Matricola n. 0000772140

ALMA MATER STUDIORUM
UNIVERSITA' DI BOLOGNA

SCUOLA DI 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

Index

Introduction

The DNA is the macromolecule responsible to bring the genetic instructions necessary for the cell growth, development, functioning and reproduction. In the human being, most of the DNA is located in the nucleus and during the mitotic phase of the cell cycle the molecule is reorganized in 22 pairs of autosomal chromosomes and two sexual chromosomes.

The high level of DNA packing is necessary both for reduce its size and to prevent damages, since it constitutes the phenotype of the cell.

In the normal population, it is common to find genes that are functional even if they have a change in the DNA sequence. In particular, the variation of a single nucleotide is the most common type of genetic variation. Among them, the Single Nucleotide Polymorphism (SNP) is a variation of a single nucleotide which is observed in at least 1% of the population.

Gregor Johann Mendel was the first to hypothesize the transmission of the somatic traits by genotypic background. In this view a Mendelian trait, which correspond to a change in a single locus, can cause a disease. The information about the relationship between genotype and phenotype are collected in *OMIM* (Online Mendelian Inheritance in Man) database which is available online (*https://www.omim.org/*).

If the genome replication process is affected by an error, it can be retained in the daughter cells. Although the majority of the errors resulting in genome variations do not have a pathological effect, nevertheless, combination of multiple variations can affect gene networks resulting in complex disorders like asthma, diabetes, obesity, infertility, hypertension and heart disease. Among the complex disorders, cancer is one of the most common cause of death in the western society. It is caused by an

accumulation of genetic variants that provide to the cells selective advantages. Although large scale sequencing experiments allowed to detect a huge amount of genetic variations in tumor cells, the role of most of them in the insurgence and progression of the disease is still unknown. Thus, the main aim of this thesis consists in the detection and prioritization of cancer causing genes and variations in Colon Adenocarcinoma.

I Definition of cancer and its causes.

Cancer is a neoplastic disorder which is caused by abnormal cell growth. It is based on a genetic changes that can potentially originate in each cell of the organism; usually, the causative affected genes are proto-oncogenes, tumor suppressor genes and DNA fixing genes.

The proto-oncogenes are normally active in a normal cell and they handle the cell growth and division. When aberrations switch them to oncogenes they become more active than normal hindering the cell specialization.

On the other hand, the affected tumor suppressor genes (TSGs) can lose the capability to induce the death and the apoptosis of the cell since they are not able to manage the cell growth and division anymore, even if the cell is too old or if the DNA is irreparably damaged.

The DNA fixing genes are normally able to repair errors in the DNA of a normal cell; if mutated, they can not avoid the cell to accumulate mutations which might have pathogenic effect. The large variety of mutated genes makes the cancer highly heterogeneous.

According to the newest version of the reference genome (hg38), Ensembl website (Aken *et al.*, 2016) maps a typical tumor-suppressor gene, TP53, among the positions 7,661,779-7,687,550 in the chromosome 17. The function of TP53 reported in UniProt website (The UniProt Consortium, 2017) consists in inducing growth arrest or apoptosis. An example of proto-oncogenes is KRAS which is located in position 25,204,789-25,250,936 of the chromosome 12 (hg38). As reported in UniProt, it promotes oncogenic events by inducing transcriptional silencing of tumor suppressor genes.

Outside of the human body, environmental factors (as smoke) and oncovirus (as hepatitis B virus) and electromagnetic fields (as ultraviolet rays) can generate aberrations in the normal duplication of the DNA producing Single Nucleotide Variants (SNVs) and Structural Variations (SVs) (Casás-Selves and Degregori, 2011).

There are six different effects resulting from a SNVs in the protein coding region:

- Synonymous variant: a SNV which does not change the protein coding residue.
- Nonsynonymous variant: a SNV which results in an amino acid change in the coded protein. In this case the new residues can have the same physicochemical properties or not.
- Frameshift: it brings a totally new translation of the next part of the sequence just shifting the reading frame. It can be caused by an insertion or a deletion when the added or deleted nucleotides are not multiple of three.
- Stop-gain: the modification adds a new stop codon (TAG, TAA or TGA) that truncates the coded protein.
- Stop-loss: the modification changes a stop codon and a new portion pieces of sequence is added to the protein.

The SVs characterize the modification of the structure in one or more chromosomes. The typical variations includes insertions, deletions, duplications, inversions, translocations and copy-number variants.
The tissue in which the modifications generate a carcinogenic cell identifies the type of the tumor, which is further characterized by the kind of altered cells. From more than 100 typologies of cancer, the National Institute of Health (NIH) classified them in:

- Carcinoma (about the skin cells)
- Sarcoma (about connective or supportive tissues)
- Leukaemia (about blood cells)
- Lymphoma (about immune system)
- Multiple myeloma (about plasma cells)
- Melanoma (about melanocytes)
- Brain and spinal cord tumors (about central nervous system)

The most common type of cancers are the carcinomas and, when it origins in glandular cells, it is called adenocarcinoma. Aging, obesity, lack of physical activity and other environmental factors make the colon one of the best candidates in the development of the Colon Adenocarcinoma (COAD), which is the focus of this project.

Nevertheless, it is a very rare event that a normal cell switches to a carcinogenic state but, when it occurs the tumor is generated by the following mechanisms (Hanahan and Weinberg, 2011):

- sustaining proliferative signaling
- evading growth suppressor
- resisting cell death
- enabling replicative immortality
- inducing angiogenesis
- activating invasion and metastasis
- deregulating cellular energetics
- avoiding immune destruction.

Usually, when the cancer is identified, the first step is to proceed with the biopsy, a procedure in which the abnormal tissue is partially extracted by a pathologist that classifies it by means of a microscope. Depending on the carcinogenic mechanism, it is possible classify the tumor in different stages:

- stage 1: the cell shape is changing
- stage 2 and 3: the closest tissues are violated and it is spreading either in blood vessels or lymph nodes.
- stage 4: one or more cancer cells are settled in a new tissue like a metastasis.

The knowledge of the cancer cell stage suggests one of the three main strategies to treat the disease: surgery, chemotherapy and radiotherapy.

The surgery is the most useful if the cancer is not spread and it has

circumscribed edges. However, in most of the cases, a tumor is discovered when it becomes symptomatic and, often, spread.

The chemotherapy consists to inject in the organism toxic agents that kill all the cells with high replication level. Nevertheless, several kind of cells in the human body, like the bone marrow and hair follicles ones, have this characteristic so they are attacked too. This therapy usually can reduce the original cancer but the survival cells tend to become immune to the used drugs (McNerney *et al.*, 2017).

The radiotherapy kills the cancer cells by an high level of ionizing radiations. Especially if the margins of the tumor are not delimited, normal cells can be hit by the rays and new tumors could develop.

None of these therapies are always effective, for this reason cancer research is focusing on the development of new treatments based on the genotypic profile of the patients.

With the advances of the Next-Generation Sequencing (NGS) technology the amount of data that was generated became enormous than the low-throughput sequencing.

During the last few years, The Cancer Genome Atlas (Chang *et al.*, 2013) consortium sequenced many human exomes from tumor patients. These studies allowed to identify a huge amount of genetic variants, the majority of which might not be related to the insurgence and progression of the disorder.

Nevertheless, several mutations in common among the subjects are discovered. In particular, the variants detected with higher frequency are more likely to be causative and therefore are classified as potential “driver” mutations.

One of the key challenges in cancer genomics is the detection of the pattern of mutations among the large and heterogeneous variety of genetic changes typical of each kind of cancer (Vogelstein *et al.*, 2013). Accordingly, the main aim of the cancer genomics research consists in the identification of

the smallest subset of mutations needed to switch from a normal to a carcinogenic cell. The discovery of the causative changes should make easier a creation of the *ad hoc* therapy, the aim of the personalized medicine.

II Detection of the genetic variations.

Starting from 2004, the sequencing procedure developed by Frederick Sanger was made faster by the massive parallel sequencing technologies produced by Solexa, Roche, Life Technologies and Illumina companies.
Their machines use the sequencing-by-synthesis or sequencing-by-ligation approaches that allow to reduce the sequencing time from weeks to days.
The Next Generation Sequencing (NGS) machines differ for the nucleotide matching, the input reads length, the output depth coverage and, not less important, the cost per run.
The first step of the sequencing consists in the preparation of the libraries which is the input of the machine. This process is followed by a data analysis step to reconstruct the possible sequence. The raw data is represented in FASTQ format in which each read nucleotide is associated to a score *S*:

$$S = -\log_{10} P \tag{1}$$

where *P* is the probability to have a particular nucleotide in a position in the sequence. The quality check is the second step: an average quality per read position is computed to discriminate good against poor data. The third pre-processing step is a filtering and trimming procedure. It allows to select the set of read with the higher quality reducing the coverage but even the noise.
Algorithms based on the Burrow-Wheeler transform allow to map and align the read against a reference genome to reconstruct the whole DNA ordering each read in the right position. The final output is stored in a BAM format that is the binary version of the SAM format.
An additional filter considers the elimination of the repeated reads that increase the coverage but do not give more information. The located mismatches are classified as possible variants or not by means of several

programs that take advantage of statistical models. Broad Institute developed GATK (McKenna *et al.*, 2010) and MuTect (Cibulskis *et al.*, 2013); in the end, the potential variants are stored in a BCF file, the binary version of the human-readable and scriptable VCF (Variant Call Format) format file.

VCF format contains an header with the information of the data. The strings of the header start with a two hash characters (##). There are different versions of VCF format and it is specified in the first row; the last released version is the 4.2. The header row (marked by ##SAMPLE) contains also the information about the platform used to perform the variant calling. The last raw of the header, with a single hash (#), contains the identifiers of the columns by which the data are represented in the file. Each mutation is represented by chromosome, position, dbSNP id, reference base, alternative base, quality, filter, info, format. After the column number nine, it is included a column for each sample represented in the file.

Among the typical information present in the formatted fields there are: the genotype (GT), the read depth for position (DP), the read depth supporting the alleles (AD), the average base quality (BQ), fraction of reads supporting each reported alternative allele (FA). If a certain value is not available, it is substituted by a ".".

These information are used to perform a statistical analysis and select the most reliable mutations in a given genome. If a mutation is selected, it is marked with "PASS" in filter field.

The genotype representation is a combination of single digits associated to the reference and possible alternative alleles. The digit is a number that can be 0 if it matches the reference allele or any other number for the alternative alleles. In this case, the digit matches the order of the alleles in the fifth column.

Both copies of the genome can be phased or unphased when the alleles are separated with the symbols "|" and "/" respectively.

III Genome variant databases

The low cost for sequencing a genome with the new NGS machines promoted the development of large-scale sequencing projects. Most of these studies focused on the detection of the genetic variants among the individuals in different populations both in the whole genome and in the exome genome.

1000 Genomes Project

The 1000 Genomes Project aim was to collect the most genetic variants with frequencies higher than 1% coming from healthy human beings (Auton *et al.*, 2015). This data, released on the 02/05/2013, is available online at the address http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

The 1000 Genomes Project cohort consists of 2504 individuals from 26 populations.

From the website it is possible to download 24 files, one for each chromosome, one file for the mitochondrial DNA, and for each of these files it is available a corresponding tbi file that contains the indexes to speed up the search of the mutations in the file.

The files are represented in the VCF format version 4.1 and the annotation was performed by GATK software. The sequencing platforms used in this study are ABI SOLiD and Illumina with a minimum coverage of 4x. For each mutation is reported the genotype about the 2504 individuals. No other value is specified in FORMAT column.

In this project the mitochondrial DNA was excluded from the analysis and the counted unique mutations are 84,801,880. All the variants are distributed among the individuals according to the Fig. 1. The Fig. 2 shows how the 18,831 protein coding genes are distributed among the samples.

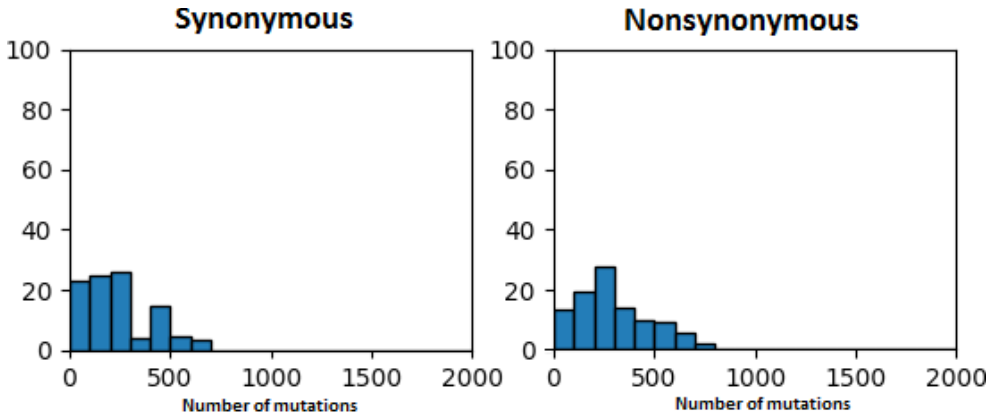


Fig 1. Distributions of mutations in the 1000 Genomes Project dataset. The x-axis represents the number of variants while in the y-axis is present the percentage of clusters of samples that have a certain amount of mutations. On the left column are reported the synonymous mutations while in the other one the nonsynonymous. The upper limit of the x-axis, representative of the number of mutations, is set to 2000 to make the plots comparable.

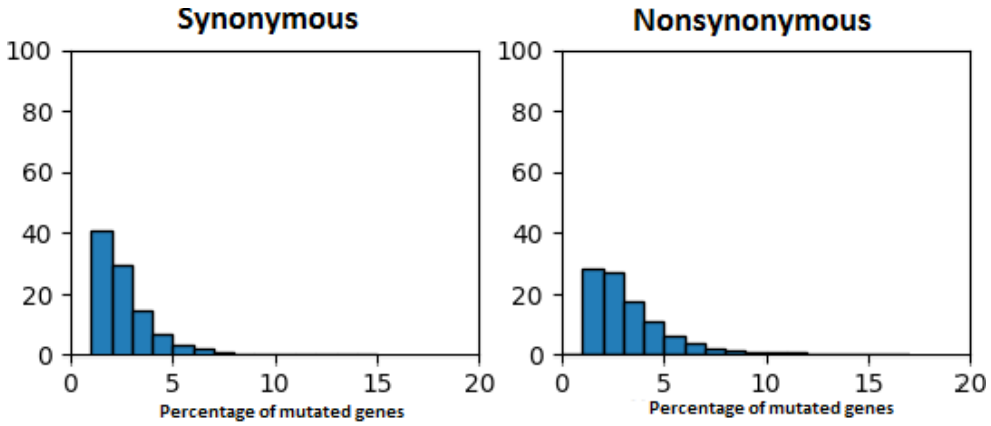


Fig 2. Distributions of mutated genes among the samples about synonymous and nonsynonymous mutations in 1000 Genomes dataset. Both of the axis represent a percentage. The x-axis gives the amount of genes that are mutated for a percentage indicated in the y-axis.

There are so few genes that are mutated more than in 20% of the samples so they are excluded by the plot in order to do a appreciable representation.

ESP and ExAc datasets

Exome Sequencing Project (ESP, http://evs.gs.washington.edu/EVS/) and Exome Aggregation Consortium (ExAC) dataset collect information about exome genomes coming from 6,503 and 60,706 individuals respectively (Lek *et al.*, 2016). However, they are not directly used but they are included in gnomAD dataset.

GnomAD

One of the major problems managing DNA data is the amount of data collected and stored in non-homogeneous way.

The goal of Genome Aggregation Database (gnomAD) is to aggregate and harmonize all the genome sequencing data coming from large-scale sequencing projects.

In the release of the 27/02/2017, the version 2.0, it contains 123,136 exome sequences and 15,496 whole genome sequences from unrelated individuals of different population (Table 1) coming from datasets like 1000 Genomes Project, ESP and ExAC.

ClinVar

It is a freely collection of report linking medically important variants to phenotype (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965032/). It takes its information from dbSNP, dbVAR and MedGen (Landrum *et al.*, 2014). The version used in the project is of the 05/09/2017.

COSMIC

The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a regularly updated dataset in which are collected the impact of all the found somatic mutations in human cancer *(Forbes et al., 2015)*. It is a manually curated database in which the mapping of the genes is made by means of the analysis of the profile of the mutations of each gene among the observed

samples. If a gene is frequently mutated in a specific position it is classified like oncogene whereas, if the mutations are spread among the length of the gene this is classified as a tumor suppressor gene (Fig. 3). The number of genes present in the release of the 25/05/2017 are 372s from which 123 are TSGs, 127 are oncogenes, 45 are both TSGs and oncogenes, 18 are TSGs and fusion genes, 54 are oncogenes and fusion genes and 5 are all of them.

Table 1. List of data collected in gnomAD divided by different populations.

Population	Description	Genomes	Exomes	Total
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other	491	2,743	3,234
Total		15,496	12,314	138,632

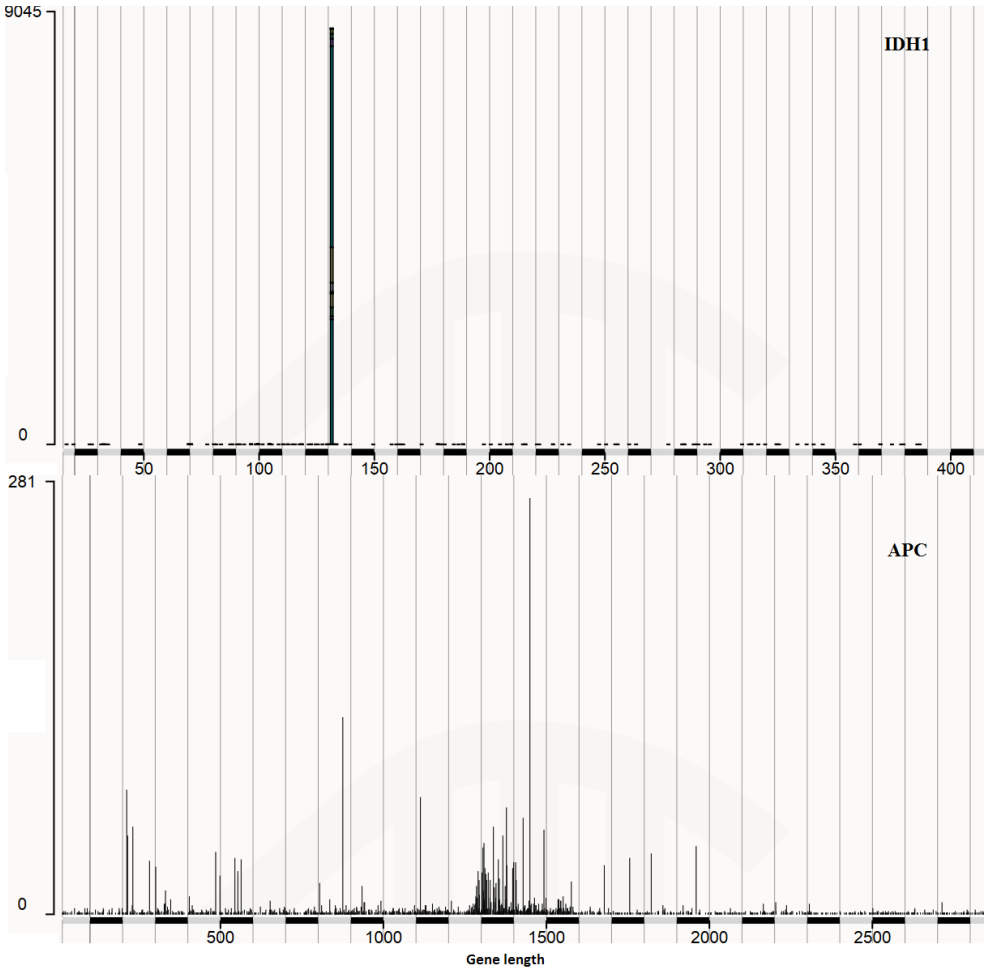


Fig 3. COSMIC mutation profile for a typical oncogene (IDH1, on the top panel) and a tumor suppressor gene (APC, on the bottom panel). In the x axis is represented the gene length and the y axis the number of mutated samples.

IV Methods for genome variant annotation.

In the last ten years there has been an increasing effort in the study of characterization of cancer. The amount of available data and the large variability of the variants associated to the disease require the use of specific tools that allow to quickly retrieval general and specific information about the analysed individuals. Furthermore, methods based on statistical and machine learning approaches were developed to extract the main features for the classification of disease from the available data.

VCFTools

VCF files can be large in size. It is not easy to retrieval the huge amount of information collected inside. Thus, the 1000 Genomes Project designed the program package VCFTools (Danecek *et al.*, 2011) to conduct several complex operations on VCF files using few simple commands.

ANNOVAR

ANNOVAR was the most used software in the first part of the project to annotate genetic variants (Wang *et al.*, 2010). Using the first five fields in each VCF record (chromosome, start position, end position, reference nucleotide, observed nucleotides), ANNOVAR can perform gene-based annotation, filter-based annotation and region-based annotation.

