Bioinformatic analysis of next-generation sequencing data

Master`s Thesis
Bioinformatics Masters
Degree Programme,
Institute of Biomedical Technology
University of Tampere, Finland
Tommi Rantapero
May, 2012

ACKNOWLEDGEMENTS

This work has been done in the Genetic Predisposition to Prostate Cancer group lead by Prof. Johanna Schleutker in the Institute of Biomedical Technology, University of Tampere. I would like to thank Prof. Johanna Schleutker for giving me the opportunity to work with this interesting project. Her guidance and support has been crusial for the success of this project.

I would also like to thank my other supervisor Prof. Mauno Vihinen for his guidance. I learned a lot about discussions with you. In retrospect, I should have consulted you much more often than I did during my thesis work. I would also like to thank Adjunct Prof. Csaba Ortutay for reviewing my thesis. Your ideas and comments have influenced a lot to my thesis work. I would like to thank also other members of the staff responsible for the master´s program in bioinformatics for your good work.

I owe a lot of graditute for Ayodeji E. Olatubosun and Jouni Väliaho for helping me with PON-P. You always had the time to answer my questions which I am very thankful of. I also like to thank all the Prof. Schleutker´s group members for their assistance during my thesis work. They really made me feel as part of the group. Special thanks go to Daniel Fischer for dedicating time to read my thesis. The comments you gave me helped a lot in my effort to improve the more mathematical sections of my thesis.

Last but not least I would like to thank my family, friends and Heli. Your trust in me and the effusive support that I have received from you has led me to the point where I am now. I could not have managed to finish my thesis without you.

May 2012

Tommi Rantapero

MASTER`S THESIS

Place:	UNIVERSITY OF TAMPERE Bioinformatics Masters Degree Programme, Institute of Biomedical Technology Tampere, Finland
Author:	Tommi Rantapero

Title:	Bioinformatic analysis of next-generation sequencing data
Pages:	59 + appendices
Supervisors:	Prof. Johanna Schleutker, Prof. Mauno Vihinen
Reviewers:	Adjunct. Prof. Csaba Ortutay, Prof. Johanna Schleutker
Time:	May, 2012

Abstract

Backround and aims: In a recent linkage study involving 69 Finnish HPC (Hereditary prostate cancer) families, a novel prostate cancer susceptibility locus 2q37.3 was found (Cropp et al. 2011). In addition a signal from 17q21-22, found in a previous study, was confirmed. To further study these loci the families showing the strongest linkage were selected for targeted high-throughput sequencing in FIMM (Finnish Institute for Molecular Medicine). The aim of this study was to utilize bioinformatics methods to assess the variant data produced by the FIMM high-throughput sequencing pipeline in order to find potential candidates predisposing to prostate cancer

Methods: The variants were annotated utilizing an in house Python program and a local database constructed of resources including annotation tracks from UCSC Genome browser, Ensemble, microRNA.org and Vista. To evaluate the pathogenicity of the variants, three tolerance predictor programs were used: Mutation Taster, PolyPhen-2 and PON-P. These results were used to construct a list of candidate genes and variants. To find prostate cancer associated genes two databases DDPC, and COSMIC were used. To further study the relationship of the prostate cancer associated genes and candidate genes a gene ontology and pathway enrichment analysis was conducted for the prostate cancer gene set using WebGestalt2.

Results: As a result of pathogenicity prediction 155 pathogenic mutations were found. These variants were distributed to 101 genes of which four are associated to prostate cancer based on previous research.

Conclusion: In conclusion bioinformatics methods seem to be efficient in prioritizing variants for experimental validation. In addition, these methods can provide insights of how the pathogenic variants can predispose to cancer.

PRO-GRADU TUTKIELMA

Paikka: Tampereen Yliopisto
Bioinformatics Masters Degree Programme,
Institute of Biomedical Technology
Tampere, Suomi
Tekijä: Tommi Rantapero
Otsikko: Bioinformatic analysis of next-generation sequencing data
Sivumäärä: 59 + liitteet
Supervisors: Prof. Johanna Schleutker, Prof. Mauno Vihinen
Reviewers: Adjunkti Prof. Csaba Ortutay, Prof. Johanna Schleutker
Time: Toukokuu, 2012

Tiivistelmä

Työn tausta ja tavoitteet: Viimeaikoina tehdyssä kytkentä-analyysissä, jossa oli mukana 69 suomalaista eturauhassyöpä–perhettä, havaittiin uusi eturauhassyöpään kytkeytynyt alue 2q37.3. Tämän lisäksi aikaisemmassa tutkimuksessa havaittu signaali 17q21-22:sta vahvistettiin. Perheet, joilla havaittiin voimakkain kytkentä näihin alueisiin, valittiin sekvensoitavaksi FIMM:iin (Molekulaarisen lääketieteen Instituutti). Sekvensointi tehtiin hyödyntämällä uuden sukupolven kohdistettua sekvensointi-menetelmää. Tässä pro-gradu tutkielmassa on tarkoituksena analysoida sekvensoinnin tuottamaa variantti-informaatiota ja priorisoida potentiaalisia variantteja jatkotutkimuksia varten hyödyntäen bioinformatiikan menetelmiä.

Menetelmät: Varianttien annotaatiossa käytettiin hyödyksi paikallista tietokantaa ja python ohjelmointikielellä luotuja skriptejä. Paikallinen tietokanta luotiin yhdistämällä informaatiota UCSC:n genomi-selaimen, EnsEMBL:en, MicroRNA.org:in sekä Vistan tietokannoista. Varianttien patogeenisuuden arvioimisessa käytettiin kolmea ennustavaa ohjelmaa, jotka olivat Mutation Taster, PolyPhen-2 sekä PON-P. Tämän analyysin perusteella valittiin kandidaattigeenit sekä variantit tarkempaa tarkastelua varten. Kandidaattigeenejä verrattiin niihin geeneihin, joiden on havaittu aikaisempien tutkimusten perusteella olevan yhteydessä eturauhassyöpään. Näiden geenien määrittämiseksi käytettiin kahta tietokantaa, jotka olivat DDPC ja COSMIC. Vertailua varten, saadulle eturauhassyöpägeenien joukolle tehtiin Geeni Ontologia termi-ja Pathway-analyysi WebGestalt-2 ohjelmalla.

Tulokset: Patogeenisuus analyysin tuloksena havaittiin kaikkiaan 155 patogeenisiksi ennustettua varianttia, jotka jakautuivat 101 geeniin. Näistä geeneistä neljä on ennestään yhdistetty eturauhassyöpään.

Yhteenveto: Bioinformatiikan menetelmät vaikuttavat tehokkailta varianttien priorisoinnissa sekä antavat viitteitä niistä mekanismeista, joihin varianttien kyky altistaa syövälle perustuu.

CONTENTS

Abbreviations

aaPSEC	**amino acid Position SpEsifiC score**
ANN	**Artificial Neural Network**
APC	**Anaphase Promoting Complex**
BLAST	**Basic Local Alignment Tool**
BLOSUM	**BLOcks of Amino Acid Substitution Matrix**
bwa	**Burrows-Wheeler aligner**
CCDS	**Consensus Coding Sequence**
CI	**Conservation Index**
COSMIC	**Catalogue of Somatic Mutation in Cancer**
EJC	**Exon Junction Complex**
FIMM	**Finnish Institute for Molecular Medicine**
GO	**Gene Ontology**
GOSS	**Gene Ontology Similarity Score**
GWAS	**Genome-Wide Association Study**
HapMap	**Haplotype Map**
HGMD	**Human Gene Mutation Database**
HGNC	**Hugo Gene Nomeclature Commitee**
HPC	**Hereditary Prostate Cancer**
HRPC	**Hormone Refractory Prostate Cancer**
KEGG	**Kyoto Encyclopedia of Genes and Genomes**
LBS	**Locus Specific Databases**
LOH	**Loss Of Heterozygosity**
MAF	**Minor Allele Frequency**
MAP	**Maximum A Posteriori**
MCC	**Matthews Correlation Coefficient**
MCM	**Mini Chromosome Maintenance**

MSA	**Multiple sequence alignment**
NGS	**Next-Generation Sequencing**
NMD	**Nonsense Mediated Decay**
nsSNP	**non-synonymous Single Nucleotide Polymorphism**
OMIM	**Online Mendelian Inheritance in Man**
PASS	**Polyandenylation signal site sequences**
PMD	**Protein Mutation Database**
PON-P	**Pathogenic-or Not Pipeline**
PRCA	**Prostate cancer**
PSIC	**Position Specific Independent Counts**
RBF	**Radial Basis Kerner**
RI	**Reliablity Index**
SNP	**Single nucleotide polymorphism**
snSNP	**synonymous Single Nucleotide Polymorphism**
SNV	**Single nucleotide variant**
subPSEC	**substitution Position Spesific score**
SVM	**Support Vector Machine**
UCSC	**University of Santa Cruz**
VCP	**Variant Calling Pipeline**

1.1 Introduction

Prostate cancer (PRCA) is the most common cancer type among men in well developed countries such as Fnland (American Cancer Society 2012, Finnish cancer registry 2007). It has been shown that the risk of PRCA entails a significant genetic component (D.J. Schaid 2004). In cancer genetics, genome-wide association studies (GWAS) and linkage analysis have been used to localize regions and variants associated to cancer susceptibility. GWAS has been used to screen large population for common variants associated to cancer having low-penetrance whereas linkage analysis has been used to discover rare variants which are highly penetrant. During the past decades GWAS and linkage studies have revealed several novel prostate cancer loci (O. Fletcher and R. Houlston 2010).

The development of next-generation sequencing (NGS) technology has provided a new valuable tool in cancer genetics. The greater coverage provided by the new technology has led to significantly more reliable discovery of variants in the genome compared to traditional Sanger sequencing (S.C. Schuster 2008). During past years next-generation sequencing has been applied in several studies to find novel cancer associated variants in loci discovered previously in linkage studies (S. Saarinen et al. 2011, Y.P. Mossé et al. 2008).

In a recent genome wide linkage study, involving 69 Finnish HPC families, a novel PRCA locus 2q37.3 was found and another previously discovered signal from 17q21-22 was verified (Cropp et al. 2010). The families having the strongest signals from these loci were selected for targeted Next-generation-sequencing (NGS) in Finnish Institute of Molecular Medicine (FIMM).

Since sequencing studies produce a large number of variant data, the validation of all variants using experimental methods such as genotyping would be a laborious and expensive task. Therefore, methods to that can be used to highlight variants, which have the potential to predispose to PRCA, are needed. Bioinformatics provide many methods to gain knowledge of the variants which can be used predict their clinical consequences. In this study a selection of these methods are utilized.

1.2 Aims of the study

The aims of this study include:

- Learn about standard file formats used to store sequencing data
- Construct scripts for efficient manipulation of variat data-files
- Learn how to utilize databases to extract knowledge
- Learn to use and interpret the results of pathogenicity predictors and Gene Ontology term and enrichment analysis software
- Analyze the variant data captured by the FIMMs sequencing and variant calling pipeline using approariate bioinformatics methods to prioritize variants for validation with genotyping

2. Literature review: The prediction of pathogenic variants in cancer research using tolerance predictors

Variants can be classified based on their position in the genome, the type of the alteration which they induce at the DNA level, and the effect of the variant in the protein level. Variants that are located in regions which are flanking genes and other coding elements, such as microRNAs, are called non-genic or intergenic variants. As they do change the sequences of genes, also the gene products remain unchanged. However, non-genic variants may alter the regulation of genes if located in the regulatory sites of the genome.

Variants located in genes can be divided into two categories: coding and non-coding. The non-coding variants are located either in the untranslated (UTRs) or in the intronic regions. Although not changing the primary structure of gene products directly, they can alter the splicing pattern of the mRNA, which may result in an alternative gene product. Non-coding mutations can also have effects on gene regulation and to the stability and translation of the mRNA product. The coding variants are located in the exonic regions of the genes which are retained in the mature mRNAs after the intronic parts have been spliced off from the pre-mRNA. Since the exons define protein primary sequence, coding variants have the potential to change the primary structure of the protein directly.

Variants can be also classified into different categories based on their effects on the DNA-level. Insertion and deletions of bases in the DNA sequence are generally referred as "indels"[2] whereas single nucleotide exchanges are referred as SNPs (Single nucleotide polymorphisms). The "SNPs", occurring in the coding regions of genes can be further classified, based on their effect at the protein level, to synonymous SNPs and non-synonymous SNPs. Synonymous SNPs do not lead to the change in amino acid sequence contrary to non-synonymous SNPs, which can be further classified into two different types: missense variants and nonsense variants. Missense variants change an amino acid to another whereas nonsense variants introduce a stop codon leading to a truncated protein product (J.Thusberg and M. Vihinen, 2009).

[1]In some context the term non-coding may also refer to regions that are outside genes.
[2]The use of the term "indel" may also refer to changes where one or more bases have been deleted and inserted in the same positions.

In the search for variants which causes diseases such as cancer, variants in the coding regions of genes are considered more interesting since they are more likely to alter the protein products of genes, which in turn might lead to drastic effects on the phenotype. Nonsense variants are probably regarded as the most damaging since they alter the length of the protein product, which might result to the loss of normal function of proteins. In addition, insertion or deletions in the coding regions of genes are in many cases damaging since they are likely to introduce a frameshift in the coding sequence. Frameshifts can change the protein product significantly depending on the location of the variant in the gene (J. Hu and P.C Ng, 2012).

The consequences of missense variants are much harder to predict compared to nonsense variants and indels. Therefore, the development of methods to assess the effect of missense variants has been a major subject of research in the field of bioinformatics during the past decade. Today, there are many tools available which can predict the consequences of missense variants for protein structure and function. These programs can predict effects on specific features such as stability, localization, disorder and the aggregation propensity of proteins. (J.Thusberg and M. Vihinen, 2009)

Furthermore, programs have been developed that evaluate the pathogenicity of mutations. These so called tolerance predictors evaluate the effects of mutations on the phenotype by assessing the changes that are caused by the alterations at the DNA level and to a greater extend at the protein level. In order to predict the effects of variants, the tolerance predictors consider many features: including evolutionary conservation, changes in the physico-chemical characteristics of the amino acids, the sequence environment of the affected amino acid and alteration in structural properties of proteins (J.Thusberg and M. Vihinen, 2009)

Tolerance predictors can be divided into three categories based on the method used in the prediction. Evolutionary based methods apply the phylogenetic information derived from multiple sequence alignments of related protein sequences to evaluate the probability of pathogenicity. The Bayesian methods apply Bayesian statistics to infer the pathogenicity of a variant based on a set of known examples of pathogenic and neutral variants. Machine learning methods are based on classifier algorithms trained to distinguish between pathogenic and neutral mutations. In a similar fashion

to Bayesian methods, sets of known examples of pathogenic and neutral variants are used to train the classifier. (J.Thusberg and M. Vihinen, 2009)

Most of the tolerance predictors only consider the effects of missense variants. However, Mutation Taster and the most recent version of SIFT can also evaluate the effects of indels (Schwarz JM et al. 2010; J. Hu and P.C Ng, 2012). Furthermore, Mutation Taster can assess the effects of non-coding variants making it the most versatile program in use at the moment.

2.1 Evolutionary conservation based methods

2.1.1 SIFT

Sorting Intolerant From Tolerant (SIFT) is a simple software which utilizes only evolutionary information to evaluate whether the mutation is likely to be tolerated or not. The prediction is based on calculating the normalized probabilities of all possible amino acid substitutions for each amino acid position. The probabilities are obtained from a multiple alignment sequence alignment (MSA) which is constructed of the mutated protein sequence and its homologs. The sequences for the MSA are either defined by the user or SIFT itself. If the user does not give the sequences for MSA construction, SIFT searches similar sequences for the given protein sequence from SWISS-PROT, SWISS-PROT/TrEMBL, or the non-redundant protein databases of NCBI (P.C Ng and S. Henikoff, 2001) to construct the MSA.

SIFT output is the normalized probability that the mutation is tolerated. SIFT considers the variant to be either tolerated or non-tolerated based on this normalized probability. If the probability of tolerance is under 0.05, the variant is considered to be non-tolerated; otherwise the mutation is considered to be tolerated (P.C NG and S. Henikoff, 2001).

2.1.2 Panther

Similar to SIFT, Panther predicts the pathogenicity of missense mutations based on the knowledge of evolutionary conservation of the amino acids. The evolutionary information is derived from MSAs constructed of homologs which are retrieved from the PANTHER library of protein families. The selection of protein sequences is done by comparing the query sequence to Hidden Markov Model-profiles for each protein

family. The best matching profile is selected and the substitution position specific score (subPSEC) is calculated for the variant. The subPSEC score is determined first by calculating the amino acid position specific scores (aaPSEC scores) which represent the likelihood of a single amino acid at a specific position (P.D.Thomas et al. 2003). Formally, the score can be presented as follows:

eq. 1 $$aaPSEC(a,i,j) = ln\left[\frac{P_{\text{aij}}}{P_{\text{max}}}\right],$$

where, P_{aij} represents the probability of amino acid a at position i, given a HMM j and P_{max} is the maximum probability observed at position i.

The Score of 0 means that the amino acid is the most evolutionary conserved in that position. The smaller the aaPSEC score, the smaller the likelihood of observing the amino acid in that particular position becomes. The aaPSEC scores for amino acids a and b are used to calculate the subPSEC score for the amino acid substitution from a to b as follows:

eq. 2 $$subPSEC(a,b,i,j) = -[aaPSEC(a,i,j) - aaPSEC(b,i,j)] = -ln\left[\frac{P_{\text{aij}}}{P_{\text{bij}}}\right]$$

The subPSEC score represents the difference in the probability of observing the wild type amino acid and the mutant amino acid b. The score is interpreted such that as the score decreases, the likelihood of pathogenicity of the amino acid substitution increases. Panther differs from the other tolerance predictors in the sense that the cut off value that separates the pathogenic from the non-pathogenic mutations is user defined. However, the developers of Panther suggest a cut off value of -3 (P.D. Thomas et al. 2003).

2.2 Bayesian methods based tolerance predictors

2.2.1 Naïve Bayesian classifier

The naïve Bayesian classifier assigns data, which is represented by so called "feature vectors", to classes. The elements of the vectors represent the values of the features used by the classifier. In order to be able to assign data to a class, the classifier has to be trained with a training set. The training set consists of feature vectors for which the class is known. Based on the training set a statistical model which aims to describe the data is constructed (I. Pop, 2006).

The naïve Bayesian classifier is based on a conditional probability model described by Bayes' theorem. The Bayesian theorem states that the probability of a feature vector V belonging to a particular class C can be determined by first calculating the product of prior probability that an arbitrary feature vector belongs to class C and the likelihood of observing a particular feature vector V given that this feature vector belongs to class C. This product is then divided by the probability of observing this particular feature vector from any class (I. Pop, 2006). Mathematically, this model can be formulated as follows:

eq. 3 $$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1,..,F_n|C)}{P(F_1,..,F_n)},$$

where C is a variable representing the class of the prediction, and the F_i ($1 \leq i \leq n$) represents the values of the feature vector V. This equation can be rewritten by applying the joint probability rule:

eq. 4 $$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1|C)P(F_2|C,F_1)P(F_3|C,F_1,F_2)\dots P(F\mathrm{n}|C,F_1,\dots,F_{n-1})}{P(F_1,..,F_n)},$$

Since the naïve Bayesian classification model assumes the features to be independent, the equation 4 can be rewritten as follows:

eq. 5 $$P(C|F_1, \dots, F_n) = \frac{1}{P(F_1,..,F_n)} p(C) \prod_{i=1}^{n} P(F_i|C)$$

The class prior probability can be estimated from the training data using either the relative frequencies of observed classes or alternatively assuming equal probabilities for each class. The feature distributions can be approximated using some well-defined distributions such as Gaussian distribution or the parameters can be estimated using non-parametric modeling.

The probability model described here can be implemented in data classification by the addition of a decision rule. The most common decision rule is the maximum a posteriori decision rule (MAP), which assigns the data to the class which is the most probable given the data. This rule can be formulated as follows:

eq. 6 $$classify(F_1, \dots, F_n) = argmax_c \left[P(C = c) \prod_{i=1}^{n} P(F_i|C)\right]$$

The assumption of independence of features is most often invalid. However, if the dependencies of features are evenly distributed in each class, the bias effects caused by the dependent features cancel each other out. (H. Zhang 2004).

2.2.2 PolyPhen-2

PolyPhen-2 predicts the effects of missense variants and it is based on a Naïve Bayesian classifier. PolyPhen-2 consists of two prediction models which have been trained using one of two training sets: *HumVar* or *HumDiv*. The *HumVar* variant dataset consists of 3155 SNPs annotated in SwissProt which have been associated with mendelian diseases and 6321 neutral SNPs. *HumDiv* contains 13032 variants causing human disease from SwissProt and 8946 human SNPs that have not been associated with diseases (I.A. Adzhubei et al. 2010).

PolyPhen-2 makes the prediction based on the evolutionary conservation of the sequence position being affected, the physico-chemical characteristics of the amino acids involved in the substitution, the sequence environment of the mutation site and the structural features being affected by the mutation. The sequence based features are evaluated by first searching and selecting orthologous and paraloguous sequences for the protein sequence using the Basic Local Alignment Tool (BLAST) followed by the construction of multiple sequence alignment using Multiple Alignment using Fast Fourier Transform (MAFFT) program. To improve the accuracy of the prediction, the MSA is refined using Leon software.

From the constructed MSA eight sequence based features derived. To of the most essensential features are considered by the PolyPhen-2 are the Position Specific Independent Counts score (PSIC) for the wild-type residue and the difference between the PSIC-scores of wild type residue and the mutant residue. The PSIC score represents the likelihood of an amino acid to occur at a specific position in the protein sequence. The likelihood of given amino acid to occur at a specific position is based on the observed counts of different amino acid residues and the relatedness of the sequences in the MSA.

Other features determined from the MSA include the alignment depth at the position of mutation, the sequence identity of the closest homologue having an amino acid

residue differing from the wild-type residue and the congruency of mutant residue. The congruency mutant residue to the MSA is calculated as follows.

- All the amino acid residues that have been observed at the mutation site in the alignment the sequence identity of the analyzed protein and the closest homolog where the amino acid residue is observed is determined.
- The products of the sequence identities and the probability of the substitution of each amino acid residue to the mutant residue are calculated. The probabilities are based on the substitution rates in Blocks of Amino Acid Substitution matrix (BLOSUM).
- Finally, the maximum value of these products is taken as the congruency of the mutant amino residue.

In addition to the sequence based features, PolyPhen-2 considers also two physico-chemical features being affected by the variant: the change in the amino acid volume and hydrophobic characteristics. Moreover, PolyPhen-2 checks if the mutation changes the CpG context of the DNA-sequence. Furthermore, the program evaluates three structural features. These features include the crystallographic B-factor of the amino acid position, the surface area accessibility of the wild-type amino acid residue and the PFAM-domain annotation associated to the site of mutation.

Polyphen-2 classifies variants two into one of three categories: benign, possibly damaging and probably damaging, based on the probability of pathogenicity given by the classifier. The mutation is considered benign if the probability of pathogenicity is under 0.15. The mutation is considered possibly pathogenic if the probability of pathogenicity is over 0.15 and under 0.85, and probably pathogenic when the probability of pathogenicity is over 0.85. In addition, Polyphen-2 gives the estimated true positive and false positive rates.

2.2.3 Mutation Taster

Mutation Taster is a prediction tool capable of analyzing synonymous, non-synonymous and non-coding SNPs. In addition, the program is able to assess small indels limited up to 12 bases in length. Mutation Taster has three different prediction models for different types of variants: *Without_aae* is designed for the synonymous and non-coding variants which do lead to amino acid substitution but might have an

effect to the splicing pattern of the transcript, *Simple_aae* is for missense variants and *complex_aae* for variants causing more complex effect such as frameshifts or truncated protein products (J.M. Schwarz. et al. 2010).

Mutation Taster utilizes a Naïve Bayesian classifier which has been trained with variant data gathered from several resources. The dataset containing neutral variants is a selection of annotated SNPs and Indels from dbSNP. The selection of the SNPs is based on population frequencies in Haplotype Map (HapMap) which means that in order to be selected in the neutral dataset frequencies of all three genotypes had to be at least 10% in at least one population. This filtering procedure ensures that rare variants which might potentially cause rare diseases are excluded.

Due to the fact that the HapMap set does not contain Indels, the selection indels is based on the genotype frequencies. As a criterion for the selection, at least two different genotypes have to be found among the populations. The polymorphism dataset contains 515 263 SNPs and 8 162 Indels in total. The disease associated variant dataset has been gathered from the Online Mendelian Inheritance in Man (OMIM), Human Gene Mutation Database (HGMD) and literature. It consists of 42 989 point mutations and 14 067 indels in total.

The features that have been selected for the classifier include: Evolutionary conservation of the affected site, splice site changes, loss of protein features, changes in the amount of mRNA and length of the protein.

The evolutionary conservation of the mutation site is analyzed by first constructing a multiple sequence alignment of ten homologous sequences from different species including chimp, rhesus macaque, mouse, cat, chicken, claw frog, puffer fish, zebra fish, fruit fly and worm, using bl2seq. Based on the MSA, the Mutation Taster assigns the position of the amino acid in the sequence to one of the three different categories: all identical, conserved or non-conserved.

Mutation Taster makes use of third party splice site prediction software NNSplice to predict if alterations in the genomic sequence will lead to alternative splicing. NNSplice analyzes 60 bases around the mutation site comparing wild type sequence to the mutated sequence. The program can predict if the mutation affects an existing splicing site making it stronger, weaker or completely lost. In addition NNSplice is

able to determine if the mutation activates an additional splice site. If the prediction score given by the NNSplice is 0.5 or higher, the Mutation Taster considers the mutation to alter splicing.

The Mutation Taster evaluates the changes in the amount of mRNA by investigating if the variant has effects on the kozak consensus sequence or the poly-adenylation signal. The kozak consensus sequence is a small sequence which initiates the translation the mRNA to protein and is located upstream of the start codon and ending +4 downstream of the first base of the start codon. The sequence has two highly conserved bases purine (R) and guanine (G) in positions -3 and +4 respectively. The Mutation Taster checks if the mutation makes changes to these conserved bases leading to possible alterations in the initiation of translation which in turn affects the amount of the mRNA.

Mutation Taster uses polyadq to predict if the mutation site is located within a polyadenylation signal site (J.E. Tabaska and M.Q. Zhang, 1999). The most common polyadenylation signal sites in human genes consist of six base sequences (hexamers). The most common hexameric sequence is AAUAAA. The other sequences are single nucleotide variants of this sequence (E. Wahle and W. Keller 1996; D.F. Golgan and J.L. Manley, 1997). Alterations in the polyadenylation signal site sequences (PASS) are suggested to predispose the mRNA to non-spesific degradation thus affecting the stability of the mRNA (G. Edwalds-Gilbert et al. 1997).

To predict if the variants changes protein features, Mutation Taster utilizes a database constructed of SwissProt protein features (A. Bairoch and R. Apweiler, 1996; V. Junker et al. 1999). Mutations can affect protein features either directly by changing the amino acid sequence within a region having a particular feature or indirectly via introduction of a termination codon, frameshift or altered splicing.

Moreover, Mutation Taster tests if the protein sequence is elongated, truncated or likely to undergo nonsense mediated decay (NMD). The protein sequence is elongated if the variant changes the stop codon to another codon. On the other hand, in case the variant induces a premature stop codon, this will lead to a truncated protein product. (J.M. Schwarz. et al. 2010)

NMD is a mechanism that prevents the translation of truncated protein products. The main component of NMD pathway is the exon junction complex (EJC) which is located approximately 20-24 nucleotides upstream of the last splice junction (H. Le Hir et al. 2000). During normal translation ribosome displaces EJC and continues translation until stop codon is reached. However, if the ribosome encounters a premature stop codon, the translation ends and EJC remains bound triggering the NMD (L.E. Maquat and G.G. Garmichael 2001). The Mutation Taster evaluates if the mutation is likely to cause nonsense mediated decay by setting the NMD border to -50 base pairs from the last intron-exon boundary. If the premature stop codon occurs on the 5´-side of this border, the mutation is likely to cause NMD. (J. Lykke-Andersen et al. 2000)

The Mutation Taster classifies the variant in one of two classes: polymorphism or pathogenic based on the probability of pathogenicity. If the probability is under 0.5 the variant is classified as polymorphism and otherwise pathogenic. In addition to the actual classification, Mutation Taster gives also a p-value which reflects the security of the prediction. (J.M. Schwarz. et al. 2010)

2.3 Machine learning based tolerance predictors

2.3.1 Random forest classifier

Random forest classifier is based on classification and regression trees (CART). Classification trees are decision trees which assign vectorial data into classes. The elements of the vectors represent the attributes which are used by the trees to classify the data. An example of a classification tree is illustrated in Figure 1.

The random forest algorithm grows a vast number of classification trees in a recursively manner. New data is assigned to classes based on majority vote which means that data is assigned to the class which is supported by the majority of trees. The trees are grown such that for each tree N number of samples from the training set is randomly chosen with replacement, where N is the number of samples in the training set. The samples that are not selected are used to estimate the error of the classification. This principle is known as bagging (L. Breinman 2001).

At each node the best attribute and the rule based on this attribute is determined. This is done by first selecting a random subset of all attributes. The size of this subset is

held constant during the forest growing. Next, for each attribute the most optimal rule is determined. The best attribute is then selected from the the subset of attributes for which the most optimal rule has been selected. The combination of the best attribute having the most optimal rule is defined as the best split at this given node.

