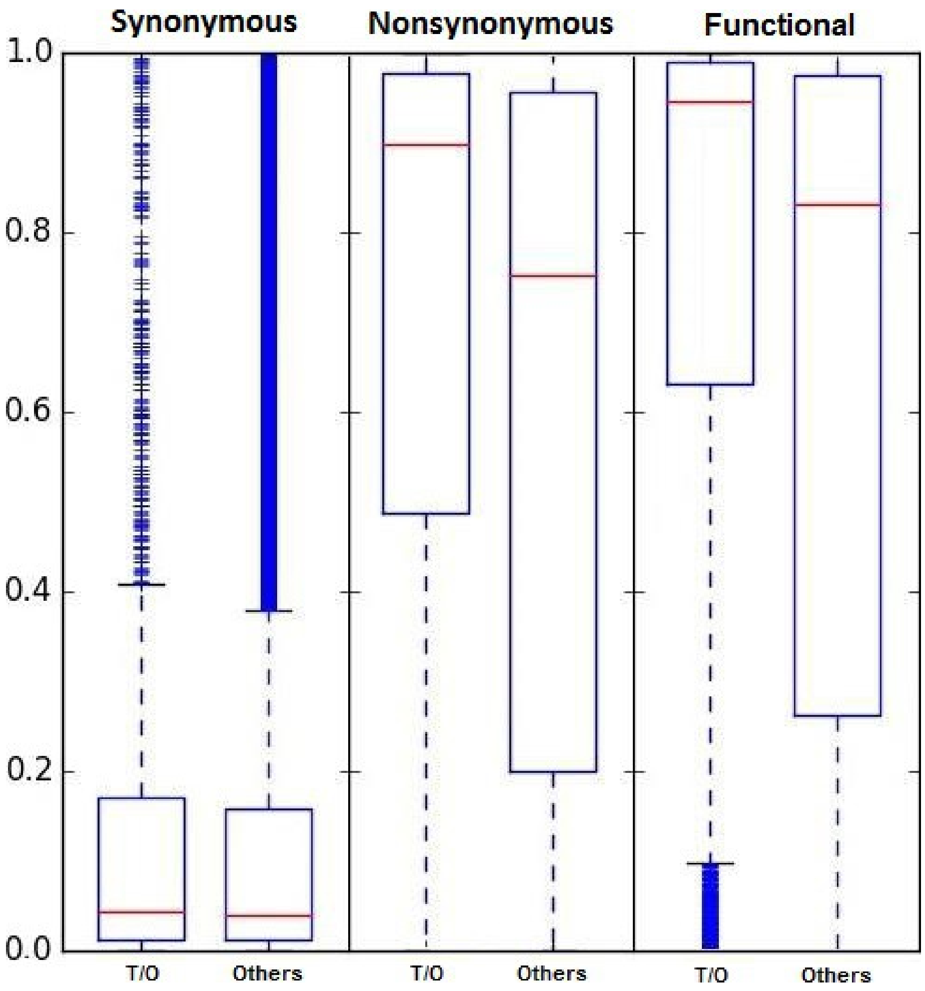
- Occurrence
- *Germline* or *somatic* annotation
- COSMIC Cancer Census

Four classes are analysed:

- *Benign* **vs** *pathogenic* variants
- *Germline* **vs** *somatic* variants
- *TSG* **vs** *oncogene*
- *TSG and oncogene* **vs** *all the remaining genes*