This tool is composed by six perl scripts. Among them *annotate_variation.pl, convert2annovar.pl* and *table_annovar.pl* were used for this thesis project.

The script *table_annovar.pl,* combined with a filter-based annotation, takes in input a VCF file and it produces a tab-delimited VCF file in which each column is involved in a set of annotations.

ContrastRank

ContrastRank is a gene prioritization approach for ranking cancer driver genes. The genes are scored comparing the rate of rare variants (allele frequency < 0.5%) detected in tumor samples with those in normal and 1000 Genomes (Tian *et al.*, 2014).

PhD-SNP^g

It is the extension of *PhD-SNP* (Predicting human Deleterious SNP) algorithm which aim is to predict the impact insurgence of human SNVs by a training procedure on a subset variants from Swiss-Prot which includes neutral and deleterious polymorphisms for the homo sapiens (Capriotti *et al.*, 2006). *PhD-SNP^g* is a machine learning binary classifier based on support-vector machines (Capriotti and Fariselli, 2017). The train data has been extracted from the Clinvar dataset (version January 2016).
Giving in input a VCF file, *PhD-SNP^g* returns back an output file including in each row a column for the prediction (benign or pathogenic), the score associated to the prediction, the false discovery rate, the *PhyloP100* score (of the mutated position), the average *PhyloP100* score (Pollard *et al.*, 2010) of the five nucleotides centered of the mutated site.
In this project we used the standalone version of *PhD-SNP^g*.

V Cancer genome samples from Colon Adenocarcinoma.

Cancer is a disease characterized by an high level of heterogeneity. To perform a robust statistical analysis for discovering the common features among the carcinogenic cells a large set of samples is needed. The Cancer Genome Atlas consortium (TCGA) is one of the biggest databanks in terms of collected data about cancer illnesses (https://goo.gl/qCVq7x). It includes 33 types of cancer selected by more than 11,000 patients collected in 2.5 petabyte of data. The data are publicly available under compressed VCF files using version 4.1.

Several groups are organized by area of focus and provide information for the TCGA; in this project the data are provided by the Baylor College of Medicine and the Broad Institute.

Baylor College of Medicine

Situated at Houston, in Texas, the Baylor College of Medicine (BCoM) was one of the centers strongly involved in the sequencing of the cancer genome.

The BCoM released the sequencing data of 220 patients affected by colon adenocarcinoma (COAD). For each of them, the genotype of the normal and cancer cells were extracted and analyzed by an Illumina platform.

The mutations are characterized, according to the FORMAT column, by GT, DP, AD, BQ, SS, SSC, MQ60.

The FILTER field specifies that the variant calling annotation was made by *CARNAC* (Consensus And Repeatable Novel Alterations in Cancer) software, a logic based on tumor and normal coverage, tumor variant count, mapping quality, allele fraction, strand, variant read position, base quality.

It counts 41,634,730 mutations for which 11,343,465 are unique (27.25%) and they are collected in 18,295 protein coding genes. From the total

amount of mutations, 8,697,089 have the PASS filter (20.89%). All the mutations are distributed among the individuals according to the Fig. 4 while in the Fig. 5 is shown the distribution of the genes among the samples.

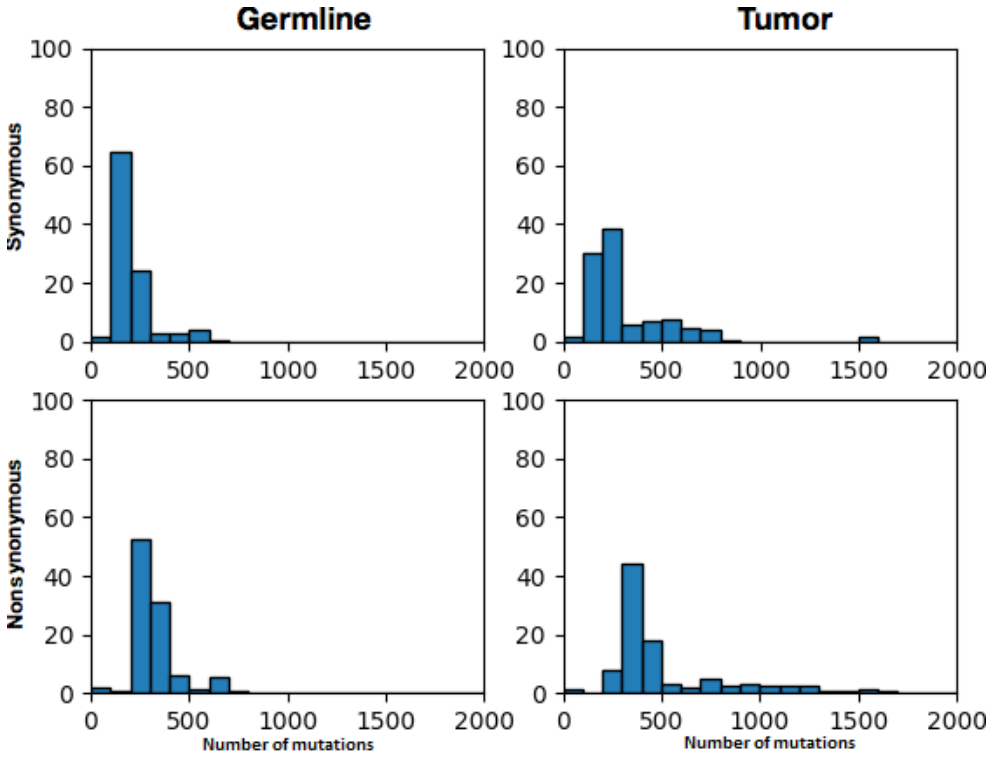


Fig 4. Distributions of mutations in the BCoM dataset. In the x-axis, is represented the amount of mutations while in the y-axis is present the percentage of clusters of samples that have a certain amount of mutation. In toppest row are reported the synonymous mutations while in the other one the nonsynonymous. The upper limit of the x-axis, representative of the number of mutations, is set to 2000 to make the plots comparable.

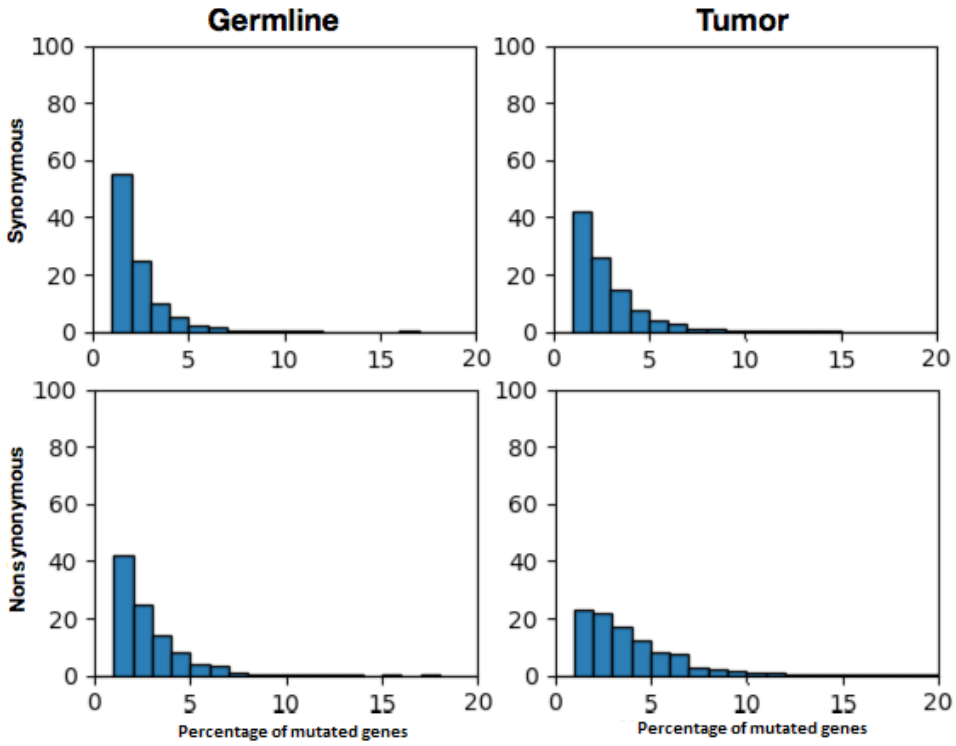


Fig 5. Distributions of mutated genes among the samples in the BCoM dataset. Both of the axis represent a percentage. The x-axis gives the amount of genes that are mutated for a percentage indicated in the y-axis. This meaning there are a lot the genes that are mutated in few people in germline (left column) while there are many more mutated genes among the patients in the tumor case (right column).

Both in synonymous and nonsynonyous datasets there are few genes that are mutated more than in 20% of the samples. For a graphical reason they are excluded from the plot.

Broad Institute

Situated in Cambridge, Massachusetts, the Broad Institute is among the main sequencing center in US and one of the leading institution involved in the TCGA project.

The colon adenocarcinoma dataset released by the Broad Institute is obtained from 456 patients for which a pair of normal and tumor samples were sequenced by an Illumina platform. The mutations in the VCF file are reported with a FORMAT field including GT, AD, DP, FA, MQ0, SS, SSC. The FILTER field specifies that the variant calling annotation was made by two algorithms. MuTect, an algorithm that, after 4 steps, can detect with high level of sensitivity and low level of specificity the somatic variations taking advantage of two bayesian classifiers. An other filter is realized by oxoG3 used to remove possible OxoG (Oxidation of Guanine to 8-oxoguanine) artifacts.

The dataset count 26,976,326 mutations for which 1,154,316 are unique (4.28%) and they are collected in 18,222 protein coding genes. From the amount of total mutations, only the somatic ones can have the PASS filter and they are counted for 1,271,763 (4.71%). The distribution of the variants and the mutated genes across individuals are reported in Fig. 6 and Fig. 7, respectively.

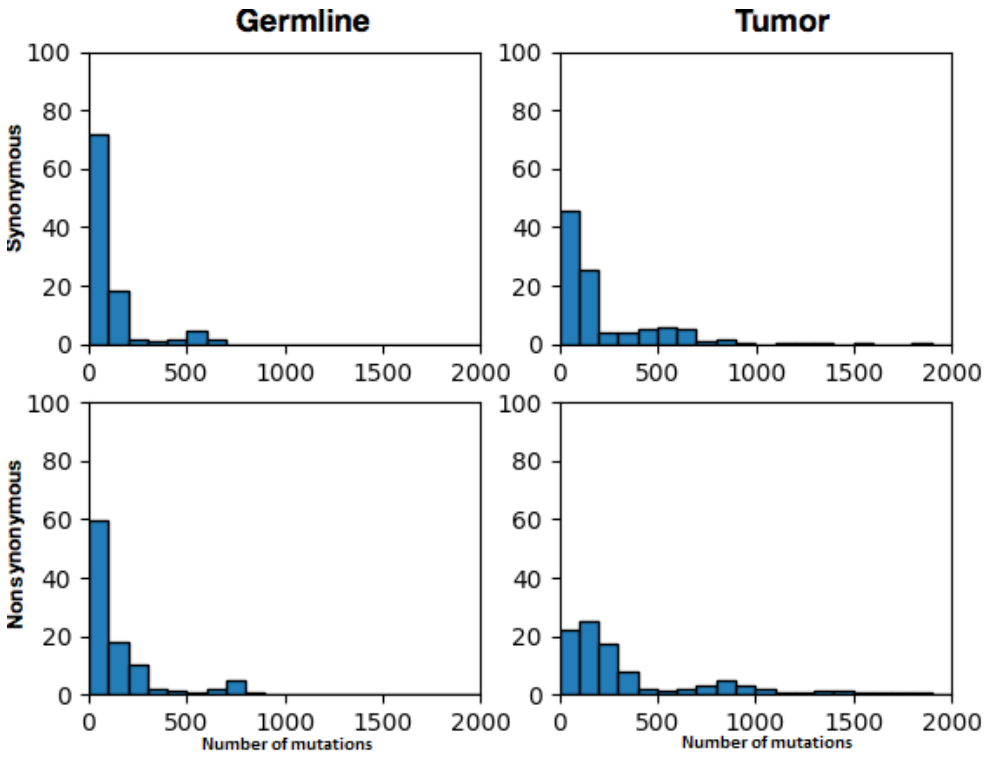


Fig 6. Distributions of mutations in the Broad dataset. In the x-axis, is represented the amount of mutations while in the y-axis is present the percentage of clusters of samples that have a certain amount of mutation. On the topp panels are reported the synonymous variants while nonsynonymous variants are reported in the bottom panels. For a graphical reason the upper limit of the x-axis, representative of the number of mutations was set to 2,000.

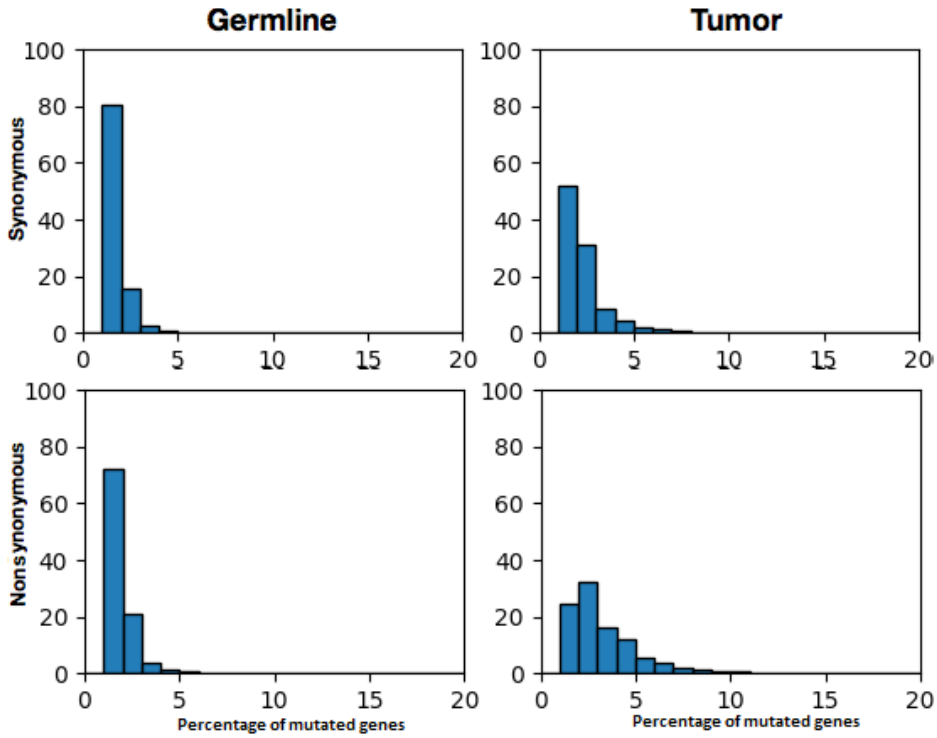


Fig 7. Distributions of mutated genes among the samples in the Broad dataset. Both of the axis represent a percentage. The x-axis gives the amount of genes that are mutated for a percentage indicated in the y-axis. This meaning there are a lot the genes that are mutated in few people in germline (left column) while there are many more mutated genes among the patients in the tumor case (right column).
In both synonymous and nonsynonyous datasets there are few genes that are mutated more than in 20% of the samples. For a graphical reason these genes were excluded from the plot.

VI Extraction and annotation of the cancer genome variants.

The dataset that used in this analysis are the BCoM, Broad and 1000 Genomes ones.

First of all it is necessary to retrievel the allele frequency of each mutation these set using *ANNOVAR*. The script *table_annovar.py* can be used to match the mutations collected in the input file against other databases to adding more information by means of the option filter. The used dataset that collect the allele frequencies information are gnomAD, 1000 Genomes, ExAC and ESP. More datasets are used in order to have the possibility to appreciate each mutation among different populations. Furthermore, the script adds, for each mutation, the information about the the involved gene, the changed residue in that position and the impact of the substitution.

Filters for selecting high-confidence variants

In the analysis of the genomic data, there are several important tasks among them the variant calling. Depending on the laboratory, the extraction and the storage of a DNA sequence can be made in different ways, by different NGS machinery and variant calling criteria. Evaluating different parameters like average base quality and number of reads, the institutes can select a mutation assigning “PASS” on the FILTER field of the VCF file.

The first filter excludes all the mutations that are no signed with “PASS”.

The variant calling of the Broad dataset was filtered by MuTect algorithm, designed to catch just the somatic mutations, so it is needed a sort of logic for accepting also the germline ones; each mutation of this set for which the parameters respect the threshold DP>=10, FA>=0.05, BQ>=30 are considered PASS mutations.

Since the idea of *ContrastRank* is based on the rarest mutations, the next

filter is based on the mutation rate. The allele frequencies obtained by *ANNOVAR* for each mutation are compared among the used databases and the highest one is chosen as representative for that mutation because, some set can have figure out a variation not present in other populations.

The chosen threshold is the same used in *ContrastRank* that is representative for the rare derived alleles with a frequency at most of 0.5% in a population (Fig. 8) (Khurana *et al.*, 2013). Since gnomAD is not a collection of somatic mutations, the most of the matches fail: in that case the allele frequency for that mutation is set to 0.0%.

In summary the BCoM dataset counts now 359,345 mutations for which 179,595 are unique (49.98%). The unique mutations that did not match are 61,351 (34.16%).

Broad dataset counts now 609,439 mutations for which 323,705 are unique (53.12%). The unique mutations that did not match are 173,914 (53.73%).

The 1000 Genomes mutations have all "PASS" filter but the other filters reduce the mutation to 2,109,801 for which 845,489 are unique across the samples (40.07%).

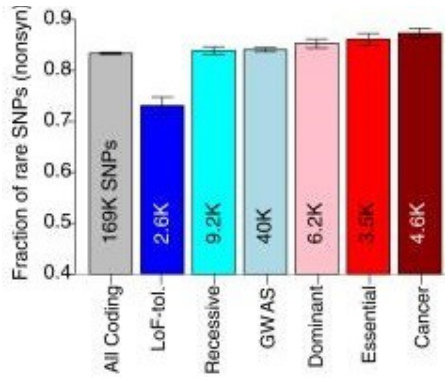


Fig 8. Fraction of rare SNPs among different categories from the paper *Integrative annotation of variants from 1092 humans: application to cancer genomics* (Khurana *et al.*, 2013). Here is considered like rare the derived allele frequency (DAF) <0.5.

If each patient in the TCGA datasets is collected in a file, the 1000 Genomes dataset subdivides the file by chromosomes which include the genotype of all the 2504 individuals.

About the TCGA, two files are created for each sample, one for germline and one for both the germline and the somatic mutations. So, whenever the genotype representation is different to 0/0, is is included in the file. The two possible situation are 0/1 and 1/1.

The situation about the 1000 Genomes dataset is slightly different. One file for individual is created but the mutation is added only if the genotype representation is different than 0|0. It is possible have more than two possible cases (0 or 1) because for each mutation, the individuals can have different alternative alleles and, each of them, is represented with a different number. The number of mutations in all the file are 2,109,801, from which 845,489 are unique.

Classifying the mutation effect

After processing the VCF files data all the information to performing the ranking procedure are available. First of all, a list of all the genes that are mutated is extracted.

The Table 2 shows the percentage of the mutations and its impact respect the totality in the three datasets.

Table 2. Amount of mutations among the three main used dataset divided by their effect

	Syn	Nsyn	FR	FD	SG	SL	Total
BCoM	19.44%	73.36%	1.30%	3.93%	1.94%	0.03%	359,345
Broad	17.76%	74.40%	1.86%	3.63%	2.31%	0.04%	609,439
1000G	40.21%	58.20%	0.17%	0.38%	0.99%	0.05%	2,109,801

The functional effects, in the column, stay for synonymous (Syn), nonsynonymous (Nsyn), frameshift insertion (FR), frameshift deletion (FD), stopgain (SG) and stoploss (SL). In the row, the dataset are Baylor College of Medicine (BcoM), Broad Institute (Broad) and 1000 Genomes project (1000G).

Defining like “functional” all the mutations that can have a negative impact on the stability of the protein or modify significantly its sequence (nonsynonymous, frameshift insertion, frameshift deletion, stopgain, stoploss), three ranked list are computed for each set: one for the synonymous, one for the nonsynonymous and one for the functional mutations.

Prioritizing cancer genes

For each gene is computed the two-tailed p-value and the odd-ratio using the Fisher's Exact test. Nevertheless, since the distribution is not symmetrical, even the one-tailed p-value was computed and, a priori, was decided to get the value coming from the greater part of the distribution.

To better visualize the data, a rank for each gene is computed by calculating the $-\log_{10}(\text{p-value}_{1\text{tail}})$.

Considering the number of mutations in a normal tissue as control value, the classes of the contingency table are represented by the number of times in which a gene is mutated in normal and the tumor samples.

In this way, the genes that have situation very dissimilar in the table among the tumor and the normal situation have a low p-value and an high score.

The same score is computed substituting the normal control class with the 1000 Genomes dataset. The analysis are computed for both the BCoM and Broad datasets.

To test the strength of the method the Kendall-Tau and Spearman correlations are computed for comparing the obtained ranking lists.