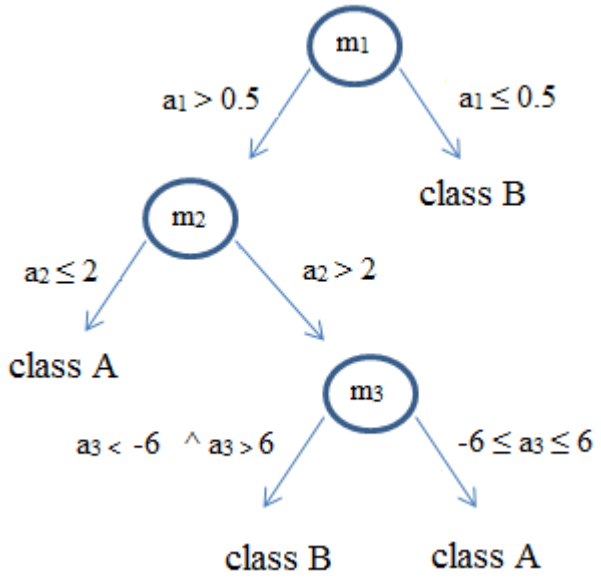

Figure 1. An example of a decision tree. The decision tree consists of three nodes denoted as m_1, m_2 and m_3. At each node the data is split based on a rule associated to that node and the attribute associated to the vectors denoted as a_1, a_2 and a_3. In the terminal nodes the class is assigned for the vector.

The best split is determined using node impurity as the measure of optimality. One of the most commonly used node impurity measure is the gini impurity which is defined by the gini index. To calculate the gini-index, first the estimated probabilities of samples to be assigned to a particular sample k ∈ K described by eq. 7

eq. 7 $$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

where, m denotes the node, x_i is the vector class to classified and y_i denotes the class of x_i, R_m denotes the set of all samples that have been partitioned to m, N_m denotes the number of samples in R_m, and the k denotes the class of the sample.

The gini index is calculated using the estimated class probabilities as follows:

eq. 8 $$Gini\ index_m = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk})$$

The Gini-index is calculated for each possible value attribute which defines the rule how the the samples are split according to a particular attribute. The best rule for a

given attribute is the one having the smallest Gini-index value. For each attribute the best rule is determined. Next, the best attribute for splitting is selected such that the attribute of which best rule has the smallest gini-index is selected for splitting. The tree is grown by adding new nodes which are used to split the samples until some stopping criterion is reached. After this trainging step the random forest can be used to classify new data.

2.3.2 Support Vector Machine classifier

Support vector machine (SVM) is a machine learning based method which can be used in data classification. The classification is based on a hyperplane or a set of hyperplanes in high-dimensional space. The hyperplane is used to separate data, represented as points in space, into classes. The separation of the hyperplane and the nearest data points on each side of the hyperplane defines the margins. The hyperplane is selected such that the margin is maximized (C-H. Hsu et al. 2003). The principle of maximum separation and definition of margins in two dimensional space are illustrated in Figure 2.

If a hyperplane can be set such that the data points are completely separated into two classes the data is said to be linearly separable. This represents the simplest case of data classification problem and can be solved using linear SVMs. The classification function can be represented as the dot product of the data point and the normal vector of the hyperplane, and the sum of constant b. The function can get values of either -1 or 1 which represent the two classes. Formally the classification function can be presented as follows.

eq. 9 $$f(x) = \langle w, x \rangle + b \,,$$

where $\langle w, x \rangle$ is the dot product of the data point x and the normal vector of the hyperplane w and b is a parameter which together with w defines the offset of the hyperplane from the origin.

In many cases the data points are not linearly separable. In this case the data points are mapped in to a higher-dimensional space called the feature space using a transformation function. The purpose of this function is to transform the data in such way that it is linearly separable.

The classification function is then written as follows:

eq. 10 $$f(x) = \langle \phi(w), \phi(x) \rangle + b,$$

where ϕ is the transformation function from lower dimension to higher dimension

The transformation of data is computationally expensive since each element of the vectors has to be transformed before the product of two vectors can be calculated. This problem can be solved using kerner functions as transformation functions. For the kerner functions it holds that:

eq. 11 $$K(\langle x, y \rangle) = \langle K(x), K(y) \rangle$$

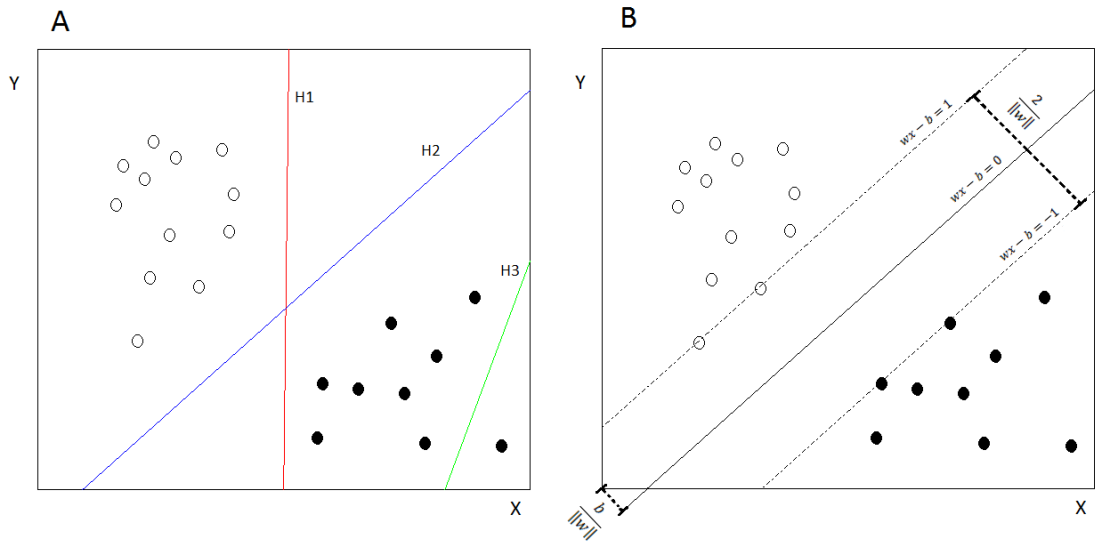

Figure 2. A) The maximum separation principle. The blue line is the best separator since the distance to the nearest point is the longest while the green line is the worst since it is not separating the white data points from the black ones. B) The margins of a separator. In two dimensional space, the margins can be defined as lines parallel two the separator which goes through the nearest data points to the separator also known as the support vectors.

The use of kerner function reduces the number of computations needed since it can be applied after the calculation of the dot product of two vectors. When the kerner function K is applied to the equation 10 it can be rewritten as follows:

eq. 12 $$f(x) = K\langle (w), (x) \rangle + b$$

Some of the most common kerner functions used in SVMs include the polynomial homogenous (eq. 13) and inhomogenous functions (eq. 14), Gaussian radial basis function (eq. 15) and the hyberbolic tangent (eq. 16).

eq. 13 $$K(x_i, x_j) = (x_i \cdot x_j)^d$$

eq. 14 $$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

eq. 15 $$K(x_i, x_j) = e^{(-\gamma \| xi - xj \|)}, \gamma > 0$$

eq. 16 $$K(x_i, x_j) = \tanh(\kappa x_i \cdot y_j + c)$$

2.3.3 Artificial Neural Networks

Artificial neural networks (ANN) mimic the activity of biological neuronal networks. They can be used to in various applications which include data classification. ANNs consist of layers of nodes which are connected to each other to form a network. The nodes consist of three components: inputs, activation function and output (F.E. Ahmed, 2005). The node architecture is illustrated in Figure 3.

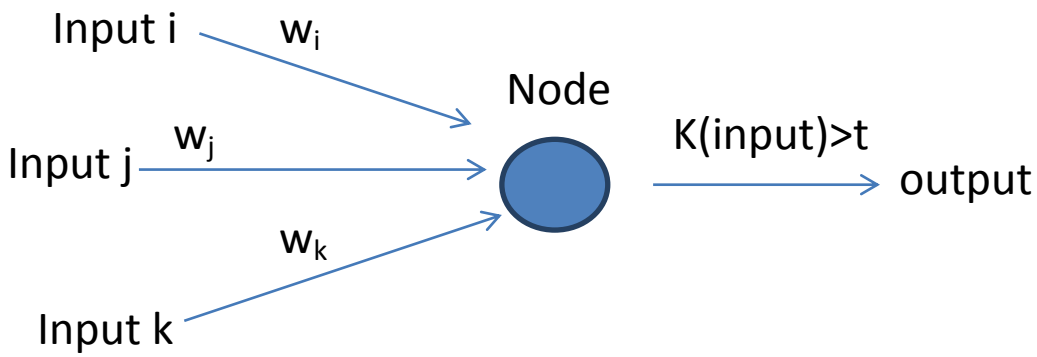

Figure 3. A schematic presentation of a node. In this figure node has three inputs i, j and k with weights w_i, w_j and w_k respectively. The inputs are processed by the node using the activation function K. If the threshold of activation t is reached the node is activated and the signal is transmitted forward.

The inputs which bring the signal to the nodes correspond to the synapses of biological neurons. The strength of inputs coming from different neurons is modified by application of weight for each input. The values of the inputs are processed by the activation function. If the value of function reaches to a certain threshold the node will be activated and otherwise it will remain non-active. If the node is activated the signal will be relayed forward to become an input of the connected node.

The activation function is formally presented in equation. 17.

eq. 17 $$f(x) = K(\sum_i w_i x_i),$$

where K is the activation function, w_i represents weigth and x_i represents the value of input i

The networks can have different topologies. In multilayer perceptrons, typically used for data classification, the nodes are organized to an input layer, one or more hidden layers and an output layer. The input values are first entered to the network through the nodes of the input layer which process the input values and transmit the signals to the nodes in the hidden layer. From the hidden layers the signals are finally transmitted to the nodes in the output layer. These nodes transform their input to the output of the network.

The networks can be either feedforward or recurrent. In the feedforward networks the signal is transmitted only to one direction unlike in recurrent networks in which the signal can proceed in both directions. In data classification the feedforward networks are more commonly used. A simple model of feedforward ANN with two hidden layers is illustrated in Figure 4.

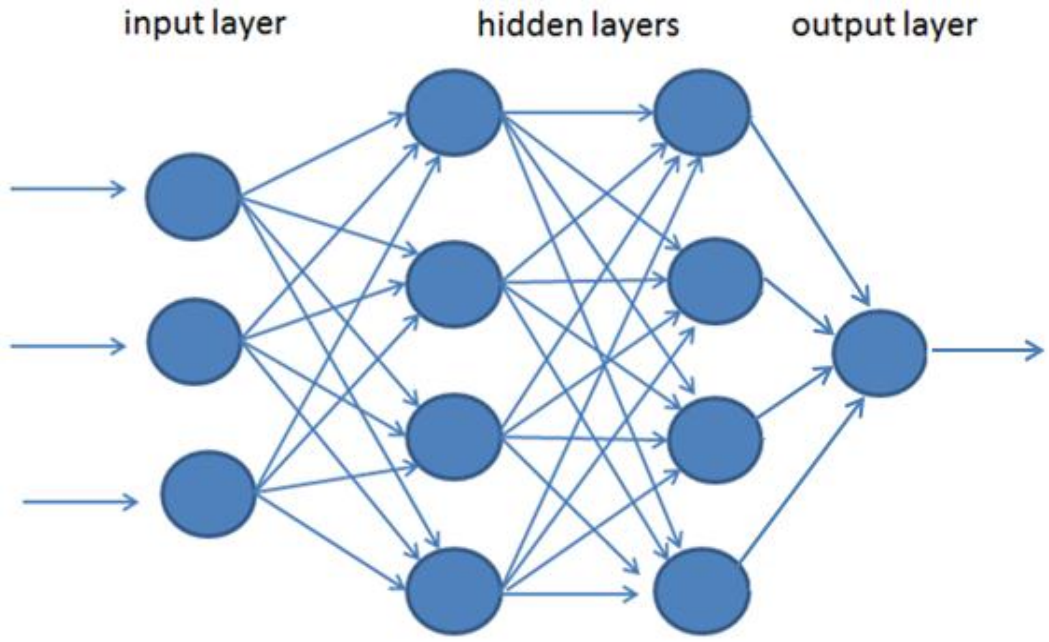

Figure 4. A schematic representation of a feedforward ANN with two hidden layers. The blue circles represent the nodes and the blue arrows represent the connected nodes and the direction of signaling.

The artificial networks can be trained using different methods. In the process of training the weights are adjusted for each node to attain an optimal network function. All different learning methods aim to minimize the value of a cost function which is a measure of the distance between the current network function and the optimal network function.

The networks used for classification are trained using supervised learning method. The training set can be represented as pairs (x, y), where x is a vector for input values and y denotes the class for x. The aim of supervised learning is to find a network function F such that $F: X \rightarrow Y$. The optimality of F is evaluated by the cost function which is usually the mean-squared error. To minimize the value of the cost function the weights are adjusted using the backpropagation algorithm. The training using backpropagation consists of two phases. In the first phase input values are feeded in to the network and the error in the output is determined. In the second phase weights are adjusted stepwise such that in the first step the error observed in the output layer nodes is minimized. This procedure continues layer by layer until all the weights are adjusted.

2.3.4 PON-P (Pathogenic-Or Not-Pipeline)

PON-P is a metatool which aims to overcome the limitations of individual pathogenicity prediction programs by combining several programs to predict the pathogenicity of variants. This pipeline is suggested to improve the reliability of the pathogenicity prediction and also gives a more comprehensive view on the effects of variants on the functional and structural level. The programs used by PON-P can be divided into two categories: Tolerance predictors and tools that predict the effects of the mutations to spesific structural and functional features of proteins. (A. Olatubosun et al. 2012)

The selection of tolerance predictors consists of eigth individual programs: SIFT, Panther, PolyPhen, PolyPhen-2, nsSNPanalyzer, PhD-SNP, SNAP, SNPs&GO and PON-P´s own tolerance predictor. The PON-P predictor utilizes a random forest classifier trained with 14,610 pathogenic missense variants retrieved from PhenCode database, IDbases and 16 individual Locus Specific Databases (LSDBs) and 17,393 neutral variants in dbSNP. The PON-P predictor considers eight features, which are

based on the output values of PhD-SNP, Polyphen-2, SIFT, SNAP and I-mutant-3. These features are listed in Table 1.

Table 1. List of features selected for PON-P predictor. In this table the feature name and it´s descrption are shown

Feature name	Description
PHDSNP_PRED	PHDSNP prediction
PHDSNP_REL	PHDSNP relibility
POL_PPH2_PROB	Polyphen2 classifier probability
SIFT_PROB	SIFT normalized probability
SNAP_PRED	SNAP prediction
SNAP_REL	SNAP reliability
SNAP_E_ACC	SNAP expected accuracy
IM_DDG	ddG value predicted by I-mutant

The features have been selected from a larger set by first constructing a random forest classifier including all features. During the process of training those features which affected the least to the accuracy of the prediction were discarded from the set after which the random forest classifier training was repeated using the obtained optimal subset. The PON-P classifies the variant in to one of three categories: neutral, unclassified and pathogenic. In addition, PON-P gives an estimate of the reliability of the prediction.

The structural and functional properties affected by the variation which are evaluated by PON-P include stability, aggregation, disorder and localization. In Table 2 all programs included in PON-P are listed.

Table 2 .The complete list of programs in PON-P. Table shows the name and the function of the program. In addition the website of each program is shown.

Program	Function	Website
SIFT	Tolerance prediction	http://sift.jcvi.org/
Panther	Tolerance prediction	http://www.pantherdb.org/tools/csnpScoreForm.jsp
Polyphen	Tolerance prediction	http://genetics.bwh.harvard.edu/pph/
PolyPhen-2	Tolerance prediction	http://genetics.bwh.harvard.edu/pph2/
nsSNPanalyzer	Tolerance prediction	http://snpanalyzer.uthsc.edu/
PhD-SNP	Tolerance prediction	http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm
SNAP	Tolerance prediction	http://rostlab.org/services/snap/
SNPs&GO	Tolerance prediction	http://snps-and-go.biocomp.unibo.it/snps-and-go/
Automute	Stability prediction	http://proteins.gmu.edu/automute/
Cupsat	Stability prediction	http://cupsat.tu-bs.de/
Dmutant	Stability prediction	http://sparks.informatics.iupui.edu/hzhou/mutation.html
Foldx	Stability prediction	http://foldx.crg.es/
I-mutant3	Stability prediction	http://gpcr.biocomp.unibo.it/~emidio/I-Mutant3.0/old/IntroI-
Mupro	Stability prediction	http://www.ics.uci.edu/~baldig/mutation.html

Table 2 continued.

Scide	Stability prediction	http://www.enzim.hu/scide/
SCpred	Stability prediction	http://www.enzim.hu/scpred/
SRide	Stability prediction	http://sride.enzim.hu/
iPTREE	Stability prediction	http://210.60.98.19/IPTREEr/iptree.htm
Aggrescan	Aggregation prediction	http://bioinf.uab.es/aggrescan/
Waltz	Aggregation prediction	http://waltz.vib.be/
Tango	Aggregation prediction	http://tango.crg.es/
DisProt	Disorder prediction	http://www.disprot.org/
FoldIndex	Disorder prediction	http://bip.weizmann.ac.il/fldbin/findex
FoldUnfold	Disorder prediction	http://antares.protres.ru/ogu/ogu.cgi
GlobPlot	Disorder prediction	http://globplot.embl.de/
IUPred	Disorder prediction	http://iupred.enzim.hu/
metaPrDos	Disorder prediction	http://prdos.hgc.jp/cgi-bin/meta/top.cgi
PrDos	Disorder prediction	http://prdos.hgc.jp/cgi-bin/top.cgi
PreLink	Disorder prediction	http://genomics.eu.org/spip/PreLink
RONN	Disorder prediction	http://www.bioinformatics.nl/~berndb/ronn.html
Spritz	Disorder prediction	http://distill.ucd.ie/spritz/
PRObalizer	Localization prediction	http://bioinf.uta.fi/PRObalizer/
WoLF-PSORT	Localization prediction	http://wolfpsort.org/

2.3.5 PhD-SNP

PhD-SNP is an SVM-based method which has been trained using human variant data from Swiss-Prot. The training set constitutes of 8241 neutral and 12 944 pathogenic variants. The SVM classifier has been constructed using LIBSVM software. The tranformation function used to map the data to feature space is the radial basis kernel (RBF) function. (E. Capriotti et al. 2006)

The predictor considers 44 input values. The first 20 components are reserved for the indication the amino acid substitution and the next 20 components encode the sequence environment of the variant site. The four remaining components encode the sequence profile information. The first and the second of these components encode the frequencies wild-type and mutant residues observed in multiple sequence alignment built on the basis of blast search against uniref90 database. The third component encodes for the number of aligned sequences covering the mutation site and the fourth component is the conservation index (CI).

The output values of the predictor range from 1 (neutral) to 0 (disease related) and the threshold has been set to 0.5. In addition the reliability index (RI) is determined for the prediction.

The reliability index is calculated as follows:

eq. 18 $$RI = 20 \times |Out - 0.5|,$$

where Out is the output value of the predictor.

2.3.6 SNPs&GO

SNPs&GO is a more recently developed tolerance predictor created by the developers of PhD-SNP. SNPsGO considers knowledge of Gene Ontology-term (GO) in addition to information about evolutionary conservation, sequence profile and sequence environment. The predictor is a SVM classifier trained with selected set of annotated variants retrieved from Swiss-Prot. The training set consists of 33 762 mutations observed in humans of which 16 330 are associated to diseases and 17 432 are considered to be neutral. All of the unclassified variants were excluded from the training set. Similarly to PhD-SNP the classifier has been constructed using LIBSVM software implementing the radial basis kernel (RBF) function (R. Calabrese et al. 2009)

The classifier considers 52 input values. Twenty components are reserved to indicate the amino acid substitution and another 20 components encode the sequence environment of the variant site. The sequence environment of the variant site constitutes of the mutant residue and eight adjacent amino acids taken from both sides of the variant site. Five input values encode the features of sequence profile. These include the frequencies of wild type and mutant residues observed in the sequence alignment, the coverage of the alignment at the position of the mutation, the conservation index and the last input value represents whether the sequence profile is present or absent.

The next five input values represent PANTHER output given the amino acid substitution. These features consist of the disease related probability of the substitution, probabilities of the wild type and mutant residues, Number of Independent Counts and the presence or absence of PANTHER output. The last two values encode the GO-information. The first value indicates the GO Log-odd score and the second value indicates the presense or absence of GO Log-odd score. Similar to PhD-SNP the output values of the predictor range from 1 (neutral) to 0 (disease related) and the threshold has been set to 0.5. In addition the RI is also given as output

2.3.7 SNAP

SNAP is a machine learning based method which makes its predictions based on a trained neural network. The predictor has been trained with a set variants of containing 40 641 non-neutral and 14 334 neutral mutations which were retrieved from the protein mutation database (PMD). The distinction between to neutral and non-neutral variants is based on the annotation data. To increase the number of neutral mutations 26 840 neutral pseudo mutants were constructed based on Swiss-Prot database. The neural network consists of 150 input and 50 hidden nodes. The features selected for the predictor are listed in Table 3. (Y. Bromberg and R. Burkhard, 2007)

Table 3. Features selected for SNAP predictor. The table shows the name of the feature and it´s description.

Feature name	**Description**
Explicit PSI-BLAST frequency profile	Represents the degree of conservation of the amino acid substituted
Relative solvent accessibility	Information about the relative solvent accessibility predicted by PROFacc
Secondary structure	Information about the secondary structure predicted by PROFsec
Sequence-only predictions of 1D structure	The change induced by the amino acid substitution to the predicted secondary structure and relative solvent accessibility predicted by PROFacc and PROFsec.
Pfam information	Pfam-information related to the mutation site including presense of domains and the model scores for the domains. The model scores include information about the conservation of the amino acid being substituted and whether the mutation improves or weakens the fit to the pfam-model.
PSIC scores	The Position specific independent count score
Residue flexibility	The change in the flexibility predicted by PROFbval
Transition frequencies	Represents the likelihood of a given mutation. The probability is based on the frequencies of amino acid triplets in the protein of PDB and UniProt.
Sequence environment	A window of five amino acids selected such that two residues flanking the mutated on both sides are considered in addition to mutant residue.

SNAP has two output nodes which are interpreted as the probabilities of variant being neutral or pathogenic.

SNAP also gives an estimate of the reliability of the prediction indicated by the reliability index formally presented as follows:

eq. 19 $$RI = integer\left(\frac{|output_{neutral} - output_{non-neutral}|}{10}\right),$$

where $output_{neutral}$ is the value given by the neutral output node and $output_{non\text{-}neutral}$ is the value given by the non-neutral output node. The RI ranges from 0, indicating the lowest possible reliability, to 9 indicating the highest possible reliability.

2.3.8 CanPredict

CanPredit is a tolerance predictor designed to distinguish between driver mutations from passenger mutations. The driver mutations initiate the cells transformation to cancer cells whereas passenger mutations occur during the progression of cancer but do not participate in this process. (J.S. Kaminker et al. 2007)

CanPredict is based on machine learning approach that predicts whether variant is a driver or a passenger mutation. The training set for the program consists of variants that can be classified into four categories: common polymorphisms (non-disease causing), mendelian disease causing, complex disease causing and cancer driver variants. The set of common polymorphisms contains 5747 variants having minor allele frequency greater than 20 %. The variants have been retrieved from dbSNP. The mendelian disease variant set contains 11456 mutations and has been retrieved from SwissProt database. The complex disease variant set consists of 27 variants which have been gathered from the previous work of the developers of CanPredict. The cancer driving mutations contains 1091 variants which have been retrieved from Catalogue of Somatic Mutation in Cancer database (COSMIC).

The predictor of CanPredict is a random forest classifier which has been constructed by implementing randomForest 4.5-16 package for R. CanPredict classifies the variants to three categories: likely cancer, likely non-cancer or not determined. The predictor considers three features: SIFT score, the PFAM-based logR E-value and the Gene Ontology Similarity Score (GOSS). The PFAM-based logR E-value describes how well a peptide sequence fits on to a profile constructed of a PFAM model. Variuants occurring in a region matching to a particular PFAM profile can either improve or impair the fit to the profile which can be assumed to have an effect on the

function of the protein. The GOSS measures the similarity of a gene to cancer associated genes. The GOSS score for a gene g is calculated as follows:

eq. 20 $$GOSS_g = \sum_{t \in T_g} \log_2 \left(\frac{x_t^{cancer}}{x_t^{non-cancer}} \right),$$

where T is the set of all GO-terms associated to gene g, x_t^{cancer} and $x_t^{non-cancer}$ are the number of occurences of the term t in genes associated to cancer and genes not associated to cancer respectively.

2.3.9 CHASM

Cancer-Spesific High-throughtput annotation of Somatic Mutations (CHASM) is another prediction program that attempts to identify cancer driver mutations. The driver mutation training set consists of 2488 missense variants that have been shown to cause oncogenic transformations. The variant data is based on findings of resequencing studies of breast, colorectal and pancreatic tumor and the COSMIC database. The passenger mutation dataset has been constructed of 4500 synthetically created mutations. The CHASM predictor is a random forest classifier which assigns the variants to be either passenger or driver mutations. The classifier has been created using PARF software (H.Carter et al. 2009).

The predictor considers 49 features in the prediction process. The features include:

- Changes in the physico-chemical properties
- The solvent accessibility of the wild-type amino acid residue
- Evolutionary conservation
- The sequence environment of the mutation site
- Substitution scores obtained from amino acid substitution matrices
- Substitution frequencies based on variant databases
- The presense of known protein domains in the site of mutation
- Structural features

All features used by the CHASM predictor are described in Table 20.

2.4 Tolerance predictors in cancer research

SIFT is undeniably the most frequently used tolerance predictor used for in silico analysis of SNPs in cancer research. Search from pubmed yielded over thirty publication related to cancer research reporting the use of SIFT. Many in silico studies of variants obtained from databases have utilized SIFT. In one of the studies, SIFT was used to assess the effects of missense variants affecting protein involved in steroid hormone metabolism which has been associated to cancer pathogenesis. In this study, 31 predicted pathogenic missense variants were discovered by SIFT (M.M. Johnson et al. 2005). In another study SIFT was used to assess 65 missense variants which have been reported in BRCA1. Twenty eight of these variants were deleterious according to SIFT (R. Rajasekaran et al. 2007). In a third study, deleterious variants related to leukemia in *MLL* were studied using SIFT. As a result, 10 missense variants were predicted to be pathogenic (C.G.P.Doss et al. 2009). Moreover, SIFT was used to evaluate variants in *IGF1R* which has been associated to breast and prostate cancer. In this study out of 32 nsSNPs evaluated with SIFT, 6 were predicted to be pathogenic (S.A. De Alencar et al. 2010). Furthermore, SIFT was used to study missense variants in BRCA1 and BRCA2 reported in a database in which data has been collected from French breast and ovarian cancer families (S. Caputo et al. 2012).

In addition to SNPs obtained from databases, SIFT has been as an additional tool to predict the consequences of missense variants along with experimental assays. In a mutational analysis of ovarian cancer cell lines, three predicted pathogenic variants in *B-Raf* and one *MEK1* were discovered using SIFT. The predicted pathogenic missense variant p.D67N in *MEK1* was further studied in an experimental assay. As a result it was shown that this variant increases *MEK1*:s kinase activity promoting it´s role as an oncogene (A.L. Estep et al. 2007). In another study, six predicted pathogenic missense variants in FGFR4 were discovered using SIFT. Two of these variants were shown to increase the autophosphorylation of FGR4 which leads to increased cell proliferation, invasion and metastatic potential rhabdomyosarcoma cells (J.G. Taylor et al. 2009). In a third study missense variants in *CSNK2A1P*, which has been associated to lung cancer, small cell lung cancer and leukemia, were studied using SIFT. One missense variant p.I133T in *CSNK2A1P* was found which was predicted to be pathogenic according to SIFT. In a cell growth assay this variant was shown to increase the proliferation of NIH3T3 cells. In addition, it was shown in this study that the presense

of p.I133T increases the degradation tumor suppressor protein PML (M.S. Hung et al. 2010). Furthermore, SIFT was used in studies related to gastric cancer, pancreatic cancer, and metastasis of prostate cancer and breast cancer (C.M. Robbins et al. 2010, L. Ding et al. 2010, J.D. Holbrook et al. 2011, L.J. Barber et. 2011).

SIFT has also been used in two studies related to cancer genetics. SIFT was used to assess familial colon cancer germline variants found in *MET*. As a result one predicted pathogenic missense variant p.T992I was discovered (D.W. Neklason et al. 2011). In another study, SIFT was used to evaluate germline variants related to familial breast cancer (G.T. Toh et al. 2008). As a result, three predicted pathogenic variants were discovered in BRCA1 and BRCA2.

SIFT has been also used frequently in epidemiological research. SIFT was used to study missense variants which might be predisposing to prostate cancer. SIFT was used to assess the effects of one variant p.Y424H located in CHEK2 which was found in more than one of the families involved in the study. SIFT predicted the variant to be pathogenic. However, in a functional assay using Saccharomyces cerevisiae as a model organism, this variant was not shown to alter the *CHEK2:*s function (M.D. Tischkowitz 2008). In addition to prostate cancer, SIFT was used in population studies related to several other cancer types including ovarian, breast, skin and colon cancer (H. Nan et al. 2008, F. Gu et al. 2008, M.D, L.A. Dong et al. 2009, L.M. Dong et al. 2008, L.L. Christensen et al.2008, T.V. Tavtigian et al. 2009, P.T. Campbell et al. 2009, J.A. Doherty et al.2010).