Prediction score

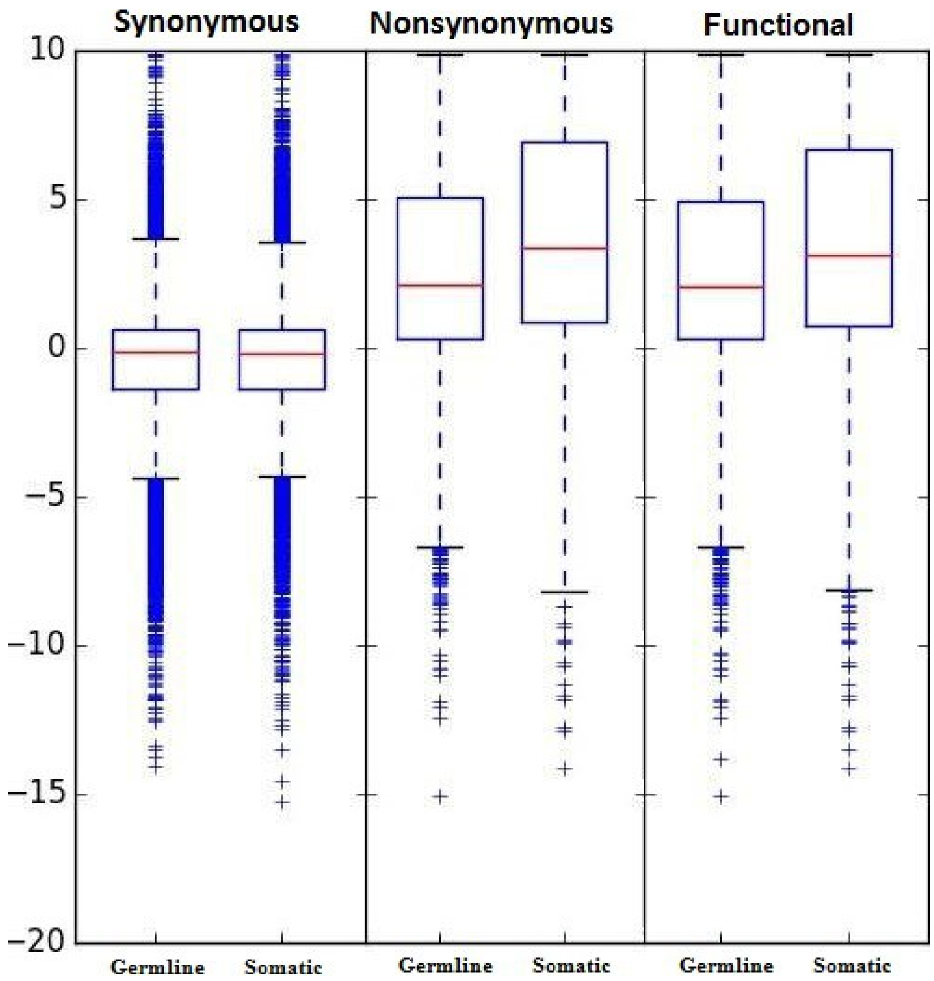
PhD-SNP^g predicts a large fraction of synonymous variants as benign and most of the functional variants are pathogenic.



P-value=1.9e-1 p-value=2.4e-59 p-value=3.8e-87

Conservation score

A large fraction of synonymous variants occurrs in genomic regions less conserved than functional variants.