Predicting the impact of genetic variants with PhD-SNP^{g}

In the second part of the project it is tested a machine learning predictor based on the nucleotide conservation against all the mutations in the sets. So, all the mutations of the TCGA datasets are collected to become respectively two input files for *PhD-SNP^{g}*. In a further test, variants with no

impact are filtered out from the calculation of the functional prioritization scores. Thus, among all the mutations in our datasets we removed all the variants predicted as "benign" by *PhD-SNP*^{g}. Conversely, for the calculation of synonymous prioritization score, all the synonymous variants predicted as "pathogenic" were removed.

The Table 3 shows all the differences among the input and the output datasets. The final BCoM dataset is composed by 16,334 mutations while the Broad one by 28,962.

The new datasets were used in our the algorithm to perform the same statistics for the calculation of the gene prioritization score.

After including *PhD-SNP*^{g} predictions, the mutations are classified as germline and somatic. A somatic variant has been observed at least once in a tumor sample and not in its matching normal. All the variants that occurs both in normal and matching tumors are classified as germline.

The second level of information is obtained matching the mutated gene against the COSMIC Cancer Census dataset where genes are classified as TSG, oncogene, fusion or their combination.

Finally, they are analysed according to different aspects: prediction score, conservation and occurrence focusing on four categories: "benign" vs "pathogenic", "germline" vs "somatic", "tumor suppressor genes" vs "oncogenes", "tumor suppressor genes and oncogenes" vs "all the remaining genes"

For each of them is computed a representative plot, an histogram or a boxplot, for synonymous, nonsynonymous and functional categories, in both the TCGA sets.

Table 3. Prediction of all the TCGA mutations by PhD-SNP^{g}.

		Total	Syn	NSyn	FD	FI	SG	SL
BCoM	Input	287,940	54,711	216,690	7,656	3,195	5,586	102
	Output	287936	-	216686	-	-	-	-
	Path.	54.08%	9.45%	62.32%	98.79%	98.44%	84.98%	56.86%
BROAD	Input	520,371	90,716	393,332	15,310	8,449	12,351	213
	Output	520,339	90,711	393,306	15,309	-	-	-
	Path.	57.36%	9.48%	64.93%	99.23%	99.21%	87.39%	61.97%

The mutations are divided by their impact. Syn, Nsyn, FD, FI, SG, SL respectively are synonymous, nonsynonymous, frameshift deletion, frameshift insertion, stop gain, stop loss number of mutations. The "-" symbol in output means that the mutation in output are the same that in input, in size. In the second row of the output is reported the percentage of the amount of the pathogenic mutations for each category (Path.).

VII Statistical methods for cancer gene prioritization.

Dealing with big data, the lack of perfect gold standard and complete knowledge of biological processes, the presence of error (like in variant calling) make statistical inference indispensable in cancer research.

Thus, the main idea of *ContrastRank* was to test whether the amount of patients with mutated genes across tumor samples are significantly different from the fraction of mutated control samples from the healthy individuals or tissues. To perform our analysis, the statistical significance was calculated by the Fisher's Exact test. This test was preferred to the binomial distribution because it can be calculated even when no patients mutated samples are present in the control dataset (background mutation rate equal to 0%.). Fisher's Exact test is calculated from the contingency tables, a matrix, usually 2x2, with n+1 rows and columns, where n is the number of subgroups of the population and the extra field is for the sum of the frequency distribution of that subproperties. By convention, in the first row there is the analysed category while in the second row the control one.

The test is called exact because it is possible realize an exact computation of the significance of the deviation from the null hypothesis. The contingency is verified if the proportion of the values not vary significantly among columns and rows, so that the variables are not independent; if it is not verified, the variables are independent. By the contingency matrix two values are calculated: the odds-ratio and the p-value. Since the probability is defined like the number of time for which an event is observed over all the possible events:

$$Prob = \frac{number\ of\ observed}{total} \tag{2}$$

the odd is defined like the amount of occurrences for the event over the time in which the occurrences are not verified

$$Odd = \frac{number\ of\ occurrences}{numer\ of\ not\ occurrences} = \frac{number\ of\ occurrences}{1 - numer\ of\ occurrences} \tag{3}$$

The odds-ratio is defined, like the term says, as the ratio of the odds of the classes and, in this case:

$$Odd\ of\ mutated = \frac{a}{c} \tag{4}$$

$$Odd\ of\ non\ mutated = \frac{b}{d} \tag{5}$$

$$OddsRatio = \frac{odd\ of\ mutated}{odd\ of\ non\ mutated} = \frac{a}{c} / \frac{b}{d} = \frac{ad}{bc} \tag{6}$$

In the case in which the classes are perfect balanced the odds-ratio is equal to one; the complete domain of the odds-ratio goes from 0 to $+\infty$.

The p-value, that is defined as the probability of obtaining a result "more extreme" than the observed one, assuming the null hypothesis (H_0) as true, indicates a statistically significant difference between groups and, in this test, it is computed following the hypergeometric distribution:

$$p = \frac{\binom{a+b}{b}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!\,a!\,b!\,c!\,d!} \tag{7}$$

The range of the p-value, instead, goes from 0 to 1.

For understanding the statistical significance of the p-value is chosen a cut-off value α that is usually 5%.

When the p-value is under the set threshold, it is possible conclude that the event is statistical significant and the null hypothesis can be rejected. In the opposite case, the event is not statistical significant respect the control data so the null hypothesis cannot be rejected. However, in the first part of this project, there is not interest to the null hypothesis but the p-value is used for ranking the list of genes in the set.

It is possible compute three different p-values. The typical situation involves the computation of the two-tailed p-value; however, this is a good situation if the control distribution tends to follow the Gaussian curve. For the data that will be analysed here, this is not always true so the one-tailed p-value is calculated. Since the interest is focusing on the genes which are significantly more mutated in case than control, the p-value is calculated for the upper part of the tail.

Once the ranked genes list is computed by means of p-value outcomes, it is interesting observe the differences among the lists by means of the correlation, a class of the statistical relationships that measures the association of two random variables returning a correlation coefficient.

The correlation coefficient provides an estimation of the similarity between two samples. Assuming linearity, the maximum similarity corresponds to maximum correlation coefficient (+1) whereas a correlation coefficient 0 represents the minimum level of similarity. The null hypothesis corresponds to the uncorrelation between the two dataset.

In this project, the correlation coefficient rho (ρ) from Spearman's and tau (τ) from Kendall-Tau correlation are analyzed because they are more robust to the outliers. Both of the tests are nonparametric.

Both of them return similar results, but the first test, based on the deviations of the data, gives usually higher coefficient that the second one, based on discordant pairs.

To provide an intuitive representation of the results, several plots are drawn. The boxplot and the histogram plots were used to represent different

distributions.

To evaluate the nature of the distributions, the nonparametric Kolmogorov-Smirnov test is used since it is sensitive for changing both in location and in shape of the two empirical distribution. More the p-value is near 0, more the two distributions are dissimilar. Thus, when the p-value is under 5%, it is possible reject the hypothesis that the two distributions come from the same experiment (null hypothesis).

VIII Results.

The results of *ContrastRank* in the original paper bring a classification about three typologies of cancer from which one is the COAD.

The algorithm of this thesis, carrying some modifications, was at the beginning used to emulate it; the Table 4 shows the first five genes, with relative scores, obtained by the two algorithms about the nonsynonymous category. Even if the scores are slightly far, due to the different way in which the p-value was computed and the way in which the allele frequencies was retrieved, all the first five genes are in common; the only difference is the change in position among KRAS and TP53 genes. The correlations among the lists show high score both with Spearman (0.713) and Kendall-Tau (0.551) procedures.

Since the Broad dataset was not used in the original paper, it is not directly compared. However, it is compared against the classification of the BCoM dataset with the new method.

Table 4. First five genes of the BCoM ranked lists that have nonsynonymous effects.

	ContrastRank	**Thesis project**
# of genes	18,537	17,005
1st gene	KRAS	TP53
score	72.62	35.17
2nd gene	TP53	KRAS
score	63.73	31.43
3rd gene	PIK3CA	PIK3CA
score	39.43	20.02
4th gene	BRAF	BRAF
score	29.89	10.16
5th gene	RYR2	RYR2
score	12.90	9.38

The scores of *ContrastRank* project (center column) are computed using a binomial distribution whereas the ones of this thesis project (righter column) was computed by means of Fisher's Exact test. All of them and rounded to the second digit. All the first five genes are in common among the methods. The first row count the amount of genes in each dataset.

The Table 5 shows not only the Broad ranked list against the BCoM one about the nonsynonymous but even the scores obtained considering only the synonymous mutations and the ones obtained including all the functional effects.

To observe the differences among the three ranked lists and the strength of the method, both Spearman and Kendall-Tau correlations are computed in the following ways: firstly, for both of the TCGA datasets, in the same set among synonymous against nonsynonymous and functional, secondly, among the two datasets about the same effect; the result are in Table 6. Since to compute the correlations the list have to be of the same size, the genes not present in both of the sets are excluded.

Table 5. First five genes of the TCGA ranked lists.

	BCoM			**Broad**		
	SYN	**NSYN**	**FUNC**	**SYN**	**NSYN**	**FUNC**
# of genes	14,813	17,005	17,405	16,703	18,012	18,191
1st gene	DCHS2	TP53	APC	TTN	KRAS	APC
score	3.37	35.17	52.46	20.09	67.76	116.30
2nd gene	CRMP1	KRAS	TP53	MUC16	TP53	TP53
score	3.37	31.43	45.53	12.20	63.05	90.94
3rd gene	ZFHX4	PIK3CA	KRAS	FAT3	PIK3CA	KRAS
score	3.15	20.02	31.43	11.59	46.87	67.76
4th gene	TTN	BRAF	PIK3CA	PCDH17	TTN	PIK3CA
score	3.11	10.16	20.39	9.74	33.20	47.26
5th gene	PIK3CG	RYR2	BRAF	OBSCN	RYR2	TTN
score	2.75	9.38	11.17	8.32	24.28	39.12

Divided by the two different set, for each set it is computed the score of the first five genes using the Fisher's Exact test among the classes tumor and normal. The three different list are obtained considering only the synonymous mutation (SYN), only the nonsynonymous (NSYN) and all the mutations excluding the synonymous ones (FUN). For each category, the score is computed as $-\log_{10}(\text{p-value}_{1\text{tail-greater}})$ and rounded to the second digit. The first row count the amount of genes in the dataset.

For both of the TCGA datasets is analysed the behavior of the scores coming from the nonsynonymous mutations lists as represented in Fig. 9. There are not represented the score higher than 30 for appreciable reasons; anyway, just two scores are excluded from the BCoM and three from Broad dataset.

Table 7 lists the scores of the genes which are not displayed in the figure. The most amount of genes have a score lower than 10 for the first dataset and less than 20 for the second one.

Table 6. List of scores computed by different correlation methods among different classes of TCGA datasets.

Set	Method	1st class	2nd class	Score
BCoM	Spearman	Synonymous	Non synonymous	0.346
	Spearman	Synonymous	Functional	0.360
	Kendall-Tau	Synonymous	Non synonymous	0.239
	Kendall-Tau	Synonymous	Functional	0.248
Broad	Spearman	Synonymous	Non synonymous	0.533
	Spearman	Synonymous	Functional	0.536
	Kendall-Tau	Synonymous	Non synonymous	0.374
	Kendall-Tau	Synonymous	Functional	0.376
BcoM vs Broad	Spearman	Synonymous	Synonymous	0.678
	Spearman	Non synonymous	Non synonymous	0.732
	Spearman	Functional	Functional	0.747
	Kendall-Tau	Synonymous	Synonymous	0.496
	Kendall-Tau	Non synonymous	Non synonymous	0.540
	Kendall-Tau	Functional	Functional	0.553

Both of the Spearman and Kendell-Tau correlations were used to analyse the behavior of the lists. In particular, firstly for the BCoM set and than for the Broad one, they were computed among the list of the synonymous against the nonsynonymous and than against the functional mutations. In the the end the same classes of effects were analysed comparing directly the two datasets. The scores are rounded to the third digit.

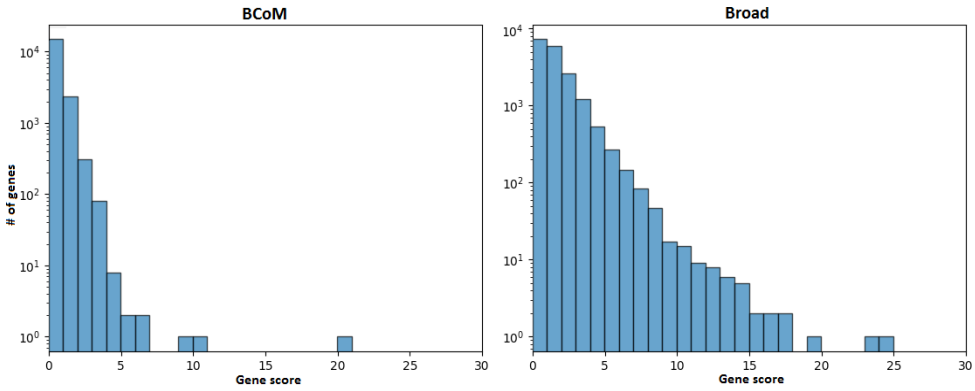


Fig 9. Distribution of the scores coming from the ranking lists of the nonsynonyomous variants for the BCoM and the Broad. The x-axis, represents the prioritization scores of the gene. The y-axis is count the amount of genes with that score.

Table 7. Score coming out from algorithm for nonsynonymous mutations for genes with score higher than 10 for the BCoM and higher than 30 for the Broad.

BCoM (Score > 10)	Broad (Score > 30)
APC: 35.17	APC: 67.76
TP53: 31.43	TP53: 63.05
PIK3CA: 20.02	PIK3CA: 46.87
BRAF: 10.16	TTN 33.20

Here are represented just the genes with score higher than 10 for the BCoM and than 30 for the Broad dataset.

The results obtained by using the 1000 Genomes dataset like control class it is not directly reported because the huge amount of samples present in the set makes most of the scores approximately 0 so the computation of the scores become not appreciable. Since the smallest dataset was the BCoM ones, a different strategy was adopted: randomly, 220 samples from the Broad and the 1000 Genomes were chosen and they are used to compare both BCoM and Broad scores against the 1000 Genomes ones.

The shuffling was repeated 100 times. In the end, among all the shuffles, to each gene is assigned the mean one-tailed greater p-value. For each shuffle, the scores are computed against the normal control samples in order to

make a comparison. Obviously, the randomization in the BCoM against the normal control does not produce any difference while in the Broad small fluctuations are observed in the ranking.

For each category and both datasets, the first five genes in the ranked lists against the 1000 Genomes control samples are illustrated in the Table 8.

Table 8. First five genes of the TCGA ranked lists after 1000 Genomes comparison.

	BCoM			Broad		
	SYN	**NSYN**	**FUN**	**SYN**	**NSYN**	**FUN**
# of genes	14,983	17,160	17,536	9,962	14,139	14,923
1st gene score	KRTAP5-5 38.08	PRB4 81.10	ZNF806 114.48	ZNF806 20.25	KRAS 32.87	APC 45.40
2nd gene score	MUC6 34.46	MUC6 77.06	CBWD6 103.93	COX10 5.68	FRG1 31.11	TP53 40.83
3rd gene score	SPATA3 34.16	KRTAP9-1 71.63	PRB4 80.95	IFNA10 5.56	TP53 27.67	KRAS 33.26
4th gene score	KRT4 32.94	KIAA0040 56.92	MUC6 77.70	KRT6B 5.36	CBWD1 26.18	FRG1 32.69
5th gene score	DSPP 32.79	HGC6.3 38.21	KRTAP9-1 71.40	KRTAP5-5 4.74	ZNF806 22.92	CBWD1 26.79

Divided by the two different set, for each set it is computed the score of the first five genes using the Fisher's Exact test among the classes tumor and normal and 1000 Genomes. The set were shuffled in subset of 220 elements and each time was computed the p-value for each gene. At the end, the mean of the p-value was considered to compute the score. The three different list are obtained considering only the synonymous mutation (SYN), only the nonsynonymous (NSYN) and all the mutations excluding the synonymous ones (FUN). For each category, the score is computed as $-\log_{10}(\text{p-value}_{1\text{tail-greater}})$ and rounded to the second digit. The first row count the amount of genes in the dataset.

The last step is a normalization procedure in which, for each ranked gene with both and normal and 1000 Genomes sample as control, is assigned the minimum of its scores. The final ranking list is reported in the Table 9.

Table 9. First five genes of the TCGA ranked lists after normalization procedure.

	BCoM			**Broad**		
	SYN	**NSYN**	**FUN**	**SYN**	**NSYN**	**FUN**
# of genes	14,813	17,005	17,405	4,215	9,723	11,075
1st gene	NEFH	TP53	APC	MLLT3	KRAS	APC
score	1.23	31.51	52.46	1.60	32.35	45.40
2nd gene	CUZD1	KRAS	TP53	NUDT	TP53	TP53
score	1.11	31.43	41.46	1.48	27.67	40.83
3rd gene	OR4L1	PIK3CA	KRAS	ALG13	PIK3CA	KRAS
score	0.99	20.02	31.43	1.44	21.33	32.81
4th gene	ADAMTS2	BRAF	PIK3CA	MAML2	BRAF	PIK3CA
score	0.97	9.51	20.39	1.31	8.76	21.30
5th gene	ZFHX4	RYR2	BRAF	XPOT	RYR2	BRAF
score	0.90	7.11	10.13	1.24	6.79	8.82

For each gene in common among the lists obtained using the normal samples and the 1000 Genomes samples, the minimum score is assigned.

In the analysis performed using the *PhD-SNP*^{g} predictions, it is possible to observe little changes in the gene ranking (Table 10). Indeed some genes, for which no pathogenic variants are predicted, were ranked in a lower position. The scores reported in Table 11 shows low correlation for each comparison between synonymous and ranking lists based on different type of variants (nonsynonymous, functional).

Table 10. First five genes of the TCGA ranked lists after the prediction by PhD-SNP^{g}.

	BCoM			**Broad**		
	SYN	**NSYN**	**FUNC**	**SYN**	**NSYN**	**FUNC**
# of genes	14,340	15,070	16,023	16,703	18,012	18,191
1st gene	CRMP1	TP53	APC	TTN	KRAS	APC
score	3.37	33.84	58.18	18.98	66.46	123.66
2nd gene	TTN	KRAS	TP53	MUC16	TP53	TP53
score	3.12	31.00	44.05	11.71	62.57	90.69
3rd gene	ZFHX4	PIK3CA	KRAS	FAT3	PIK3CA	KRAS
score	2.77	20.01	31.00	11.10	46.87	66.46
4th gene	OBSCN	BRAF	PIK3CA	PCDH17	TTN	PIK3CA
score	2.75	10.16	20.39	8.87	31.72	47.26
5th gene	DCHS2	RYR2	BRAF	OBSCN	FAT4	TTN
score	2.72	8.51	11.17	8.23	20.80	39.06

Divided by the two different set, for each set it is computed the score of the first five genes using the Fisher's Exact test among the classes tumor and normal. The three different list are obtained considering only the synonymous mutation (SYN), only the nonsynonymous (NSYN) and all the mutations excluding the synonymous ones (FUN). For each category, the score is computed as $-\log_{10}(\text{p-value}_{1\text{tail-greater}})$ and rounded to the second digit. The first row count the amount of genes in the dataset. For the calculation of the prioritization score the synonymous variants predicted pathogenic and the functional mutations predicted benign are not considered.

To estimate the effect of the allele frequency threshold on of the gene prioritization score, the previous scores (allele frequency threshold 0.5%) are compared with those obtained with thresholds 0.1%, 1% and 10%. Table 12 shows the Kendall-Tau and Spearman correlation scores among the genes with a score higher or equal then 3 (in common with 0.5% list) in the gene list obtained from functional variants in the BCoM dataset.

Table 11. Correlation score between different classes of TCGA datasets after the prediction by PhD-SNP[g].

Set	Method	1st class	2nd class	Score
BCoM	Spearman	Synonymous	Non synonymous	0.304
	Spearman	Synonymous	Functional	0.309
	Kendall-Tau	Synonymous	Non synonymous	0.011
	Kendall-Tau	Synonymous	Functional	0.011
Broad	Spearman	Synonymous	Non synonymous	0.486
	Spearman	Synonymous	Functional	0.484
	Kendall-Tau	Synonymous	Non synonymous	0.040
	Kendall-Tau	Synonymous	Functional	0.041
BcoM vs Broad	Spearman	Synonymous	Synonymous	0.620
	Spearman	Non synonymous	Non synonymous	0.716
	Spearman	Functional	Functional	0.734
	Kendall-Tau	Synonymous	Synonymous	0.073
	Kendall-Tau	Non synonymous	Non synonymous	0.116
	Kendall-Tau	Functional	Functional	0.125

Both of the Spearman and Kendall-Tau correlations were used to analyse the behavior of the lists. In particular, for the BCoM set and then for the Broad one, they were computed among the list of the synonymous against the nonsynonymous and then against the functional mutations. In the the end the same classes of effects were analysed comparing directly the two datasets. The scores are rounded to the third digit.