Four cancer-related studies have utilized CHASM to predict the effects of SNPs. The developers of CHASM used the program to evaluate the effects of 607 missense variants obtained from the tumour samples of 21 glioblastoma patients and 963 missense variants obtained by sequencing of 24 pancreatic cancer patients in order to discover cancer driving variants. As a result 49 predicted cancer driver variants obtained from the glioblastoma samples and 56 predicted cancer driver variants from the samples obtained from the pancreatic cancer samples (H. Carter et al. 2009, 2010). In addition CHASM was used to study missense variants discovered in ovarian tumour samples. CHASM predicted 122 of these variants to be cancer drivers (Cancer Genome Atlas Research Network, 2011). Furthermore, CHASM was used to assess

SNPs located in genes related to childhood medulloblastoma (D.V. Parsons et al. 2011)

The next most frequently used tolerance predictors in cancer research are PolyPhen-2, Mutation Taster and CanPredict. In a study conducted by two samples representing different tumour types MSS and MSI from two colorectal cancer patients were sequenced. The variant data was assessed with Mutation Taster and PolyPhen, which discovered 45 potentially pathogenic variants in MSS tumour sample and 359 potentially pathogenic variants in MSI tumour sample. Two missense variants p.W487R and p.E502G, which were located in *BMPR1A*, were chosen for a functional assay. Both variants were shown to impair the normal function of *BMPR1A* (B. Timmermann et al. 2010). In another study, Mutation Taster was used for find association of selected *BRCA2* variants to pancreatic cancer (L.J. Barber et al. 2011).

In a case-control study conducted by F. Le Calves-Kelm et al. 2011 potentially breast cancer predisposing mutations in *CHEK2* were evaluated in silico using PolyPhen-2 and SIFT. Polyphen-2 prediction yielded 10 possibly pathogenic and 18 probably pathogenic missense variants whereas SIFT predicted 23 of the missense variants to be pathogenic according to score threshold 0.05. In another study PolyPhen-2, SNP&GO and SIFT were used to evaluate a SNP (rs17632542) located in KLK3 which might be predisposing to prostate cancer. PolyPhen-2 and SNPs&GO predicted rs17632542 to be neutral whereas SIFT suggested the variant to be pathogenic (H. Parikh et al. 2011).

CanPredict was used to study mutations in tumour suppressor gene *DAPK3* for potential loss of function mutations. Three variants p.T112M, p.D161N and p.P216S predicted to be cancer drivers by CanPredict were evaluated in a phenotypic assay. The results of this assay showed that all of the three mutant type proteins lost the ability to regulate cell survival and proliferation (J. Brognard et al. 2011). In another study CanPredict was used to discover activating germline mutations, located in tyrosine kinase genes, which were screened from 94 acute myeloid leukemia patients (M.H.Tomassson et al. 2008). Furthermore, CanPredict and SIFT were used to assess 3 mutations in ADAM12 discovered in breast cancer cells (E. Dyczynska et al.2008).

Tolerance predictors seem to be quite frequently used in cancer research. SIFT was the first tolerance predictor that was created which probably explains the reason for

the substantially large number of studies were it has been used. On the other hand many of the programs that have been discussed in this review have been published very recently and therefore the number studies that have used these programs is small.

2.5 Comparison of the performance of tolerance predictors

SIFT and Panther are the simplest methods which rely only on the evolutionary information obtained from a multiple sequence alignment. The performance of more complex methods have been frequently compared to SIFT and Panther (E. Capriotti, Y.Bromberg and B.Rost, 2007; R.Calabrese et al.2009; I.A. Adzhubei et al.2010). Based on these studies the more complex prediction software perform better than SIFT and Panther in general.

However, in a benchmark study were several prediction software were compared, SIFT and Panther even outperformed methods that use also structural data in addition to evolutionary conservation information (J.Thusberg et al. 2010). These findings suggest that addition of structural features in the prediction model does not improve the performance of the predictor. In this study the performance of nine tolerance predictors (SIFT, Panther, nsSNPAnalyzer, PhD-SNP, Polyphen, PolyPhen-2, MutPred, SNAP and SNP&GO) were evaluated using different metrics including Matthews correlation coefficient (MCC) which is commonly used to describe the overall performance of classifiers. As a result SNP&GO and MutPred were found to be the best performing predictors having MCCs of 0.65 and 0.63 respectively. The next best performing predictors were Panther, SNAP having MCCs of 0.53, 0.47 respectively. Polyphen2a[1] and PhD-SNP had both MCCs of 0.43. Similarly Polyphen2b[2] and Polyphen1a[3] had both MCCs of 0.39. The worst performing predictors were Polyphen1a[4], SIFT and nsSNPAnalyzer which had MCCs of 0.39, 0.30 and 0.19 respectively.

[1] A binary classification model of Polyphen-2 such that probably pathogenic variants are considered pathogenic while possibly pathogenic and benign variants are considered neutral.
[2] A binary classification model of Polyphen-2 such that probably pathogenic variants and possibly pathogenic are considered pathogenic while benign variants are considered neutral.
[3] A binary classification model of Polyphen-1 such that probably pathogenic variants are considered pathogenic while possibly pathogenic and benign variants are considered neutral.
[4] A binary classification model of Polyphen-1 such that probably pathogenic variants and possibly pathogenic are considered pathogenic while benign variants are considered neutral.

In a study conducted by the developers of the Mutation Taster, the accuracies of seven methods including Mutation Taster, SNAP, Polyphen-1, PolyPhen-2[5], Panther and PMut (J.M. Schwarz et al. 2010). 2000 variants obtained from dbSNP were selected as a test set. As a result, Mutation Taster outperformed all other programs having an accuracy of 85.8 %. Polyphen1, PolyPhen-2_HumVar and PolyPhen-2_HumDiv methods performed moderately well having accuracies of 76.0%, 80.7% and 80.2% respectively. SNAP, PMut and Panther were the worst performing programs having accuracies of 68.5%, 65.4% and 50.8% respectively. This study also revealed the limitations of the methods to generate a prediction for variants. These limitations arise usually from the size of the analyzed protein. The Mutation Taster was the only method able to give predictions to all the variants included in the test set. Among the other programs, the number of non-predicted variants varied from 5 to 610.

In recent study the eigth tolerance predictors including Mutation Taster, MutPred, nsSNPanalyzer, PhDSNP, PolyPhen, SIFT, SNAP and SNP&GO were evaluated with a dataset collected from 168 variants located in a mismatch repair gene *MMR*. As a result, the best tolerance predictor based on MCC was nsSNPanalyzer having MCC of 0.61 and the second best was SNPs&GO with MCC of 0.59. Next best performing tolerance predictors were PhD-SNP, PolyPhen, MutPred and SNAP having the MCCs of 0.58, 0.45, 0.43 and 0.39 respectively. The worst performing methods were Mutation Taster and SIFT having MCCs of 0.37 and 0.36 respectively.

Tolerance predictors have been widely used in cancer research but how well do "regular" tolerance predictors perform against cancer specific methods? In cancer research the aim is either to find cancer predisposing mutations in the germline or distinguishing cancer driving mutations from passengers in the somatic cells. Neutral mutations in the germline selected in the training sets of tolerance predictors are generally considered as variants not affecting protein function and having high MAFs. It has been hypothesized that the passenger might have different characteristics than neutral variants because unlike the neutral variants; they might have effects on the function of proteins (H.Carter et al. 2009). CHASM and SNAP have been trained using pseudo mutations instead of common polymorphisms with high MAFs. Therefore, these methods should perform better in distinguishing driver mutations

[5] Both prediction models: *HumVar* and *HumDIv* were include in this study

from passengers. Furthermore, due to the fact that not all genes are associated to cancer it follows that not all pathogenic mutations are cancer promoting. The ability to distinguish cancer associated mutation have been enhanced in CanPredict by addition the GOSS feature which describes the how well the GO terms associated to given gene resemble to those frequently occurring in cancer associated genes.

In a study conducted by the developers of CHASM, five programs: CHASM, CanPredict, KinaseSVM, polyphen and SIFT were compared to evaluate their performance using three tests sets including the training set of CHASM and two sets of mutations located in TP53 and EFGR (H. Carter et al. 2009). The performance was measured calculating the precision and recall. As a result CHASM and CanPredict preformed significantly better in distinguishing cancer driving mutation from passengers than SIFT and Polyphen. CHASM outperformed all the other methods in having better precision and recall almost in all categories. As an exception KinaseSVM scored a better recall with the training set of CHASM and CanPredict had a better score with the TP53 mutation set. However, since KinaseSVM and CanPredict were not able to give prediction to all mutations in the datasets the results are not directly comparable. In conclusion this study suggests that the cancer specific tolerance predictors perform better in distinguishing driver mutations from passengers and CHASM is currently the best method available for this purpose.

2.6 Selection of tolerance predictor for variant data analysis

The performance is one of the most important criterion when selecting a tolerance predictor for the assessment of the pathogenicity of variants. Currently, there is not sufficient number of studies to rank all the tolerance predictors in terms of their performance. The results of these studies are also controversial due to the fact that different test sets tend to produce different results.

One of the major issues in benchmark testing is the lack of standardized tests sets. When comparing the results of different benchmark test it is clear that the selection of the variants for the test set has a great impact on the results. Another issue is the variety of different performance metrics used in the studies. Having many different metrics makes the comparison of the results of different benchmark test harder and on the other hand considering many different metrics to evaluate the performance makes

the ranking of tolerance predictors more complex. The solution for this problem would be selecting a single performance measure which could describe the overall performance accurately.

MCC can be considered to be the most robust performance measure since it is not affected by the bias that is is induced by the fact that different program have differences in the number of variants that they can assess (Baldi et al. 2000). However, MCC cannot be still considered as the absolute measure of performance. For example, a method having high MCC might have a low sensitivity. Therefore, if a program is selected solely based on it´s high MCC it might lead to a situation where the number of false negatives is high.

However, some conclusion of the performance of tolerance predictors can be drawn from the benchmark studies done so far. It seems that in general the methods that utilize multiple features tend to perform better than SIFT. On the other hand the performance of Panther, which also relies only on evolutionary information, seems to vary depending on the test set. In addition, machine learning methods utilizing random forest classifiers seem to perform better than methods utilizing SVMs, Bayesian or neural network classifiers. This observation is consistent with the results obtained from comparison of the performance SVMs and random forest classifier using the same training set and features (B. Li et al. 2009, H. Carter et al.2009).

Although being an important factor, the performance alone cannot be the only criterion when selecting a tolerance predictor for variant priorization. The type and size of the variant set to be analyzed sometimes might limit the selection of methods. Most of the tolerance predictors are able to analyze missense variants. Currently, only Mutation Taster and the most recent version of SIFT can assess indels. Since Mutation Taster is also able to analyze non-coding variants it is the most versatile program available at the moment.

Another matter to be considered, when selecting a tolerance predictor, is the format of variants data. Most of the tolerance predictors including CanPredict, Panther, PON-P, PhD-SNP, SNPs&GO and SNAP require the variant data to be submitted as amino acid substitutions and some of the methods even require the protein sequences to be submitted. However, variant data produced by many experimental methods is encoded as base substitutions. When analyzing variant data from genome wide studies the

variant data produced can be huge making the mapping of variants from DNA level to protein level a laborious task. Some methods such as SIFT, PolyPhen-2 Mutation Taster have intrinsic mapping modules which make the analysis of variant data efficient.

When working with large variant sets also another issue has to be considered. All tolerance methods such as Panther, PhD-SNP, SNAP and SNPs&GO do not support batch queries which makes the analysis of huge datasets inefficient. The batch sizes also vary significantly among the methods. PON-P limits the query to 10 protein sequences and 100 variants while other methods including Mutation Taster, SIFT and PolyPhen-2 accept almost unlimited number of variants.

In conclusion choosing a tolerance predictor for variant data analysis is not a straightforward task as it might seem. The choice of method does not depend on the performance alone. For a large variant set, methods that support batch queries are the only reasonable choices. Moreover, methods that are able to give a prediction to a large variety of variants are more useful than those which have a limited ability to assess variants. None of the methods cover all of these qualities thus there is no method that could be considered as the best for all purposes. Therefore, several methods should be chosen for the assessment of pathogenic variants which can complement each other.

3. Materials and Methods

3.1 Sample selection for sequencing

Based on the linkage analysis of Cropp et al. 2011, 21 families were selected for targeted re-sequencing in FIMM of which 12 families had shown a strong linkage to 2q37 and 6 families had a strong signal coming from 17q21-22. In addition, 3 families showing linkage to both regions were sequenced. 65 of the sequenced individuals were diagnosed with PRCA while 5 were unaffected. The sequenced families are listed in Table 4.

Table 4. The list of sequenced samples in FIMM.

family id	sequenced region	affected individuals	non-affected individuals
066	17q21-22	3	0
069	2q37	2	0
097	17q21-22	4	0
106	2q37	2	0
283	2q37	2	0
308	2q37	5	0
359	2q37 and 17q21-22	3	2
362	2q37	2	0
374	2q37	3	0
375	17q21-22	3	0
386	2q37	3	0
399	2q37	3	0
400	2q37 and 17q21-22	3	0
401	17q21-22	4	1
402	2q37	3	0
414	17q21-22	3	1
421	2q37 and 17q21-22	3	0
427	17q21-22	5	1
429	2q37	3	0
438	2q37	3	0
443	2q37	3	0

3.2 Targeted re-sequencing in FIMM

The DNA samples were sequenced using Illumina Solexa Analyzer IIx sequencing platform and NimbleGen´s array probes were used to capture the target regions. The reads were aligned and variant were called using an in house variant calling pipeline (VCP).

In the first step duplicated reads were removed from the raw sequencing data (A.M Sulonen et. al 2011). Reads were aligned to genome with bwa (Li H. and Durbin R,

2009) using the most recent human genome build hg19 as the reference. Paired end anomalies were detected and reported in GFF-format. Circos-software was used to visualize the paired end anomalies. (A.M Sulonen et. al 2011) Variant calling was done using the SAM tools for SNPs and small indels, and Pindel was used for the detection of larger insertion and deletions (H. Li et al. 2009; Ye K et al. 2009).

The VCP- pipeline produces variant files in VCF-file format. In addition to genomic coordinates, alleles, allele frequencies and quality measures the final output files include annotation data from EnsEMBL, dbSNP build 130 and 1000 genomes database (P.Flicek et al. 2011; S.T Sherry et al. 2001; The 1000 Genomes Consortium, 2010).

3.3 The bioinformatics workflow for variant data analysis

In order to prioritize variants, discovered by the FIMMs VCP-pipeline, for further experimental study, a bioinformatics workflow was developed. As the first step in the bioinformatics workflow variant data from FIMM-variant files was filtered to remove those variants which were not located in the genomic region of interest. The remaining variants are annotated using a local database. The variants located in genes were submitted to tolerance predictors to assess their pathogenicity. Next, those variants predicted to be pathogenic were selected for further investigation. The genes containing pathogenic variants are selected to form the candidate set. Subsequently, these genes were compared to a set known PRCA associated genes obtained from databases. In addition, Gene Ontology terms (GO) and biological pathways were determined for the candidate genes. In the next stage of the worklflow a PRCA gene set was constructed for GO term and pathway enrichment analysis. The GO-terms and pathways associated to the candidate genes were compared to PRCA related GO-terms and pathways obtained from the enrichment analysis of the PRCA gene set. Finally, based on the characteristics of candidate variants and gene specific information of the candidate genes potential variants are selected for genotyping. The bioinformatics workflow is summarized in Figure 5.

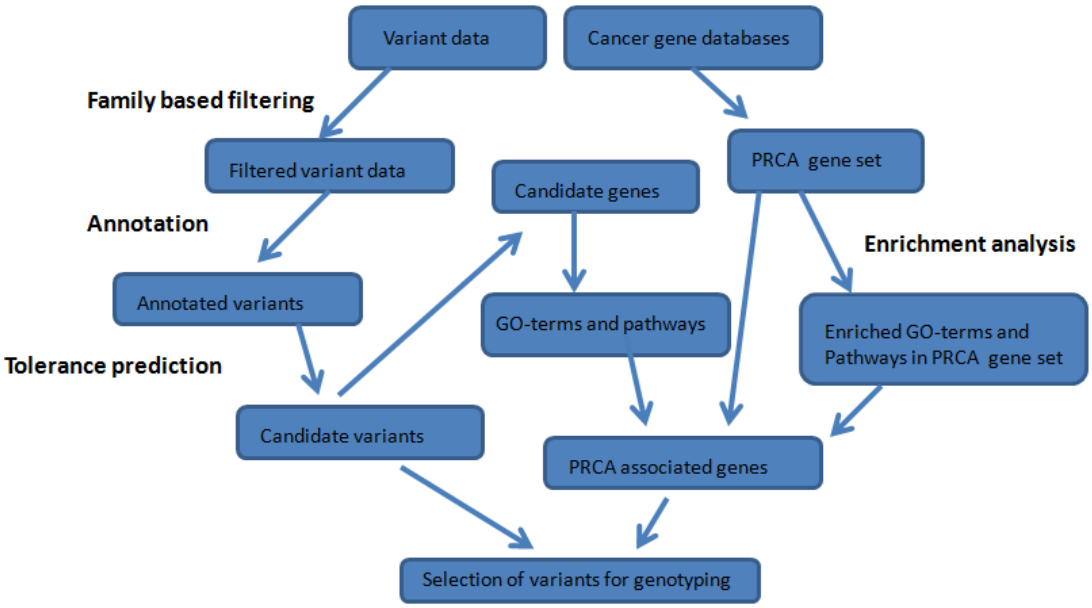

Figure 5. The bioinformatics workflow. In the first step the variant data is filtered based on the affection status of the samples. Remaining variants are annotated and tolerance prediction is made for each variant. The variants predicted to be pathogenic are selected for further investigation. The genes which the pathogenic variants are targeting are selected to for the candidate gene set. The candidate set is compared to set of known PRCA associated genes and pathway information and GO-terms are retrieved for each gene in the candidate gene set. These GO-terms and pathways are compared to a list of pathways and GO-terms enriched in PRCA set. Based on the characteristics of candidate variants and similarity of the candidate genes to known PRCA genes the most potential variants are selected for further experimental research.

3.4 Variant data filtering and the construction of the local annotation database

In the first step of the bioinformatics workflow variants located outside of the regions of interest were filtered out from the variants discovered by the FIMMs VCP-pipeline. Moreover, from the remaining variants only those which where common to all affected family members were selected for further study. In order to gain more knowledge of the variants called by the VCP-pipeline, a local database was constructed from a selection of annotation tracks in UCSC genome browser (K.R Rosenbloom et al. 2010) and other resources including MicroRNA.org (D Betel et al. 2008), EnsEMBL (P.M Flicek et al. 2011) and VISTA (A Visel et al. 2007). All tracks and resources used to build the annotation database are listed in Table 5.

The UCSC genome broser tracks were retrieved using the table browser application in the UCSC genome browser web page (D. Karolchik et al. 2004). The datasets from MicroRNA.org, EnsEMBL and VISTA were retrieved from the ftp-sites except for the gene symbols, and descriptions corresponding to EnsEMBL gene ids and dbSNP-

ids (build 135) for known variants. These datasets were retrieved from EnsEMBL using EnsMART (Kasprzyk A et al. 2004)

Table 5. Tracks and resources used to construct the local database used for annotation of variants. The table shows the track/ resource name and it´s description.

Track/resource	Description	Url
EnsEMBL-genes	Gene transcripts and miRNA known to EnsEMBL	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ensGene.txt.gz
Sno/miRNA	miRNA and SnoRNA	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/wgRNA.txt.gz
TS-miRNA sites	miRNA binding sites	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/targetScanS.tx.gz
MicroRNA.org	miRNA binding sites	http://www.microrna.org/microrna/home.do
Tnx-factor chip	Transcription binding sites	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/
TFBS-conserved	Conserved transcrption binding sites	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/tfbsConsSites.txt.gz
Poly(A)	Polyadenylation signal sites	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/PolyDB.txt.gz
CpG-island	CpG islands	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz
DNAse-clusters	DNaseI sensitive sites	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/
VISTA	Enhancer regions	http://enhancer.lbl.gov/

3.5 Description of datasets selected for the local database

The genomic coordinates for genes were obtained from the EnsEMBL-genetrack obtained from the UCSC genome browser. The EnsEMBL-genes track contains the start and end coordinates of genes and also more detailed information such as the coordinates of exons, introns, UTR-regions and the consensus coding sequences (CCDSs) of all transcripts known to EnsEMBL. In order to broaden the variety of coding elements incuded in the database, coordinates of predicted miRNAs and SnoRNAs provided by Sno/MiRNA-track were retrieved. The data for this track has been collected from miRBase and snoRNABase (S. Griffiths-Jones et al. 2006; L. Lestrade and M.J Weber 2006).

To determine if variants are located in gene regulatory sites a selected set of UCSC regulatory tracks were downloaded. TS-miRNA sites -track contains predicted regulatory target sites for conserved miRNA-families in the 3´UTR regions of genes. The predictions for this track were made with TargetScanHuman 5.1 (B.P. Lewis et al. 2005; A. Grimsom et al. 2007; R.C. Friedman et. al. 2009). In addition to TS-miRNA sites track another dataset from MicroRNA.org was used as a source for miRNA-binding site coordinates (D. Betel et. al. 2008). The MicroRNA database

consists of miRNA binding-sites predicted by MIRANDA-software (A.J Enright et al. 2003). The database is divided into four sections: Conserved binding sites with good mirSRV score, non-conserved binding sites with good mirSRV score, conserved binding sites with non-good mirSRV score and non-conserved binding sites with non-good mirSRV score. Only the datasets having good mirSRV scores were retrieved and used in the annotation of variants.

Tnx-factor chip combines the results of 5 ChIP-seq assays which have been done using several different cell lines and transcription-factor targeting antibodies (G.M. Euskirchen et. al. 2007; M.E. Hudson and M. Snyder, 2006). TFBS-conserved track has been created by Matt Weirauch and Brian Raney at The UCSC contains the genomic coordinates and the score of transcription binding sites conserved in multiple sequence alignment done that has been constructed of human, rat and mouse (G.E.Liu et al. 2008). The profile matrices have been obtained from TRANSFAC (V. Matys et al. 2006). The alignment scores have been calculated using tfloc (Transcription Factor binding site LOCator) which has been developed in the University of Pensylvania and modified by Matt Weirauch from the UCSC (G.E. Liu et al. 2008).

Poly-A-signaling sites were studied using Poly (A)-track. This track was built using data retrieved from poly-adenylation signaling database Poly_DB (H. Zhang et al. 2005). CpG-island and DNase-clusters tracks were used to predict if mutations are located in promoter regions or other possible regulatory sites. CpG-island track contains information about the CpG rich areas also known as CpG-islands which have been associated to transcription start sites, promoters and are also common targets of epigenetic regulation (M. Gardiner-Garden and M. Frommer, 1987). DNAse-clusters track shows DNaseI hypersensitive areas in the human genome which are commonly associated to regulatory regions in general. The data has been collected from DNaseI treatment assays involving a large variety of cell types (P.J. Sabo et al. 2004, 2006).

The coordinated for known enhancer were retrieved from Vista A. (Visel et al. 2007) Enhancers are cis-regulatory elements which can regulate gene expression acting from a distance. The dataset has been generated based on a comparative genome analysis in which human enhancer activity has been studied in transgenic mice (A.Visel et al. 2007).

3.6 Annotation of variants with Python scripts

To obtain information of the variants their coordinates were matched to the coordinates of genes and other functional elements in the local database using Python scripts. The Python scripts were done using Python 3. Annotations from the original from the original FIMM VCP-pipeline were also included in the annotated variant files.

3.7 Pathogenicity prediction

The variants located in known genes were selected for pathogenicity prediction. The variants were submitted to the tolerance predictors as batch queries. The input files were prepared from the annotated variant files. Mutation Taster requires short sequence snippets covering several bases around the variant and the EnsEMBL-id of the transcript targeted by the variant. The variant files for Mutation Taster were constructed using a Python script designed to automatically extract the correct sequence snippet from the genomic sequence and insert the variant and the wild-type allele in the sequence. The script extracts also the EnsEMBL transcript-ids from the annotated variant files. The genomic sequences for the genes were retrieved from the EnsEMBL gene track from UCSC using the table browser application. The input files were submitted to the Mutation Taster server and the results were automatically analyzed using Perl-scripts provided by the developers of the Mutation Taster.

The variant files for PolyPhen-2 batch query were also prepared using a Python script. The script extracts the genomic position, mutant and the wild-type allele from the annotated variant file. PolyPhen-2 batch queries were submitted to the polyphen2 bgi-website manually. Both prediction models were used in pathogenic prediction.

The PON-P requires two separate input files. The first file contains the protein sequences and the second file contains the variants encoded as amino acid substitutions. The protein sequences for PON-P were retrieved manually from UniProt. The file containing the amino acid substitutions was prepared using PolyPhen-2 output files. In addition to PON-P´s own tolerance predictor three other tolerance predictors were selected for prediction including SNAP, SIFT and PhD-SNP.

3.8 Construction of candidate and PRCA gene sets

Based on the results of tolerance prediction, all genes containing pathogenic variants were listed. This list will be further referred as the candidate gene set. To discover if this list has genes associated to prostate cancer based on previous research; the candidate gene set was compared to a set of known prostate cancer associated genes. This set was constructed searching two databases for genes associated to prostate cancer: COSMIC and Dragon Database of genes associated with Prostate Cancer (DDPC) (S.A. Forbes et al. 2011, M. Magungo et al. 2011). This list is referred furher as the PRCA gene set.

3.9 Gene Ontology-term and pathway enrichment analysis for PRCA gene set

In enrichment analysis genes or proteins from a given set, which is usually an end result of a high-throughput experiment such as a microarray study, are divided into different categories based on a property of interest. In Gene Ontology enrichment analysis genes are divided in to categories based on the GO terms associated to these genes whereas in pathway enrichment analysis the categories are represented by biological pathways associated to the genes. The categories that are over- or underrepresented in a given gene set can be determined by comparison to a reference set which is usually either the whole set of known genes or a set of genes in a microarray chip. The over- and underrepresented genes are determined using statistical tests (D. Duncan et al. 2010).

To elucidate important pathways and typical characteristic of genes involved in PRCA, Gene Ontology term (GO) and pathway enrichment analysis was conducted for the prostate cancer gene set. Both enrichment analyses were made using WebGestalt2 (D. Duncan et al. 2010). In both analyses genome was used as the reference set and the P-value was adjusted for multiple hypothesis testing using BH-method (Benjamini -Hochberg). The minimum number of genes in a category was set to 2 and the significance threshold was set to 0.01. The pathway enrichment analysis was done against all three available pathway databases: Pathway Commons (E.G. Cerami et al. 2011), Wikipathways (A.R. Pico et al. 2008) and KEGG (M. Kanehisa and S. Goto, 2000).

3.10 Search for Gene Ontology terms and pathways for candidate genes

To compare the characteristics between the candidate genes and prostate cancer associated genes, GO-terms and pathways related to candidate genes were retrieved. The GO-terms for candidate genes were retrieved with EnsMart (Kasprzyk A et al. 2004) and the pathways related to the candidate genes were retrieved from Pathway commons, KEGG and Wikipathways.

4. Results

4.1 Variant statistics

After the family based filtering and annotation of the variants discovered by the FIMM´s variant calling pipeline, the following distributions of variants in chromosome 2 and 17 were obtained respectively (Table 6 and Table 7).

Table 6. In this table the distribution of variants in chromosome 2 is shown. The table shows the family an the number of all variants discovered, variants in genes, coding SNPS, coding Indels,missense mutation and nonsense mutations.

Family	Total number of variants	Variants located in genes	Non-coding variants	Coding variants	Coding SNPS	Coding Indels	snSNPS	Missense SNPS	Nonsense SNPS
69	8764	2903	2703	200	179	21	157	22	0
106	6042	1946	1815	131	109	22	101	8	0
283	5290	1641	1544	97	87	10	81	6	0
308	5893	2209	2058	151	136	15	118	18	0
359	6106	1902	1707	195	180	15	162	18	0
362	8957	3081	2908	173	157	21	140	17	0
374	4110	1558	1450	108	97	11	80	17	0
386	5507	2082	1967	115	105	10	95	10	0
399	5742	1817	1698	119	106	13	94	12	0
400	6505	2213	2062	151	137	14	122	15	0
402	6981	2481	2298	183	168	15	150	18	0
421	5245	1686	1574	112	106	6	93	13	0
429	6089	2447	2296	151	137	14	123	14	0
438	3515	1219	1140	79	73	6	66	7	0
443	7201	2565	2394	171	155	16	138	17	0

Table 7. In this table the distribution of variants in chromosome 17 is shown. The table shows the family an the number of all variants discovered, variants in genes, coding SNPS, coding Indels, missense mutation and nonsense mutations.

Family	Total number of	Variants located in	Non-coding variants	Coding SNPs	Coding Indels	snSNPS	Missense SNPs	Nonsense SNPs
66	16680	9421	8762	659	110	556	100	3
97	10009	5641	5151	490	67	406	81	3
359	14743	8418	7815	603	94	508	92	3
375	14263	8342	7794	548	87	464	82	2
400	15295	8696	8053	643	93	544	96	3
401	9311	5354	4934	420	49	351	67	2
414	12192	6855	6355	500	69	415	83	2
421	13356	7711	7192	519	65	433	84	2
427	9491	5695	5293	402	58	333	66	3

4.2. Pathogenicity prediction results

4.2.1 Non-synonymous single nucleotide polymorphisms

The non-synonymous single nucleotide polymorphisms (nsSNPs) consist of two types of variants: missense and nonsense variants. The missense variants were evaluated using Mutation Taster, PolyPhen-2 and PON-P. In addition to PON-Ps own tolerance predictor, three additional tolerance predictors PHD-SNP, SNAP and SIFT, included in PON-P, were applied in the analysis. The nonsense mutations were discovered using Mutation Taster and PolyPhen-2.

Based on the results of Mutation Taster, PolyPhen-2 and PON-P, 20 missense variants discovered in chromosome 2 were predicted pathogenic at least by one of three predictors. The prediction results for these variants are shown in Table 8.