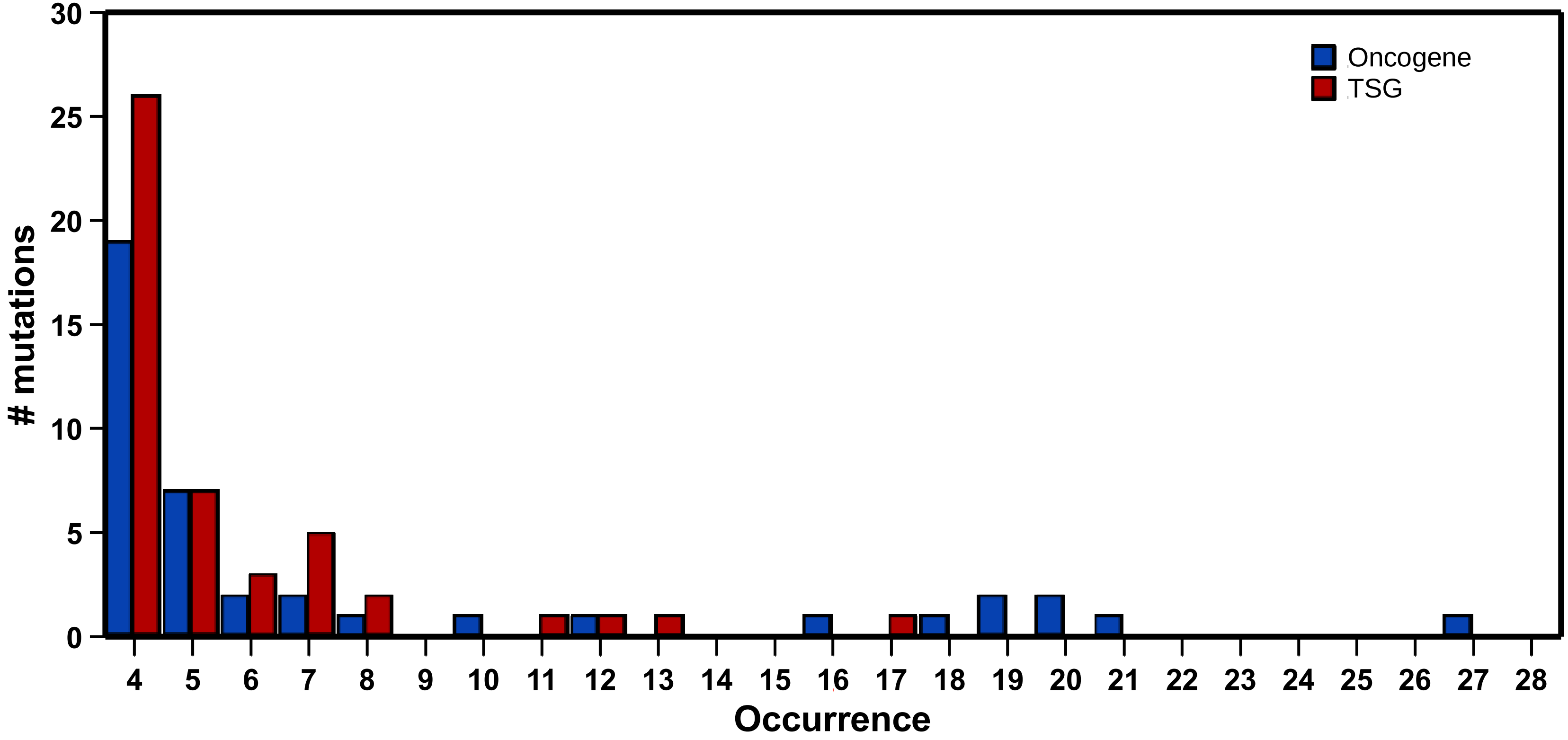
P-value=1.2e-2

P-value=0.0

P-value=0.0

Variation occurrence

Oncogene variants tend to have higher occurrence than TSG variants. The distribution shows the number of TSG and Oncogene variants with occurrence ≥ 4.



TSG variants: 2,880

Oncogene variants: 2,300

Outline

- Introduction
- Databases
- Method
- Result
- Variant interpretation
- Conclusions and future perspectives

Conclusions

- Our gene prioritization method is robust. Using alternative scoring schemes, the top ranking genes are shown in similar order.

- The method is weakly dependent on the variant calling procedure. The order of the top ranking genes from Broad and BCoM datasets are similar.

- The use of functional variants allows to detect cancer associated genes not found considering only the nonsyonymous variants (APC).

- Variant interpretation predictions support the hypothesis that the functional mutations are more likely to be pathogenic than synonymous variants.

Future perspectives

- Integrate the gene prioritization (ContrastRank) and variant interpretation (PhD-SNP^g) scores for estimating disease risks.
- Include gene expression level to select the subset of variants that are significantly expressed.
- Estimate the impact of genetic variants at network level including information from protein-protein interaction and gene pathways.

Thank you!

Questions?