Table 12. Correlation scores about the common genes at different threshold with score threshold of 3.

	0.1%	1%	10%
# of genes	294	134	73
Spearman	0.71	0.90	0.84
Kendall-Tau	0.54	0.81	0.75

The scores coming out the Spearman (S) and the Kendall-Tau (KT) correlations between the ranked list of common genes with score higher or equal than 3 in the functional category of the BCoM dataset. The ranked list obtained using the allele frequency threshold of 0,1%, 1% and 10% are compared with the ranked list of the project with the allele frequency threshold set to 0.5%. The first row count the number of common genes among the list of the project and the selected list.

The scores are computed without the score filter of 3 (Table 13). Obviously, the more higher is the threshold, the more genes are included in the list. Nevertheless, the effect of threshold it is not significant in the range of 0.1% and 1% since the correlation scores against the 0.5% value indicatively highlight a strong correlation. However, if the threshold grows, the amount of mutations not involved in the disease increases.

Table 13. Correlation scores about the common genes at different threshold.

	0.1%	1%	10%
# of genes	17,066	17,405	17,405
Spearman	0.98	0.99	0.93
Kendall-Tau	0.91	0.94	0.80

The scores coming out the Spearman (S) and the Kendall-Tau (KT) correlations between the ranked list of all common genes in the functional category of the BCoM dataset. The ranked list obtained using the allele frequency threshold of 0,1%, 1% and 10% are compared with the ranked list of the project with the allele frequency threshold set to 0.5%. The first row count the number of common genes among the list of the project and the selected list.

An other consideration is made by means of ClinVar. This database classifies the mutations in different ways among which benign or pathogenic. All the mutations involved in the change of a nucleotide against just an alternative allele and classified like benign or pathogenic are selected. Since not only the mutations have a reported allele frequency, if it is not present it is considered like 0.0%. Using the same previous thresholds, four contingency tables are computed with the classes pathogenic and benign against the features of the thresholds. The Fisher's Exact test is used to compute the p-value but for all the tables, that are strongly unbalanced, it is 0 so the values of the table and the relative $\log_{10}$(odd-ratio) (LOR) are reported in the Table 14. The consideration made before is still good even if, using just ClinVar, the best separation occurs with an allele frequency of 0.1% even if the LOR it is not significantly distant to the 0.5% and 1% frequencies.

Table 14. Log Odd-Ratio scores of four contingency table at different threshold about ClinVar

	0.5%	0.1%	1%	10%
P-U	13,345	13,305	13,347	13,369
P-O	43	83	41	19
B-U	4,057	2,184	5,149	8,140
B-O	9,193	11,066	8,101	5,110
LOR	2.85	2.91	2,71	2,65

The contingency table is realized by the classes pathogenic and benign of the mutations in ClinVar in which just a nucleotide is reported like having just an alternative allele. The different used threshold, in the first row, divide the characterize the category of the contingency table with P-U (pathogenic and frequency under the threshold), P-O (pathogenic and frequency under the threshold), B-U (benign and frequency under the threshold), B-O (benign and frequency under the threshold). The LOR row stay for Log Odd-Ratio ($\log_{10}$(odd-ratio)) and it was preferred to the p-value because it was 0 for each table.

The Table 15 and the Table S22 are reported the mutations or the genes, after the prediction, classified in the four category:

1) benign vs pathogenic
2) germline vs somatic
3) TSG vs oncogene
4) TSG and oncogene vs all the remaining genes

For each category, for both of the TCGA datasets and for the synonymous, nonsynonymous and functional mutations, the plot is computed. Depending on the categories, many of them result pretty similar so in the main text are reported the most representative ones: one plot about the category 4 of the score (Fig. 10), one plot for the conservation in category 1 (Fig. 11) and one in category 2 (Fig. 12) and one about the occurrence in the category 4 (Fig. 13). It is quite intuitive the behavior of the classes in the different categories just looking at the Table 16 and the Table S23 that show the p-values for each of the classes in the plot computing by the Kolmogorov-Smirnov test for the BCoM and the Broad set, respectively.

Table 15. Amount of mutations or genes involved in the distribution computed after PhD-SNP^g prediction about the BCoM set.

Category	Effect	1st class	2nd class	Total
1	Synonymous	90%	10%	54,711
	Nonsynonymous	37%	63%	108,343
	Functional	33%	67%	124,882
2	Synonymous	60%	40%	54,711
	Nonsynonymous	49%	51%	108,343
	Functional	45%	55%	124,882
3	Synonymous	56%	44%	1,206
	Nonsynonymous	58%	42%	2,334
	Functional	62%	38%	2,957
4	Synonymous	3%	97%	54,711
	Nonsynonymous	3%	97%	108,343
	Functional	3%	97%	124,882

The number in column "Category" are "benign vs pathogenic" (1), germline vs somatic (2), TSG vs oncogene (3), TSG and oncogene vs all the remaining genes (4). The 1st class column represent the percentage of gene in the first class for each category while 2nd class column is about the second class. The total column shows how many values are into the distribution, both for synonymous, nonsynonymous and functional effects.

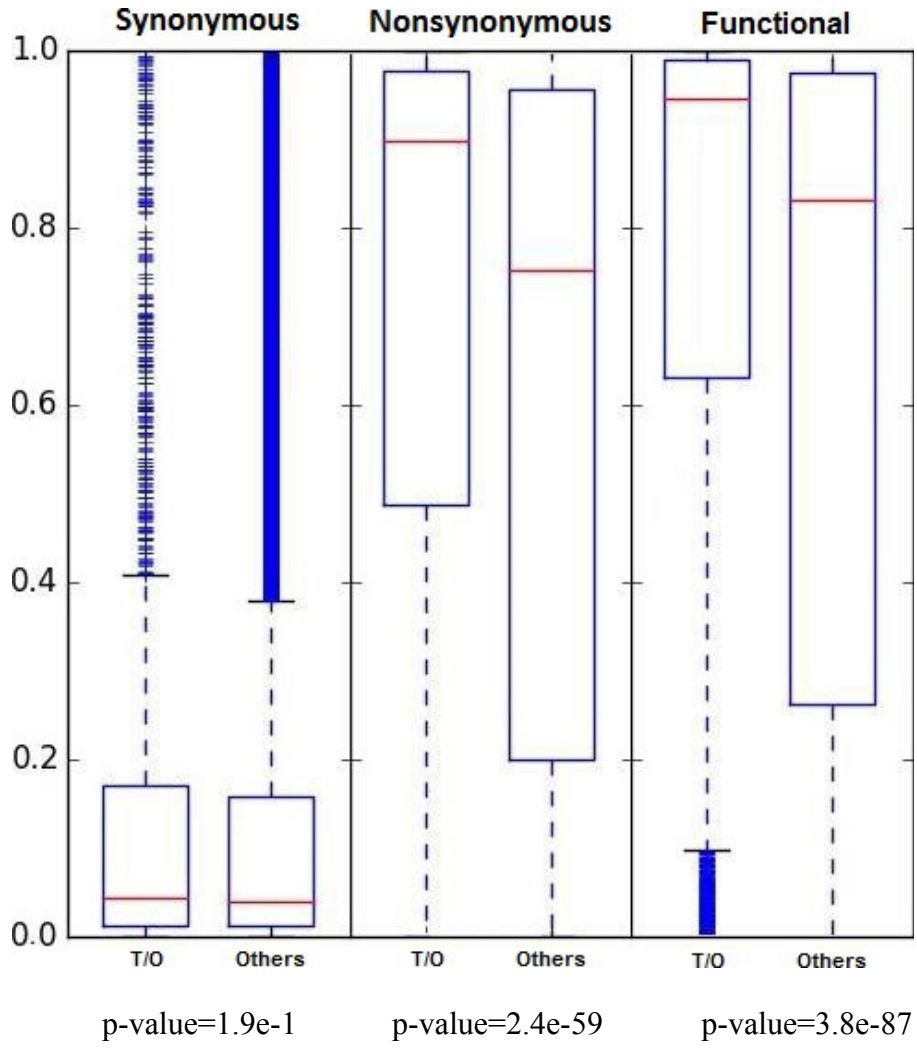


Fig 10. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the TSGs and oncogenes ("T/O", tthe left plot in each panel) and all the genes not mapped in COSMIC dataset or mapped using other terms ("Others", the right plot in each panel). The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

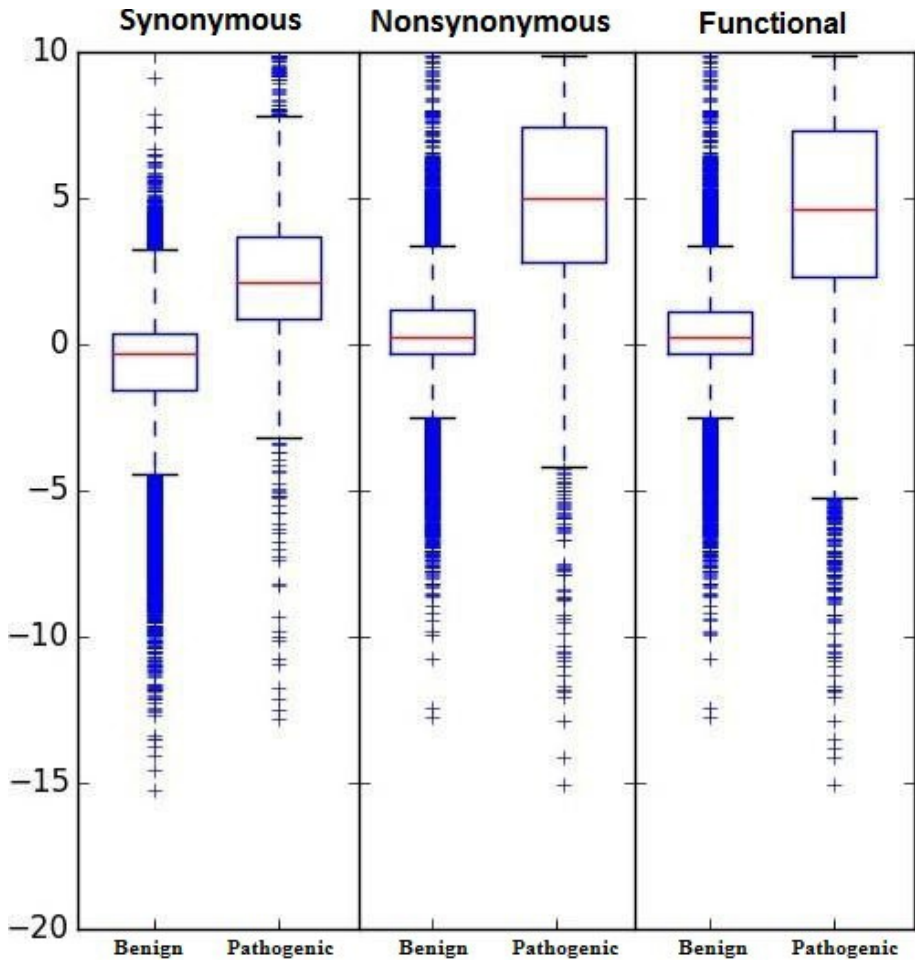


Fig 11. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the benign (the left plot in each panel) and the pathogenic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

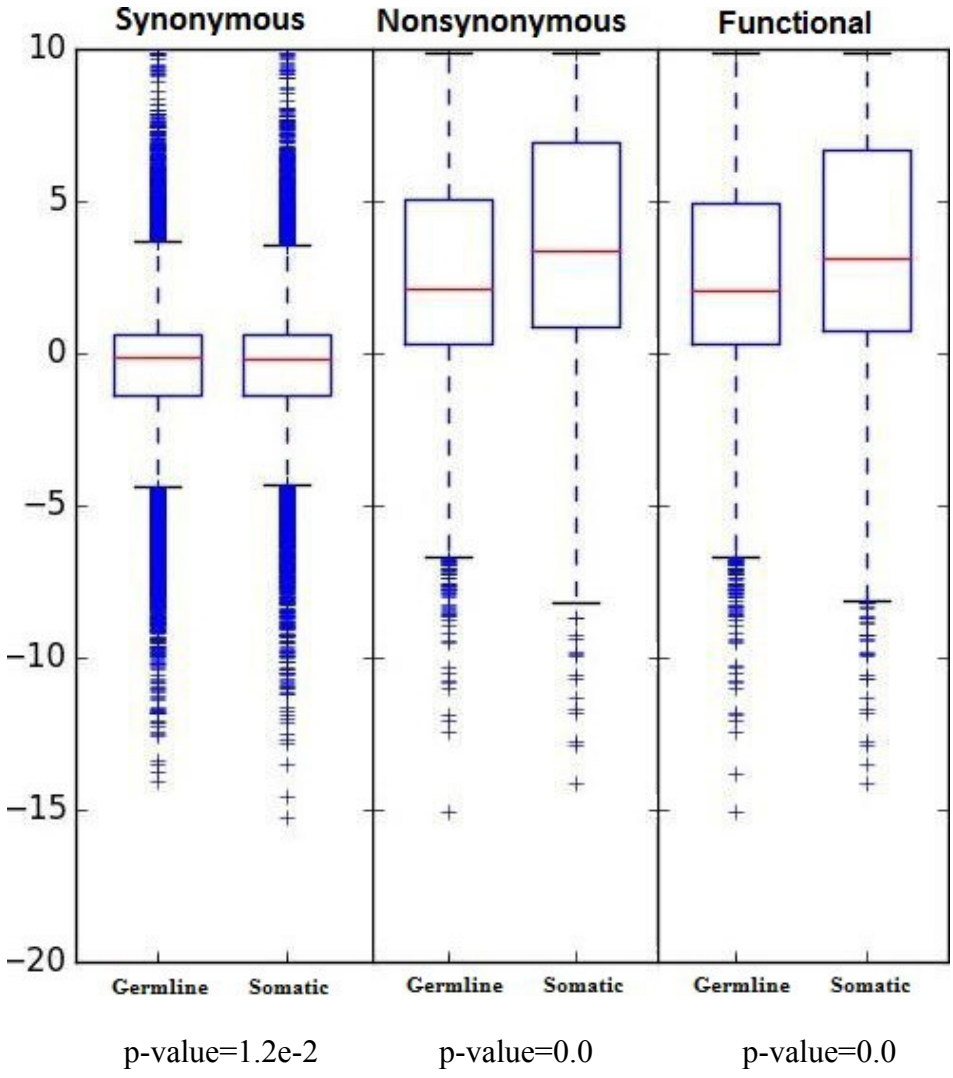


Fig 12. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the germline (the left plot in each panel) and the somatic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov betweenthe distributions of the two classes.

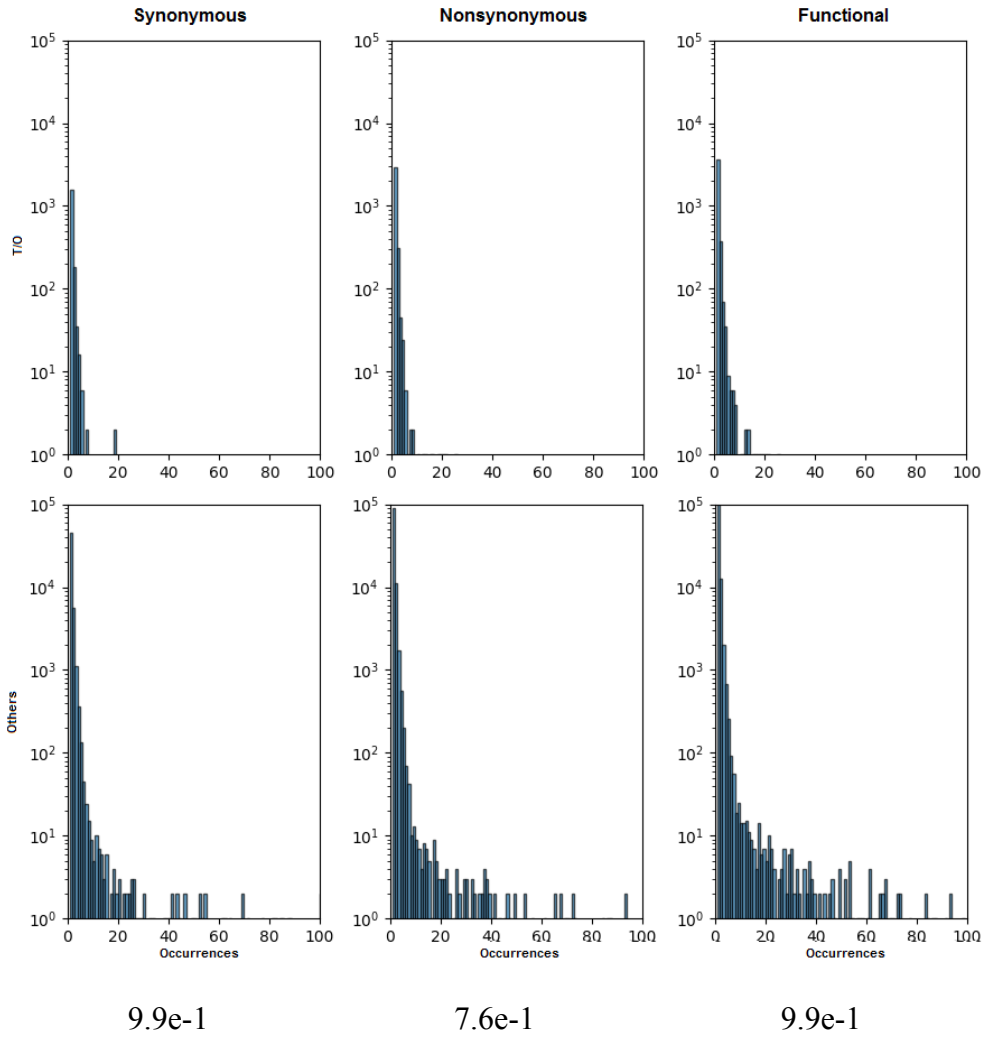


Fig 13. Histograms of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the TSGs and oncogenes ("T/O", the plots in the first row) and all the genes not mapped in COSMIC dataset or mapped using other terms ("Others", the plots in the second row). To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

Table 16. P-value of distribution about the prediction of PhD-SNP^g in BCoM dataset.

Prediction	Category	Effect	p-value
Score	1	Synonymous	**0.0**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	2	Synonymous	**8.5^{-3}**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	3	Synonymous	6.8^{-1}
		Nonsynonymous	9.2^{-1}
		Functional	**4.7^{-3}**
	4	Synonymous	1.9^{-1}
		Nonsynonymous	**2.4^{-59}**
		Functional	**3.8^{-87}**
Conservation	1	Synonymous	**0.0**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	2	Synonymous	**1.2^{-2}**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	3	Synonymous	9.5^{-1}
		Nonsynonymous	1.9^{-1}
		Functional	7.5^{-2}
	4	Synonymous	**6.9^{-8}**
		Nonsynonymous	**1.7^{-68}**
		Functional	**6.9^{-82}**
Occurrence	1	Synonymous	3.3^{-1}
		Nonsynonymous	**6.1^{-12}**
		Functional	**1.2^{-10}**
	2	Synonymous	**9.4^{-128}**
		Nonsynonymous	**5.8^{-113}**
		Functional	**3.1^{-131}**
	3	Synonymous	9.9^{-1}
		Nonsynonymous	6.2^{-1}
		Functional	9.4^{-1}
	4	Synonymous	9.9^{-1}
		Nonsynonymous	7.6^{-1}
		Functional	9.9^{-1}

P-value computed by Kolmogorov-Smirnov test for the PhD-SNP^g prediction scores. The number in column "Category" are "benign vs pathogenic" (1), germline vs somatic (2), TSG vs oncogene (3), TSG and oncogene vs all the remaining genes (4). The 1st class column represent the percentage of gene in the first class for each category while 2nd class column is about the second class. In bold are highlighted the distribution that reject the null hypothesis on a threshold of 0.05.

IX Discussion.

The *ContrastRank* algorithm computes the p-value used for the ranking using a binomial distribution. The reason for which it is preferred the Fisher's Exact test is due to the possibility to assign a value equal to 0 to a class if no sample for a particular category is included. Since it is not possible to assign a background probability 0 in a binomial distribution, a value of 0.005 was used. Table 4, where are present the top ranked genes using the two methods, shows exactly the same prioritized genes even if the TP53 and KRAS are reversed. Furthermore, not even the presence of a large set of databases used to catch the allele frequencies has strong influence. The method can be considered strong under this point of view.

As well, the ranking of the nonsynonymous mutations computed in the Broad dataset, shown in Fig. 5, gives even a strong independence from the variant calling strategy used by the institute, since, the most ranked genes are in the same order of the *ContrastRank* list, but the TTN. The difference with the BCoM dataset, in terms of score, is due to the different size of the two datasets. Indeed, the Broad dataset is more than double of the BCoM one.