The predicted pathogenic non-synonymous variants in chromosome 2 were distributed to 13 genes. Six variants of the 20 missense variants were found in *COL6A3* of which two were novel according to the latest dbSNP build. Three genes *hGC_1642047*, *LOC151174* and *TRAF3IP1* had two predicted pathogenic variants which were all known to dbSNP. The other variants predicted to be pathogenic were located at *OR6B2*, *SCLY*, *OR6B3*, *PRR21*, *LRRFIP1*, *KLHL30*, *MLPH*, *ESPNL* and *ASB18*. All of these variants are known to dbSNP. Nonsense variants were not found in chromosome 2.

In chromosome 17 there were 49 missense variants predicted pathogenic at least by one of the three tolerance predictors. In addition two nonsense variants were found. The prediction results for these variants are shown in Table 9. The non-synonumous variants in chromosome 17 were distributed to 40 genes. The nonsense variants were located in *C17orf57* and *CDC27* which also had 6 missense variants. *HAP* and *KRT32* both had 3 predicted pathogenic missense variants. In *BRCA1*, *JUP*, *CC6B*, *GSDMA* and *GSDMB* there were 2 predicted pathogenic missense variants were discovered in each gene. Of the 49 missense variants 6 were novel ones. These variants are located i*n GSDMA*, *GSDMB*, *TBX21*, *MSL1*, *ADAM11* and *AP2B1*.

Table 8. Predicted pathogenic nsSNPs in chromosome 2.

Coordinate	RS-number	Variant type	Gene name	Mutation Taster	PolyPhen-2	PON-P	PHD-SNP	SIFT	SNAP
234775346	rs13384181	missense	hCG_1642047	N	P	N	N	P	N
234775675	rs213544	missense	hCG_1642047	N	P	P	P	N	P
238247734	rs36104025	missense	COL6A3	P	P	UV	N	P	P
238280504	-	missense	COL6A3	P	N	N	N	P	N
238283511	-	missense	COL6A3	P	P	UV	P	P	P
238289980	rs113897824	missense	COL6A3	P	P	N	N	N	N
238289984	rs112010940	missense	COL6A3	P	P	UV	P	P	P
238427251	rs3751107	missense	MLPH	N	P	UV	N	N	N
238668783	rs3213869	missense	LRRFIP1	N	P	N	N	P	N
238990388	rs3210400	missense	SCLY	N	N	UV	N	P	P
239039970	rs78076311	missense	ESPNL	N	P	N	N	N	N
239049928	rs112962843	missense	KLHL30	N	N	UV	N	N	P
239133980	rs7572285	missense	LOC151174	N	P	no data	no data	no data	no data
239134063	rs7584376	missense	LOC151174	N	P	no data	no data	no data	no data
239237388	rs61742338	missense	TRAF3IP1	P	N	N	N	P	N
239237953	rs12464423	missense	TRAF3IP1	N	P	N	N	P	N
240969312	rs61730690	missense	OR6B2	N	P	P	P	P	P
240981375	rs6732185	missense	PRR21	N	N	N	N	P	P
240984789	rs12465491	missense	OR6B3	N	N	UV	P	P	P

Table 9. Predicted pathogenic nsSNPs in chromosome 17.

Coordinate	RS-number	Variant type	Gene name	Mutation Taster	Polyphen-2	PON-P	PHD-SNP	SIFT	SNAP
32647831	rs1133763	missense	CCL8	N	P	UV	N	N	P
32957114	rs56879769	missense	TMEM132E	P	N	N	N	N	N
33269648	rs2230553,	missense	CCT6B	N	N	UV	P	P	P
33286664	rs2230552	missense	CCT6B	P	P	UV	P	P	P
33520965	rs8079507	missense	AMAC1	N	P	UV	N	P	P
33771996	rs72483216	missense	SLFN13	N	P	UV	P	P	P
33951431	-	missense	AP2B1	P	N	N	N	N	N
34871721	rs2306595	missense	MYO19	N	P	P	P	N	P
34893655	rs61755368	missense	PIGW	P	P	UV	P	P	N
35311114	rs115760333	missense	AATF	P	N	N	N	N	N
35743010	rs1714987	missense	C17orf78, ACACA	N	N	N	N	P	P
35981285	rs34337635	missense	DDX52	P	N	N	N	P	N
36889559	rs2879097	missense	CISD3	N	N	UV	P	P	P
38062217	rs2305479	missense	GSDMB	N	P	UV	N	N	P
38062422	-	missense	GSDMB	N	P	P	P	P	P
38127835	-	missense	GSDMA	N	P	UV	P	P	N
38131187	rs56030650	missense	GSDMA	N	N	UV	N	N	P
38282536	-	missense	MSL1	P	N	N	N	P	N
38324554	rs41283419	missense	CASC3	P	N	N	N	N	N
38416827	rs142659099	missense	WIPF2	P	N	N	N	N	N
38450248	rs4135012	missense	CDC6	P	N	UV	N	P	P
38555188	rs61732514	missense	TOP2A	P	N	UV	N	N	N
38928014	rs9898164	missense	KRT26	N	P	UV	P	P	P
39525750	rs61729509	missense	KRT32	N	P	N	P	P	N
39577215	rs8071814	missense	KRT37	N	N	no data	no data	no data	no data
39619115	rs2071563	missense	KRT32	N	P	UV	P	P	P
39622068	rs2071561	missense	KRT32	N	P	UV	P	P	P
39637244	rs743686	missense	KRT35	N	P	no data	no data	no data	no data
39659913	rs9891361	missense	KRT13	N	N	N	N	P	P

Table 9 continued.

Coordinate	RS-number	Variant type	Gene name	Mutation Taster	Polyphen-2	PON-P	PHD-SNP	SIFT	SNAP
39884065	rs35612698	missense	HAP1	N	N	UV	N	P	P
39884583	rs4796693	missense	HAP1	N	N	UV	N	N	N
39890876	rs4796604	missense	JUP, HAP1	N	N	no data	no data	no data	no data
39925713	rs41283425	missense	JUP	P	N	UV	P	P	N
39983808	rs1046404	missense	NT5C3L	N	N	N	N	P	P
40722029	rs665268	missense	MLX	N	P	no data	no data	no data	no data
41244435	rs16941	missense	BRCA1	N	N	no data	no data	no data	no data
41246481	rs1799950	missense	BRCA1	N	N	no data	no data	no data	no data
42745180	rs77416189	missense	C17orf104	N	P	UV	N	P	P
42850243	-	missense	ADAM11	P	N	N	N	N	N
43111558	rs117298907	missense	DCAKD	P	P	UV	P	P	N
45234298	rs62077263	missense	CDC27	P	N	N	N	N	N
45234343	rs3208659	missense	CDC27	P	N	N	N	N	N
45234360	rs62077264	nonsense	CDC27	P	N	no data	no data	no data	no data
45234403	rs75184508	missense	CDC27	P	N	N	N	N	N
45234420	rs78493795	missense	CDC27	P	N	N	N	N	N
45234430	rs78072949	missense	CDC27	P	N	N	N	N	N
45468858	rs118004742	nonsense	C17orf57	P	N	no data	no data	no data	no data
45820022	-	missense	TBX21	P	N	UV	P	P	P
47246163	rs7224888	missense	B4GALNT2	N	N	P	P	P	P
47246956	rs61743617	missense	B4GALNT4	N	N	N	N	P	P
47293906	rs2233369	missense	ABI3	N	P	UV	N	N	N

4.2.2. Indels

The total number of pathogenic indels predicted by Mutation Taster was 30. Twenty four of these indels are located in chromosome 17 and 5 in chromosome 2. The five predicted pathogenic indels in chromosome 2 were distributed to five genes. Two predicted pathogenic indels (rs74521182 and rs72316729) are located in *SCLY*. Since *SCLY* and *UBEF2F* overlap, rs72316729 is also located in *UBE2F*. The three other indels are also known to dbSNP and they are located in *AGAP1*, *IQCA1* and *CXCR7*.

The 24 predicted pathogenic indels located in chromosome 17 were distributed to 19 genes. Four of these indels are novel and they are located in *MYO19, SMARCE, MRPL10* and *MBTD1*. In addition to the novel indel in *MRPL10* this gene has also another predicted pathogenic indel (rs34919891) which is known to dbSNP. Of the remaining 19 known indels, three are located in *HOXB3.* Two in predicted pathogenic indels were discovered in *ACACA* and *HEXIM1*. The complete list of predicted pathogenic indels is shown in Table 10.

Table 10. Predicted pathogenic indels in chromosome 2 and 17.

Chromosome	Position	dbSNP	Gene	Genotypes
17	34100350	rs58543174	MMP28	+C/+C
17	34863729	-	MYO19	-G/*,*/-G
17	35696820	rs67231825	ACACA	-AAAG/*,*/-AAAG
17	35766563	rs150239106	ACACA	-A/-A,-A/*,*/-A
17	38804135	-	SMARCE1	-T/*,*/-T
17	38858134	rs11309872	KRT24	-A/-,-A/*
17	41150464	rs11305686	RPL27	-G/-G
17	41196821	rs33947868	BRCA1	-TTT/-TTT,-TT/*,-T/-T
17	42263978	rs5820525	C17orf65	*/-G, -G/*,-G/*
17	42884847	rs66761765	GJC1	*/-AAAAG, -AAAAG/*
17	43192549	rs75518897	PLCD3	+C/+C
17	43226472	rs111687345	HEXIM1	+TT/+TT,+TT/+T,+T/*,*/+T
17	43226477	rs7216041	HEXIM1	-C/*,-C/-C
17	45438886	rs10538163	C17orf57	-AGTG/-AGTG,*/-AGTG,-AGTG/*
17	45900828	rs34919891	MRPL10	*/-AA
17	45906639	-	MRPL10	+GAAGGAAG/+GAAG,+GAAG/+GAAG
17	46630569	rs10554930	HOXB3	-ACA/*,*/-ACA,-ACA/-ACA
17	46631064	rs66599671	HOXB3	+A/+A,+A/+AA,+A/*,*/+A
17	46651512	rs138436281	HOXB3	+T/*,*/+T
17	47014651	rs5820737	SNF8	-T/*,*/-T,-T/-T
17	48172255	rs16942045	PDK2	-AGGAT/-AGGAT,*/-AGGAT,-AGGAT/*
17	48445674	rs111380254	MRPL27	-CTGGTCAG/*,*/-CTGGTCAG
17	48912895	rs113159761	WFIKKN2	-T/*,-T/-TT,-T/-T
17	49268944	-	MBTD1	-T/-TT,-TT/-T,-T/-T
2	236761414	rs142341634	AGAP1	+CAGG/*,+CAGG/+CAGG,*/+CAGG,-AA/-
2	237247036	rs111440161	IQCA1	+A/+A,*/A,*/+T
2	237478329	rs34276711	CXCR7	*/-GT,+T/*
2	238969572	rs74521182	SCLY	+G/+G
2	239002480	rs72316729	SCLY/UBE2F	-TTG/-TTG,*/-TTG,-TTG/*,-A/-A

4.2.3 Noncoding single nucleotide polymorphisms

The total number of predicted pathogenic variants located in non-coding regions of genes predicted by Mutation Taster was 51. Sixteen of these variants were found in chromosome 2 and 35 in chromosome 17. Only one predicted pathogenic non-coding variants in chromosome 2 was novel. It is located in *COPS8* in which also a known predicted pathogenic non-coding variant (rs61433497) was found. Of the known 15 predicted pathogenic non-coding variants, three were discovered in *RAB17*. Two predicted pathogenic variatns were found in *RBM4* and *ASB1*. The remaining 7 variants were found in *COL6A3*, *MLPH*, *LRRFIP1*, *RAMP1*, *ESPNL*, *MYEOV2* and *OTOS*.

In chromosome 17, 4 novel and 31 known predicted pathogenic non-coding variants were discovered. The novel variants were found in four genes: *STARD3*, *ORMDL3*, *NAGLU* and *OSBL7*. The known variants were distributed to 24 genes. Three of these variants were located in *ACACA*. *SLFN12*, *MRPL10*, *SNF8*, *SCL35B1*. Two keratine

proteins coding genes: *KRT15* and *KRT31* had two predicted pathogenic non-coding variants. The complete list of pathogenic non-coding variants is shown in Table 11.

Table 11. Predicted pathogenic synonymous single nucleotide polymorphisms in chromosomes 2 and 17.

Chromosome	Positions	dbSNP	Gene	Reference	Variant
17	33477242	rs80100968	UNC45B	G	A
17	33760082	rs3106577	SLFN12	T	C
17	33760211	rs1838150	SLFN12	C	T
17	33849646	rs4796095	RP11-1094M14.3	A	G
17	34270582	rs117265188	LYZL6	C	T
17	34344961	rs111879665	CCL23	C	T
17	34923498	rs17138347	GGNBP2	A	G
17	35295230	rs111599710	LHX1	G	A
17	35696804	rs58654829	ACACA	G	A
17	35702299	rs77402427	ACACA	G	A
17	35766475	rs72828246	ACACA	A	G
17	36003420	rs35498905	DDX52	G	C
17	37817230	-	STARD3	G	A
17	37886986	rs3809717	C17orf37	C	A
17	38020419	rs1453559	IKZF3	T	C
17	38064469	rs11078928	GSDMB	T	C
17	38082684	-	ORMDL3	C	T
17	38804179	rs7342818	SMARCE1	C	A
17	39346622	rs79470847	KRTAP9-1	T	A
17	39553811	rs6503629	KRT31	G	T
17	39553833	rs79496913	KRT31	G	A
17	39677699	rs56389952	KRT15	T	C
17	39678160	rs4796672	KRT15	C	T
17	40688016	-	NAGLU	T	C
17	42035360	rs62080323	PYY	G	C
17	44039820	rs63750529	MAPT	G	A
17	45897088	-	OSBPL7	G	C
17	45906757	rs62076131	MRPL10	T	C
17	45906869	rs62076132	MRPL10	C	T
17	46894377	rs3959442	TTLL6	A	C
17	47022270	rs2270574	SNF8	C	T
17	47022413	rs4793999	SNF8	A	G
17	47286380	rs55680377	GNGT2	G	A
17	47395288	rs8082083	ZNF652	C	T
17	47783602	rs11552685	SLC35B1	G	A
17	47785064	rs9908959	SLC35B1	C	A
2	237994045	rs61433497	COPS8	T	G
2	237994232	-	COPS8	C	T
2	238322885	rs7599762	COL6A3	G	C
2	238395813	rs17524101	MLPH	C	T
2	238499318	rs2292875	RAB17	A	G
2	238499319	rs57423896	RAB17	T	C
2	238499528	rs78523256	RAB17	T	G
2	238643907	rs3769078	LRRFIP1	G	A
2	238707464	rs1529805	RBM44	C	T
2	238707957	rs62196080	RBM44	C	T
2	238767662	rs3754699	RAMP1	C	G
2	239008923	rs148770428	ESPNL	C	T
2	239335473	rs474478	ASB1	C	T
2	239344663	rs11904390	ASB1	T	A
2	241075809	rs13406410	MYEOV2	C	T
2	241083967	rs6730518	OTOS	C	A

4.3 Genes and loci associated to prostate cancer

The database search using three databases revealed 703 genes and loci associated to prostate cancer based on experimental findings. Twenty of these genes were located in the sequenced regions. Nineteen of these genes were located in chromosome 17 and one in chromosome 2. Predicted pathogenic variants were found in four of these genes namely ACACA, BRCA1, JUP and CDC6. All PRCA associated genes located in 17q21-22 and 2q37.3 are listed in Table 12.

Table 12. Genes associated to prostate cancer located within the sequenced regions.

Chromosome	Gene Name	Description
17	ACACA	acetyl-CoA carboxylase alpha
17	BRCA1	breast cancer 1, early onset
17	CCL2	chemokine (C-C motif) ligand 2
17	CCL5	chemokine (C-C motif) ligand 5
17	CDC6	cell division cycle 6 homolog (S. cerevisiae)
17	CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)
17	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2
17	ETV4	ets variant 4
17	HNF1B	HNF1 homeobox B
17	HOXB13	homeobox B13
17	HSD17B1	hydroxysteroid (17-beta) dehydrogenase 1
17	IGFBP4	insulin-like growth factor binding protein 4
17	JUP	junction plakoglobin
17	PHB	prohibitin
17	RARA	retinoic acid receptor, alpha
17	RPL19	ribosomal protein L19
17	STAT3	signal transducer and activator of transcription 3
17	STAT5A	signal transducer and activator of transcription 5A
17	STAT5B	signal transducer and activator of transcription 5B
2	TRPM8	transient receptor potential cation channel, subfamily M, member 8

4.4 Gene ontology enrichment analysis for PRCA set

From the original prostate cancer gene containing 703 genes, 660 could be mapped based on the HGNC to symbols to WebGestalt-2. These 660 genes were used in the enrichment analysis. The resulting enriched gene ontology terms, divided into three domains, are are shown in Tables 13, 14 and 15 respectively. In all of the three domains 40 enriched GO-terms were found.

Table 13. Enriched GO-Slim terms belonging to the domain of biological process.

GO-slim term id	GO-slim term	Enriched terms having the ancestoral GO-slim term
GO:0016265	death	7
GO:0008219	cell death	6
GO:0008283	cell proliferation	3
GO:0007275	development	3
GO:0030154	cell differentiation	2
GO:0008152	metabolism	2
GO:0007154	cell communication	2
GO:0009719	response to endogenous stimulus	2
GO:0009653	morphogenesis	1
GO:0009605	response to external stimulus	1
GO:0007165	signal transduction	1
GO:0006950	response to stress	1

The enriched terms belonging to the biological process can be divided to 12 categories based on the GO-slim classification. GO-slim terms can be described as the ancestor nodes in the three structure of the Gene Ontology. GO-slim terms were retrieved using CateGOrizer software (Z-L HU et al. 2008). The enriched GO-slim terms belonging to the biological process domain are shown in Table 13 and the complete list of enriched Gene Ontology terms is shown in Table 21.

Based on the GO-slim classification, the terms belonging to the domain of molecular function can be also divided to 13 categories. The enriched GO-slim terms are shown in Table 14 and the complete list of enriched Gene Ontology terms is shown in Table 22.

Table 14. Enriched GO-Slim terms belonging to the domain of molecular function.

GO-slim term id	GO-slim term	Number of child nodes to GO-slim term
GO:0005488	binding	22
GO:0005515	protein binding	19
GO:0003824	catalytic activity	9
GO:0016740	transferase activity	9
GO:0016301	kinase activity	7
GO:0005102	receptor binding	7
GO:0004672	protein kinase activity	6
GO:0004871	signal transducer activity	6
GO:0004872	receptor activity	3
GO:0003677	DNA binding	1
GO:0030234	enzyme regulator activity	1
GO:0003676	nucleic acid binding	1

The enriched terms belonging to the domain of cellular component can be divided to 11 categories based on the GO-slim classification**.** The enriched GO-slim terms are

shown in Table 15 and the complete list of enriched Gene Ontology terms is shown in Table 23.

Table 15. Enriched GO-Slim terms belonging to the domain of cellular component

GO-slim term id	GO-slim term	Number of child nodes to GO-slim term
GO:0005623	cell	30
GO:0005622	intracellular	14
GO:0005737	cytoplasm	9
GO:0016023	cytoplasmic membrane-bound vesicle	5
GO:0005576	extracellular region	5
GO:0005886	plasma membrane	5
GO:0005634	nucleus	4
GO:0005654	nucleoplasm	1
GO:0005615	extracellular space	1
GO:0005578	extracellular matrix (sensu Metazoa)	1
GO:0005829	cytosol	1

4.5 GO terms associated to candidate genes

The Gene Ontology terms found for candidate genes were compared to enriched gene ontology terms in the prostate cancer gene set. The GO terms that were enriched in the prostate cancer gene set and also associated to genes in the candidate gene set are listed in Tables 16, 17 and 18.

Table 16. Result for GO-term (biological process) comparison between candidate gene set and PRCA gene set.

Gene ontology term	Gene Ontology ID	Associated genes
anatomical structure morphogenesis	GO:0009653	KRT35, LHX1
apoptosis	GO:0006915	AATF, BRCA1, GSDMA, MAPT
cell differentiation	GO:0030154	GGNBP2, LHX1, UNC45B
cell proliferation	GO:0008283	CDC27, PYY
multicellular organismal development	GO:0007275	ASB1, GGNBP2, HOXB3, TBX21, UNC45B
positive regulation of cellular metabolic process	GO:0031325	ACACA, MLX
regulation of apoptosis	GO:0042981	BRCA1
regulation of cell proliferation	GO:0008284	BRCA1, PLCD3
response to organic substance	GO:0010033	BRCA1
response to stimulus	GO:0050896	OR6B2, OR6B3
response to stress	GO:0006950	CASC3
signal transduction	GO:0007165	CCL23, CCL8, PDK2, PLCD3, RAB17

Table 17. Result for GO-term (molecular function) comparison between candidate gene set and PRCA gene set.

Gene ontology term	Gene Ontology ID	Associated genes
protein binding	GO:0005515	JUP, C17orf37, CASC3, CCL8, CDC27, HAP1, HOXB3, IKZF3, LHX1, LRRFIP1, MLPH, MYO19, ORMDL3, OSBPL7, PLCD3, RAB17, SMARCE1, TOP2A, TRAF3IP1, UBE2F, WFIKKN2, ZNF652, CDC6, COL6A3, COPS8, CXCR7, ESPNL, GJC1, IQCA1, KLHL30, KRT13, KRT15, MAPT
transcription regulator activity	GO:0004871	HOXB3
binding	GO:0005488	C17orf37, JUP, UNC45B, GSDMB
enzyme binding	GO:0019899	CASC3, MAPT, TOP2A
hormone activity	GO:0005179	PYY
identical protein binding	GO:0042802	C17orf37, CASC3, MAPT
kinase binding	GO:0016301	CDC6
protein complex binding	GO:0032403	PDK2
protein heterodimerization activity	GO:0046982	IKZF3, PDK2, TOP2A
protein homodimerization activity	GO:0042803	JUP, PDK2, TOP2A
protein kinase activity	GO:0004672	PDK2
protein kinase binding	GO:0019901	JUP
sequence-specific DNA binding	GO:0043565	LHX1, TOP2A
signal transducer activity	GO:0004871	PLCD3, GNGT2
transcription activator activity	GO:0003713	LHX1
transcription factor binding	GO:0008134	SNF8

Table 18. Result for GO-term (cellular component) comparison between candidate gene set and PRCA gene set.

Gene ontology term	Gene Ontology ID	Associated genes
axon	GO:0030424	MAPT
cell projection	GO:0042995	MAPT
cytoplasm	GO:0005737	AATF, ABI3, ACACA, AGAP1, BRCA1, C17orf37, CASC3, CCT6B, CDC27, CDC6, COPS8, GSDMA, GSDMB, HAP1,JUP, KRT24, LRRFIP1, MAPT, MLPH, MLX, MYO19, NT5C3L, PDK2, PLCD3, SCLY, SNF8, STARD3,TOP2A, TRAF3IP1, TTLL6, WIPF2
cytoplasmic membrane-bounded vesicle	GO:0016023	GGNBP2,HAP1
cytosol	GO:0005829	AATF, ACACA, AP2B1, C17orf37, CDC27, CDC6, JUP, MAPT, RPL27, SCLY, SNF8
extracellular matrix	GO:0031012	COL6A3, MMP28, CCL23, CCL8, GNGT2, LYZL6, MMP28, OTOS, PYY, WFIKKN2, RAMP1
integral to plasma membrane	GO:0005887	C17orf37, RAMP1
membrane fraction	GO:0005624	PLCD3
microsome	GO:0005792	B4GALNT2, SLC35B1
neuron projection	GO:0043005	HAP1
nucleoplasm	GO:0005654	BRCA1, CDC27, TOP2A
nucleus	GO:0005634	AATF, BRCA1, CASC3, COPS8, DDX52, HAP1, HOXB3, IKZF3, JUP, LHX1, LRRFIP1, MAPT, MBTD1, MLX, MLS1, PDK2, SMARCE1, SNF8, TBX21, TOP2A, ZNF652
plasma membrane	GO:0005886	ADAM11, AP2B1, BRCA1, C17orf37, CXCR, GJC1, GNGT2, JUP, MAPT, OR6B2, OR6B3, RAB17, RAMP1
proteinaceous extracellular matrix	GO:0005578	MMP28
soluble fraction	GO:0005625	ACACA, PLCD3, PYY, SMARCE1

4.6 Pathway enrichment results for PRCA set and the pathways associated to candidate genes

The enriched pathways in in KEGG, Wikipathways and Pathway Commons in the PRCA set are listed in Tables 24, 25 and 26, respectively. The pathways found for candidate genes were compared to enriched pathways in the prostate cancer gene set. The pathways which are associated to candidate genes and also found enriched in prostate gene set are listed in Table 19.

Table 19. Pathways associated to candidate genes which are also enriched in the PRCA gene set.

Pathway	Adjusted P-value	Database	Genes
Pathways in cancer	$8{,}05\times10^{-126}$	KEGG	JUP,CDC6
Glypican pathway	$1{,}98\times10^{-79}$	Pathway Commons	JUP,TBX21
Glypican 1 network	$6{,}21\times10^{-75}$	Pathway Commons	JUP, TBX21
IFN-gamma pathway	$2{,}13\times10^{-65}$	Pathway Commons	JUP, TBX21
TRAIL signaling pathway	$7{,}11\times10^{-61}$	Pathway Commons	TBX21
Regulation of cytoplasmic and nuclear SMAD2/3 signaling	$2{,}61\times10^{-59}$	Pathway Commons	TBX21
Regulation of nuclear SMAD2/3 signaling	$2{,}61\times10^{-59}$	Pathway Commons	TBX21
TGF-beta receptor signaling	$2{,}61\times10^{-59}$	Pathway Commons	TBX21
Plasma membrane estrogen receptor signaling	$1{,}26\times10^{-51}$	Pathway Commons	JUP, TBX21
Class I PI3K signaling events	$4{,}88\times10^{-47}$	Pathway Commons	JUP, TBX21
Sphingosine 1-phosphate (S1P) pathway	$6{,}75\times10^{-46}$	Pathway Commons	JUP, TBX21
IL2-mediated signaling events	$6{,}82\times10^{-46}$	Pathway Commons	TBX21
Endothelins	$3{,}01\times10^{-44}$	Pathway Commons	JUP, TBX21
BMP receptor signaling	$3{,}58\times10^{-42}$	Pathway Commons	TBX21
TGFBR	$3{,}75\times10^{-39}$	Pathway Commons	AP2B1, CDC27
LPA receptor mediated events	$7{,}88\times10^{-39}$	Pathway Commons	MAPT
Cytokine-cytokine receptor interaction	$1{,}05\times10^{-38}$	KEGG	CXCR7, CCL8, CCL23
Chemokine signaling pathway	$2{,}31\times10^{-38}$	KEGG	GNGT2
Acute myeloid leukemia	$4{,}26\times10^{-38}$	KEGG	JUP
p38 MAPK signaling pathway	$1{,}46\times10^{-33}$	Pathway Commons	TBX21
Signalling by NGF	$7{,}49\times10^{-33}$	Pathway Commons	AP2B1, AATF
Cell cycle	$3{,}96\times10^{-32}$	KEGG	CDC6
Regulation of p38-alpha and p38-beta	$1{,}3\times10^{-30}$	Pathway Commons	TBX21
Cell cycle	$2{,}86\times10^{-30}$	Wikipathways	CDC6
Syndecan-1-mediated signaling events	$1{,}34\times10^{-25}$	Pathway Commons	TBX21
Role of Calcineurin-dependent NFAT signaling in lymphocytes	1,94×10-25	Pathway Commons	TBX21
Apoptosis	$3{,}45\times10^{-24}$	Wikipathways	MAPT
IL12-mediated signaling events	$1{,}8\times10^{-23}$	Pathway Commons	TBX21
Apoptosis	$2{,}68\times10^{-23}$	KEGG	MAPT
DNA damage response	$9{,}92\times10^{-23}$	wikipathways	BRCA1
Signaling events mediated by VEGFR1 and VEGFR2	$3{,}65\times10^{-22}$	Pathway Commons	JUP, TBX21
Class I PI3K signaling events mediated by Akt	$8{,}79\times10^{-22}$	Pathway Commons	JUP
TNF alpha/NF-kB	$3{,}38\times10^{-21}$	Pathway Commons	SMARCE1
Signaling by EGFR	$3{,}1\times10^{-20}$	Pathway Commons	AP2B1
Thyroid cancer	$6{,}41\times10^{-20}$	KEGG	CDC6
IGF1 pathway	$2{,}88\times10^{-19}$	Pathway Commons	JUP, BRCA1, TBX21
S1P1 pathway	$2{,}81\times10^{-18}$	Pathway Commons	JUP, TBX21
Signaling by Aurora kinases	$3{,}89\times10^{-18}$	Pathway Commons	BRCA1
Aurora A signaling	$2{,}16\times10^{-16}$	Pathway Commons	BRCA1
Syndecan-4-mediated signaling events	$4{,}49\times10^{-15}$	Pathway Commons	BRCA1
Signaling by GPCR	$1{,}27\times10^{-14}$	Pathway Commons	RAMP1
FOXA1 transcription factor network	$4{,}42\times10^{-14}$	Pathway Commons	BRCA1

Table 19 continued.

Pathway	Adjusted P-value	Database	Genes
FOXA transcription factor networks	$5{,}4\times10^{-14}$	Pathway Commons	BRCA1
mTOR signaling pathway	$5{,}42\times10^{-14}$	KEGG	JUP, BRCA1,TBX21
PDGFR-beta signaling pathway	$1{,}28\times10^{-13}$	Pathway Commons	JUP, BRCA1,TBX21
Metabolic pathways	$5{,}62\times10^{-13}$	KEGG	PIGW, NAGLU
Glypican 3 network	$8{,}31\times10^{-13}$	Pathway Commons	BRCA1
Apoptosis	$1{,}15\times10^{-12}$	Pathway commons	MAPT
Axon guidance	$1{,}88\times10^{-12}$	KEGG	AP2B1
IL27-mediated signaling events	$9{,}35\times10^{-12}$	Pathway Commons	TBX21
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	$1{,}32\times10^{-11}$	KEGG	JUP
Vascular smooth muscle contraction	$3{,}45\times10^{-11}$	KEGG	RAMP1
Insulin Pathway	$4{,}86\times10^{-11}$	Pathway Commons	JUP, BRCA1, TBX21
Canonical Wnt signaling pathway	$5{,}86\times10^{-11}$	Pathway Commons	BRCA1
Cell Cycle, Mitotic	$6{,}1\times10^{-11}$	Pathway Commons	CDC27,TOP2A
Reelin signaling pathway	$3{,}07\times10^{-10}$	Pathway Commons	MAPT
Cell Cycle Checkpoints	$5{,}4\times10^{-10}$	Pathway Commons	CDC6,CDC27
G2/M Checkpoints	$7{,}88\times10^{-10}$	Pathway Commons	CDC6
p75 NTR receptor-mediated signalling	$2{,}97\times10^{-09}$	Pathway Commons	AATF
Calcineurin-regulated NFAT-dependent transcription in	$3{,}03\times10^{-09}$	Pathway Commons	TBX21
Endocytosis	$5{,}79\times10^{-09}$	KEGG	AGAP1,AP2B1,SNF8
Axon guidance	$2{,}23\times10^{-08}$	Pathway Commons	AP2B1
Diabetes pathways	$4{,}02\times10^{-08}$	Pathway Commons	RPL27
Alzheimer's disease	$5{,}96\times10^{-08}$	KEGG	MAPT
E2F mediated regulation of DNA replication	$7{,}49\times10^{-08}$	Pathway Commons	CDC6
G1/S Transition	$1{,}31\times10^{-07}$	Pathway Commons	CDC6
DNA Repair	$1{,}34\times10^{-07}$	Pathway Commons	BRCA1
G2/M DNA damage checkpoint	$1{,}66\times10^{-07}$	Pathway Commons	CDC6
ATM mediated response to DNA double-strand break	$2{,}14\times10^{-07}$	Pathway Commons	BRCA1
Homologous Recombination Repair	$2{,}14\times10^{-07}$	Pathway Commons	BRCA1
Homologous recombination repair of replication-independent	$2{,}14\times10^{-07}$	Pathway Commons	BRCA1
IL12 signaling mediated by STAT4	$2{,}14\times10^{-07}$	Pathway Commons	TBX21
Double-Strand Break Repair	$7{,}72\times10^{-07}$	Pathway Commons	BRCA1
Huntington's disease	$6{,}44\times10^{-06}$	KEGG	HAP1,AP2B2
Ubiquitin mediated proteolysis	$1{,}05\times10^{-05}$	KEGG	UBEF2
Host Interactions of HIV factors	$2{,}7\times10^{-05}$	Pathway Commons	AP2B1
Apoptotic cleavage of cellular proteins	$3{,}76\times10^{-05}$	Pathway Commons	MAPT
Apoptotic execution phase	0,0002	Pathway Commons	MAPT
NRAGE signals death through JNK	0,0002	Pathway Commons	AATF
ATM mediated phosphorylation of repair proteins	0,0004	Pathway Commons	BRCA1
DNA Replication Pre-Initiation	0,0004	Pathway Commons	CDC6
Fatty acid biosynthesis	0,0004	wikipathways	ACACA
M/G1 Transition	0,0004	Pathway Commons	CDC6
Recruitment of repair and signaling proteins to double-strand	0,0004	Pathway Commons	BRCA1
S Phase	0,0004	Pathway Commons	CDC6
Lysosome	0,0005	KEGG	NAGLU
Activation of ATR in response to replication stress	0,0008	Pathway Commons	CDC6
Activation of the pre-replicative complex	0,0008	Pathway Commons	CDC6
mTOR signaling pathway	0,0013	Pathway Commons	JUP, BRCA1,TBX21
Assembly of the pre-replicative complex	0,0015	Pathway Commons	CDC6
Regulation of DNA replication	0,0021	Pathway Commons	CDC6
Cell death signalling via NRAGE, NRIF and NADE	0,0027	Pathway Commons	AATF
EGFR downregulation	0,0032	Pathway Commons	AP2B1
Fatty acid biosynthesis	0,0045	KEGG	ACACA
Synthesis of DNA	0,0063	Pathway Commons	CDC6
The role of Nef in HIV-1 replication and disease pathogenesis	0,0074	Pathway Commons	AP2B1

5. Discussion

5.1 Assessment of methods used in this study

The aim of this study was to find variants potentially predisposing to prostate cancer. To accomplish this aim first family based information was used to capture the most interesting variants. At this stage only those variants which were found in every affected family member were selected for further assessment. This approach significantly reduces the amount of variants to be analyzed in the next steps which suggests that this is an effective strategy for priorization of variants. However, those variants filtered out in this step should still be analyzed since in many cases the disease predisposing variants do not segregate perfectly.

The use of tolerance predictors seems an efficient method based on comparison of the total number of variants to the number of variants predicted to be pathogenic. However, based on the observation of the results of different tolerance predictors it is clear that the classification of variants to pathogenic and non-pathogenic using tolerance predictors is not a straight forward task. Since the output of SIFT, SNAP and PolyPhen-2 is evaluated by PON-Ps tolerance predictors as features it is not meaningful to compare the results given by these programs to PON-P. Therefore, greatest emphasis should be placed on the results of Mutation Taster and PON-P when considering which of the variants are pathogenic.

Based on the pathogenicity prediction done for the variants in this study Mutation Taster is far more sensitive than PON-P. This is consistent with the results of a recent benchmark study (H. Ali et al. 2012). Mutation Taster predicted 6 variants in chromosome 2 and 22 in chromosome 17 to be pathogenic whereas PON-P predicted 2 pathogenic variants in chromosome 2 and 3 pathogenic variants in chromosome 17. In addition, it is notable that Mutation Taster does not agree on any predicted pathogenic variant predicted by PON-P and vice versa.

The reason for the differences in the prediction results is likely due to a very different selection of features utilized by these two predictors. Mutation Taster lacks many features that are commonly used to predict the effects of missense varian including physico-chemical properties of amino acids being changed, the sequence environment of the variant position and structural features present at the position of the variant.

PON-P, on the other hand, has all of these features included. Therefore, it is likely that PON-P performs better when predicting the effects of missense variants. On the other hand, Mutation Taster also considers other types of effects that variant might have on the transcriptional level neglected by most tolerance predictors. This leads to a conclusion that these two predictors complement each other in a way which makes the discovery of pathogenic variants more effective.

5.2 Elucidation of potentially PRCA predisposing variants

Although the tolerance predictors are effective in reducing the number of potentially pathogenic variants, the number of candidate variants is still too large for genotyping. Therefore, additional criteria are needed to reduce the list of number of candidate variants. Allele frequency is a common criterion used to evaluate the pathogenicity of variants. Rare variants having allele frequencies lower than 0.05 are generally considered more interesting than common variants. Rare variants are more likely to be associated to diseases because of the evolutionary drive towards purifying selection. In other words, evolutionary conserved genomic positions have significant role in functionally important regions of the genome. When these positions are subjected to mutations they likely have effects on the phenotype. In addition to rare variants also novel variants are considered interesting because it is likely that they have a low allele frequency.

In this study 6 rare and 5 novel variants were found in chromosome 2. In chromosome 17 there were 14 rare and 15 novel variants. Limiting the variants to rare and novel would lead to reasonable amount of variants for genotyping. However, the sole use of allele frequency is arguable because the data is in many cases inadequate in terms of population sample size and the selection of populations available. In addition, it should be noted that common variants are also known to predispose to diseases. The current hypothesis is that rare variants would have more significant in the pathogenesis of common diseases being the drivers whereas common variants would have a modifying effect (E.T. Cirulli and D.B.Goldstein, 2010).

In this study the prior knowledge of the genes gathered from cancer related databases. In the candidate set four genes *CDC6, JUP, BRCA1* and *ACACA* were found to be associated to prostate cancer based on previous studies.

The variants located in these four genes are naturally interesting and therefore should be included in the set of variants to be genotyped.

CDC6 (cell division cycle 6 homolog) encodes a protein cdc6 which as a part of pre-replicative complex initiates the DNA replication by incorporation of MCM (mini chromosome maintenance) proteins into DNA (A. Bueno and P. Russell, 1992). Previous research has associated the overexpression of cdc6 to several types of cancer including cervical, brain and non-small cell lung cancer (G.H.Williams et al. 1998; S. Ohta et al. 2001; L. Bonds et al. 2002; P. Karakaidos et al. 2004; N. Murphy et al. 2005). Furthermore, in a study conducted by S. Gonzales et al. 2006, cdc6 was found to be a repressor of locus INK4/ARF which encodes three tumour supressors: $p15^{INK4b}$, $p16^{INK4b}$ and ARF (S. Gonzalez 2006). This finding may suggest that in addition of being a biomarker for cancer cdc6 might be oncogenic.

Surprisingly unlike in other cancer types studied previously, cdc6 was shown to be down-regulated in the aggressive form of prostate cancer (L.D. Robles et al. 2002). According to L.D. Robles et al. 2002 the reason for this discrepancy might arise from the dual role of cdc6. Previous studies have shown that cdc6 not only promotes cell proliferation but also prevents the occurrence of multiple replication events during one cell cycle. Furthermore the down regulation of cdc6 has been shown to cause genomic aberrations in the daughter cells during the cell cycle (C.V. Bruschi et al. 1995, R.S. Williams et al. 1997)

Although cdc6 may have different roles in the pathogenesis of different cancer types it is clear that the deregulation of cdc6 resulting in the loss of cell cycle control seems to cause malignant growth of cells. One predicted pathogenic variant was found in *CDC6*. This variant (rs4135012) is especially interesting because of its very low allele frequency 0.0078 reported in dbSNP.

JUP (Junction plakoglobin) encodes a cytoplasmic protein γ-catenin involved in the formation of two types of submembranous plaques: desmosomes and intermediate junctions (M. Mathur et al. 1994; K.A. Knudsen and M.J. Wheelock, 1992; H. Aberle et al.1995). γ-catenin also acts in cell signaling as a mediator of Wnt signal transduction. The Wnt-pathway is involved in cell differentiation related to embryonic development and tumorigenesis (R.H. Giles et al. 2003). In a study conducted by H. Shiina et al. 2005 the role of γ-catenin in the initiation, progression and metastasis of

prostate cancer was studied. As a result, *JUP* was found to be significantly down-regulated in prostate cancer cell lines due to epigenetic regulation and LOH (Loss of heterozygosity) events. Furthermore, mutations observed in HRPCs (hormone refractory prostate cancer) were associated to γ-catenin accumulation in the nucleus where it can activate the expression of Bcl-2 which in turn promotes cell growth by inhibiting apoptosis (S. Hakimelahi et al. 2000).

These findings suggest that γ-catenin has multiple roles in the progression of prostate cancer. In this study two predicted pathogenic missense variants rs4796604 and rs41283425 were found in *JUP*. Of these two SNPs, rs41283425 seems more prominent having very low allele frequency of 0.0027. rs4796604 is a common variant having allele frequency of 0.52 which suggests that this variant is less likely to be associated to prostate cancer susceptibility.

BRCA1 (Breast cancer 1 early onset) encodes a protein which has an essential role as a tumour suppressor by maintaining genomic stability. *BRCA1* responds to DNA-damage conducting multiple cellular events including ubiquitinylation of proteins, DNA damage repair and induction of cell cycle arrest through activation of the expression of p21 which is a key player in cell cycle regulation (R.A. Venkitaraman, 2002). Numerous studies have shown that mutations impairing *BRCA1* predispose to sporadic and familial breast, ovarian and pancreatic cancer (Y. Miki et al.1994, L.H. Castilla et al.1994, A.A. Langston et al. 1996, D.F. Easton et al 2007, W. Al-Sukhni et al. 2008). The association of germline mutations in *BRCA1* and prostate cancer susceptibility has been investigated in several studies. Some of the studies have showed an association of mutations in *BRCA1* to increased risk of prostate cancer (J. P. Struewing et al. 1995; D.Thompson and D.F Easton 2002; J.A.Douglas et al.2007; Cybulski et al.2008)

In this study three potentially pathogenic variants rs1799950, rs16941 and rs33947868 were found in *BRCA1*. The variants rs1799950 and rs16941 are missense variants and rs33947868 is a deletion. According to dbSNP allele frequency data rs1799950 is a rare variant having allele frequency of 0.0283 contrary to rs16941 which a common variant having allele frequency of 0.303. Based on this observation rs1799950 seems a more likely candidate than rs16941. The deletion rs33947868 has no genotype data making it incomparable to rs1799950 and rs16941.

ACACA is a gene which encodis an enzyme ACCA (acetyl-CoA carboxylase).ACCA acts in lipogenesis catalyzing the carboxylation of acetyl-CoA to malonyl-CoA (F. Lopéz-Casillas et al. 1988). In a study where *ACACA* was silenced with RNAi inhibition of growth and apoptosis of prostate cancer cells were observed. ACCA has been also shown to interact with BRCA1. BRCA1 inhibits ACCA by preventing its phosphorylation thus decreasing the fatty acid synthesis (K. Moreau et al. 2006). Based on these findings mutations altering function of ACCA as well BRCA1 might abrupt the tumour suppression characteristic of these proteins leading to predisposition to cancer. Six potentially pathogenic variants were found in *ACACA*. Four of the variants rs1714987, rs58654829, rs77402427 and rs72828246 are common variants according to dbSNP genotype data. One of the SNPs, rs1714987 is a missense mutation while the other mutations are non-coding located in the 5-UTR region of *ACACA*. The remaining two variants rs67231825 and rs150239106 are both deletions having no genotype data available.

In addition to the genes previously known to be associated to prostate cancer also the results of the Gene Ontology and pathway enriched analysis are can be utilized when limiting the number of genes for variant selection. The knowledge gained from these analyses can be used to characterize genes involved in prostate cancer. The genes in the candidate set having the features similar to those in the prostate cancer set are likely to harbour variants that predispose to cancer.

The use of GO-terms in characterization of genes involved in a pathogenesis of a disease such as cancer is problematic. GO-terms are most often too general to be associated to a specific disease. However in the list of enriched GO terms associated to both candidate and prostate cancer genes, four terms stands out: apoptosis, cell proliferation, regulation of cell proliferation and cell differentiation. *AATF, GSDMA, MAPT* and previously mentioned *BRCA1* are associated to apoptosis which has been considered one of the most essential cellular processes affected in cancer.

GSDMA (Gasdermin A) is mainly expressed in epithelium of stomach, esophagus, mammary gland, skin and gastric cells whereas *MAPT* is expressed almost solely in the neuronal cells (N. Saeki 2007; R.L. Neve et al.1986). The fact that these two genes have not been reported to be expressed or are expressed in low quantities in prostate cells makes variants locates in these genes unlikely to be prostate cancer

predisposing. Contrary to *GSDMA* and *MAPT, AATF* is also expressed in the prostate tissue. *AATF* is a transcription factor having two distinct roles. *AATF* promotes cell proliferation by inhibiting apoptosis (G. Page et al. 2000). On the other hand, similarly to *BRCA1, AATF* responds to DNA damage by activating the expression p53 inducing cell cycle arrest (T.D. Halazonetis and J. Bartek, 2006). The role in the inhibition of cellular growth suggests that *AATF* has a role as a tumour suppressor. Therefore, defects in AATF might be predisposing to cancer. In this study one pathogenic variant (rs115760333) was observed in *AATF*. This variant has a low reported allele frequency 0.001 which further suggests that it is truly a pathogenic variant.

In addition to *CDC6* mentioned earlier *CDC27, PLCD3, BRCA1, TOP2A* and *PYY* are also involved in the process of cell proliferation making variants located in these genes interesting. In addition to rs1799950 and rs4135012, located in BRCA1 and CDC6 respectively, mentioned earlier there is one variant in *TOP2A* which is very rare. This variant (rs61732514) has the minor allele frequency of 0.002.

In addition to the known prostate cancer associated genes *CDC6* and *BRCA1, CDC27* seems particularly interesting because of its essential role in TGF-β signaling. The loss of this pathway has been recurrently observed in human tumours (P.M. Siegel and J. Massague 2003; J. Massague 2008; D.C.Clarke and X.Liu 2008). In recent study conducted by L. Zhang et al. 2011 showed that mutations in *CDC27* that prevent its activation through phosphorylation leading to the inhibition of the anaphase promoting complex (APC/cyclosome). The APC complex degrades SnoN which is a co-suppressor of TGF-β signaling responsive genes promoting cell cycle arrest (Y. Wan et al. 2001).

In this study six variants predicted to be pathogenic were found in *CDC27*. These variants include ones that have been observed in all families. This makes the gene especially interesting since it might indicate that predisposition to prostate cancer in these families could be caused by variants in *CDC27* alone. The six variants are known to dbSNP but lacking genotype data making the priorization of these variants a hard task. However, one variant rs62077264 seems interesting since it is a nonsense variant.

5.3 Future perspectives

The analysis of variant data has mainly consentrated on the coding regions of genes and intronic variants thus leaving a great amount of variants located in the regulatory regions out from consideration. These variants should not be ignored since they might have a significant role in cancer predisposition by changing the expression of genes having oncogenic or tumour suppressive properties. Therefore, future studies should aim to reveal those variants affecting the transcription factor binding- or miRNA binding sites or sites subjected to epigenetic regulation.

At the moment the the analysis of RNA-expression profiles of a subset of the families incuded in this study is currently ongoing. Combining the gene expression data from the RNA-profiling with the annotated variant data obtained from this study offers a great opportunity to find associations between variants and gene regulation.

The currently existing workflow can be also further improved. There is more annotation data available in UCSC and other databases than was used in this study and the volume of this data is constantly increasing. In addition, the workflow could be improved by adding more tolerance predictors in the pathogenicity analysis. Good choices for this workflow would be CHASM and CanPredict because they are specifically designed for cancer research.

6. Conclusions

In conclusion the aims of this study were met at least in some extent. With the methods used in this study it was possible to extract a set containing fewer than 150 variants from thousands of variants discovered in the sequencing, by using annotation data and pathogenicity prediction. From this set variants could be even further prioritized using cancer gene databases and enrichment analysis. Bioinformatics methods truly offer efficient tools for priorization of variants. However, bioinformatics methods currently available are still unable to detect all possible effects of variants. Therefore, many variants that might be associated to diseases are be neglected. There is still much work to be done to evaluate efficiently the numerous effects that the variants may have on gene regulation. The increasing variant data obtained from high-throughput sequencing provides a great challenge to the development of bioinformatics methodology. It is likely that the finding of genotype-phenotype association will remain as the bottleneck of the analysis of NGS data for the years to come. However, as the sequencing technology develops it high-throughput sequencing becomes more affordable. This will lead to a rapid discovery of novel variants. The increasing variant data can be used to develop new tools for variant effect prediction and also improve the quality of population data which can be further used to make the priorization of variants more efficient.

7. References

Aberle, H., Bierkamp, C., Torchard, D., Serova, O., Wagner, T., Natt, E., Wirsching, J., Heidkämper, C., Montagna, M. and Lynch, H.T. (1995). The human plakoglobin gene localizes on chromosome 17q21 and is subjected to loss of heterozygosity in breast and ovarian cancers. *PNAS*, 92 (14), pp. 6384-6388.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nature methods,* 7 (4), pp. 248-249.

Ahmed, F.E. (2005). Artificial neural networks for diagnosis and survival prediction in colon cancer. *Molecular cancer,* 4, pp. 29.

Ali H, Olatubosun A, Vihinen M. (2012). Classification of mismatch repair gene missense variants with PON-MMR. *Hum Mutat*. Apr; 33 (4) :642-50.

Al-Sukhni W, Rothenmund H, Borgida Ae, Zogopoulos G, O'shea Am, Pollett A, Gallinger S. (2008). Germline BRCA1 mutations predispose to pancreatic adenocarcinoma. *Hum Genet. Oct*; 124 (3):271-8. Epub 2008 Sep 2.

American Cancer Society: Cancer Facts and Figures. (2012). Atlanta, Ga: American Cancer Society, 2012

Aretz, S., Uhlhaas, S., Sun, Y., Pagenstecher, C., Mangold, E., Caspari, R., Moslein, G., Schulmann, K., Propping, P. and Friedl, W., 2004. Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Human Mutat.*, 24 (5), pp. 370-380.

Ayodeji Olatubosun, Jouni Väliaho, Jani Härkönen, Janita Thusberg and Mauno Vihinen. (2012). PON-P: Integrated Predictor for Pathogenicity of Missense Variants. *Hum Mutat*. Apr 13.

Bairoch, A. and Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.*, 25 (1), pp. 31-36.

Bao, L., Zhou, M. And Cui, Y. (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.*, 33 (suppl 2), pp. W480-W482.

Barber, L.J., Rosa Rosa, J.M., Kozarewa, I., Fenwick, K., Assiotis, I., Mitsopoulos, C., Sims, D., Hakas, J., Zvelebil, M., Lord, C.J. and Ashworth, A. (2011). Comprehensive genomic analysis of a BRCA2 deficient human pancreatic cancer. *PloS one,* 6 (7), pp. e21639.

Betel D, Wilson M, Gabow A, Marks Ds, Sander C. (2008). The microRNA.org resource:targets and expression. *Nucleic Acids Res*., Jan; 36 (Database Issue): D149-53.

Bonds, L., Baker, P., Gup, C. and Shroyer, K.R. (2002). Immunohistochemical localization of cdc6 in squamous and glandular neoplasia of the uterine cervix. *APLM.* 126 (10), pp. 1164-1168.

Breiman, L. (2001). "Random Forests". *Machine Learning* 45 (1): 5–32

Brognard, J., Zhang, Y.W., Puto, L.A. and Hunter, T. (2011). Cancer-associated loss-of-function mutations implicate DAPK3 as a tumor-suppressing kinase. *Cancer Res.*, 71 (8), pp. 3152-3161.

Bruschi C.V, Mcmillan J.N, Coglievina M., Esposito M. S.. (1995). The genomic instability of yeast cdc6-1/cdc6-1 mutants involves chromosome structure and recombination. *Mol. Gen. Genet.* , 249: 1. 8-18 Nov

Brusselmans, K., De Schrijver, E., Verhoeven, G., Swinnen, J.V. (2005). RNA interference-mediated silencing of the acetyl-CoA-carboxylas-alpha gene induces growth inhibition and apoptosis of prostate cancer cells. *Cancer Res*., Aug 1; 65 (15):6719-25.

Bueno, A. and Russell, P. (1992). Dual functions of CDC6: a yeast protein required for DNA replication also inhibits nuclear division. *The EMBO journal*, 11 (6), pp. 2167-2176.

Calabrese R, Capriotti E, Fariselli P, Martelli Pl, Casadio R. (2009).Functional annotations improve the predictive score of human disease-related mutations in proteins., 2009. *Hum Mutat*., Aug 30: 1237-1244

Campbell, P.T., Curtin, K., Ulrich, C.M., Samowitz, W.S., Bigler, J., Velicer, C.M., Caan, B., Potter, J.D. and Slattery, M.L. (2009). Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle Factors. *Gut*, 58 (5), Pp. 661-667.

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474 (7353), pp. 609-615.

Capriotti, E., Calabrese, R., Casadio, R. (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.

Caputo, S., Benboudjema, L., Sinilnikova, O., Rouleau, E., Beroud, C., Lidereau, R. and French Brca Ggc Consortium. (2012). Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. *Nucleic Acids Res.*, 40 (Database issue), pp. D992-1002.

Carter, H., Samayoa, J., Hruban, R.H. and Karchin, R. (2010). Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). Cancer biology & therapy, 10 (6), pp. 582-587.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B. and Karchin, R. (2009). Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. *Cancer Res*., 69 (16), pp. 6660-6667.

Castilla, L.H., F.J. Couch, M.R. Erdos, K.F. Hoskins, K.Calzone, J. Garaber, J. Boyd, M.B. Lubin, M.L. Deshano, L.C. Brody, F.S. Collins, and B.L. Weber. (1994). Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer. *Nature Genet*., 8: 387-391.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D. and Sander, C. (2010). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*., January; 39 (Database issue): D685–D690.

Chan, P.A., Duraisamy, S., Miller, P.J., Newell, J.A., Mcbride, C., Bond, J.P., Raevaara, T., Ollila, S., Nystrom, M., Grimm, A.J., Christodoulou, J., Oetting, W.S. and Greenblatt, M.S. (2007). Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). *Human Mutat.*, 28 (7), pp. 683-693.

Chen, J. and Fang, G. (2001). MAD2B is an inhibitor of the anaphase-promoting complex. *Genes & development*, 15 (14), pp. 1765-1770.

Christensen, L.L., Madsen, B.E., Wikman, F.P., Wiuf, C., Koed, K., Tjonneland, A., Olsen, A., Syvanen, A.C., Andersen, C.L. and Orntoft, T.F. (2008). The association between genetic variants in hMLH1 and hMSH2 and the development of sporadic colorectal cancer in the Danish population. *BMC medical genetics*, 9, pp. 52.

Cirulli E, Goldstein D. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet*., Jun;11 (6):415-25.

Clarke, D.C. and Liu, X. (2008). Decoding the quantitative nature of TGF-β/Smad signaling. *Trends in cell biology*, 18(9), pp. 430-442.

Coates, R.J., Khoury, M.J. and Gwinn, M. (2008). Five genetic variants associated with prostate cancer. *NEJM.*, 358(25), pp. 2738; author reply 2741.

Cropp, C.D., Simpson, C.L., Wahlfors, T., Ha, N., George, A., Jones, M.S., Harper, U., Ponciano-Jackson, D., Green, T.A., Tammela, T.L.J., Bailey-Wilson, J. and Schleutker, J. (2011). Genome-wide linkage scan for prostate cancer susceptibility in Finland: Evidence for a novel locus on 2q37.3 and confirmation of signal on 17q21-q22. *Int. J. Cancer*, 129 (10), pp. 2400-2407.

Cybulski, C., Gorski, B., Gronwald, J., Huzarski, T., Byrski, T., Debniak, T., Jakubowska, A., Wokolorczyk, D., Gliniewicz, B., Sikorski, A., Stawicka, M., Godlewski, D., Kwias, Z., Antczak, A., Krajka, K., Lauer, W., Sosnowski, M., Sikorska-Radek, P., Bar, K., Klijer, R., Romuald, Z., Malkiewicz, B., Borkowski, A., Borkowski, T., Szwiec, M., Posmyk, M., Narod, S.A. and Lubinski, J. (2008). BRCA1 mutations and prostate cancer in Poland. *European journal of cancer prevention*: the official journal of the European Cancer Prevention Organisation (ECP), 17 (1), pp. 62-66.

De Alencar, S.A. and Lopes, J.C. (2010). A comprehensive in silico analysis of the functional and structural impact of SNPs in the IGF1R gene. *J. Biomed. Biotechnol.*, pp. 715139.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., Mclellan, M.D., Fulton, R.S., Fulton, L.L., Abbott, R.M., Hoog, J., Dooling, D.J., Koboldt, D.C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M.C., Mcmichael, J.F., Magrini, V.J., Cook, L., Mcgrath, S.D., Vickery, T.L., Appelbaum, E., Deschryver, K., Davies, S., Guintoli, T., Lin, L., Crowder, R., Tao, Y., Snider, J.E., Smith, S.M., Dukes, A.F., Sanderson, G.E., Pohl, C.S., Delehaunty, K.D., Fronick, C.C., Pape, K.A., Reed, J.S., Robinson, J.S., Hodges, J.S., Schierding, W., Dees, N.D., Shen, D., Locke, D.P., Wiechert, M.E., Eldred, J.M., Peck, J.B., Oberkfell, B.J., Lolofie, J.T., Du, F., Hawkins, A.E., O'laughlin, M.D., Bernard, K.E., Cunningham, M., Elliott, G., Mason, M.D., Thompson, D.M.,Jr, Ivanovich, J.L., Goodfellow, P.J., Perou, C.M., Weinstock, G.M., Aft, R., Watson, M., Ley, T.J., Wilson, R.K. and Mardis, E.R. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464 (7291), pp. 999-1005.

Dixit, A., Yi, L., Gowthaman, R., Torkamani, A., Schork, N.J. and Verkhivker, G.M. (2009). Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PloS one*, 4 (10), pp. e7485.

Doherty, J.A., Rossing, M.A., Cushing-Haugen, K.L., Chen, C., Van Den Berg, D.J., Wu, A.H., Pike, M.C., Ness, R.B., Moysich, K., Chenevix-Trench, G., Beesley, J., Webb, P.M., Chang-Claude, J., Wang-Gohrke, S., Goodman, M.T., Lurie, G., Thompson, P.J., Carney, M.E., Hogdall, E., Kjaer, S.K., Hogdall, C., Goode, E.L., Cunningham, J.M., Fridley, B.L., Vierkant, R.A., Berchuck, A., Moorman, P.G., Schildkraut, J.M., Palmieri, R.T., Cramer, D.W., Terry, K.L., Yang, H.P., Garcia-Closas, M., Chanock, S., Lissowska, J., Song, H., Pharoah, P.D., Shah, M., Perkins, B., Mcguire, V., Whittemore, A.S., Di Cioccio, R.A., Gentry-Maharaj, A., Menon, U., Gayther, S.A., Ramus, S.J., Ziogas, A., Brewster, W., Anton-Culver, H., Australian Ovarian Cancer Study Management Group, Australian Cancer Study (Ovarian Cancer), Pearce, C.L. and Ovarian Cancer Association Consortium (Ocac). (2010). ESR1/SYNE1 polymorphism and invasive epithelial ovarian cancer risk: an Ovarian Cancer Association Consortium study. *Cancer Epidemiol Biomarkers Prev.*, 19 (1), pp. 245-250.