In spite of the most of the mutations which might have an impact on the protein function belong to the nonsynonymous category (Table 2), the rank list was computed even for the class of the functional mutations that consider the *nonsynonymous*, *frameshift insertion*, *framemshift deletion*, *stopgain* and *stoploss* variants. As expected, the ranking, for both of the datasets, is quite similar except for a new gene (APC) which showed the highest score. APC gene shows a little number of nonsynonymous modifications respect the other top ranked genes but it is characterized by the other type of functional ones.

Fig. 9 shows that most of the genes have low score and Table 7 highlights

few top ranked genes for which there is no large difference, in the mutation rate, between normal and tumor samples.

The situation is quite different for the synonymous mutations. Actually, there are no genes with high score: consequentially, below this method, it is possible conclude that the synonymous mutations should not have a strong impact on cancer development and progression; furthermore, a gene like the TTN that has an high score in the Broad dataset for the nonsynonymous mutations, which is also highly ranked in the synonymous list is more likely to have no direct relation to the disease.

Following this line of reasoning, once PhD-SNP^{g} predicted all the mutations of the TCGA sets, all the synonymous mutations predicted pathogenic and all the functional mutations predicted benign are excluded. The ranked list is computed against this new datasets and it is reported in Table 9; the top ranking genes are the same for nonsynonymous and functional variants, pretty similar even in the score, whereas the synonymous list is slightly different due to the low scores that allow large fluctuations.

The original ranked list of BCoM for the functional mutations was then matched against the COSMIC dataset. At different thresholds, a Fisher's Exact test was computed to check the ones that better separates the tumor suppressor genes and oncogenes from the other genes. The contingency table considers as classes the number of genes under and over a threshold that are mapped to the list of COSMIC like TSGs and oncogenes. Since the peak of the separation is obtained with a threshold of 1 and a score of 25.69, even the thresholds ranging from 0 to 2 and step of 0.1 are reported in Table S21. The highest score is at a 1.2 (28.34). Anyway, it is possible observe that the score has a fluctuating behavior around the maximum.

Although the better option using the largest cohort available to identify the maximum number of variants, the ranked lists for BCoM and Broad dataset were also computed using the 1000 Genomes samples instead of the control samples. After the shuffling procedure, the Table 8 shows the rank genes

lists. This table could mislead since the most of the genes have very lower score against normal samples. Following the idea that the mutation with an high allele frequency should be not considered, including the TCGA sets the most of high ranked genes obtained against the 1000 Genomes would not be present. Thus, the final score associated to each gene after the normalization procedure is the minimum among the score obtained using the normal and the one obtained using the 1000 Genomes samples and the final ranked list is reported at the Table 9 in which are highlighted with high score the already genes correctly identified.

After a statistical validation of the data based on the frequency of the genes with rare mutations, it is interesting to observe how the mutations are interpreted by a machine learning method (*PhD-SNP*^{g}) based on nucleotides conservation score (PhyloP100).

For each plot of, both the BCoM and the Broad dataset convey a similar behavior so only the BCoM ones are inserted in the main text while the other ones are in the supplementary material.

The boxplot that represents the prediction of the scores it is not so informative for category 1 (Fig. S1 and Fig. S4) since, by definition, all the benign mutations have a score lower than 0.5 (with a mean near to 0 for the synonymous and 0.1 for nonsynonymous and functional) whereas the pathogenic one is higher (with a mean near to 0.7 for synonymous and 0.9 for nonsynonymous and functional).

Less informative is also the plot of category 3 (Fig. S3 and Fig. S6) since their genes show a quite similar behavior.

Very interesting and similar among them are the plot of the categories 2 (Fig. S2 and S5) and 4 (fig 10 and S7) where the classes of “somatic” and “TSG-oncogene” and the classes “germline” and “all the remaining genes” are respectively comparable.

The algorithm assigns a very low score to all the classes of the synonymous mutations because all of them are not considered causative of the disease.

This observation supports the idea of using the synonymous mutations for calculating the background mutation rate. For both the nonsynonymous and functional mutations the “somatic” and “TSG-oncogene” classes have a mean score that is respectively higher than 0.8 and 0.9 respectively, therefore they are more likely related to cancer. For the nonsynonymous and functional mutations, the “germline” and “all the remaining genes” show a possible association to the disease but with score slightly less pronounced (~0.6 and 0.8 respectively).

The boxplots representing the conservation highlight an interesting and expected observation about the category 1 (Fig. 11 and Fig. S10): synonymous, nonsynonymous and functional variants have small conservation values with average near the 0 for the “benign” category while the “pathogenic” variants cover a large range of values with average ~2 for synonymous and 5 for nonsynonymous and functional classes for the BCoM dataset and ~2 for nonsynonymous and 5 for functional variants in the Broad dataset.

The category 2 (Fig. 12 and Fig. S11) does not present a similar situation among the two classes of mutations with an almost perfect conservation in the synonymous one against a slightly lower conservation for the nonsynonymous and functional where the mean of the somatic shows a shift of the plot.

The category 3 (Fig. S8 and Fig. S12) shows a mean close to 0 for the synonymous and ~5 for the nonsynonymous and functional classes. The typical expectations for “TSGs” and “oncogenes” is observed.

The category 4 (Fig. S9 and Fig. S13) highlights a situation similar to the category 2, like in the case of the score plot, with a mean near 0 for the synonymous for both of the classes and 5 for the “TSG-oncogene” and 3 for “all the remaining genes” classes in nonsynonymous and functional cases.

In general, for a genetic variant, the higher the frequency of observation in health subjects, the higher is the probability of not being involved in the

disease. For this reason it is interesting to analyse the occurrences of the mutations, collected in the TCGA datasets, after the *PhD-SNP*^{g} prediction.

If the category 3 of the occurrence (Fig. S16 and Fig. S19) shows similarity between the classes of synonymous, nonsynonymous and functional mutations, category 1 (Fig. S14 and Fig. S17), category 2 (Fig. S15 and Fig. S18) and category 4 (Fig. 13 and Fig. S20) have a very similar situation among "pathogenic" and "somatic and "TSG-oncogene" respectively against "benign" and "germline" and "all the remaining genes" for both the synonymous, nonsynonymous and functional mutations respectively. The first case shows a huge amount of modifications among the samples whereas, in the second case the modifications tend to be unique.

In the Fig. 13 is plotted the distribution of the occurrence for the category 4 of the BCoM dataset for the synonymous, nonsynonymous and functional mutations: the "TSGs-oncogenes" never count more than 30 occurrences of a mutation among the samples whereas "all the remaining genes" count up to 256 occurrences, too.

A different situation is observed for the functional variants in the category 1 for which some mutations occur more frequently (in the Broad dataset this is true also for the nonsynonymous).

The Table 14 shows how many mutations or genes, divided by the four categories, are involved in the distributions and how they are organized among the classes in the BCoM dataset. The organization of the Broad is reported in the Table S22. In the Table 16 and Table S23 there are computed the distance among the distributions of all the plots; the p-value computed by the Kolmogorov-Smirnov test.

Conclusions

The study of the cancer based on the analysis of Single Nucleotide Variants or small insertions and deletions can be used to prioritize genes involved in the disease but this information provides only a partial explanation of the mechanism of the disease that is by far more complex.

First of all it is know that the cancer is directly linked with the genetic expression; as well, the Structural Variants can generate chromosome rearrangement and leading to duplication, deletion, inversion or translocation of big pieces of DNA, aberrations that are hard to discover.

All these modifications can bring even small changes but the main problem is the cascade effect, especially if the altered gene is part of a network in which it is an hub.

In addition, there are technical issues that overlook the nature of the modifications which are related to the annotation.

Trying to retrieve the common alterations is very hard for problems due to the limitations of the variant calling procedures. The detection of all the possible somatic mutations is hindered by the heterogeneity of the tumor among the samples and even among the cell in the same body that can acquire the specific hallmarks during the tumorigenesis.

Nevertheless, the machine learning methods allowed us to identify several possible genetic alterations associated to cancer with higher true positive rate.

It is expected that the increasing amount of data will allow to develop more accurate algorithms for the identification of causative genetic changes of the tumor enabling the implementation of new and personalized treatment strategies.

The new approaches will include more complex layer of information based on the analysis of the mutated networks of genes and the transcription level of each gene in the specific tissues.

Bibliografy

Aken,B.L. *et al.* (2016) The Ensembl gene annotation system. *Database*, 2016, baw093.

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Capriotti,E. *et al.* (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, **22**, 2729–2734.

Capriotti,E. and Fariselli,P. (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Research*.

Casás-Selves,M. and Degregori,J. (2011) How cancer shapes evolution, and how evolution shapes cancer. *Evolution (N Y)*, **4**, 624–634.

Chang,K. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, **45**, 1113–1120.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**, 213–219.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Forbes,S.A. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, **43**, D805–D811.

Hamosh,A. (2004) Online Mendelian Inheritance in Man (OMIM), a

knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**, D514–D517.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Khurana,E. *et al.* (2013) Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science*, **342**, 1235587.

Landrum,M.J. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, **42**, D980–D985.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

McNerney,M.E. *et al.* (2017) Therapy-related myeloid neoplasms: when genetics and environment collide. *Nature Reviews Cancer*, 17, 513–527.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20, 110–121.

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**, D158–D169.

Tian,R. *et al.* (2014) ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics*, **30**, i572–i578.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**, e164–e164.

Glossary

- **Allele frequency** is the relative frequency of an allele (the variant of a gene) at a particular locus, expressed as a fraction or percentage. Specifically, it is the fraction of all chromosomes in the population that carry that allele.
- **Benign tumor** is an abnormal proliferation of cells driven by at least one mutation in an oncogene or tumor suppressor gene. These cells are not invasive (i.e., they cannot penetrate the basement membrane lining them), which distinguishes them from malignant cells.
- **Driver gene** is a gene that contains driver gene mutations (Mut-Driver gene) or is expressed aberrantly in a fashion that confers a selective growth advantage (Epi-Driver gene).
- **Driver gene mutation** is mutation that directly or indirectly confers a selective growth advantage to the cell in which it occurs.
- **Germ cells** is any biological cell that gives rise to the gametes of an organism that reproduces sexually
- **Germline mutation** is mutation that occur in a germ cell.
- **Malignant tumor** is a abnormal proliferation of cells driven by mutations in oncogenes or tumor suppressor genes that has already invaded their surrounding stroma. It is impossible to distinguish an isolated benign tumor cell from an isolated malignant tumor cell. This distinction can be made only through examination of tissue architecture.
- **Metastatic tumor** is a malignant tumor that has migrated away from its primary site, such as to draining lymph nodes or another organ.
- **Metastasis** is the process whereby cancer cells leave their tissue

of origin and establish new tumors in additional sites in the same or other organ.

- **Mutation** is a random change in genetic makeup.
- **Nonsynonymous mutation** is a mutation that alters the encoded amino acid sequence of a protein. These include missense, nonsense, splice site, translation start, translation stop, and indel mutations.
- **Oncogene** is a gene whose dysregulation or mutational activation can contribute to cancer development (i.e. cancer promoting genes). In tumor cells, it is often mutated or expressed at high levels. A gene that, when activated by mutation, increases the selective growth advantage of the cell in which it resides.
- **Oncogenic mutation** is a change in the DNA code of a normal cellular gene that creates an oncogene or tumor suppressor gene.
- **Primary tumor** is the original tumor at the site where tumor growth was initiated. This can be defined for solid tumors, but not for liquid tumors.
- **Proto-oncogene** is a normal gene that could become an oncogene due to mutations or increased expression. It codes for proteins that help to regulate cell growth and differentiation.
- **Somatic cell** is any biological cell forming the body of an organism; that is, in a multicellular organism, any cell other than a gamete, germ cell, gametocyte or undifferentiated stem cell.
- **Somatic mutation** is mutation that occur in a somatic cell, such as those that initiate tumorigenesis.
- **Telomeres** are structures at the ends of chromosomes which protect these ends. Telomeres shorten with each cell division unless maintained by an enzyme called telomerase.
- **Tumor suppressor gene** is a gene whose inactivation, such as

by mutation, can contribute to cancer development (i.e. cancer suppressive genes). It is a gene that, when inactivated by mutation, increases the selective growth advantage of the cell in which it resides.

Supplementary materials.

All the figures, tables and files that can be interesting to observe and that are not present in the main text are listed following. Furthermore, a *readme* web page at the address https://github.com/LuigiChiricosta/Detecting-cancer-causing-genes-and-variants-in-Colon-Adenocarcinoma/blob/master/README.md list all the possible information related to the project that cannot be inserted even in the supplementary section.

Supplementary figures

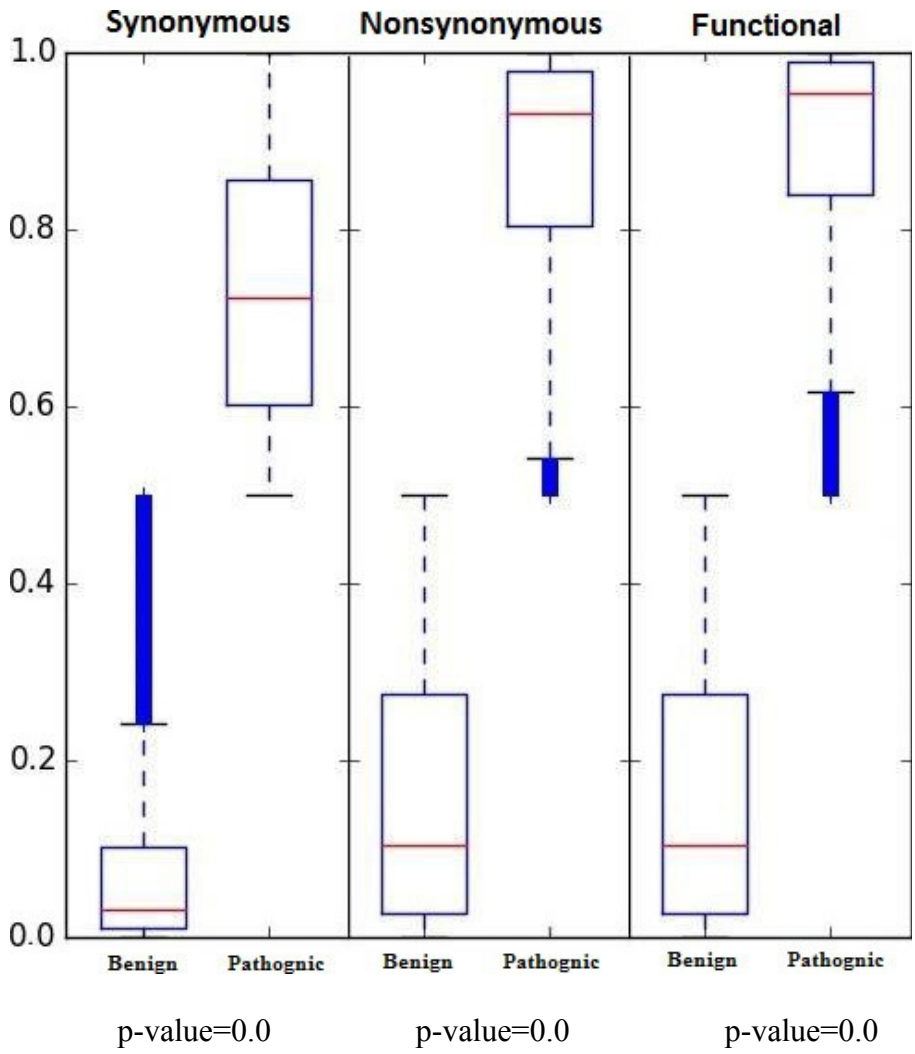


Fig S1. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the benign (the left plot in each panel) and the pathogenic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

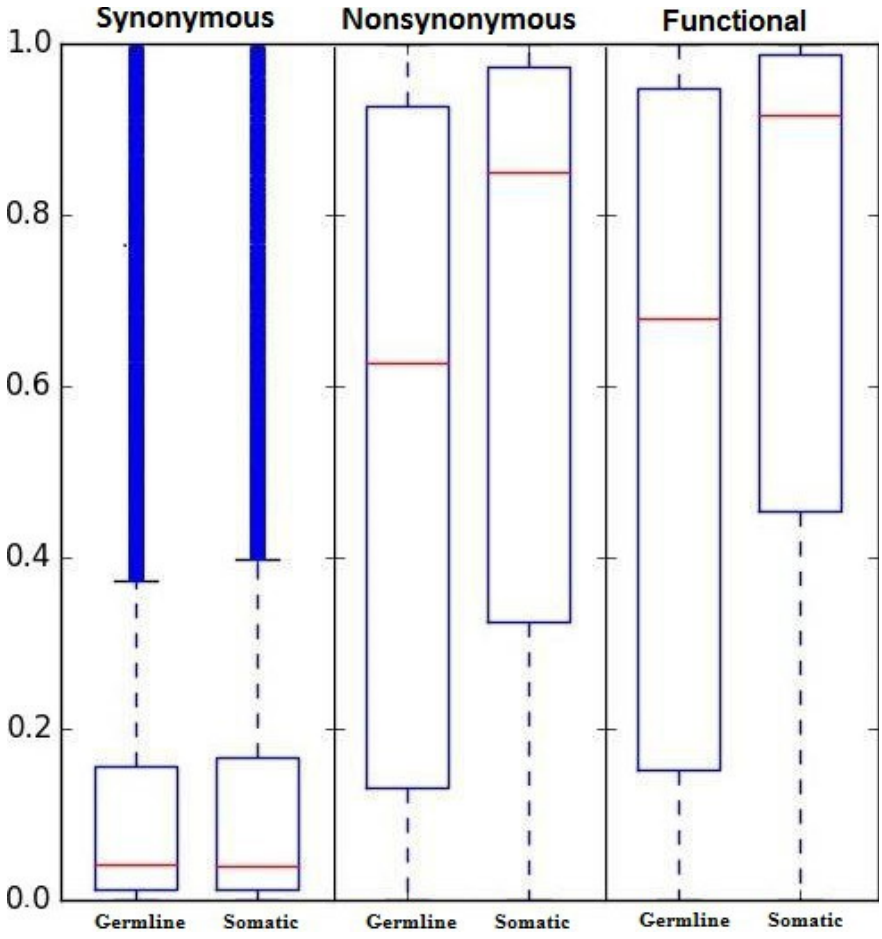


Fig S2. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the germline (the left plot in each panel) and the somatic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

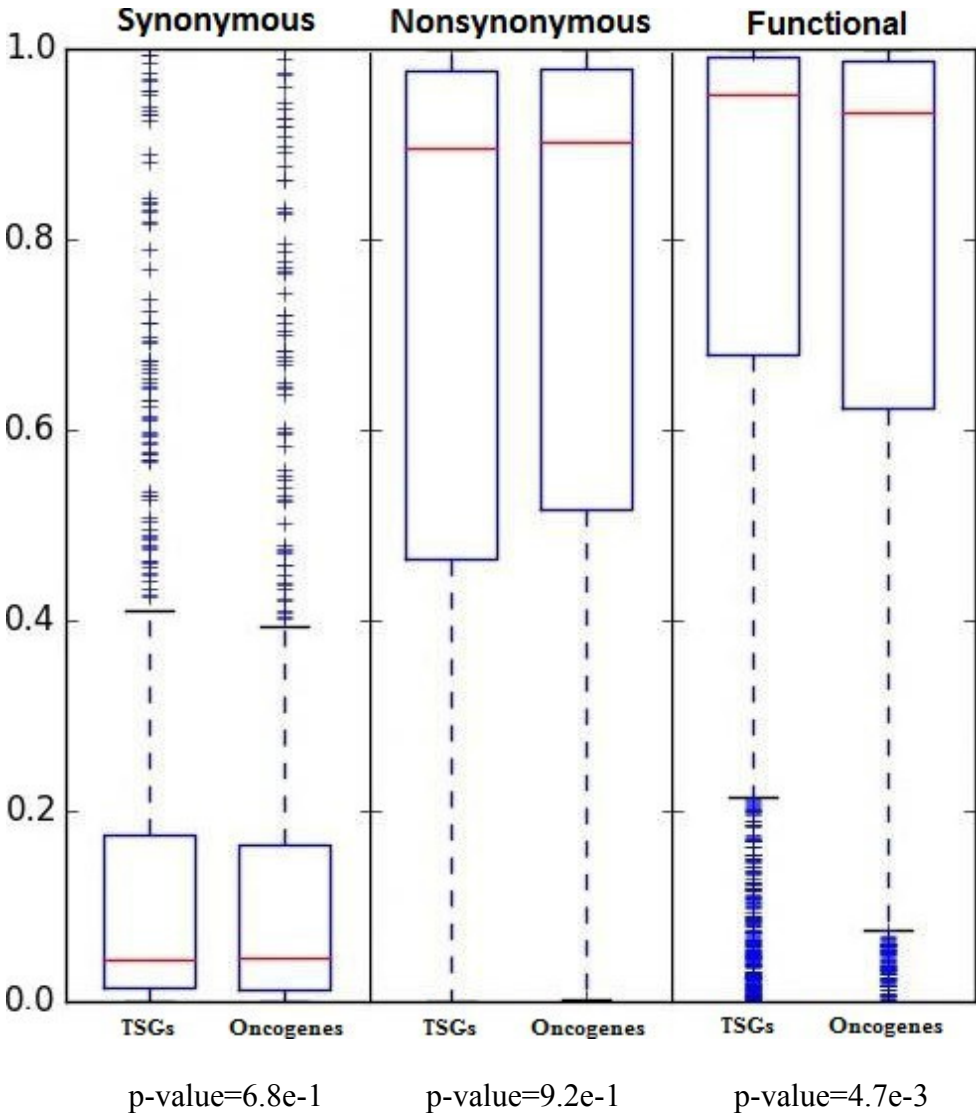


Fig S3. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the TSGs (the left plot in each panel) and the oncogenes (the right plot in each panel) like mapped in COSMIC dataset. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