Dong, L.M., Ulrich, C.M., Hsu, L., Duggan, D.J., Benitez, D.S., White, E., Slattery, M.L., Caan, B.J., Potter, J.D. and Peters, U. (2008). Genetic variation in calcium-sensing receptor and risk for colon cancer. *Cancer Epidemiol Biomarkers Prev*, 17 (10), pp. 2755-2765.

Dong, L.M., Ulrich, C.M., Hsu, L., Duggan, D.J., Benitez, D.S., White, E., Slattery, M.L., Farin, F.M., Makar, K.W., Carlson, C.S., Caan, B.J., Potter, J.D. and Peters, U. (2009). Vitamin D related genes, CYP24A1 and CYP27B1, and colon cancer risk. *Cancer Epidemiol Biomarkers Prev*, 18 (9), pp. 2540-2548.

Doss, C.G. and Sethumadhavan, R. (2009). Investigation on the role of nsSNPs in HNPCC genes--a bioinformatics approach. *J. Biomed. Sci.*, 16, pp. 42.

DOUGLAS, J.A., LEVIN, A.M., ZUHLKE, K.A., RAY, A.M., JOHNSON, G.R., LANGE, E.M., WOOD, D.P. and COONEY, K.A. (2007). Common Variation in the BRCA1 Gene and Prostate Cancer Risk. *Cancer Epidemiol Biomarkers Prev.*, 16 (7), pp. 1510-1516.

Dowdell, K.C., Niemela, J.E., Price, S., Davis, J., Hornung, R.L., Oliveira, J.B., Puck, J.M., Jaffe, E.S., Pittaluga, S., Cohen, J.I., Easton Df, Deffenbaugh Am, Pruss D, Frye C, Wenstrup Rj, Allen-Brady K, Tavtigian Sv, Monteiro An, Iversen Es, Couch Fj, Goldgar De. (2007). A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am. J. Hum. Genet.*, Nov; 81 (5): 873-83. Epub 2007 Sep 6.

Enright A.J, John B, Gaul U, Tuschl T, Sander C and Marks DS. (2003). MicroRNA targets in Drosophila. *Genome Biol*. 5 (1):R1. Epub 2003 Dec 12.

FINNISH CANCER REGISTRY. Cancer incidence and mortality in Finland: Cancer Statistics. (2007). Finnish Cancer Registry, Helsinki, Finland.

Fleisher, T.A. and Rao, V.K., 2010. Somatic FAS mutations are common in patients with genetically undefined autoimmune lymphoproliferative syndrome. *Blood*, 115 (25), pp. 5164-5169.

Duncan, D., Prodduturi, N. and Zhang, B. (2010). WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. *BMC Bioinformatics*, 11, pp. P10.

Dyczynska, E., Syta, E., Sun, D. and Zolkiewska, A. (2008). Breast cancer-associated mutations in metalloprotease disintegrin ADAM12 interfere with the intracellular trafficking and processing of the protein. *Int. J. Cancer*, 122 (11), pp. 2634-2640.

Edwalds-Gilbert, G., Veraldi, K.L. and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res*., 25 (13), pp. 2547-2561.

Estep, A.L., Palmer, C., Mccormick, F. and Rauen, K.A. (2007). Mutation analysis of BRAF, MEK1 and MEK2 in 15 ovarian cancer cell lines: implications for therapy. *PloS one,* 2 (12), pp. e1279.

Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J. and Altman, R.B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27 (13), pp. 1741-1748.

Forbes Sa, Bindal N, Bamford S, Cole C, Kok Cy, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague Jw, Campbell Pj, Stratton Mr, Futreal Pa. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.*Nucleic Acids Res*. Jan; 39 (Database issue): D945-50. Epub 2010 Oct 15.

Fletcher, O. and Houlston, R.S. (2010). Architecture of inherited susceptibility to common cancer. *Nat. Rev. Cancer,* 10 (5), pp. 353-361.

Fletcher, O., Johnson, N., Dos Santos Silva, I., Orr, N., Ashworth, A., Nevanlinna, H., Heikkinen, T., Aittomaki, K., Blomqvist, C., Burwinkel, B., Bartram, C.R., Meindl, A., Schmutzler, R.K., Cox, A., Brock, I., Elliott, G., Reed, M.W., Southey, M.C., Smith, L., Spurdle, A.B., Hopper, J.L., Couch, F.J., Olson, J.E., Wang, X., Fredericksen, Z., Schurmann, P., Waltes, R., Bremer, M., Dork, T., Devilee, P., Van Asperen, C.J., Tollenaar, R.A., Seynaeve, C., Hall, P., Czene, K., Humphreys, K., Liu, J., Ahmed, S., Dunning, A.M., Maranian, M., Pharoah, P.D., Chenevix-Trench, G., Kconfab Investigators, Aocs Group, Beesley, J., Bogdanova, N.V., Antonenkova, N.N., Zalutsky, I.V., Anton-Culver, H., Ziogas, A., Brauch, H., Ko, Y.D., Hamann, U., Genica Consortium, Fasching, P.A., Strick, R., Ekici, A.B., Beckmann, M.W., Giles, G.G., Severi, G., Baglietto, L., English, D.R., Milne, R.L., Benitez, J., Arias, J.I., Pita, G., Nordestgaard, B.G., Bojesen, S.E., Flyger, H., Kang, D., Yoo, K.Y., Noh, D.Y., Mannermaa, A., Kataja, V., Kosma, V.M., Garcia-Closas, M., Chanock, S., Lissowska, J., Brinton, L.A., Chang-Claude, J., Wang-Gohrke, S., Broeks, A., Schmidt, M.K., Van Leeuwen, F.E., Van't Veer, L.J., Margolin, S., Lindblom, A., Humphreys, M.K., Morrison, J., Platte, R., Easton, D.F., Peto, J. and Breast Cancer Association Consortium. (2010). Missense variants in ATM in 26,101 breast cancer cases and 29,842 controls. *Cancer Epidemiol Biomarkers Prev*., 19 (9), pp. 2143-2151.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., Mclaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A.,

Zadissa, A. and Searle, S.M. (2011). Ensembl 2012. *Nucleic Acids Res*. Jan; 40 Database issue: D84-D90

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D. and Kent, W.J. (2010). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.,* Jan; 39 (Database issue):D876-82. Epub 2010 Oct 18

George Priya Doss, C., Rajasekaran, R., Arjun, P. and Sethumadhavan, R. (2010). Prioritization of candidate SNPs in colon cancer using bioinformatics tools: An alternative approach for a cancer biologist. *Interdiscip Sci.*, 2 (4), pp. 320-346.

George Priya Doss, C., Rajasekaran, R. and Sethumadhavan, R. (2010). Computational identification and structural analysis of deleterious functional SNPs in MLL gene causing acute leukemia. *Interdiscip Sci.*, 2 (3), pp. 247-255.

George Priya Doss, C. and Sethumadhavan, R. (2009). Computational and structural analysis of deleterious functional SNPs in ARNT oncogene. *Interdiscip Sci*., 1 (3), pp. 220-228.

Gonzalez S, Klatt P, Delgado S, Conde E, Lopez-Rios F, Sanchez-Cespedes M, Mendez J, Antequera F, Serrano M. (2006). Oncogenic activity of Cdc6 through repression of the INK4/ARF locus. *Nature*, Mar 30; 440 (7084): 702-6.

Goto, M., Shinmura, K., Nakabeppu, Y., Tao, H., Yamada, H., Tsuneyoshi, T. and Sugimura, H. (2010). Adenine DNA glycosylase activity of 14 human MutY homolog (MUTYH) variant proteins found in patients with colorectal polyposis and cancer. *Human Mutat*, 31 (11), pp. E1861-74.

Gu, F., Qureshi, A.A., Niu, T., Kraft, P., Guo, Q., Hunter, D.J. and Han, J. (2008). Interleukin and interleukin receptor gene polymorphisms and susceptibility to melanoma. *Melanoma Res*, 18 (5), pp. 330-335.

Hakimelahi S, Parker Hr, Gilchrist Aj, Barry M, Li Z, Bleackley Rc, Pasdar M. (2000). Plakoglobin regulates the expression of the anti-apoptotic protein BCL-2.J *Biol*. *Chem*., Apr 14; 275 (15):10905-11.

Halazonetis, T. D. and Bartek, J. (2006). DNA damage signaling recruits the RNApolymerase II binding protein Che-1 to the p53 promoter. *Mol. Cell*, 24, 809-810.

Hicks, S., Wheeler, D.A., Plon, S.E. and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human Mutat*., 32 (6), pp. 661-668.

Holbrook, J.D., Parker, J.S., Gallagher, K.T., Halsey, W.S., Hughes, A.M., Weigman, V.J., Lebowitz, P.F. and Kumar, R. (2011). Deep sequencing of gastric carcinoma reveals somatic mutations relevant to personalized medicine. *J. Trans. Med*, 9, pp. 119.

Hsu, C-W; Chang, C-C; And Lin, C-J. (2003). A Practical Guide to Support Vector Classification (Technical report). Department of Computer Science and Information Engineering, National Taiwan University.

Hung, M.S., Lin, Y.C., Mao, J.H., Kim, I.J., Xu, Z., Yang, C.T., Jablons, D.M. and You, L. (2010). Functional polymorphism of the CK2alpha intronless gene plays oncogenic roles in lung cancer. *PloS one*, 5 (7), pp. e11418.

Hu J, Ng P.C. (2012). Predicting the effects of frameshifting indels. *Genome Biol*. Feb 9; 13(2): R9.

Hu ZL, Bao J, Reecy JM. CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *Online J. Bioinform*. 2008; 9: 108–112.

Ioan P. (2006). An approach of the Naive Bayes classifier for the document classification. *General Mathematics*, Vol. 14, No. 4, 135–138

Johnson, M.M., Houck, J. and Chen, C. (2005). Screening for Deleterious Nonsynonymous Single-Nucleotide Polymorphisms in Genes Involved in Steroid Hormone Metabolism and Response. *Cancer Epidemiol Biomarkers Prev.*, 14 (5), pp. 1326-1329.

Junker, V.L., Apweiler, R. and Bairoch, A. (1999). Representation of functional information in the SWISS-PROT Data Bank. *Bioinformatics*, 15(12), pp. 1066-1067.

Kanehisa M, Goto S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. Jan 1; 28 (1):27-30.

Kaminker, J.S., Zhang, Y., Watanabe, C. and Zhang, Z. (2007). CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.*, 35 (Web Server issue), pp. W595-8.

Kaminker, J.S., Zhang, Y., Waugh, A., Haverty, P.M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W.F., Bazan, J.F., Seshagiri, S. and Zhang, Z. (2007). Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer research, 67 (2), pp. 465-473.

Karakaidos, P., Taraviras, S., Vassiliou, L.V., Zacharatos, P., Kastrinakis, N.G., Kougiou, D., Kouloukoussa, M., Nishitani, H., Papavassiliou, A.G., Lygerou, Z. and Gorgoulis, V.G. (2004). Overexpression of the replication licensing regulators hCdt1 and hCdc6 characterizes a subset of non-small-cell lung carcinomas: synergistic effect with mutant p53 on tumor growth and chromosomal instability--evidence of E2F-1 transcriptional control over hCdt1.*Am. J. Pathol.*, 165 (4), pp. 1351-1365.

Karchin, R., Monteiro, A.N., Tavtigian, S.V., Carvalho, M.A. and Sali, A. (2007). Functional impact of missense variants in BRCA1 predicted by supervised learning. *PLoS computational biology*, 3 (2), pp. e26.

Karolchik D, Hinrichs As, Furey Ts, Roskin Km, Sugnet Cw, Haussler D, Kent Wj. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*., Jan 1; 32 (Database issue):D493-6.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. (2004). EnsMART: a generic system for fast and flexible access to biological data. *Genome Res*., Jan; 14 (1): 160-9.

Kloss-Brandstatter, A., Schafer, G., Erhart, G., Huttenhofer, A., Coassin, S., Seifarth, C., Summerer, M., Bektic, J., Klocker, H. and Kronenberg, F. (2010). Somatic mutations throughout the entire mitochondrial genome are associated with elevated PSA levels in prostate cancer patients. *Am. J. Human Genet*., 87 (6), pp. 802-812.

Knudsen, K.A. and Wheelock, M.J. (1992). Plakoglobin, or an 83-kD homologue distinct from beta-catenin, interacts with e-cadherin and N-cadherin. *J. Cell Biol*., 118 (3), pp. 671-679.

Kumar, P., Henikoff, S. and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4 (7), pp. 1073-1081.

Langston Aa, Malone Ke, Thompson Jd, Daling Jr, Ostrander Ea. (1996). BRCA1 mutations in a population-based sample of young women with breast cancer. *N. Engl. J. Med.* Jan 18; 334 (3):137-42.

Le Calvez-Kelm, F., Lesueur, F., Damiola, F., Vallee, M., Voegele, C., Babikyan, D., Durand, G., Forey, N., Mckay-Chopin, S., Robinot, N., Nguyen-Dumont, T., Thomas, A., Byrnes, G.B., Breast Cancer Family Registry, Hopper, J.L., Southey, M.C., Andrulis, I.L., John, E.M. and Tavtigian, S.V. (2011). Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study. *Breast Cancer Res*: BCR, 13 (1), pp. R6.

Le Hir, H., Izaurralde, E., Maquat, L.E. and Moore, M.J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *The EMBO journal*, 19 (24), pp. 6860-6869.

Leaderer, D., Hoffman, A.E., Zheng, T., Fu, A., Weidhaas, J., Paranjape, T. and Zhu, Y. (2011). Genetic and epigenetic association studies suggest a role of microRNA biogenesis gene exportin-5 (XPO5) in breast tumorigenesis. *Int. J. Mol Epidemiol. genet.* 2 (1), pp. 9-18.

Lee, W., Zhang, Y., Mukhyala, K., Lazarus, R.A. and Zhang, Z. (2009). Bi-directional SIFT predicts a subset of activating mutations. *PloS one*, 4 (12), pp. e8311.

Liu Ge, Weirauch Mt, Van Tassell Cp, Li Rw, Sonstegard Ts, Matukumalli Lk, Connor Ee, Hanson Rw, Yang J. Identification of conserved regulatory elements in mammalian promoter regions: a case study using the PCK1 promoter. (2008). *Genomics Proteomics Bioinformatics*. 2008 Dec; 6 (3-4): 129-43.

López- Casillas, F., Bai, D.H., Luo, X.C., Kong, I.S., Hermodson, M.A. and Kim, K.H. (1988). Structure of the coding sequence and primary amino acid sequence of acetyl-coenzyme A carboxylase. *PNAS*, 85 (16), pp. 5784-5788.

Lykke-Andersen, J., Shu, M. and Steitz, J.A. (2000). Human Upf Proteins Target an mRNA for Nonsense-Mediated Decay When Bound Downstream of a Termination Codon. *Cell*, 103 (7), pp. 1121-1131.

Maqungo M, Kaur M, Kwofie Sk, Radovanovic A, Schaefer U, Schmeier S, Oppon E, Christoffels A, Bajic Vb. (2011). DDPC: Dragon Database of Genes associated with Prostate Cancer. *Nucleic Acids Res*., 2011 Jan; 39 (Database issue):D980-5. Epub 2010 Sep 29.

Massagué J. (2008). TGF-beta in Cancer. *Cell.* Jul 25; 134 (2):215-30.

Mathe, E., Olivier, M., Kato, S., Ishioka, C., Hainaut, P. and Tavtigian, S.V. (2006). Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res*., 34 (5), pp. 1317-1325.

Mathur, M., Goodwin, L. and Cowin, P. (1994). Interactions of the cytoplasmic domain of the desmosomal cadherin Dsg1 with plakoglobin. *J. Biol Chem.*, 269 (19), pp. 14075-14080.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal Pa, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett Lm, Ding W, Et Al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*. Oct 7; 266 (5182):66-71.

Moreau K, Dizin E, Ray H, Luquain C, Lefai E, Foufelle F, Billaud M, Lenoir Gm, Venezia Nd. (2006). BRCA1 affects lipid synthesis through its interaction with acetyl-CoA carboxylase. *J. Biol Chem*., Feb 10; 281 (6):3172-81. Epub 2005 Dec 2.

Mosse, Y.P., Laudenslager, M., Longo, L., Cole, K.A., Wood, A., Attiyeh, E.F., Laquaglia, M.J., Sennett, R., Lynch, J.E., Perri, P., Laureys, G., Speleman, F., Kim, C., Hou, C., Hakonarson, H., Torkamani, A., Schork, N.J., Brodeur, G.M., Tonini, G.P., Rappaport, E., Devoto, M. and Maris, J.M., (2008. Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature*, 455 (7215), pp. 930-935.

Murphy, N., Ring, M., Heffron, C.C., King, B., Killalea, A.G., Hughes, C., Martin, C.M., Mcguinness, E., Sheils, O. and O'leary, J.J. (2005). p16INK4A, CDC6, and MCM5: predictive biomarkers in cervical preinvasive neoplasia and cervical cancer. *J. Clin. Pathol*, 58 (5), pp. 525-534.

Nakken, S., Alseth, I. and Rognes, T. (2007). Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. *Neuroscience*, 145 (4), pp. 1273-1279.

Nan, H., Niu, T., Hunter, D.J. and Han, J. (2008). Missense polymorphisms in matrix metalloproteinase genes and skin cancer risk. *Cancer Epidemiol Biomarkers Prev*., 17 (12), pp. 3551-3557.

Neklason, D.W., Done, M.W., Sargent, N.R., Schwartz, A.G., Anton-Culver, H., Griffin, C.A., Ahnen, D.J., Schildkraut, J.M., Tomlinson, G.E., Strong, L.C., Miller, A.R., Stopfer, J.E. and Burt, R.W. (2011). Activating mutation in MET oncogene in familial colorectal cancer. *BMC cancer*, 11, pp. 424.

Ng, P.C. and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Res*., 11 (5), pp. 863-874.

Neve Rl, Selkoe Dj, Kurnit Dm, Kosik Ks. (1986). A cDNA for a human microtubule associated protein 2 epitope in the Alzheimer neurofibrillary tangle. *Brain Res*., Nov; 387 (2):193-6.

Ohta, S., Koide, M., Tokuyama, T., Yokota, N., Nishizawa, S. and Namba, H., 2001. Cdc6 expression as a marker of proliferative activity in brain tumors. *Oncology reports*, 8 (5), pp. 1063-1066.

Page, G., Lödige, I., Kögel, D. and Scheidtmann, K.H. (1999). AATF, a novel transcription factor that interacts with Dlk/ZIP kinase and interferes with apoptosis. *FEBS letters*, 462 (1–2), pp. 187-191.

Parikh, H., Wang, Z., Pettigrew, K.A., Jia, J., Daugherty, S., Yeager, M., Jacobs, K.B., Hutchinson, A., Burdett, L., Cullen, M., Qi, L., Boland, J., Collins, I., Albert, T.J., Vatten, L.J., Hveem, K., Njolstad, I., Cancel-Tassin, G., Cussenot, O., Valeri, A., Virtamo, J., Thun, M.J., Feigelson, H.S., Diver, W.R., Chatterjee, N., Thomas, G., Albanes, D., Chanock, S.J., Hunter, D.J., Hoover, R., Hayes, R.B., Berndt, S.I., Sampson, J. and Amundadottir, L. (2011). Fine mapping the KLK3 locus on chromosome 19q13.33 associated with prostate cancer susceptibility and PSA levels. *Human genetics*, 129 (6), pp. 675-685.

Parsons, D.W., Li, M., Zhang, X., Jones, S., Leary, R.J., Lin, J.C., Boca, S.M., Carter, H., Samayoa, J., Bettegowda, C., Gallia, G.L., Jallo, G.I., Binder, Z.A., Nikolsky, Y., Hartigan, J., Smith, D.R., Gerhard, D.S., Fults, D.W., Vandenberg, S., Berger, M.S., Marie, S.K., Shinjo, S.M., Clara, C., Phillips, P.C., Minturn, J.E., Biegel, J.A., Judkins, A.R., Resnick, A.C., Storm, P.B., Curran, T., He, Y., Rasheed, B.A., Friedman, H.S., Keir, S.T., Mclendon, R., Northcott, P.A., Taylor, M.D., Burger, P.C., Riggins, G.J., Karchin, R., Parmigiani, G., Bigner, D.D., Yan, H., Papadopoulos, N., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2011). The genetic landscape of the childhood cancer medulloblastoma. *Science* (New York, N.Y.), 331 (6016), pp. 435-439.

Pico Ar, Kelder T, Van Iersel Mp, Hanspers K, Conklin Br, Evelo C. (2008). WikiPathways: Pathway Editing for the People. *PLoS Biol* 6 (7): doi:10.1371/journal.pbio.0060184

Rajasekaran, R., Sudandiradoss, C., Doss, C.G.P. and Sethumadhavan, R. (2007). Identification and in silico analysis of functional SNPs of the BRCA1 gene. *Genomics*, 90 (4), pp. 447-452.

Reva, B., Antipin, Y. and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*, 39 (17), pp. e118-e118.

Robbins, C.M., Tembe, W.A., Baker, A., Sinari, S., Moses, T.Y., Beckstrom-Sternberg, S., Beckstrom-Sternberg, J., Barrett, M., Long, J., Chinnaiyan, A., Lowey, J., Suh, E., Pearson, J.V., Craig, D.W., Agus, D.B., Pienta, K.J. and Carpten, J.D. (2011). Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. *Genome Res*, 21 (1), pp. 47-55.

Robles, L.D., Frost, A.R., Davila, M., Hutson, A.D., Grizzle, W.E. and Chakrabarti, R. (2002). Down-regulation of Cdc6, a Cell Cycle Regulatory Gene, in Prostate Cancer. J. Biol. Chem., 277 (28), pp. 25431-25438.

Rosenbloom Kr, Dreszer Tr, Pheasant M, Barber Gp, Meyer Lr, Pohl A, Raney Bj, Wang T, Hinrichs As, Zweig As, Fujita Pa, Learned K, Rhead B, Smith Ke, Kuhn Rm, Karolchik D, Haussler D, Kent Wj. (2010). ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, Jan; 38 (Database issue):D620-5. Epub 2009 Nov 17.

Saarinen, S., Aavikko, M., Aittomäki, K., Launonen, V., Lehtonen, R., Franssila, K., Lehtonen, H.J., Kaasinen, E., Broderick, P., Tarkkanen, J., Bain, B.J., Bauduer, F., Ünal, A., Swerdlow, A.J., Cooke, R., Mäkinen, M.J., Houlston, R., Vahteristo, P. and Aaltonen, L.A. (2011). Exome sequencing reveals germline NPAT mutation as a candidate risk factor for Hodgkin lymphoma. *Blood*, 118 (3), pp. 493-498.

Saeki N, Kim Dh, Usui T, Aoyagi K, Tatsuta T, Aoki K, Yanagihara K, Tamura M, Mizushima H, Sakamoto H, Ogawa K, Ohki M, Shiroishi T, Yoshida T, Sasaki H. (2007). GASDERMIN, suppressed frequently in gastric cancer, is a target of LMO1 in TGF-beta-dependent apoptotic signalling. *Oncogene.* Oct 4; 26(45):6488-98. Epub 2007 Apr 30.

Saunders, C.T. and Baker, D. (2002). Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction. *J. Mol. Biol.*, 322 (4), pp. 891-901.

Savas, S., Ahmad, M.F., Shariff, M., Kim, D.Y. and Ozcelik, H. (2005). Candidate nsSNPs that can affect the functions and interactions of cell cycle proteins. *Proteins*, 58 (3), pp. 697-705.

Schwarz, J.M., Rodelsperger, C., Schuelke, M. and Seelow, D., 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*, 7 (8), pp. 575-576.

Schaid Dj. The complex genetic epidemiology of prostate cancer. (2004). *Hum. Mol. Genet.;* 13 (Spec No. 1)*:* R103–21.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. *Nature Met.* 5, 16 - 18

Shiina H, Breault Je, Basset Ww, Enokida H, Urakami S, Li Lc, Okino St, Deguchi M, Kaneuchi M, Terashima M, Yoneda T, Shigeno K, Carroll Pr, Igawa M, Dahiya R. (2005). Functional Loss of the gamma-catenin gene through epigenetic and genetic pathways in human prostate cancer. *Cancer Res*. Mar 15;65 (6):2130-8.

Siegel, P.M. and Massague, J. (2003). Cytostatic and apoptotic actions of TGF-beta in homeostasis and cancer. *Nature Rev.*.Cancer, 3 (11), pp. 807-821.

Simard, J., Dumont, M., Soucy, P.aAnd Labrie, F., 2002. Perspective: Prostate Cancer Susceptibility Genes. *Endocrinology*, 143 (6), pp. 2029-2040.

Steffensen, A.Y., Jonson, L., Ejlertsen, B., Gerdes, A.M., Nielsen, F.C. and Hansen, T.V. (2010). Identification of a Danish breast/ovarian cancer family double heterozygote for BRCA1 and BRCA2 mutations. *Familial cancer*, 9 (3), pp. 283-287.

Stehr, H., Jang, S.H., Duarte, J.M., Wierling, C., Lehrach, H., Lappe, M. And Lange, B.M. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol. Cancer*, 10, pp. 54.

Struewing, J.P., Abeliovich, D., Peretz, T., Avishai, N., Kaback, M.M., Collins, F.S. and Brody, L.C. (1995). The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. *Nature genetics*, 11(2), pp. 198-200.

Sulonen, A.M., Ellonen, P., Almusa, H., Lepisto, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A. and Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.*, 12(9), pp. R94.

Tabaska Je, Zhang Mq. Detection of polyadenylation signals in human DNA sequences. (1999). *Gene*. Apr 29; 231(1-2):77-86.

Tavtigian, S.V., Deffenbaugh, A.M., Yin, L., Judkins, T., Scholl, T., Samollow, P.B., De Silva, D., Zharkikh, A. and Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet*, 43 (4), pp. 295-305.

Tavtigian, S.V., Oefner, P.J., Babikyan, D., Hartmann, A., Healey, S., Le Calvez-Kelm, F., Lesueur, F., Byrnes, G.B., Chuang, S.C., Forey, N., Feuchtinger, C., Gioia, L., Hall, J., Hashibe, M., Herte, B., Mckay-Chopin, S., Thomas, A., Vallee, M.P., Voegele, C., Webb, P.M., Whiteman, D.C., Australian Cancer Study, Breast Cancer Family Registries (Bcfr), Kathleen Cuningham Foundation Consortium For Research Into Familial Aspects Of Breast Cancer (Kconfab), Sangrajrang, S., Hopper, J.L., Southey, M.C., Andrulis, I.L., John, E.M. and Chenevix-Trench, G. (2009). Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am. J. Hum. Genet*., 85 (4), pp. 427-446.

Taylor Jg, 6., Cheuk, A.T., Tsang, P.S., Chung, J.Y., Song, Y.K., Desai, K., Yu, Y., Chen, Q.R., Shah, K., Youngblood, V., Fang, J., Kim, S.Y., Yeung, C., Helman, L.J., Mendoza, A., Ngo, V., Staudt, L.M., Wei, J.S., Khanna, C., Catchpoole, D., Qualman, S.J., Hewitt, S.M., Merlino, G., Chanock, S.J. and Khan, J. (2009). Identification of FGFR4-activating mutations in human rhabdomyosarcomas that promote metastasis in xenotransplanted models. *J. Clin Inv.*, 119 (11), pp. 3395-3407.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*, 13 (9), pp. 2129-2141.

Thompson, D., Easton, D.F. and The Breast Cancer Linkage Consortium (2002). Cancer Incidence in BRCA1 Mutation Carriers. *J. Nat. Cancer Inst*., 94 (18), pp. 1358-1365.

Thusberg, J., Olatubosun, A. and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutat.*, 32 (4), pp. 358-368.

Timmermann, B., Kerick, M., Roehr, C., Fischer, A., Isau, M., Boerno, S.T., Wunderlich, A., Barmeyer, C., Seemann, P., Koenig, J., Lappe, M., Kuss, A.W., Garshasbi, M., Bertram, L., Trappe, K., Werber, M., Herrmann, B.G., Zatloukal, K., Lehrach, H. and Schweiger, M.R., (2010). Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PloS one*, 5 (12), pp. e15661.

Tischkowitz, M.D., Yilmaz, A., Chen, L.Q., Karyadi, D.M., Novak, D., Kirchhoff, T., Hamel, N., Tavtigian, S.V., Kolb, S., Bismar, T.A., Aloyz, R., Nelson, P.S., Hood, L., Narod, S.A., White, K.A., Ostrander, E.A., Isaacs, W.B., Offit, K., Cooney, K.A., Stanford, J.L. and Foulkes, W.D. (2008). Identification and characterization of novel SNPs in CHEK2 in Ashkenazi Jewish men with prostate cancer. *Cancer letters*, 270 (1), pp. 173-180.

Toh, G.T., Kang, P., Lee, S.S., Lee, D.S., Lee, S.Y., Selamat, S., Mohd Taib, N.A., Yoon, S.Y., Yip, C.H. and Teo, S.H. (2008). BRCA1 and BRCA2 germline mutations in Malaysian women with early-onset breast cancer without a family history. *PloS one*, 3 (4), pp. e2024.

Tomasson, M.H., Xiang, Z., Walgren, R., Zhao, Y., Kasai, Y., Miner, T., Ries, R.E., Lubman, O., Fremont, D.H., Mclellan, M.D., Payton, J.E., Westervelt, P., Dipersio, J.F., Link, D.C., Walter, M.J., Graubert, T.A., Watson, M., Baty, J., Heath, S., Shannon, W.D., Nagarajan, R., Bloomfield, C.D., Mardis, E.R., Wilson, R.K. and Ley, T.J. (2008). Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood*, 111 (9), pp. 4797-4808.

Torkamani, A. and Schork, N.J. (2008). Prediction of cancer driver mutations in protein kinases. *Cancer Res.*, 68 (6), pp. 1675-1682.

Venkitaraman A.R., (2002). Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell*. Jan 25; 108 (2):171-82.

Wan Y, Liu X, Kirschner MW. (2001). The anaphase-promoting complex mediates TGF-beta signaling by targeting SnoN for destruction. *Mol. Cell*. Nov; 8 (5):1027-39.

Weber, G.L., Parat, M.O., Binder, Z.A., Gallia, G.L. and Riggins, G.J., 2011. Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells. *Oncotarget*, 2 (11), pp. 833-849.

Williams, G.H., Romanowski, P., Morris, L., Madine, M., Mills, A.D., Stoeber, K., Marr, J., Laskey, R.A. and Coleman, N. (1998). Improved cervical smear assessment using antibodies against proteins that regulate DNA replication. *PNAS*, 95 (25), pp. 14932-14937.

Wong, W.C., Kim, D., Carter, H., Diekhans, M., Ryan, M.C. and Karchin, R. (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, 27 (15), pp. 2147-2148.

Yue, P., Melamud, E. and Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. *BMC bioinformatics*, 7, pp. 166.

Zhang, B., Kirov, S. and Snoddy, J., 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, 33, pp. W741-8.

Zhang, J., Chiodini, R., Badr, A. and Zhang, G. (2011). The impact of next-generation sequencing on genomics. *J. Genet. Genomics*, 38 (3), pp. 95-109.

Zhang, L., Fujita, T., Wu, G., Xiao, X. and Wan, Y., 2011. Phosphorylation of the Anaphase-promoting Complex/Cdc27 Is Involved in TGF-β Signaling. *J. Biol. Chem.*, 286 (12), pp. 10041-10050.

Zhang, Zemin, Hon, Lawrence, S., Kaminker and Joshua, S. (2008). Computational Approaches for Predicting Causal Missense Mutations in Cancer Genome Projects. *Curr. Bioinform.*, 3(1), pp. 46-55.

Zhi, W., Xue, B., Wang, L., Xiao, N., He, Q., Wang, Y. and Fan, Y. (2011). The MLH1 2101C>A (Q701K) variant increases the risk of gastric cancer in Chinese males. *BMC gastroenterology*, 11 (1), pp. 133.

8. Appendices

8.2 WebGestalt2

WEB-based Gene Set analysis Toolkit v2 (WebGestalt2) is a program designed for studying gene sets (B. Zhang 2005, D. Duncan, 2010). WebGestalt provides several implementations of enrichment analysis: including Gene Ontology-term, Pathway, transcription factor binding target, MicroRNA target, protein Network Interaction Module and cytogenic band analysis

The user defined gene list is compared against a reference set which is by default the whole genome. However, the user can select another reference set from a great collection of gene sets. These sets represent sets of genes chosen for different micro-array platforms.

Statistical testing of over- and underrepresented categories are calculated as following. Let A be the category of interest. Let n, be the number of genes belonging to the gene set of interest and m is the number of genes belonging to the reference set. Let k be the number of genes in the gene set of interest that belong to category A and j be the number of genes in the reference set belonging to category A. Knowing the number of genes belonging to A in the reference set the expected value k_e can be calculated as following:

eq. 21 $$k_e = \left(\frac{j}{m}\right) \times n$$

If the $k \geq k_e$ the category is said to be enriched. The ratio of enrichment r is defined as following

eq. 22 $$r = \frac{k}{k_e}$$

To evaluate the significance of the enrichment WebGestalt2 uses two statistical tests: hypergeometric and Fischer´s exact test. Hyper geometric test is used when the two gene sets are dependent. Two sets are dependent if either set is a subset of the other one. The hypergeometric test can be formulated as follows:

eq. 23 $$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

The probability P is the probability of having at least k occurrences of genes belonging to category A in a genes set of randomly selected genes from the reference set. If the two set are independent the hyper geometric test takes an alternative form called Fisher´s exact test. Fisher´s test is described by the following equation:

eq. 24 $$P = \sum_{i=k}^{n} \frac{\binom{n}{i}\binom{m}{j+k-i}}{\binom{m+n}{j+k}}$$

In enrichment analysis several categories are tested at once which leads to multiple testing problem. To adjust the p-value for multiple hypothesis testing five distinct methods to correct the p-value are provided including BH (Benjamini-Hochberg), BY (Benjamini-Yekutieli), Bonferroni, Holm and Hommel.

8.3 Supplementary tables

8.3.1.CHASM feature list

Table 20. Features selected for the CHASM predictor.

Feature	Description
17-Way Exon conservation	The exon conservation score calculated using windows overlapping the exons in a 17-species phylogenetic alignment.
COSMIC substitution frequency	The frequencies of different types of amino acids substitutions calculated from COSMIC database (COSMIC release 38). The frequencies are normalized using the occurrence of wild type amino acid in UniProtKB database
FP30 PTM Enzyme domain	Mutation is located in a PTM Enzyme domain
PAM250 substitution score	PAM250 matrix amino acid substitution score
MJ substitution score	Miyazawa-Jernigan contact energy amino acid score substution score
FP7 DNA binding domain	Mutation is located in a DNA binding domain
VB substitution count	VB (Venkatarajan and Braun) amino acid substitution score
Positional HMM_Cons	Degree of conservation obtained from multiple sequence alignment constructed with SAM-T2K
SNPDensity-all variants	The number of all variants observed in the exon where mutation is located
Relative Entropy of HMM alignment	Shannon entropy which has been calculated at the site of mutation based on the SAM-T2K alignment
Ex substitution score	EX-matrix amino acid substitution score
HGMD2003 mutation count	The number of occurences of wt to mutant substitutions reported in HGMD

Table 20 continued.

Feature	Description
HGMD2003 mutation count	The number of occurences of wt to mutant substitutions reported in HGMD
BLOSUM substitution score	BLOSUM-matricx substitution score
Pdiff_middle	Probability of observing the wild type residue in the midlle position of the codon
Backround probability of WT residue	The frequency of observing the wildtype residue in proteins in the UniProtKB database
Backround probability of mut residue	The frequency of observing the mutant residue in proteins in the UniProtKB database
Pfirstmut	Probability of observing the mutant residue in the first position of the codon
Difference in polarity	The difference in polarity of wild type and mutant amino acids
Predicted solvent access:Intermed	Predicted solvent accessibility is predicted to be intermediate. The prediction is based on a neural network trained with Predict-2nd software
Change in hydrophobicity	The net change in the hydrophobic porperties of wild type and mutant residue
OMA alignment score	The score is a compatibility score calculated from a collection of multiple sequence alignments built using T-Coffee from orthologous protein in OMA database.
Charge change(H neutral)	The net change in the charge of wild type and mutant residue
Predicted backbone flex:Med	The backbone flexibility is predicted to be intermediate at the environment of the mutation. The prediction is based on a neural network trained with Predict-2nd software
COSMICvsHAPMAP	The frequencies of different types of amino acids substitutions calculated from COSMIC database(COSMIC release 38). The frequencies are normalized using the occurrence of the amino acid substitution in HapMap
Volume change	The net change in the volume of wild type and mutant residue
Predicted solvent access(Exposed)	Predicted solvent accessibility is predicted to be exposed. The prediction is based on a neural network trained with Predict-2nd software
Volume difference	Difference in the volume of wild type and mutant residue
Predicted solvent access(buried)	Predicted solvent accessibility is predicted to be buried. The prediction is based on a neural network trained with Predict-2nd software
FP42 RNA Binding	Mutation is located in a RNA-binding domain
FP22_REGION	Mutation is located at a region which is not defined by other subsections of regions(topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, DNA binding, nucleotide binding, coiled coil, motif, compositional bias)
P5reswt	Calculated probability of observing the mutation at the center of 5 amino acid sequence

Table 20 continued.

Feature	Description
FP17 Transmembrane	Mutation is located in a Transmembrane domain
Pfirstwt	Probability of observing the wild type residue in the first position of the codon
Region Composition G	The percentage of amino acid Glysine observed in a sequence composed of 15 amino acids surrouding the mutation site
Pmiddlemut	Probability of observing the mutant residue in the middle position of the codon
Pdiff_first	The difference in the probabilities of observing the wild type residue and the mutant residue in the first position of the codon
Region_composition_P	The percentage of amino acid Proline observed in a sequence composed of 15 amino acids surrouding the mutation site
FP14 Signal Peptide Domain	Mutation is located in a Signal peptide domain
FP8 NTP Binding Domain	Mutation is located in a NTP binding domain
Predicted 2ndary Structure:Helix	Predicted secondary strucuture at the site of mutation is helix. The prediction is based on a neural network trained with Predict-2^{nd}
FP13 Propeptide Domain	Mutation is located in a propeptide domain
Predicted 2ndary Structure:Strand	Predicted secondary structure at the site of mutation is strand. The prediction is based on a neural network trained with Predict-2^{nd}
FP27 Membrane Binding DM	Mutation is located in a membrane binding domain
Difference in hydrophobicity	Difference in the volume of wild type and mutant residue
Predicted backbone flex:Low	The backbone flexibility is predicted to be low at the environment of the mutation. The prediction is based on a neural network trained with Predict-2^{nd} software
Plastwt	Probability of observing the wild type residue in the last position of the codon
Pdiff_last	The difference in the probabilities of observing the wild type residue and the mutant residue in the last position of the codon
FP16 Domain containing variants	Mutation is located in a domain containing variants
Grantham substitution score	The substitution score calculated using Grantham metrics

8.3.2 Gene Ontology enrichment analysis results for PRCA gene set

Table 21. Gene Ontology terms enriched in prostate cancer geneset belonging to the domain of biological process. The threshold for significance is 0.01.

Biological process	Gene Ontology ID	adjusted P-value
cell proliferation	GO:0008283	5.63×10^{-85}
positive regulation of biological process	GO:0048518	1.06×10^{-76}
positive regulation of cellular process	GO:0048522	6.84×10^{-75}
response to chemical stimulus	GO:0042221	2.58×10^{-72}
regulation of cell proliferation	GO:0042127	6.81×10^{-70}
negative regulation of biological process	GO:0048519	1.51×10^{-64}

Table 21 continued.

Biological process	Gene	adjusted P-value
organ development	GO:0048513	2.22×10^{-64}
negative regulation of cellular process	GO:0048523	3.44×10^{-61}
system development	GO:0048731	3.56×10^{-61}
anatomical structure development	GO:0048856	4.11×10^{-58}
multicellular organismal development	GO:0007275	7.52×10^{-57}
regulation of cell death	GO:0010941	1.50×10^{-54}
developmental process	GO:0032502	1.54×10^{-54}
regulation of programmed cell death	GO:0043067	4.80×10^{-54}
regulation of apoptosis	GO:0042981	8.46×10^{-54}
response to external stimulus	GO:0009605	1.73×10^{-51}
programmed cell death	GO:0012501	2.95×10^{-51}
apoptosis	GO:0006915	4.83×10^{-51}
death	GO:0016265	1.85×10^{-50}
response to organic substance	GO:0010033	3.46×10^{-50}
cell death	GO:0008219	5.00×10^{-50}
regulation of multicellular organismal process	GO:0051239	1.06×10^{-47}
response to hormone stimulus	GO:0009725	8.80×10^{-47}
response to endogenous stimulus	GO:0009719	6.62×10^{-46}
regulation of developmental process	GO:0050793	2.40×10^{-45}
multicellular organismal process	GO:0032501	2.38×10^{-43}
regulation of biological quality	GO:0065008	1.48×10^{-42}
response to stimulus	GO:0050896	6.58×10^{-42}
positive regulation of metabolic process	GO:0009893	8.60×10^{-42}
response to stress	GO:0006950	3.56×10^{-41}
positive regulation of cellular metabolic process	GO:0031325	2.63×10^{-40}
anatomical structure morphogenesis	GO:0009653	5.29×10^{-40}
cell differentiation	GO:0030154	2.59×10^{-39}
biological regulation	GO:0065007	7.63×10^{-38}
regulation of cellular process	GO:0050794	1.48×10^{-37}
positive regulation of cell proliferation	GO:0008284	2.44×10^{-37}
signal transduction	GO:0007165	3.34×10^{-37}
cellular developmental process	GO:0048869	3.50×10^{-37}
cell communication	GO:0007154	8.66×10^{-37}
regulation of cell differentiation	GO:0045595	1.28×10^{-36}

Table 22. Gene Ontology terms enriched in prostate cancer geneset belonging to the domain of molecular function. Threshold for significance is 0.01

Molecular function	Gene	Adjusted P-value
protein binding	GO:0005515	1.74×10^{-42}
receptor binding	GO:0005102	1.24×10^{-31}
enzyme binding	GO:0019899	1.82×10^{-19}
receptor signaling protein activity	GO:0005057	1.44×10^{-18}
transcription activator activity	GO:0003713	2.13×10^{-17}
protein kinase activity	GO:0004672	9.79×10^{-17}
protein dimerization activity	GO:0046983	5.44×10^{-14}
kinase binding	GO:0019900	1.20×10^{-13}
sequence-specific DNA binding	GO:0043565	1.20×10^{-13}
molecular transducer activity	GO:0060089	2.65×10^{-13}
phosphotransferase activity, alcohol group as acceptor	GO:0016773	2.65×10^{-13}
transcription factor binding	GO:0008134	2.65×10^{-13}
signal transducer activity	GO:0004871	2.65×10^{-13}
transcription regulator activity	GO:0004871	4.38×10^{-13}
identical protein binding	GO:0042802	5.43×10^{-13}
Binding	GO:0005488	1.05×10^{-11}
kinase activity	GO:0016301	1.40×10^{-11}
growth factor activity	GO:0008083	1.95×10^{-11}
protein kinase binding	GO:0019901	2.17×10^{-11}

Table 22 continued.

Molecular function	Gene	Adjusted P-value
transcription factor activity	GO:0003712	3.44×10^{-11}
transmembrane receptor protein kinase activity	GO:0019199	2.38×10^{-10}
cytokine receptor binding	GO:0005126	2.50×10^{-10}
ligand-dependent nuclear receptor activity	GO:0030374	3.93×10^{-09}
transferase activity, transferring phosphorus-containing groups	GO:0016772	6.83×10^{-09}
protein serine/threonine kinase activity	GO:0004674	7.28×10^{-09}
protein tyrosine kinase activity	GO:0004713	1.30×10^{-08}
steroid hormone receptor activity	GO:0003707	1.36×10^{-08}
growth factor binding	GO:0019838	2.25×10^{-08}
enzyme inhibitor activity	GO:0004857	3.72×10^{-08}
protein complex binding	GO:0032403	4.62×10^{-08}
protein heterodimerization activity	GO:0046982	8.17×10^{-08}
transmembrane receptor protein tyrosine kinase activity	GO:0004714	1.07×10^{-07}
cytokine activity	GO:0005125	1.75×10^{-07}
receptor signaling protein serine/threonine kinase activity	GO:0004702	2.33×10^{-07}
protein homodimerization activity	GO:0042803	2.33×10^{-07}
SMAD binding	GO:0046332	4.16×10^{-07}
peptide binding	GO:0042277	5.00×10^{-07}
hormone activity	GO:0005179	1.08×10^{-06}
hormone receptor binding	GO:0051427	1.98×10^{-06}
transforming growth factor beta receptor binding	GO:0005160	1.98×10^{-06}

Table 23. Gene Ontology terms enriched in prostate cancer geneset belonging to the domain of cellular component. Threshold for significance is 0.01.

Cellular component	Gene	Adjusted P-value
extracellular region part	GO:0044421	1.12×10^{-29}
extracellular space	GO:0005615	2.62×10^{-28}
extracellular region	GO:0005576	2.99×10^{-20}
cytoplasm	GO:0005737	4.81×10^{-15}
cell fraction	GO:0000267	1.66×10^{-13}
cell surface	GO:0009986	3.45×10^{-13}
plasma membrane part	GO:0044459	4.61×10^{-13}
nucleoplasm	GO:0005654	7.89×10^{-13}
cytosol	GO:0005829	2.63×10^{-12}
plasma membrane	GO:0005886	7.48×10^{-12}
membrane-enclosed lumen	GO:0031974	8.73×10^{-11}
intrinsic to plasma membrane	GO:0031226	5.56×10^{-10}
vesicle	GO:0031982	7.41×10^{-10}
organelle lumen	GO:0043233	7.41×10^{-10}
insoluble fraction	GO:0005626	7.59×10^{-10}
integral to plasma membrane	GO:0005887	7.59×10^{-10}
nuclear lumen	GO:0031981	4.54×10^{-09}
platelet alpha granule lumen	GO:0031093	5.22×10^{-09}
membrane raft	GO:0045121	6.11×10^{-09}
membrane-bounded vesicle	GO:0031988	8.10×10^{-09}
extracellular matrix	GO:0031012	9.58×10^{-09}
cytoplasmic membrane-bounded vesicle lumen	GO:0060205	1.20×10^{-08}
nuclear part	GO:0044428	1.20×10^{-08}
cytoplasmic vesicle	GO:0031410	1.22×10^{-08}
platelet alpha granule	GO:0031091	1.29×10^{-08}
vesicle lumen	GO:0031983	2.05×10^{-08}
cytoplasmic part	GO:0044444	3.33×10^{-08}
axon	GO:0030424	5.36×10^{-08}
membrane fraction	GO:0005624	5.55×10^{-08}
neuron projection	GO:0043005	6.04×10^{-08}
vesicular fraction	GO:0042598	7.89×10^{-08}

Table 23 continued.

Cellular component	Gene	Adjusted P-value
cell projection	GO:0042995	1.02×10^{-07}
external side of plasma membrane	GO:0009897	1.02×10^{-07}
microsome	GO:0005792	1.49×10^{-07}
cytoplasmic membrane-bounded vesicle	GO:0016023	2.27×10^{-07}
nucleus	GO:0005634	2.28×10^{-07}
intracellular organelle lumen	GO:0070013	2.28×10^{-07}
soluble fraction	GO:0005625	9.58×10^{-07}
secretory granule	GO:0030141	9.58×10^{-07}
proteinaceous extracellular matrix	GO:0005578	2.68×10^{-06}

8.3.3 Pathway enrichment results for PRCA gene set

Table 24. The complete list of the pathway enrichment analysis(KEGG) results for prostate cancer geneset.

KEGG pathways	Number of genes	Adjusted P-Value
Pathways in cancer	116	8.05×10^{126}
Focal adhesion	55	9.78×10^{52}
Prostate cancer	42	2.05×10^{51}
Pancreatic cancer	37	3.52×10^{47}
Chronic myeloid leukemia	36	1.47×10^{44}
Colorectal cancer	36	2.39×10^{42}
MAPK signaling pathway	51	1.04×10^{39}
Cytokine-cytokine receptor interaction	50	1.05×10^{38}
Chemokine signaling pathway	44	2.31×10^{38}
Acute myeloid leukemia	30	4.26×10^{38}
Small cell lung cancer	33	1.62×10^{37}
ErbB signaling pathway	33	6.11×10^{37}
Bladder cancer	25	1.58×10^{34}
Non-small cell lung cancer	27	1.92×10^{34}
Neurotrophin signaling pathway	35	9.11×10^{34}
Melanoma	29	9.38×10^{34}
Cell cycle	34	3.96×10^{32}
Jak-STAT signaling pathway	36	9.83×10^{32}
Endometrial cancer	24	1.11×10^{29}
Renal cell carcinoma	25	1.78×10^{27}
Glioma	24	7.85×10^{27}
Regulation of actin cytoskeleton	36	2.05×10^{26}
Adherens junction	24	8.39×10^{25}
Toll-like receptor signaling pathway	26	2.02×10^{24}
TGF-beta signaling pathway	24	2.11×10^{23}
Insulin signaling pathway	28	2.68×10^{23}
Apoptosis	24	2.68×10^{23}
T cell receptor signaling pathway	25	2.55×10^{22}
Wnt signaling pathway	27	6.70×10^{21}
Progesterone-mediated oocyte maturation	22	8.71×10^{21}
Adipocytokine signaling pathway	20	1.82×10^{20}
Thyroid cancer	15	6.41×10^{20}
GnRH signaling pathway	22	3.56×10^{19}
p53 signaling pathway	19	8.99×10^{19}
NOD-like receptor signaling pathway	18	2.67×10^{18}
VEGF signaling pathway	19	6.36×10^{18}
Prion diseases	14	1.06×10^{16}
Fc epsilon RI signaling pathway	18	2.91×10^{16}
Epithelial cell signaling in Helicobacter pylori infection	17	3.75×10^{16}
Melanogenesis	19	2.12×10^{15}
B cell receptor signaling pathway	17	2.14×10^{15}

Table 24 continued.

KEGG pathways	Number of genes	Adjusted P-Value
Leukocyte transendothelial migration	19	3.35×10^{14}
Natural killer cell mediated cytotoxicity	20	4.55×10^{14}
mTOR signaling pathway	14	5.42×10^{14}
Neuroactive ligand-receptor interaction	25	3.04×10^{13}
Metabolic pathways	52	5.62×10^{13}
Gap junction	16	7.67×10^{13}
Axon guidance	18	1.88×10^{12}
ECM-receptor interaction	15	3.83×10^{12}
Type II diabetes mellitus	12	7.21×10^{12}
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	14	1.32×10^{11}
Vascular smooth muscle contraction	16	3.45×10^{11}
Basal cell carcinoma	12	5.13×10^{11}
Metabolism of xenobiotics by cytochrome P450	13	6.22×10^{11}
Hedgehog signaling pathway	12	6.22×10^{11}
Long-term potentiation	13	6.22×10^{11}
Fc gamma R-mediated phagocytosis	14	3.62×10^{10}
Calcium signaling pathway	18	3.77×10^{10}
Complement and coagulation cascades	12	7.70×10^{10}
Long-term depression	12	9.02×10^{10}
RIG-I-like receptor signaling pathway	12	1.05×10^{09}
Intestinal immune network for IgA production	10	5.25×10^{09}
Endocytosis	17	5.79×10^{09}
Drug metabolism - other enzymes	10	6.24×10^{09}
Hypertrophic cardiomyopathy (HCM)	12	8.45×10^{09}
Cytosolic DNA-sensing pathway	10	1.58×10^{08}
Dilated cardiomyopathy	12	2.06×10^{08}
Tight junction	14	2.27×10^{08}
Oocyte meiosis	13	2.56×10^{08}
Androgen and estrogen metabolism	9	2.95×10^{08}
Notch signaling pathway	9	4.36×10^{08}
Alzheimer's disease	15	5.96×10^{08}
PPAR signaling pathway	10	1.15×10^{07}
Hematopoietic cell lineage	11	1.20×10^{07}
Drug metabolism - cytochrome P450	10	1.70×10^{07}
Dorso-ventral axis formation	6	2.29×10^{06}
Huntington's disease	13	6.44×10^{06}
Cell adhesion molecules (CAMs)	11	8.02×10^{06}
Ubiquitin mediated proteolysis	11	1.05×10^{05}
Viral myocarditis	8	1.89×10^{05}
Amyotrophic lateral sclerosis (ALS)	7	1.92×10^{05}
Allograft rejection	6	2.83×10^{05}
Pathogenic Escherichia coli infection	7	3.84×10^{05}
Retinol metabolism	7	6.48×10^{05}
Cysteine and methionine metabolism	5	0.0002
Arachidonic acid metabolism	6	0.0003
Lysosome	8	0.0005
Fatty acid metabolism	5	0.0006
C21-Steroid hormone metabolism	3	0.0008
Autoimmune thyroid disease	5	0.0015
Base excision repair	4	0.0024
Tryptophan metabolism	4	0.0039
Fatty acid biosynthesis	2	0.0045
Graft-versus-host disease	4	0.0046
Type I diabetes mellitus	4	0.0054
Caffeine metabolism	2	0.0059
Mismatch repair	3	0.0061
Phosphatidylinositol signaling system	5	0.0070

Table 24 continued.

KEGG pathways	Number of genes	Adjusted P-Value
Glutathione metabolism	4	0.0081
Pentose and glucuronate interconversions	3	0.0092
Synthesis and degradation of ketone bodies	2	0.0094
Cysteine and methionine metabolism	5	0.0002
Arachidonic acid metabolism	6	0.0003
Lysosome	8	0.0005
Fatty acid metabolism	5	0.0006
C21-Steroid hormone metabolism	3	0.0008
Autoimmune thyroid disease	5	0.0015
Base excision repair	4	0.0024
Tryptophan metabolism	4	0.0039
Fatty acid biosynthesis	2	0.0045
Graft-versus-host disease	4	0.0046
Type I diabetes mellitus	4	0.0054
Caffeine metabolism	2	0.0059
Mismatch repair	3	0.0061
Phosphatidylinositol signaling system	5	0.0070
Glutathione metabolism	4	0.0081
Pentose and glucuronate interconversions	3	0.0092
Synthesis and degradation of ketone bodies	2	0.0094

Table 25. The complete list of the pathway enrichment analysis(Wikipathways) results for prostate cancer geneset.

WIKIPATHWAYS	Number of genes	Adjusted P-Value
Androgen Receptor Signaling Pathway	45	4.45×10^{50}
IL-3 Signaling Pathway	42	1.43×10^{48}
EGFR1 Signaling Pathway	50	8.16×10^{48}
IL-6 Signaling Pathway	40	6.64×10^{46}
Focal Adhesion	49	1.13×10^{45}
Adipogenesis	42	8.47×10^{43}
TGF-beta Receptor Signaling Pathway	44	1.35×10^{42}
DNA damage response (only ATM dependent)	37	1.93×10^{42}
MAPK signaling pathway	41	1.60×10^{37}
IL-2 Signaling Pathway	32	2.26×10^{37}
IL-5 Signaling Pathway	30	4.55×10^{36}
B Cell Receptor Signaling Pathway	39	2.35×10^{35}
Senescence and Autophagy	28	1.25×10^{34}
Wnt Signaling Pathway NetPath	34	1.27×10^{34}
IL-4 signaling Pathway	28	6.29×10^{34}
Id Signaling Pathway	25	9.57×10^{32}
Cell cycle	29	2.86×10^{30}
IL-7 Signaling Pathway	23	3.75×10^{30}
TGF Beta Signaling Pathway	24	9.65×10^{30}
Alpha6-Beta4 Integrin Signaling Pathway	26	5.92×10^{29}
Insulin Signaling	33	1.93×10^{27}
ErbB signaling pathway	22	2.77×10^{27}
Delta-Notch Signaling Pathway	26	1.71×10^{26}
TNF-alpha/NF-kB Signaling Pathway	34	3.84×10^{26}
Endochondral Ossification	23	2.29×10^{25}
Toll-like receptor signaling pathway	26	2.82×10^{24}
Apoptosis	24	3.45×10^{24}
Toll-like receptor signaling pathway - mir	28	2.76×10^{23}
DNA damage response	22	9.92×10^{23}
estrogen signalling	22	3.69×10^{22}
Integrin-mediated cell adhesion	24	5.76×10^{22}
Wnt Signaling Pathway and Pluripotency	23	7.40×10^{21}
Regulation of Actin Cytoskeleton	26	1.84×10^{20}
Kit Receptor Signaling Pathway	19	4.16×10^{19}

Table 25 continued.

WIKIPATHWAYS	Number of genes	Adjusted P-Value
EPO Receptor Signaling	14	4.57×10^{19}
Wnt Signaling Pathway	18	1.16×10^{18}
IL-9 Signaling Pathway	13	7.93×10^{18}
G1 to S cell cycle control	18	1.36×10^{17}
Myometrial Relaxation and Contraction Pathways	24	2.89×10^{17}
T Cell Receptor Signaling Pathway	22	1.56×10^{16}
Nuclear Receptors	14	3.31×10^{16}
AMPK signaling	17	5.14×10^{16}
Signaling of Hepatocyte Growth Factor Receptor	13	1.39×10^{15}
p38 MAPK Signaling Pathway (BioCarta)	13	2.17×10^{15}
Osteopontin	9	3.84×10^{15}
Selenium	16	3.23×10^{13}
EBV LMP1 signaling	10	5.45×10^{13}
metapathway biotransformation	21	1.39×10^{12}
SIDS Susceptibility Pathways	14	1.53×10^{12}
Serotonin HTR1 Group --> FOS Pathway	11	1.92×10^{12}
GPCRs, Class A Rhodopsin-like	24	2.08×10^{12}
MAPK Cascade	10	1.39×10^{11}
FAS pathway and Stress induction of HSP regulation	11	1.45×10^{11}
NLR proteins	7	3.33×10^{11}
Nuclear receptors in lipid metabolism and toxicity	10	5.76×10^{11}
Peptide GPCRs	13	7.85×10^{11}
Serotonin Receptor 4/6/7 -> NR3C signaling	8	2.58×10^{10}
Matrix Metalloproteinases	9	6.06×10^{10}
Ovarian Infertility Genes	9	8.28×10^{10}
cytochrome P450	11	2.88×10^{09}
G Protein Signaling Pathways	13	3.86×10^{09}
Serotonin Receptor 2 -> ELK-SRF/GATA4 signaling	7	4.30×10^{09}
Complement and Coagulation Cascades KEGG	10	5.49×10^{09}
Hypertrophy Model	7	2.37×10^{08}
Osteoclast	6	4.52×10^{08}
Estrogen metabolism	7	4.66×10^{08}
Cytokines and Inflammatory Response (BioCarta)	7	4.66×10^{08}
Inflammatory Response Pathway	7	3.48×10^{07}
Notch Signaling Pathway	8	5.21×10^{07}
BMP signalling and regulation	5	7.77×10^{07}
Type II interferon signaling (IFNG)	8	1.16×10^{06}
Oxidative Stress	7	1.52×10^{06}
NOD pathway	7	3.10×10^{06}
Calcium Regulation in the Cardiac Cell	12	3.10×10^{06}
One Carbon Metabolism	6	3.28×10^{06}
Prostaglandin Synthesis and Regulation	6	7.66×10^{06}
Benzo(a)pyrene metabolism	4	7.87×10^{06}
Steroid Biosynthesis	4	7.87×10^{06}
Hedgehog Signaling Pathway	5	2.03×10^{05}
Signal Transduction of S1P Receptor	5	3.16×10^{05}
Osteoblast	4	4.11×10^{05}
Toll-like receptor signaling pathway	26	2.82×10^{24}
Apoptosis	24	3.45×10^{24}
Toll-like receptor signaling pathway - mir	28	2.76×10^{23}
DNA damage response	22	9.92×10^{23}
estrogen signalling	22	3.69×10^{22}
Integrin-mediated cell adhesion	24	5.76×10^{22}
Wnt Signaling Pathway and Pluripotency	23	7.40×10^{21}
Regulation of Actin Cytoskeleton	26	1.84×10^{20}
Kit Receptor Signaling Pathway	19	4.16×10^{19}
EPO Receptor Signaling	14	4.57×10^{19}
Wnt Signaling Pathway	18	1.16×10^{18}

Table 25 continued.