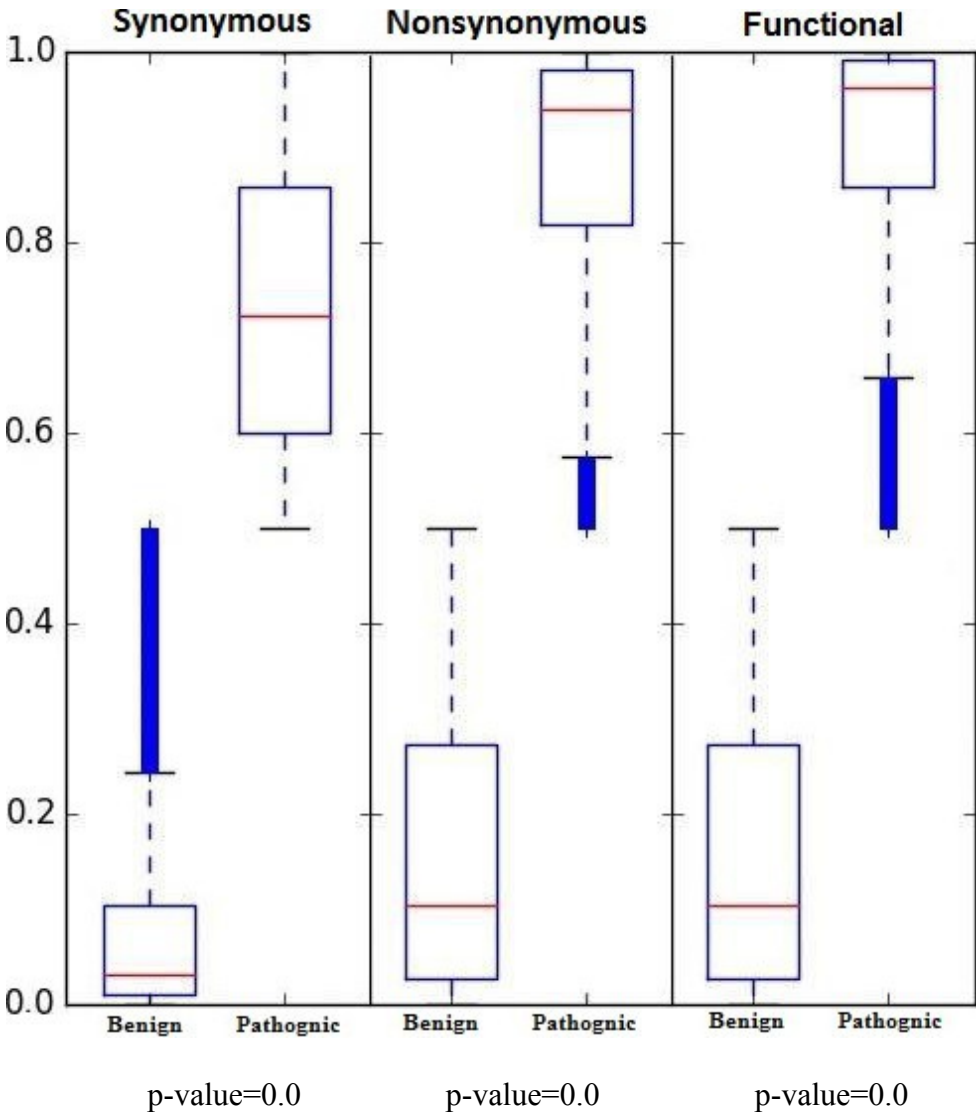


Fig S4. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the benign (the left plot in each panel) and the pathogenic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

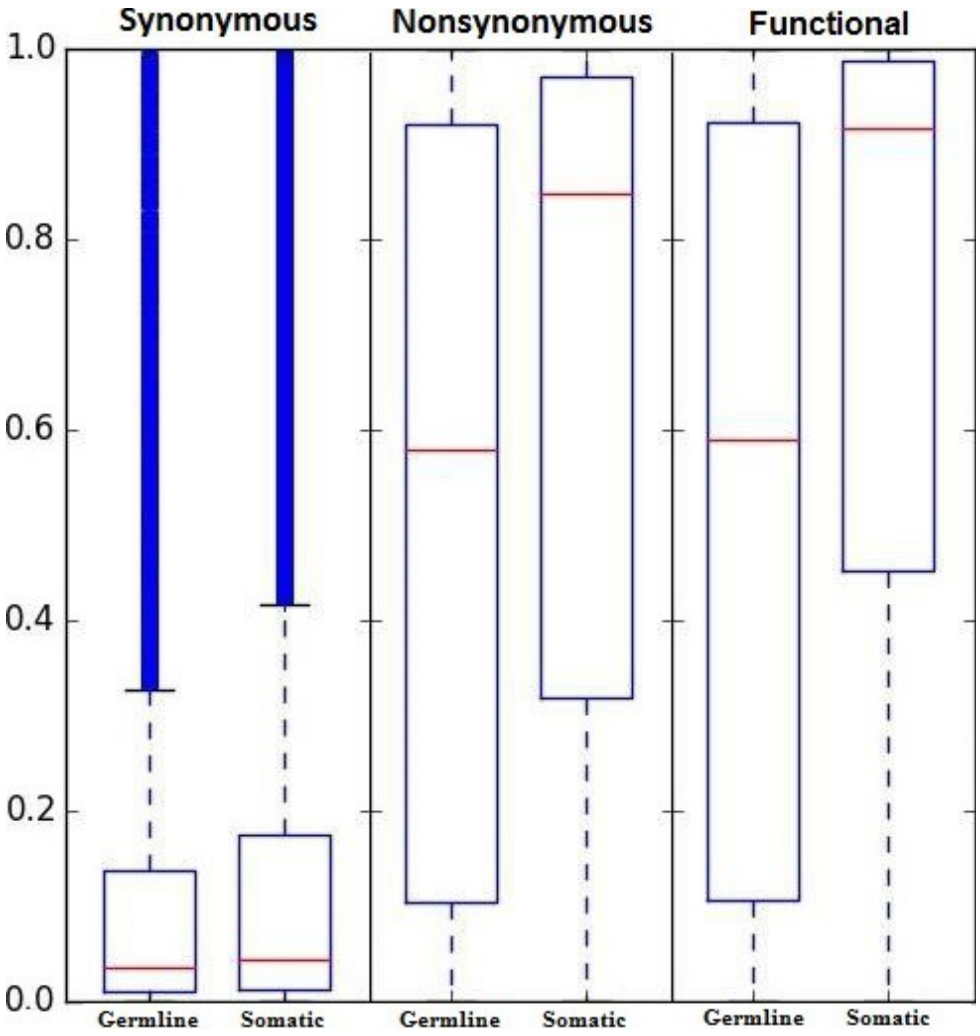


Fig S5. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP[g], is computed among the germline (the left plot in each panel) and the somatic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

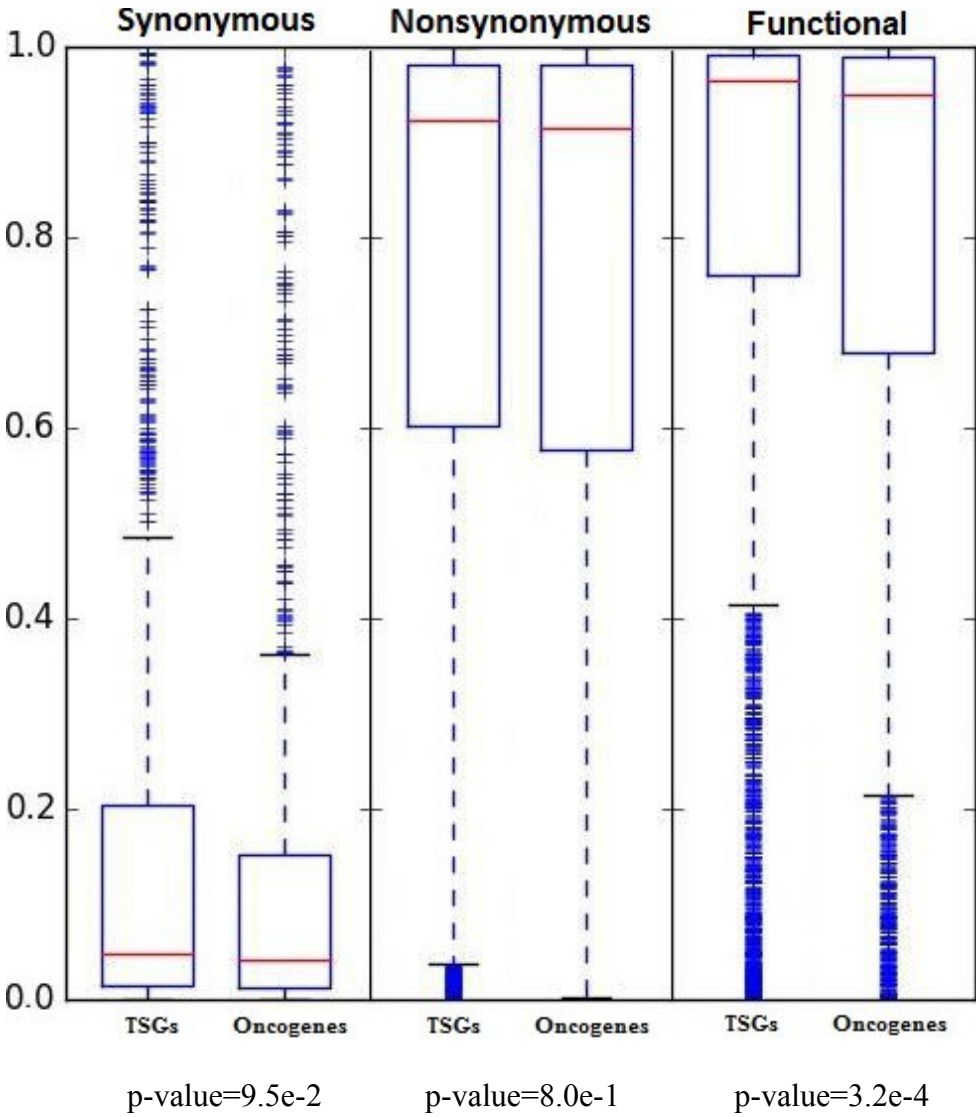


Fig S6. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the TSGs (the left plot in each panel) and the oncogenes (the right plot in each panel) like mapped in COSMIC dataset. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

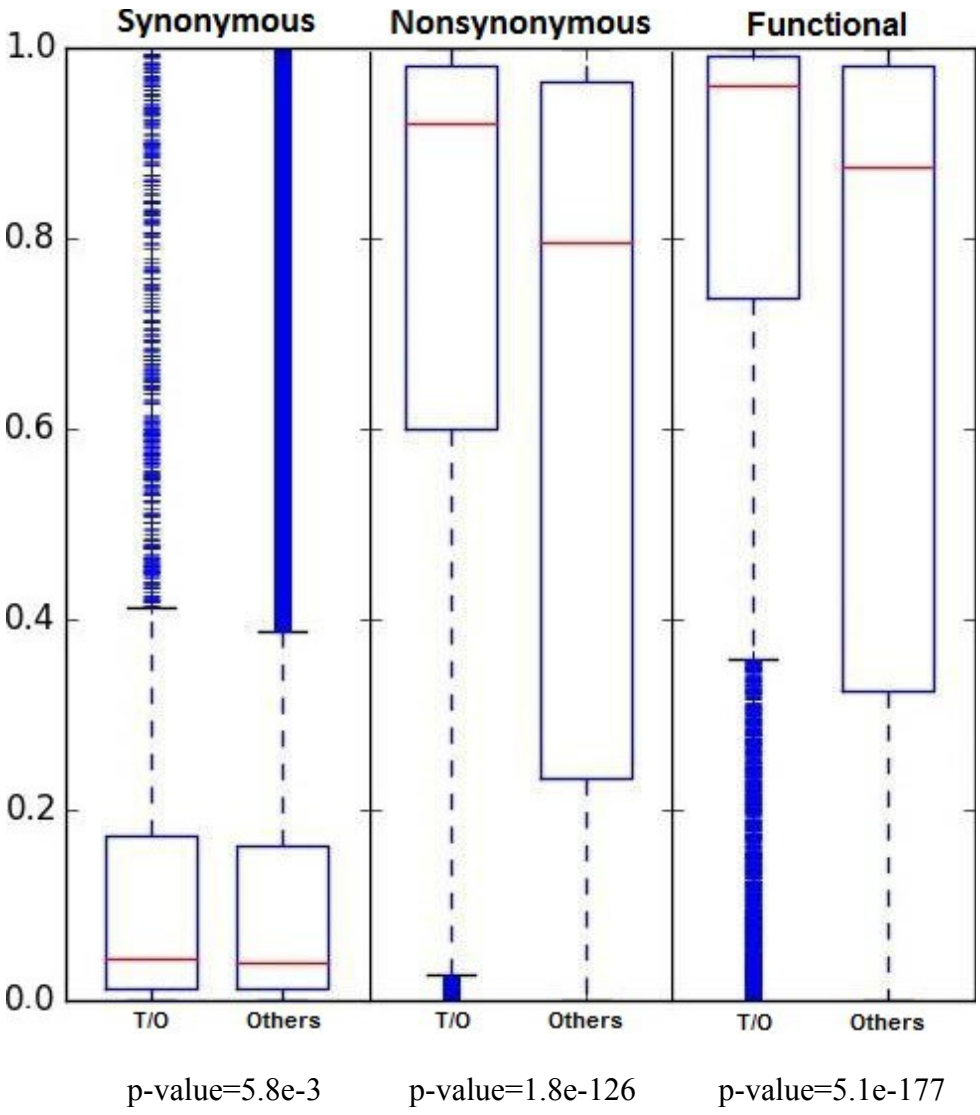


Fig S7. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the score, predicted by PhD-SNP^g, is computed among the TSGs and oncogenes ("T/O", the left plot in each panel) and all the genes not mapped in COSMIC dataset or mapped using other terms ("Others", the right plot in each panel). The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

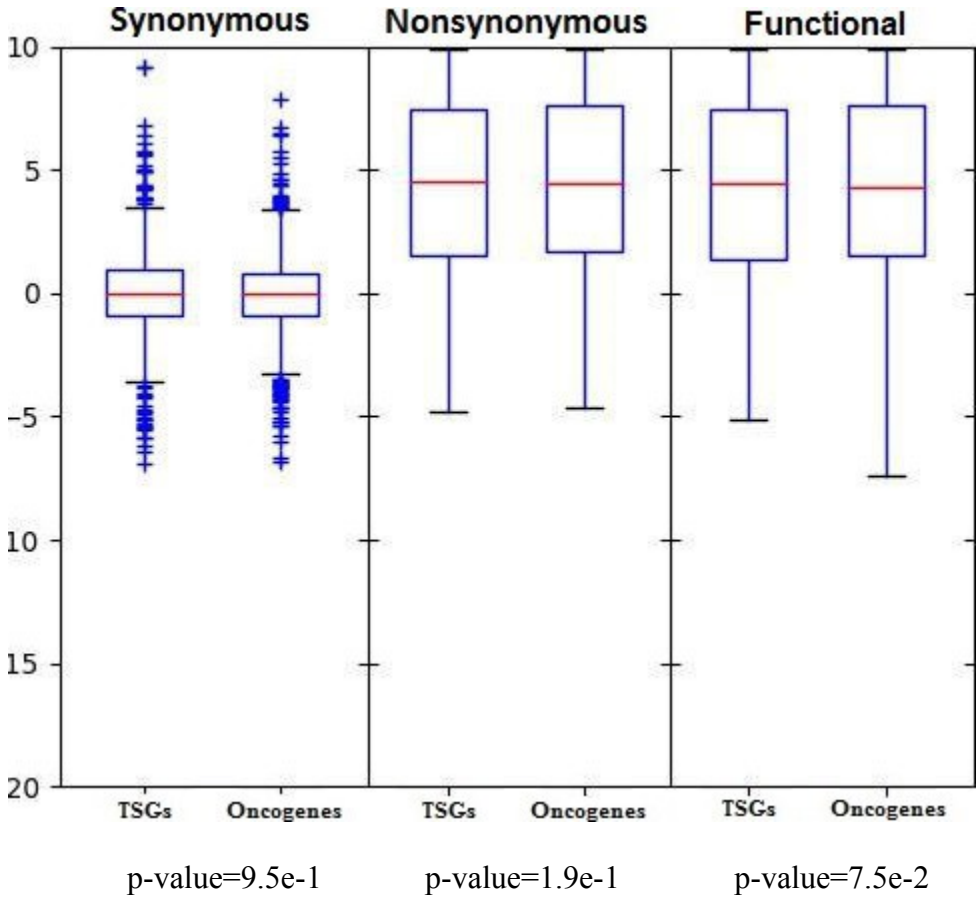


Fig S8. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the TSGs (the left plot in each panel) and the oncogenes (the right plot in each panel) like mapped in COSMIC dataset. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

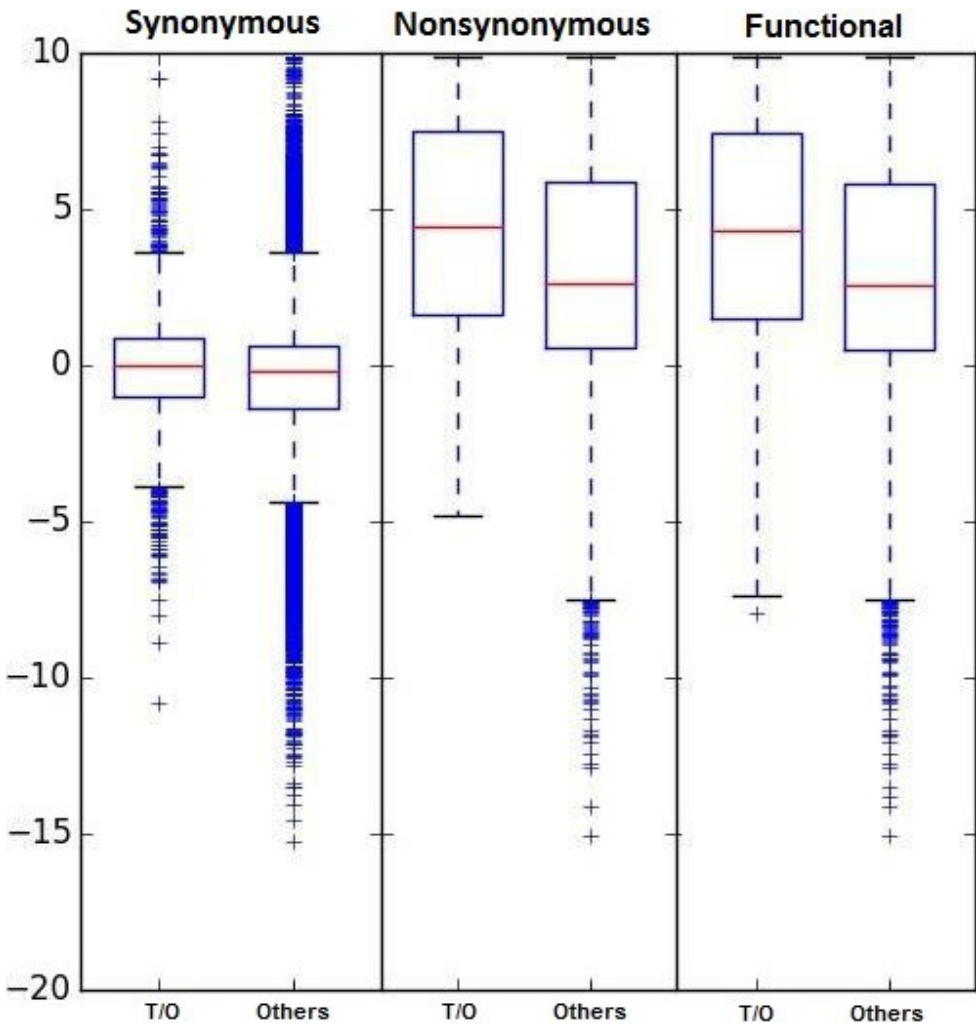


Fig S9. Boxplot of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP[g], is computed among the TSGs and oncogenes ("T/O", the left plot in each panel) and all the genes not mapped in COSMIC dataset or mapped using other terms ("Others", the right plot in each panel). The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