WIKIPATHWAYS	Number of genes	Adjusted P-Value
IL-9 Signaling Pathway	13	7.93×10^{18}
G1 to S cell cycle control	18	1.36×10^{17}
Myometrial Relaxation and Contraction Pathways	24	2.89×10^{17}
T Cell Receptor Signaling Pathway	22	1.56×10^{16}
Nuclear Receptors	14	3.31×10^{16}
AMPK signaling	17	5.14×10^{16}
Signaling of Hepatocyte Growth Factor Receptor	13	1.39×10^{15}
p38 MAPK Signaling Pathway (BioCarta)	13	2.17×10^{15}
Osteopontin	9	3.84×10^{15}
Selenium	16	3.23×10^{13}
EBV LMP1 signaling	10	5.45×10^{13}
metapathway biotransformation	21	1.39×10^{12}
SIDS Susceptibility Pathways	14	1.53×10^{12}
Serotonin HTR1 Group --> FOS Pathway	11	1.92×10^{12}
GPCRs, Class A Rhodopsin-like	24	2.08×10^{12}
MAPK Cascade	10	1.39×10^{11}
FAS pathway and Stress induction of HSP regulation	11	1.45×10^{11}
NLR proteins	7	3.33×10^{11}
Nuclear receptors in lipid metabolism and toxicity	10	5.76×10^{11}
Peptide GPCRs	13	7.85×10^{11}
Serotonin Receptor 4/6/7 -> NR3C signaling	8	2.58×10^{10}
Matrix Metalloproteinases	9	6.06×10^{10}
Ovarian Infertility Genes	9	8.28×10^{10}
cytochrome P450	11	2.88×10^{09}
G Protein Signaling Pathways	13	3.86×10^{09}
Serotonin Receptor 2 -> ELK-SRF/GATA4 signaling	7	4.30×10^{09}
Complement and Coagulation Cascades KEGG	10	5.49×10^{09}
Hypertrophy Model	7	2.37×10^{08}
Osteoclast	6	4.52×10^{08}
Estrogen metabolism	7	4.66×10^{08}
Cytokines and Inflammatory Response (BioCarta)	7	4.66×10^{08}
Inflammatory Response Pathway	7	3.48×10^{07}
Notch Signaling Pathway	8	5.21×10^{07}
BMP signalling and regulation	5	7.77×10^{07}
Type II interferon signaling (IFNG)	8	1.16×10^{06}
Oxidative Stress	7	1.52×10^{06}
NOD pathway	7	3.10×10^{06}
Calcium Regulation in the Cardiac Cell	12	3.10×10^{06}
One Carbon Metabolism	6	3.28×10^{06}
Prostaglandin Synthesis and Regulation	6	7.66×10^{06}
Benzo(a)pyrene metabolism	4	7.87×10^{06}
Steroid Biosynthesis	4	7.87×10^{06}
Hedgehog Signaling Pathway	5	2.03×10^{05}
Signal Transduction of S1P Receptor	5	3.16×10^{05}
Osteoblast	4	4.11×10^{05}
Tamoxifen metabolism	5	4.65×10^{05}
GPCRs, Other	8	6.25×10^{05}
Tryptophan metabolism	6	8.30×10^{05}
VandyConte::Blakely Network	5	0.0001
Hypothetical Network for Drug Addiction	5	0.0001
Apoptosis Modulation by HSP70	4	0.0001
Small Ligand GPCRs	4	0.0001
Glucocorticoid & Mineralcorticoid Metabolism	3	0.0003
Eicosanoid Synthesis	4	0.0003
Aflatoxin B1 metabolism	3	0.0003
Fatty Acid Biosynthesis	4	0.0004
G13 Signaling Pathway	5	0.0004
Blood Clotting Cascade	4	0.0005

Table 25 continued.

WIKIPATHWAYS	Number of genes	Adjusted P-Value
Nifedipine	2	0.0007
Diurnally regulated genes with circadian orthologs	5	0.0008
Keap1-Nrf2	3	0.0012
Folic Acid Network	4	0.0012
Fatty Acid Beta Oxidation	4	0.0019
Irinotecan Pathway	3	0.0021
Retinol metabolism (BiGCaT, NuGO)	4	0.0023
GPCRs, Class B Secretin-like	3	0.0050
ACE Inhibitor Pathway	2	0.0063

Table 26. The complete list of the pathway enrichment analysis (Pathway commons) results for prostate cancer geneset.

Pathway commons pathways	Number of genes	Adjusted P-Value
Glypican pathway	95	1.98×10^{-79}
Glypican 1 network	89	6.21×10^{-75}
IFN-gamma pathway	77	2.13×10^{-65}
TRAIL signaling pathway	70	7.11×10^{-61}
Proteogylcan syndecan-mediated signaling events	56	1.23×10^{-60}
Regulation of cytoplasmic and nuclear SMAD2/3 signaling	66	2.61×10^{-59}
TGF-beta receptor signaling	66	2.61×10^{-59}
Regulation of nuclear SMAD2/3 signaling	66	2.61×10^{-59}
TNF receptor signaling pathway	64	7.49×10^{-59}
Plasma membrane estrogen receptor signaling	53	1.26×10^{-51}
Class I PI3K signaling events	54	4.88×10^{-47}
Sphingosine 1-phosphate (S1P) pathway	42	6.75×10^{-46}
IL2-mediated signaling events	41	6.82×10^{-46}
Endothelins	44	3.01×10^{-44}
EGFR1	43	2.57×10^{-42}
BMP receptor signaling	47	3.58×10^{-42}
IL1-mediated signaling events	46	3.62×10^{-41}
Syndecan-2-mediated signaling events	34	1.26×10^{-40}
p75(NTR)-mediated signaling	42	1.08×10^{-39}
TGFBR	39	3.75×10^{-39}
LPA receptor mediated events	34	7.88×10^{-39}
Signaling events mediated by PTP1B	26	5.37×10^{-35}
p38 MAPK signaling pathway	38	1.46×10^{-33}
Integrins in angiogenesis	28	2.00×10^{-33}
Signalling by NGF	34	7.49×10^{-33}
Hemostasis	38	3.25×10^{-32}
AndrogenReceptor	29	1.07×10^{-31}
Neurotrophic factor-mediated Trk receptor signaling	29	3.76×10^{-31}
Regulation of p38-alpha and p38-beta	34	1.30×10^{-30}
HIF-1-alpha transcription factor network	26	7.53×10^{-30}
Trk receptor signaling mediated by PI3K and PLC-gamma	24	1.64×10^{-29}
TCR signaling in naive CD4+ T cells	32	3.34×10^{-29}
Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	27	5.93×10^{-29}
FGF signaling pathway	22	2.71×10^{-27}
Angiopoietin receptor Tie2-mediated signaling	21	4.81×10^{-26}
TRKA signalling from the plasma membrane	25	1.04×10^{-25}
Retinoic acid receptors-mediated signaling	20	1.17×10^{-25}
Syndecan-1-mediated signaling events	23	1.34×10^{-25}
Formation of Platelet plug	26	1.94×10^{-25}
Role of Calcineurin-dependent NFAT signaling in lymphocytes	25	1.94×10^{-25}
Wnt	25	3.82×10^{-25}
IL2 signaling events mediated by PI3K	23	6.95×10^{-25}
TCR signaling in naive CD8+ T cells	27	1.87×10^{-24}
IL6-mediated signaling events	20	4.16×10^{-24}
Osteopontin-mediated events	17	4.68×10^{-24}

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
Ceramide signaling pathway	20	6.95×10^{-24}
IL12-mediated signaling events	26	1.80×10^{-23}
Signaling events regulated by Ret tyrosine kinase	22	4.95×10^{-23}
Signaling events mediated by VEGFR1 and VEGFR2	21	3.65×10^{-22}
BCR signaling pathway	21	5.25×10^{-22}
IL23-mediated signaling events	21	7.51×10^{-22}
Class I PI3K signaling events mediated by Akt	23	8.79×10^{-22}
Alpha6Beta4Integrin	19	2.16×10^{-21}
TNF alpha/NF-kB	29	3.38×10^{-21}
Signaling in Immune system	31	7.11×10^{-21}
Regulation of Telomerase	20	3.10×10^{-20}
Signaling by EGFR	20	3.10×10^{-20}
NOTCH	19	5.52×10^{-20}
Signaling events mediated by Stem cell factor receptor (c-Kit)	18	1.22×10^{-19}
Signaling events activated by Hepatocyte Growth Factor Receptor (c-Met)	18	1.81×10^{-19}
a6b1 and a6b4 Integrin signaling	15	2.88×10^{-19}
IGF1 pathway	15	2.88×10^{-19}
Platelet Activation	20	3.65×10^{-19}
S1P1 pathway	17	2.81×10^{-18}
Signaling events mediated by HDAC Class I	21	3.89×10^{-18}
Signaling by Aurora kinases	21	3.89×10^{-18}
Trk receptor signaling mediated by the MAPK pathway	14	9.35×10^{-18}
Integrin cell surface interactions	18	2.88×10^{-17}
Activation of TRKA receptors	16	2.88×10^{-17}
Presenilin action in Notch and Wnt signaling	15	9.03×10^{-17}
FoxO family signaling	14	1.48×10^{-16}
Aurora A signaling	17	2.16×10^{-16}
Signaling events mediated by the Hedgehog family	16	2.48×10^{-16}
KitReceptor	16	3.43×10^{-16}
FOXM1 transcription factor network	14	5.92×10^{-16}
S1P2 pathway	11	1.29×10^{-15}
Fc-epsilon receptor I signaling in mast cells	16	2.89×10^{-15}
S1P3 pathway	11	2.95×10^{-15}
EPO signaling pathway	13	3.85×10^{-15}
Syndecan-4-mediated signaling events	14	4.49×10^{-15}
TRKA activation by NGF	14	6.49×10^{-15}
FAS signaling pathway (CD95)	13	9.30×10^{-15}
Signaling by GPCR	24	1.27×10^{-14}
Hedgehog signaling events mediated by Gli proteins	13	1.40×10^{-14}
Response to elevated platelet cytosolic Ca++	15	1.56×10^{-14}
Exocytosis of Alpha granule	14	1.80×10^{-14}
EphrinB-EPHB pathway	14	2.50×10^{-14}
Platelet degranulation	14	3.44×10^{-14}
VEGFR3 signaling in lymphatic endothelium	11	4.39×10^{-14}
FOXA1 transcription factor network	13	4.42×10^{-14}
FOXA transcription factor networks	16	5.40×10^{-14}
NGF processing	13	6.32×10^{-14}
Grb2 events in EGFR signaling	10	9.98×10^{-14}
PDGFR-beta signaling pathway	13	1.28×10^{-13}
amb2 Integrin signaling	12	1.37×10^{-13}
Ras signaling in the CD4+ TCR pathway	12	2.07×10^{-13}
Regulation of IGF Activity by IGFBP	9	3.18×10^{-13}
a4b1 and a4b7 Integrin signaling	11	3.61×10^{-13}
Signalling to ERKs	11	3.61×10^{-13}
PDGFR-alpha signaling pathway	10	7.21×10^{-13}
Glypican 3 network	13	8.31×10^{-13}
Apoptosis	19	1.15×10^{-12}
Paxillin-independent events mediated by a4b1 and a4b7	10	2.21×10^{-12}

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
Shc events in EGFR signaling	9	3.46×10^{12}
VEGFR1 specific signals	10	6.02×10^{12}
Toll Receptor Cascades	12	8.47×10^{12}
Biological oxidations	17	9.35×10^{12}
Canonical NF-kappaB pathway	11	9.35×10^{12}
IL27-mediated signaling events	10	9.35×10^{12}
Thromboxane A2 receptor signaling	11	9.35×10^{12}
IL4-mediated signaling events	13	1.08×10^{11}
RXR and RAR hetrodimerization with other nuclear receptor	9	1.19×10^{11}
PI3K/AKT signalling	10	1.41×10^{11}
Downstream signaling in naïve CD8+ T cells	12	1.84×10^{11}
Signalling to RAS	9	2.03×10^{11}
Prolonged ERK activation events	9	2.03×10^{11}
ARMS-mediated activation	9	2.03×10^{11}
Frs2-mediated activation	9	2.03×10^{11}
IL2 signaling events mediated by STAT5	10	2.06×10^{11}
Class A/1 (Rhodopsin-like receptors)	18	3.21×10^{11}
Insulin Pathway	12	4.86×10^{11}
Canonical Wnt signaling pathway	11	5.86×10^{11}
Wnt signaling	11	5.86×10^{11}
Cell Cycle, Mitotic	24	6.10×10^{11}
Signaling mediated by p38-alpha and p38-beta	11	1.02×10^{10}
SOS-mediated signalling	8	1.05×10^{10}
Signalling to p38 via RIT and RIN	8	1.05×10^{10}
Phosphorylation of CD3 and TCR zeta chains	12	1.46×10^{10}
TCR signaling	12	1.46×10^{10}
Down-stream signal transduction	9	2.10×10^{10}
Extrinsic Pathway for Apoptosis	16	2.45×10^{10}
Intrinsic Pathway for Apoptosis	16	2.45×10^{10}
Death Receptor Signalling	16	2.45×10^{10}
Reelin signaling pathway	9	3.07×10^{10}
EPHB forward signaling	10	3.14×10^{10}
IRS-mediated signalling	10	3.14×10^{10}
Downstream TCR signaling	11	4.36×10^{10}
Signaling events mediated by PRL	8	5.40×10^{10}
Cell Cycle Checkpoints	16	5.40×10^{10}
Insulin receptor signalling cascade	10	5.41×10^{10}
IRS-related events	10	5.41×10^{10}
Insulin receptor recycling	10	5.41×10^{10}
Signaling by Insulin receptor	10	5.41×10^{10}
G2/M Checkpoints	15	7.88×10^{10}
Innate Immunity Signaling	12	8.03×10^{10}
FasL/ CD95L signaling	15	1.12×10^{09}
Signaling by PDGF	9	1.22×10^{09}
Generation of second messenger molecules	11	1.23×10^{09}
Translocation of ZAP-70 to Immunological synapse	11	1.23×10^{09}
Activation of Pro-Caspase 8	15	1.37×10^{09}
Caspase is formed from procaspase-8	15	1.37×10^{09}
Cytochrome P450 - arranged by substrate type	12	2.21×10^{09}
Phase 1 - Functionalization of compounds	13	2.31×10^{09}
p75 NTR receptor-mediated signalling	9	2.97×10^{09}
SHC-mediated signalling	7	2.97×10^{09}
p38 signaling mediated by MAPKAP kinases	7	2.97×10^{09}
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	10	3.03×10^{09}
Negative regulation of the PI3K/AKT network	8	4.34×10^{09}
TRAF6 Mediated Induction of the antiviral cytokine IFN-alpha/beta	7	5.09×10^{09}
JNK signaling in the CD4+ TCR pathway	9	6.80×10^{09}
EGFR interacts with phospholipase C-gamma	8	1.23×10^{08}

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
Cyclin D associated events in G1	14	1.35×10^{08}
G1 Phase	14	1.35×10^{08}
SHC-related events	7	1.35×10^{08}
EphrinB reverse signaling	7	2.07×10^{08}
Toll Like Receptor 3 (TLR3) Cascade	7	2.07×10^{08}
Axon guidance	8	2.23×10^{08}
NCAM signaling for neurite out-growth	8	2.23×10^{08}
MAP kinase cascade	6	2.79×10^{08}
PLC-gamma1 signalling	8	2.97×10^{08}
Peptide ligand-binding receptors	12	4.02×10^{08}
Diabetes pathways	21	4.02×10^{08}
Activation, myristolyation of BID and translocation to mitochondria	13	4.51×10^{08}
Platelet Aggregation (Plug Formation)	7	6.42×10^{08}
E2F transcriptional targets at G1/S	13	7.49×10^{08}
E2F mediated regulation of DNA replication	13	7.49×10^{08}
Caspase cascade in apoptosis	9	7.71×10^{08}
TNF signaling	13	8.20×10^{08}
Atypical NF-kappaB pathway	6	8.35×10^{08}
HIV-1 Nef: Negative effector of Fas and TNF-alpha	8	1.05×10^{07}
Formation of Fibrin Clot (Clotting Cascade)	8	1.05×10^{07}
Signaling events mediated by HDAC Class III	7	1.22×10^{07}
G1/S Transition	13	1.31×10^{07}
ID	6	1.34×10^{07}
DNA Repair	12	1.34×10^{07}
G2/M DNA damage checkpoint	11	1.66×10^{07}
Activation of BH3-only proteins	6	2.07×10^{07}
IRS activation	6	2.07×10^{07}
IL12 signaling mediated by STAT4	7	2.14×10^{07}
Homologous Recombination Repair	7	2.14×10^{07}
Homologous recombination repair of replication-independent double-	7	2.14×10^{07}
ATM mediated response to DNA double-strand break	7	2.14×10^{07}
Glucagon-type ligand receptors	8	3.02×10^{07}
Platelet activation triggers	6	4.54×10^{07}
Extrinsic Pathway	7	4.82×10^{07}
Integrin alphaIIbbeta3 signaling	6	6.45×10^{07}
Paxillin-dependent events mediated by a4b1	6	6.45×10^{07}
AKT phosphorylates targets in the nucleus	4	6.57×10^{07}
Double-Strand Break Repair	7	7.72×10^{07}
Thrombin signalling through PARs	5	8.14×10^{07}
Gab1 signalosome	5	8.14×10^{07}
Intrinsic Pathway	7	9.63×10^{07}
SMAC-mediated dissociation of IAP:caspase complexes	6	1.21×10^{06}
SMAC binds to IAPs	6	1.21×10^{06}
Cell surface interactions at the vascular wall	9	1.42×10^{06}
SHC activation	5	2.16×10^{06}
NFG and proNGF binds to p75NTR	6	2.17×10^{06}
Class B/2 (Secretin family receptors)	8	2.34×10^{06}
Hormone biosynthesis	8	2.34×10^{06}
Opioid Signalling	7	2.66×10^{06}
Calcium signaling in the CD4+ TCR pathway	6	2.80×10^{06}
RAF phosphorylates MEK	4	4.12×10^{06}
RAF activation	4	4.12×10^{06}
Activation of BAD and translocation to mitochondria	4	4.12×10^{06}
MEK activation	4	4.12×10^{06}
Regulation of Insulin Secretion by Acetylcholine	6	4.55×10^{06}
Common Pathway	6	4.55×10^{06}
Steroid hormones	5	6.67×10^{06}
Dissolution of Fibrin Clot	4	7.94×10^{06}

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
S1P4 pathway	4	7.94×10^{06}
ERK1 activation	4	7.94×10^{06}
CaM pathway	5	9.14×10^{06}
Syndecan-3-mediated signaling events	5	9.14×10^{06}
Activation, translocation and oligomerization of BAX	10	1.23×10^{05}
Ca-dependent events	5	1.24×10^{05}
Nuclear Events (kinase and transcription factor activation)	4	1.37×10^{05}
AKT phosphorylates targets in the cytosol	4	1.37×10^{05}
ERK activation	4	1.37×10^{05}
Signal attenuation	4	2.25×10^{05}
Host Interactions of HIV factors	12	2.70×10^{05}
Apoptotic cleavage of cellular proteins	6	3.76×10^{05}
Eicosanoid ligand-binding receptors	5	5.53×10^{05}
Chemokine receptors bind chemokines	5	5.53×10^{05}
SMAC-mediated apoptotic response	9	6.09×10^{05}
Alpha-synuclein signaling	5	6.71×10^{05}
Lissencephaly gene (LIS1) in neuronal migration and development	5	6.71×10^{05}
PLC beta mediated events	5	6.71×10^{05}
Release of apoptotic factors from the mitochondria	9	6.95×10^{05}
Apoptotic factor-mediated response	9	6.95×10^{05}
Activation of NOXA and translocation to mitochondria	3	7.33×10^{05}
Gap junction trafficking and regulation	3	7.33×10^{05}
Activation of PUMA and translocation to mitochondria	3	7.33×10^{05}
Vpr-mediated induction of apoptosis by mitochondrial outer membrane	9	7.34×10^{05}
Activation and oligomerization of BAK protein	9	7.87×10^{05}
G-protein mediated events	5	9.62×10^{05}
HIV Infection	13	0.0001
Signalling to ERK5	3	0.0001
Alternative NF-kappaB pathway	3	0.0001
Xenobiotics	4	0.0001
BH3-only proteins associate with and inactivate anti-apoptotic BCL-2	3	0.0002
ERK2 activation	3	0.0002
NOSTRIN mediated eNOS trafficking	3	0.0002
Apoptotic execution phase	8	0.0002
p53-Independent G1/S DNA damage checkpoint	7	0.0002
Vitamin D (calciferol) metabolism	3	0.0002
Stabilization of p53	3	0.0002
p53-Dependent G1/S DNA damage checkpoint	7	0.0002
G1/S DNA Damage Checkpoints	7	0.0002
Thrombin signalling G-protein cascades	3	0.0002
DARPP-32 events	4	0.0002
Metabolism of lipids and lipoproteins	13	0.0002
LPA4-mediated signaling events	3	0.0002
NRAGE signals death through JNK	3	0.0002
S Phase	8	0.0004
Basigin interactions	3	0.0004
ATM mediated phosphorylation of repair proteins	3	0.0004
Transport of connexons to the plasma membrane	2	0.0004
ERK/MAPK targets	3	0.0004
Aurora B signaling	5	0.0004
Regulation of gap junction activity	2	0.0004
DNA Replication	8	0.0004
Interactions of Vpr with host cellular proteins	9	0.0004
c-src mediated regulation of Cx43 function and closure of gap junctions	2	0.0004
Recruitment of repair and signaling proteins to double-strand breaks	3	0.0004
Cytochrome c-mediated apoptotic response	8	0.0004
DNA Replication Pre-Initiation	8	0.0004
Viral dsRNA:TLR3:TRIF Complex Activates RIP1	3	0.0004

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
FOXA2 and FOXA3 transcription factor networks	5	0.0004
AKT-mediated inactivation of FOXO1A	2	0.0004
eNOS activation	3	0.0004
Activation of caspases through apoptosome-mediated cleavage	8	0.0004
Regulated proteolysis of p75NTR	3	0.0004
Signaling by BMP	3	0.0004
NF-kB is activated and signals survival	3	0.0004
Acetylation	2	0.0004
eNOS acylation cycle	3	0.0004
Nef and signal transduction	3	0.0004
M/G1 Transition	8	0.0004
Signaling by TGF beta	3	0.0006
eNOS activation and regulation	3	0.0006
Vitamins	3	0.0006
Pyrimidine metabolism	5	0.0006
Metabolism of nitric oxide	3	0.0006
p130Cas linkage to MAPK signaling for integrins	3	0.0006
Activation of the pre-replicative complex	5	0.0008
Activation of ATR in response to replication stress	7	0.0008
Oligomerization of connexins into connexons	2	0.0012
G alpha (12/13) signalling events	2	0.0012
Gap junction assembly	2	0.0012
Signalling to STAT3	2	0.0012
Polo-like kinase mediated events	6	0.0012
Transport of connexins along the secretory pathway	2	0.0012
P450 Epoxidations	2	0.0012
p75NTR signals via NF-kB	3	0.0012
VEGF binds to VEGFR leading to receptor dimerization	2	0.0012
p75NTR recruits signalling complexes	3	0.0012
Toll Like Receptor 2 Cascade	3	0.0012
mTOR signaling pathway	4	0.0013
Assembly of the pre-replicative complex	7	0.0015
Phase II conjugation	4	0.0017
Regulation of Insulin Secretion	11	0.0019
Regulation of DNA replication	6	0.0021
Androgen biosynthesis	2	0.0022
Thrombin-mediated activation of PARs	2	0.0022
NGF-independant TRKA activation	2	0.0022
S1P5 pathway	2	0.0022
Gap junction trafficking	2	0.0022
Collagen adhesion via alpha 2 beta 1 glycoprotein	2	0.0022
Proteinase-activated receptor G (12/13) cascade	2	0.0022
PKA-mediated phosphorylation of key metabolic factors	2	0.0022
Cleavage of the damaged purine	2	0.0022
p38MAPK events	2	0.0022
Depurination	2	0.0022
S Phase	8	0.0004
Basigin interactions	3	0.0004
ATM mediated phosphorylation of repair proteins	3	0.0004
Transport of connexons to the plasma membrane	2	0.0004
ERK/MAPK targets	3	0.0004
Aurora B signaling	5	0.0004
Regulation of gap junction activity	2	0.0004
DNA Replication	8	0.0004
Interactions of Vpr with host cellular proteins	9	0.0004
c-src mediated regulation of Cx43 function and closure of gap junctions	2	0.0004
Recruitment of repair and signaling proteins to double-strand breaks	3	0.0004
Cytochrome c-mediated apoptotic response	8	0.0004

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
DNA Replication Pre-Initiation	8	0.0004
Viral dsRNA:TLR3:TRIF Complex Activates RIP1	3	0.0004
FOXA2 and FOXA3 transcription factor networks	5	0.0004
AKT-mediated inactivation of FOXO1A	2	0.0004
eNOS activation	3	0.0004
Activation of caspases through apoptosome-mediated cleavage	8	0.0004
Regulated proteolysis of p75NTR	3	0.0004
Signaling by BMP	3	0.0004
NF-kB is activated and signals survival	3	0.0004
Acetylation	2	0.0004
eNOS acylation cycle	3	0.0004
Nef and signal transduction	3	0.0004
M/G1 Transition	8	0.0004
Signaling by TGF beta	3	0.0006
eNOS activation and regulation	3	0.0006
Vitamins	3	0.0006
Pyrimidine metabolism	5	0.0006
Metabolism of nitric oxide	3	0.0006
p130Cas linkage to MAPK signaling for integrins	3	0.0006
Activation of the pre-replicative complex	5	0.0008
Activation of ATR in response to replication stress	7	0.0008
Oligomerization of connexins into connexons	2	0.0012
G alpha (12/13) signalling events	2	0.0012
Gap junction assembly	2	0.0012
Signalling to STAT3	2	0.0012
Polo-like kinase mediated events	6	0.0012
Transport of connexins along the secretory pathway	2	0.0012
P450 Epoxidations	2	0.0012
p75NTR signals via NF-kB	3	0.0012
VEGF binds to VEGFR leading to receptor dimerization	2	0.0012
p75NTR recruits signalling complexes	3	0.0012
Toll Like Receptor 2 Cascade	3	0.0012
mTOR signaling pathway	4	0.0013
Assembly of the pre-replicative complex	7	0.0015
Phase II conjugation	4	0.0017
Regulation of Insulin Secretion	11	0.0019
Regulation of DNA replication	6	0.0021
Androgen biosynthesis	2	0.0022
Thrombin-mediated activation of PARs	2	0.0022
NGF-independant TRKA activation	2	0.0022
S1P5 pathway	2	0.0022
Gap junction trafficking	2	0.0022
Collagen adhesion via alpha 2 beta 1 glycoprotein	2	0.0022
Proteinase-activated receptor G (12/13) cascade	2	0.0022
PKA-mediated phosphorylation of key metabolic factors	2	0.0022
Cleavage of the damaged purine	2	0.0022
p38MAPK events	2	0.0022
Depurination	2	0.0022
Sumoylation by RanBP2 regulates transcriptional repression	3	0.0022
Cyclin B2 mediated events	2	0.0022
G2/M Transition	8	0.0022
Recognition and association of DNA glycosylase with site containing an	2	0.0022
Endogenous sterols	4	0.0023
Phase 1 functionalization	3	0.0027
Cell death signalling via NRAGE, NRIF and NADE	3	0.0027

Table 26 continued.

Pathway commons pathways	Number of genes	Adjusted P-Value
EGFR downregulation	3	0.0032
Signaling events mediated by HDAC Class II	4	0.0036
Insulin-mediated glucose transport	3	0.0037
CREB phosphorylation	2	0.0037
Axonal growth stimulation	2	0.0037
Mineralocorticoid biosynthesis	2	0.0037
Signaling by VEGF	2	0.0037
VEGF ligand-receptor interactions	2	0.0037
Glucocorticoid biosynthesis	2	0.0037
Pyrimidine salvage reactions	4	0.0038
Prostanoid ligand receptors	3	0.0043
Hormone ligand-binding receptors	3	0.0050
Further platelet releasate	3	0.0050
Activation of BIM and translocation to mitochondria	2	0.0053
p75NTR regulates axonogenesis	2	0.0053
TRAIL signaling	2	0.0053
Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex	5	0.0061
Synthesis of DNA	6	0.0063
Reversible phosphorolysis of pyrimidine nucleosides by thymidine	2	0.0071
Metablism of nucleotides	6	0.0071
Inhibition of replication initiation of damaged DNA by Rb/E2F1	2	0.0071
The role of Nef in HIV-1 replication and disease pathogenesis	3	0.0074
p53-Dependent G1 DNA Damage Response	5	0.0075
Cyclin A/B1 associated events during G2/M transition	5	0.0085
Viral dsRNA:TLR3:TRIF Complex Activates TBK1	2	0.0093
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	2	0.0093
Pregnenolone biosynthesis	2	0.0093