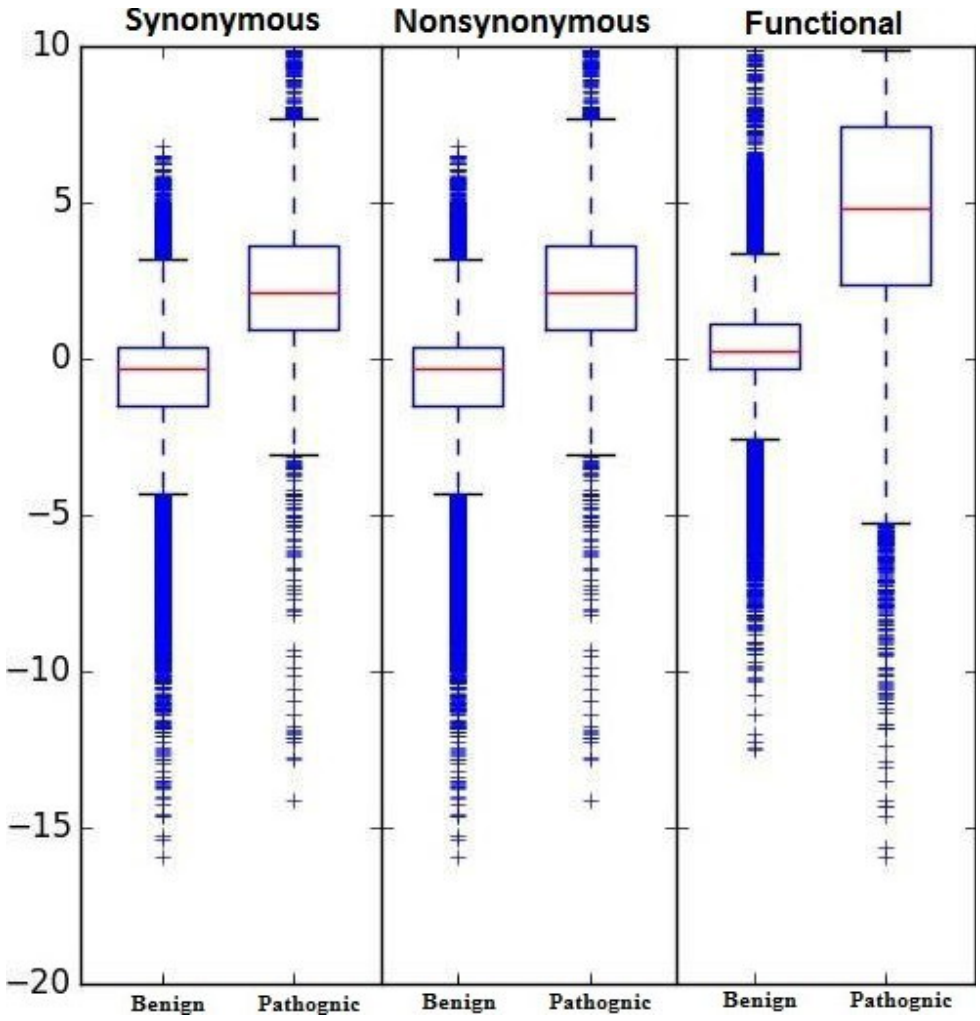


Fig S10. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the benign (the left plot in each panel) and the pathogenic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

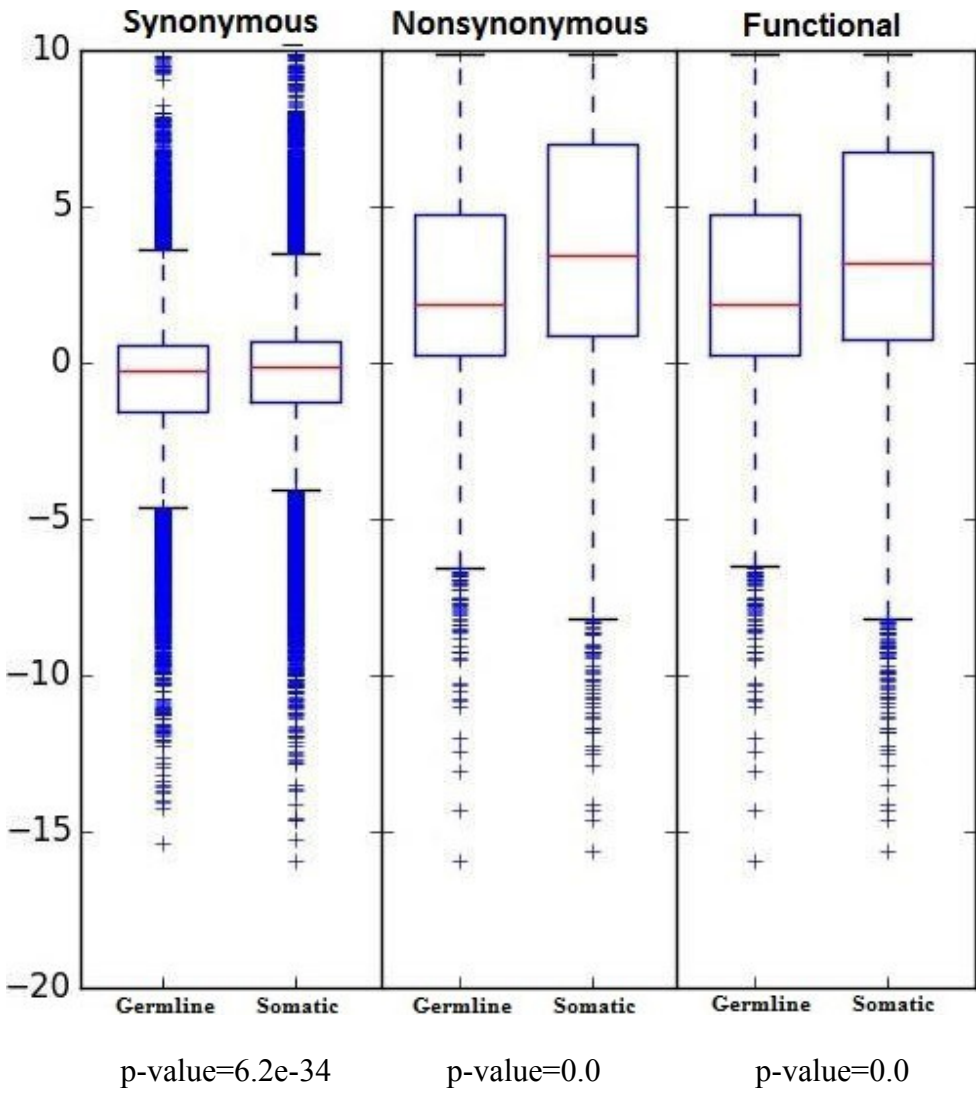


Fig S11. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the germline (the left plot in each panel) and the somatic (the right plot in each panel) mutations. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

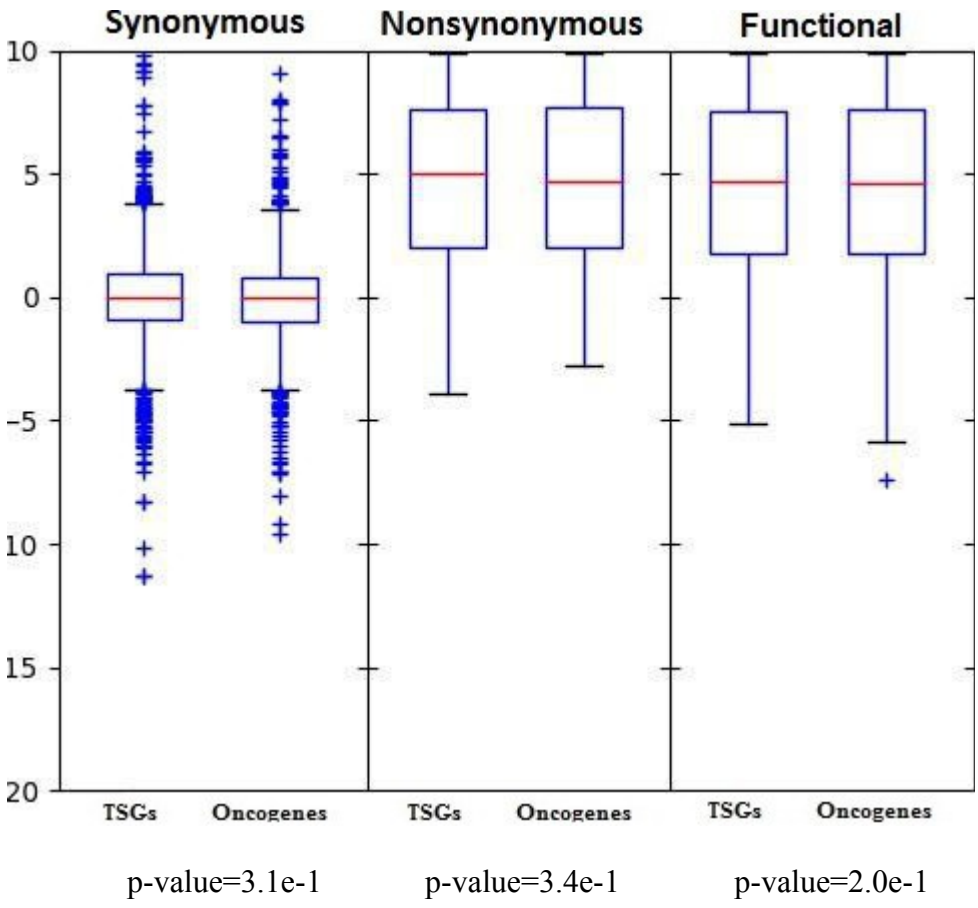


Fig S12. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the TSGs (the left plot in each panel) and the oncogenes (the right plot in each panel) like mapped in COSMIC dataset. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

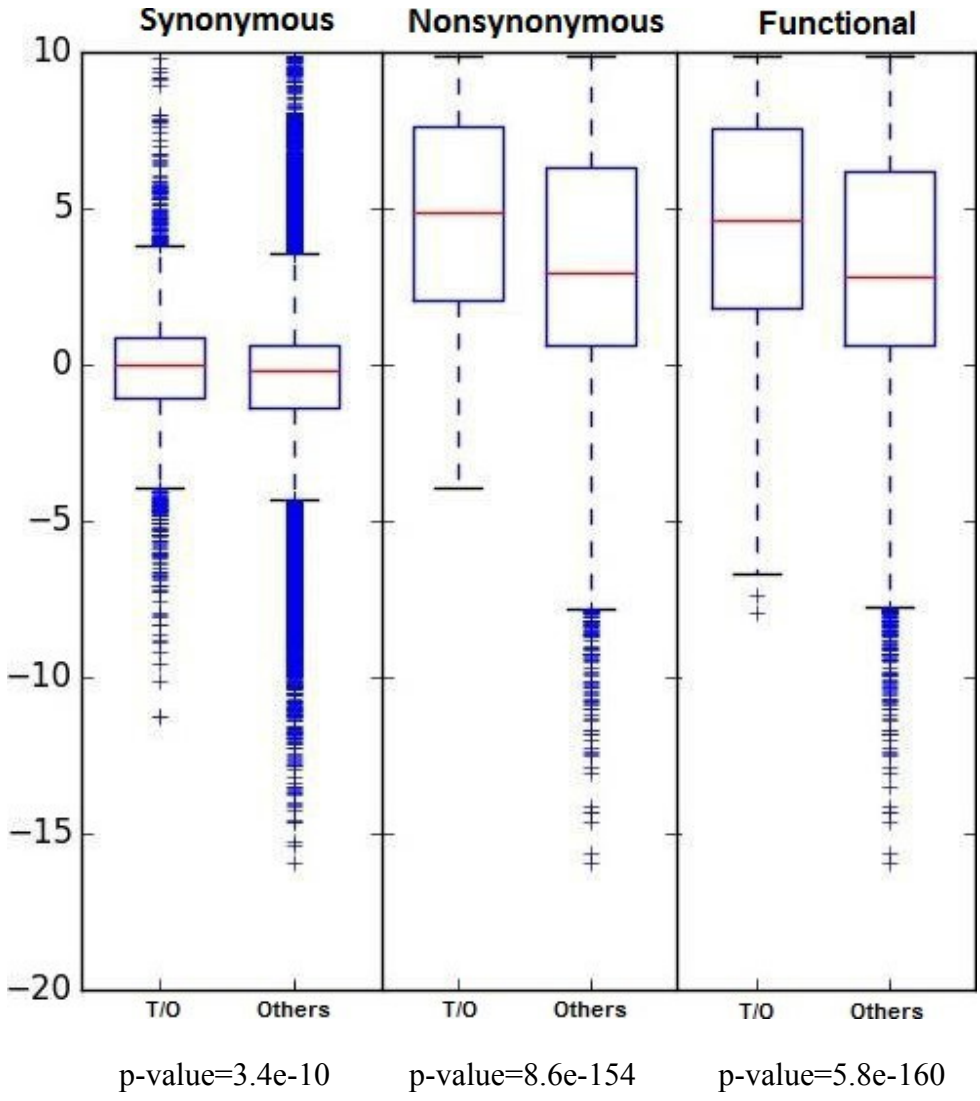


Fig S13. Boxplot of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the conservation, predicted by PhD-SNP^g, is computed among the TSGs and oncogenes ("T/O", the left plot in each panel) and all the genes not mapped in COSMIC dataset or mapped using other terms ("Others", the right plot in each panel). The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

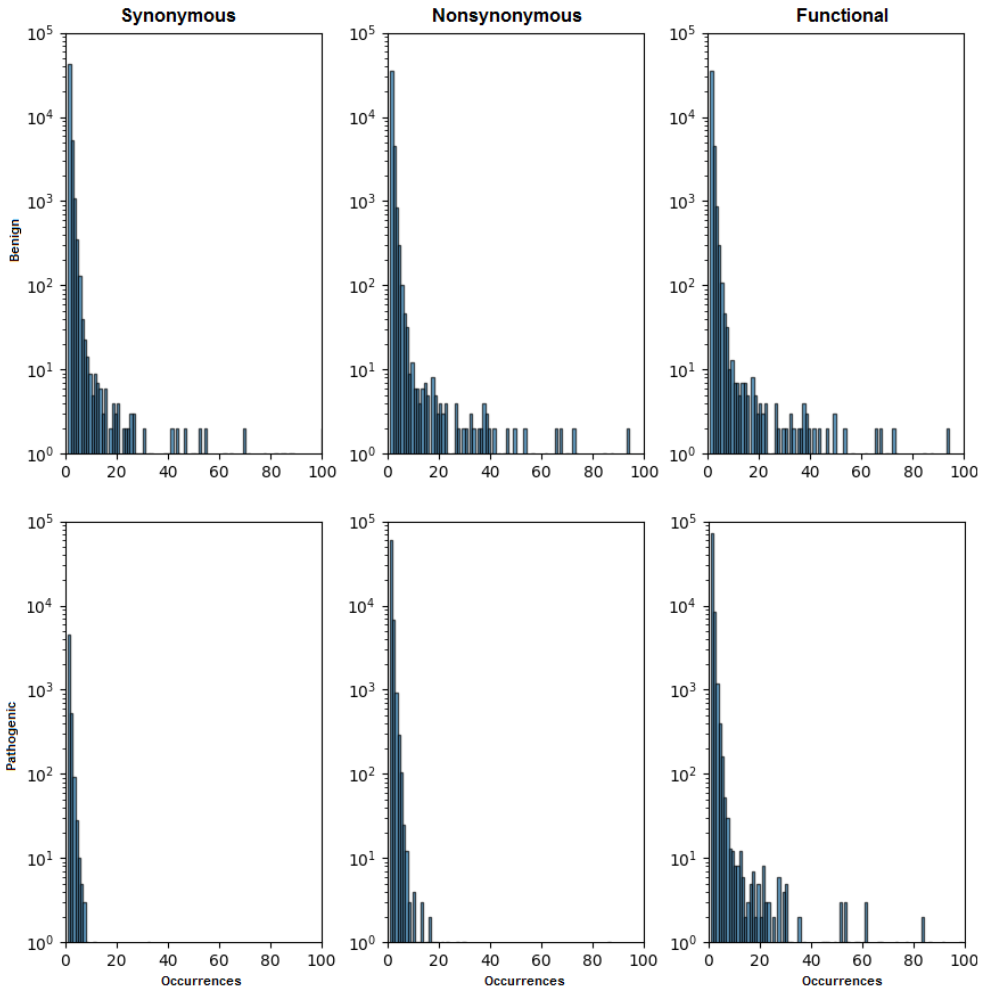


Fig S14. Histograms of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the benign (the plots in the first row) and the pathogenic mutations (the plots in the second row). To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

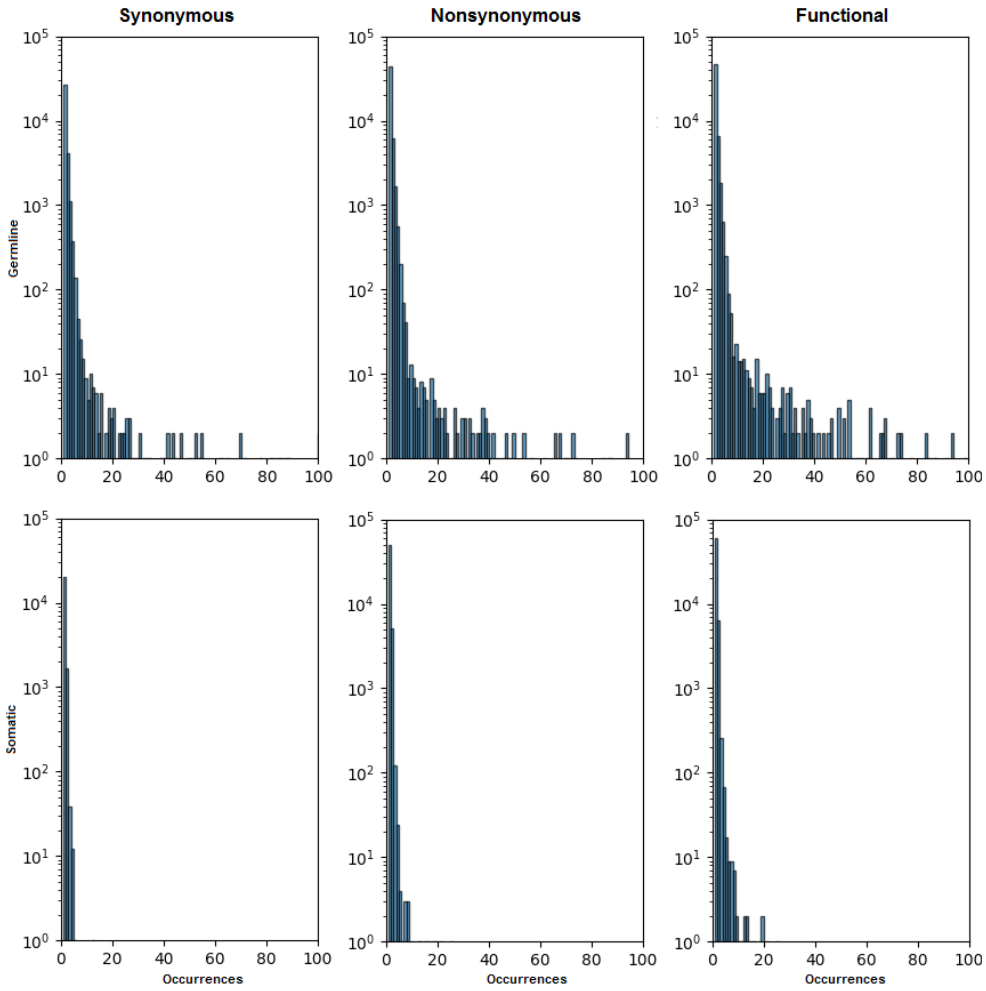
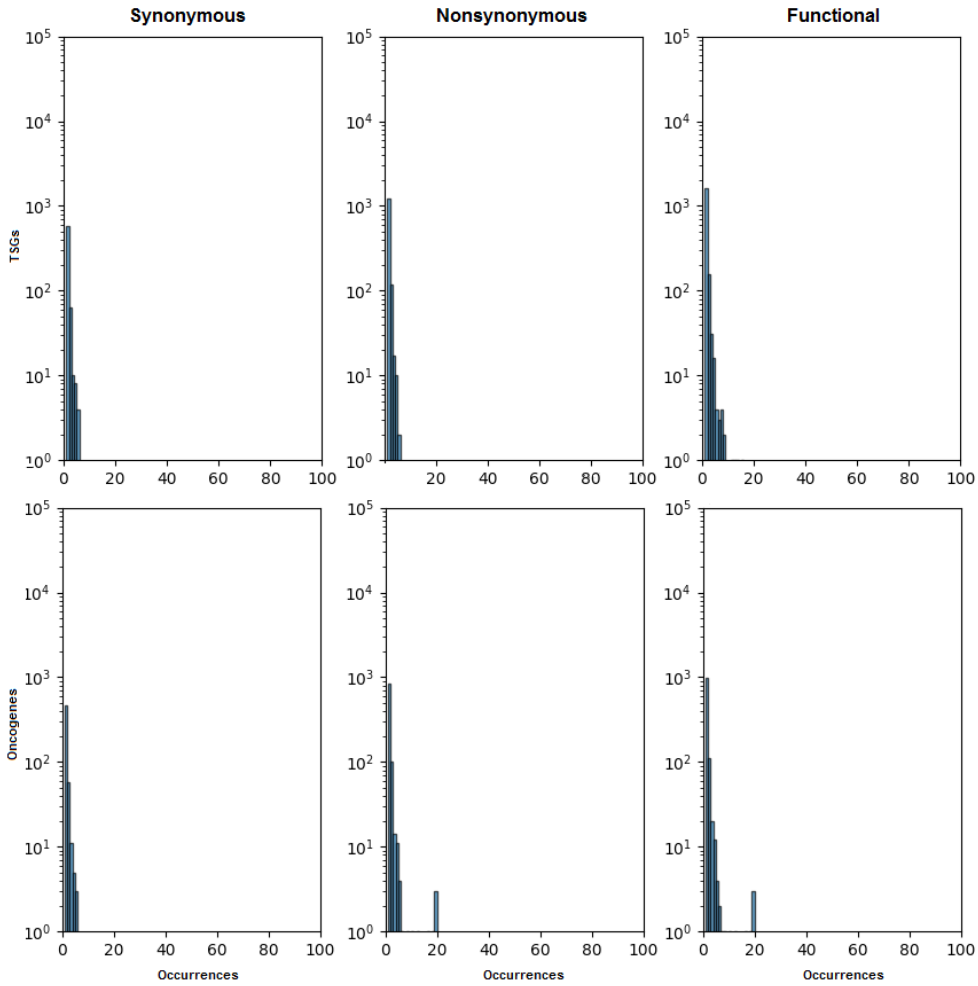
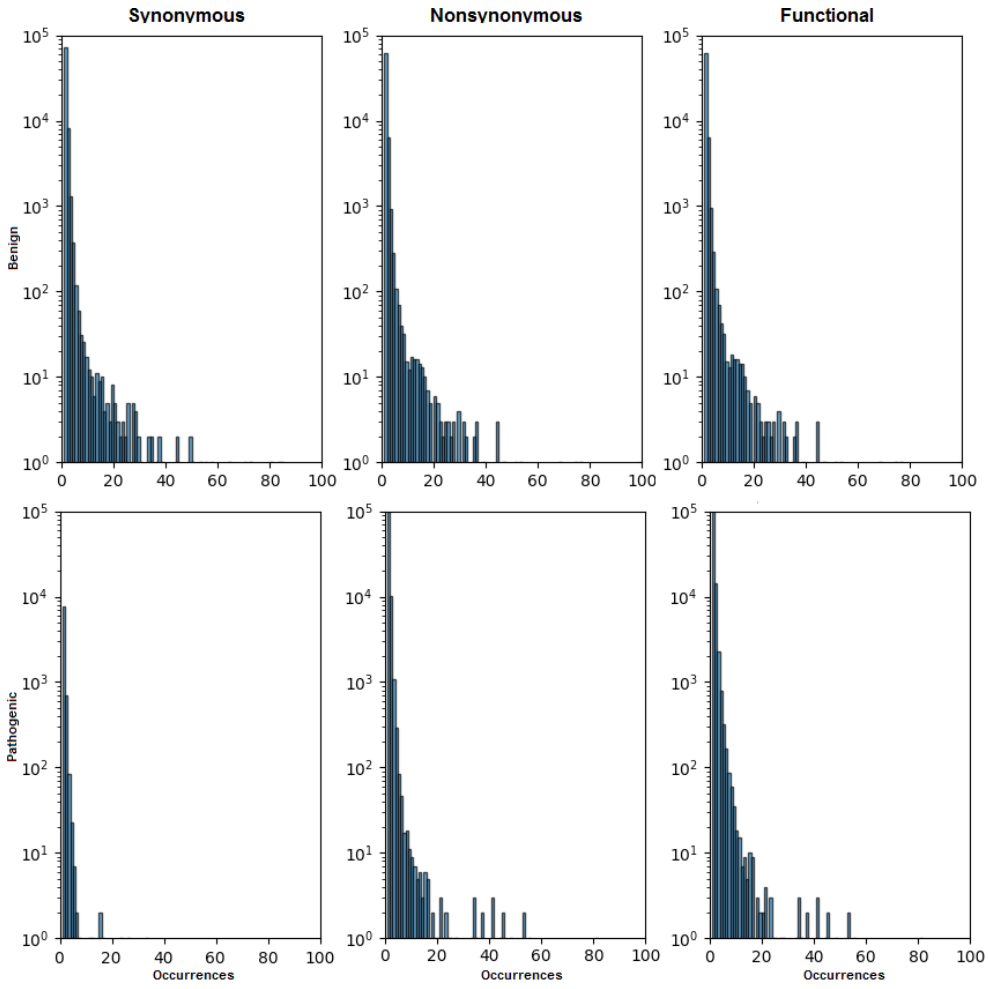


Fig S15. Histograms of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the germline (the plots in the first row) and the somatic mutations (the plots in the second row). To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.



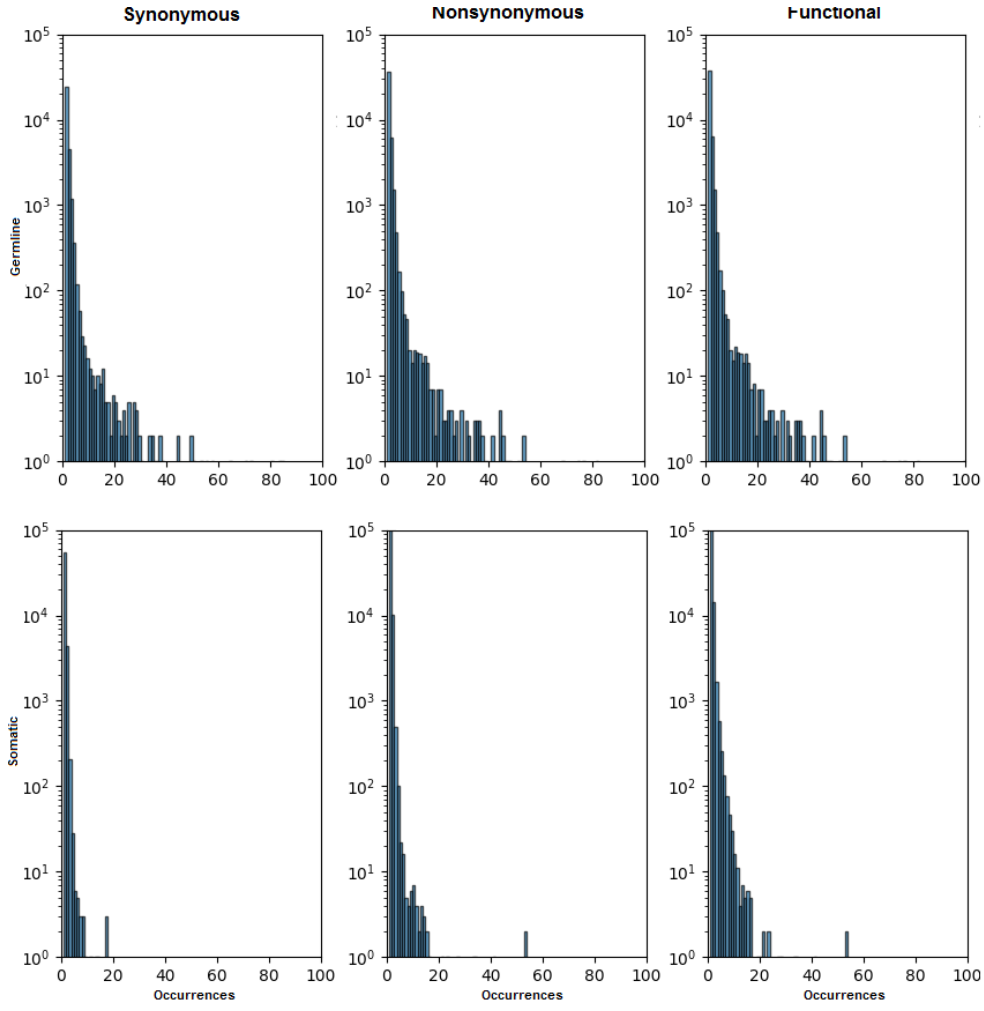
p-value=9.9e-1 p-value=6.2e-1 p-value=9.4e-1

Fig S16. Histograms of BCoM for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the TSGs (the plots in the first row) and the oncogenes (the plots in the second row) like mapped in COSMIC dataset. To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.



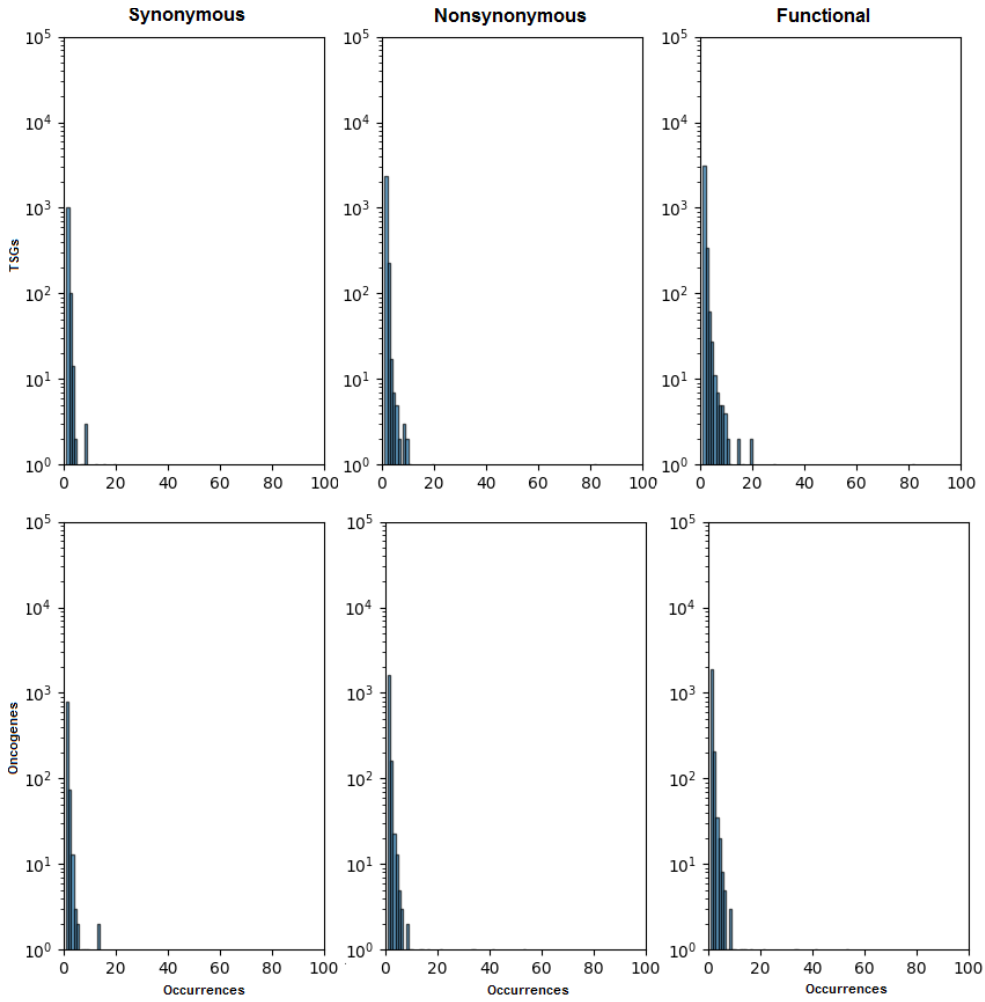
p-value=3.1e-7 p-value=3.7e-21 p-value=8.9e-1

Fig S17. Histograms of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the benign (the plots in the first row) and the pathogenic mutations (the plots in the second row). To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.



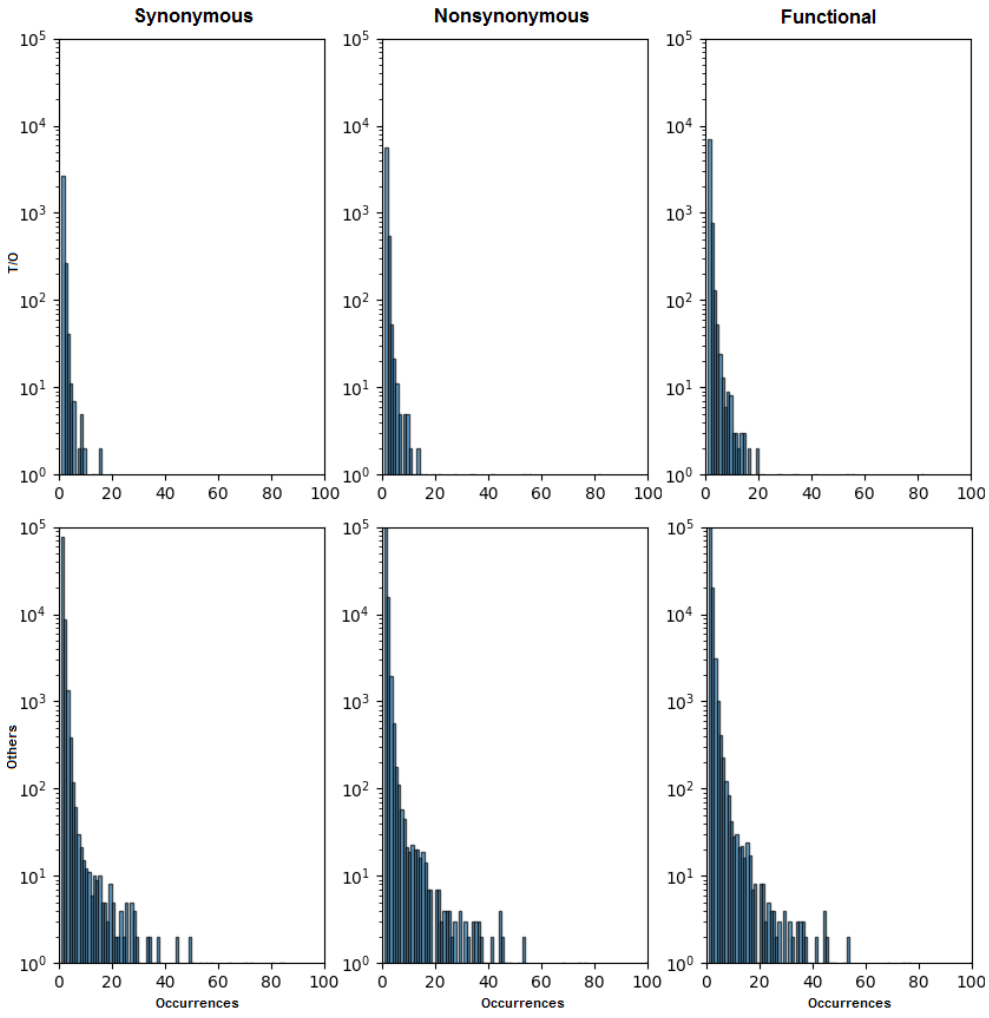
p-value=1.8e-312 p-value=0.0 p-value=5.4e-317

Fig S18. Histograms of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the germline (the plots in the first row) and the somatic mutations (the plots in the second row). To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.



p-value=1.0 p-value=9.2e-1 p-value=1.0

Fig S19. Histograms of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the TSGs (the plots in the first row) and the oncogenes (the plots in the second row) like mapped in COSMIC dataset. To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.



p-value=9.3e-1 p-value=9.9e-1 p-value=4.4e-2

Fig S20. Histograms of Broad for synonymous (left), nonsynonymous (center) and functional (right) mutations where the occurrence, among the samples in the dataset, is computed among the TSGs and oncogenes ("T/O", the plots in the first row) and all the genes not mapped in COSMIC dataset or mapped using other terms ("Others", the plots in the second row). To a good comparison, all the plot have the x-axis cut at 100, threshold that allows to exclude just few outlier values. The values under the plots are the p-values computed using the Kolmogorov-Smirnov between the distributions of the two classes.

Supplementary tables

Table S21. Table of scores obtained by means of the Fisher's Exact test among the BCoM and COSMIC datasets at different thresholds about functional category.

Threshold	O-C	O-NC	U-C	U-NC	OR	p-value	Score
2	63	653	302	16,884	5.39	3.67e-23	22.44
1.9	67	765	298	16,772	4.93	1.78e-22	21.75
1.8	77	905	288	16,632	4.91	3.69e-25	24.43
1.7	82	1,139	283	16,398	4.17	2.11e-22	21.68
1.6	90	1,256	275	16,281	4.24	1.46e-24	23.84
1.5	103	1,591	262	15,946	3.94	4.53e-25	24.34
1.4	114	1,860	251	15,677	3.83	4.32e-26	25.37
1.3	124	2,080	241	15,457	3.82	1.24e-27	26.91
1.2	**152**	**2,901**	**213**	**14,636**	**3.60**	**4.55e-29**	**28.34**
1.1	161	3,262	204	14,275	3.45	2.53e-28	27.60
1	170	3,725	195	13,812	3.23	2.03e-26	25.69
0.9	194	4,888	171	12,649	2.94	9.53e-24	23.02
0.8	203	5,699	162	11,838	2.60	2.28e-19	18.64
0.7	228	6,928	137	10,609	2.55	1.73e-18	17.76
0.6	252	8,502	113	9,035	2.37	2.94e-15	14.53
0.5	279	10,247	86	7,290	2.31	6.41e-13	12.19
0.4	303	11,916	62	5,621	2.31	7.97e-11	10.10
0.3	345	15,308	20	2,229	2.51	4.31e-06	5.37
0.2	347	15,542	18	1,995	2.48	1.66e-05	4.78
0.1	356	17,049	9	488	1.13	0.44	0.36

The contingency table used to compute the Fisher's Exact test is made by means of the classes O-C (thesis project score over the threshold and mapped in COSMIC), O-NC (thesis project score over the threshold and not mapped in COSMIC), U-C (thesis project score under the threshold and mapped in COSMIC), U-NC (thesis project score under the threshold and not mapped in COSMIC). The OR column is for the odd-ratio of the contingency table and the p-value column is the one obtained by the test (two-tailed). The score in the column Score are obtained by means $-\log_{10}$(p-value). In bold type the row with the highest score

Table S22. Amount of mutations or genes involved in the distribution computed after PhD-SNP^g prediction about the Broad set.

Category	Effect	1st class	2nd class	Total
1	Synonymous	91%	9%	90,711
	Nonsynonymous	35%	65%	196,653
	Functional	30%	70%	232,975
2	Synonymous	33%	67%	90,711
	Nonsynonymous	23%	77%	196,653
	Functional	20%	80%	232,975
3	Synonymous	57%	43%	2,027
	Nonsynonymous	59%	41%	4,429
	Functional	63%	37%	5,732
4	Synonymous	3%	97%	90,711
	Nonsynonymous	3%	97%	196,653
	Functional	3%	97%	232,975

The number in column "Category" are "benign vs pathogenic" (1), germline vs somatic (2), TSG vs oncogene (3), TSG and oncogene vs all the remaining genes (4). The 1st class column represent the percentage of gene in the first class for each category while 2nd class column is about the second class. The total column shows how many values are into the distribution, both for synonymous, nonsynonymous and functional effects

Table S23. P-value of distribution about the prediction of PhD-SNP^g in Broad dataset.

Prediction	Category	Effect	p-value
Score	1	Synonymous	**0.0**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	2	Synonymous	**4.0^{-33}**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	3	Synonymous	9.5^{-2}
		Nonsynonymous	8.0^{-1}
		Functional	**3.2^{-4}**
	4	Synonymous	**5.8^{-3}**
		Nonsynonymous	**1.8^{-126}**
		Functional	**5.1^{-177}**
Conservation	1	Synonymous	**5.8^{-3}**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	2	Synonymous	**6.2^{-34}**
		Nonsynonymous	**0.0**
		Functional	**0.0**
	3	Synonymous	3.1^{-1}
		Nonsynonymous	3.4^{-1}
		Functional	2.0^{-1}
	4	Synonymous	**3.4^{-10}**
		Nonsynonymous	**8.6^{-154}**
		Functional	**5.8^{-160}**
Occurrence	1	Synonymous	**3.1^{-7}**
		Nonsynonymous	**3.7^{-21}**
		Functional	8.9^{-1}
	2	Synonymous	**1.8^{-312}**
		Nonsynonymous	**0.0**
		Functional	**5.4^{-317}**
	3	Synonymous	1.0
		Nonsynonymous	9.2^{-1}
		Functional	1.0
	4	Synonymous	9.3^{-1}
		Nonsynonymous	9.9^{-1}
		Functional	**4.4^{-2}**

P-value computed by Kolmogorov-Smirnov test for the PhD-SNP^g prediction scores. The number in column "Category" are "benign vs pathogenic" (1), germline vs somatic (2), TSG vs oncogene (3), TSG and oncogene vs all the remaining genes (4). The 1st class column represent the percentage of gene in the first class for each category while 2nd class column is about the second class. In bold are highlighted the distribution that reject the null hypothesis on a threshold of 0.05.