UPPSALA
UNIVERSITET

UPTEC X 17024
Examensarbete 30 hp
Augusti 2017

A comparative validation of the human variant simulator SIMdrom

Sofia Ånäs

UPPSALA UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

A comparative validation of the human variant simulator SIMdrom

Sofia Ånäs

The past decade's progress in next generation sequencing has drastically decreased the price of whole genome and exome sequencing, making it available as a clinical tool for diagnosing patients with genetic disease. However, finding a disease-causing mutation among millions of non-pathogenic variants in a patient's genome, is not an easy task. Therefore, algorithms for finding variants relevant for clinicians to investigate more closely are needed and constantly developed. To test these algorithms a software called SIMdrom has been developed to simulate test data. In this project, the simulated data is validated through comparison to real genetic data to ensure that it is suitable to use as test data. Through ensuring the data's reliability and finding possible improvements, the development of algorithms for finding disease-causing mutations can be facilitated. This in-turn could lead to better diagnosing-possibilities for clinicians. When visualizing simulated data together with real genomes using principal components analysis, it clusters near it's real counterpart. This shows that the simulated data resembles the real genomes. Simulated exomes also performed well when used as a part in one of three training sets for the classifier in the Prioritization of Exome Data by Image Analysis study. Here they perform second best after an in-house data set consisting of real exomes. To conclude, the SIMdrom simulated data performs well in both parts of this project. Additional tests of its validity should include testing against larger real data sets, an improvement possibility could be to implement a simulation option for spiking in noise.

Handledare: Max Schubach
Ämnesgranskare: Torsten Günther
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X 17024

Alles hat ein Ende,
nur die Wurst hat zwei

(Allt har ett slut, bara korven har två)
Tyskt talesätt

Till mamma, pappa, mormor och farmor.

Sammanfattning

DNA är receptet som styr hur allt liv på jorden ser ut och fungerar, men även delvis vilka sjukdomar en organism riskerar att utveckla. Hos oss människor består arvsmassan av 23 kromosompar. Eftersom människor är diploida organismer betyder det att nästan alla våra gener finns i två kopior, även kallade alleler, med undantag för dem som ligger på könskromosomerna (där män har en kopia av X- och en av Y-kromosomen, medan kvinnor har två kopior av X-kromosomen). Vilken uppsättning av gener vi har beror på vad vi ärver från våra föräldrar. Denna uppsättning kallas genotyp och kan vara antingen homozygot om de två allelerna är likadana till exempel AA eller aa medan den kallas heterozygot om allelerna är olika, vilket i detta fall skulle vara Aa. Dessa alleler existerar i olika frekvenser inom olika populationer, där vissa kan vara vanligare än andra. Vissa alleler kan vara sjukdomsalstrande på så vis att de förstör, förändrar eller på annat sätt påverkar en funktion i kroppen till något negativt. Dessa allelvarianter kallas mutationer och ger alltså upphov till genetiska sjukdomar.

För att hitta vilken eller vilka mutationer hos patienter med genetiska sjukdomar som gett upphov till sjukdomen, brukar man idag ofta undersöka hela patientens genom eller exom. Att undersöka genomet innebär att läsa av, sekvensera, hela arvsmassan. En undersökning av exomet innebär däremot att endast sekvensera den del av arvsmassan som består av gener som kodar för proteiner, och därmed anses ha en direkt funktion i kroppen. Exomet utgör enbart ungefär 2% av den totala arvsmassan, men kan ändå innehålla uppemot 50 000 varianter! Det innebär alltså att det är frågan om en mycket stor informationsmängd som ska undersökas. Även om endast ovanliga allelvarianter tas med i utredningen är de för många för att enkelt kunna ögnas igenom och hitta dem som kan orsaka en sjukdom.

Här kan algoritmer, som utvecklats för att sålla bort troligtvis ofarliga varianter och prioritera dem som anses riskabla, vara av stor nytta. Dessa algoritmer kan ta in en mängd information i beräkningen för att räkna ut vilka varianter som skulle kunna orsaka skadliga förändringar hos en människa medan andra varianter kanske inte alls behöver vara farliga. För att vara säker på att dessa algoritmer fungerar som de ska och faktiskt hittar de varianter som eftersöks behöver de testköras på data som liknar den riktiga datamängden, men där den sökta mutationen är känd. På så sätt är det möjligt att ta reda på om algoritmen kan hitta mutationen eller felaktigt sållar bort den och alltså behöver utvecklas och justeras ytterligare.

I detta projekt testas och utvärderas programvaran SIMdrom som är utvecklad för att simulera genetiska data, i egenskap av genom och exom. Den simulerade datamängden kan användas som testdata för algoritmerna. Det är därför viktigt att veta att den simulerade datamängden liknar de riktiga data som algoritmerna sedan ska användas på. I det här projektet kommer den därför att jämföras med riktiga genetiska data i två olika steg, för att undersöka hur lämplig den är för det tänkta användningsområdet. Förhoppningen är att kunna säkerställa lämpligheten och i annat fall kunna föreslå utvecklingsmöjligheter för att förbättra den

simulerade datamängden. Detta för att algoritmerna i sin tur ska kunna testas på tillförlitliga data och på så sätt fortsätta utvecklas. Slutligen är syftet att dessa verktyg som algoritmerna är ska kunna hjälpa genetiker att korrekt diagnosticera sina patienter.

Table of content

Abbreviations

1KG	1000 Genomes Project
AC	allele counts
AChemi	allele count hemizygous
AC	allele count heterozygous
AChom	allele count homozygous
AF	allele frequency
AFR	African/African American
AMR	Latin American
AN	total number of alleles
AUPRC	area under precision-recall curve
AUROC	area under receiving operator characteristics curve
BOQA	Bayesian Ontology Query Algorithm
CADD	Combined Annotation Dependent Depletion
DNA	deoxyribonucleic acid
EAS	East Asian
ExAC	Exome Aggregation Consortium
FIN	Finnish
GC	genotype counts
gnomAD	Genome Aggregation Database
HPO	Human Phenotype Ontology
IRN	Iranian samples
LD	linkage disequilibrium
NFE	Non-Finnish European
NGS	next generation sequencing
OMIM	Online Mendelian Inheritance in Man
OoA	Out of Africa
PCA	principal components analysis
PEDIA	Prioritization of Exome Data by Image Analysis
PRC	precision-recall
RefSeq	Reference Sequence
ROC	receiving operator characteristics
SNP	single nucleotide polymorphism
SVM	support vector machine
VCF	variant call format

1 Introduction

The next generation sequencing (NGS) era started a decade ago, somewhere in the middle of the 2000's. Since then, sequencing technologies have continued to advance, leading to the construction of comprehensive references of the human genome and a severe cost-reduction in sequencing (Goodwin *et al.* 2016). Now, the low cost enables sequencing exomes and genomes with clinical purpose and has become a favourable tool for identifying causal variants in genetic disease (Goodwin et al. 2016, Bamshad et al. 2011). However, this is not an easy task as the human genome contains several millions of variants, most of which are not disease-causing. Even when sequencing exomes, the protein-coding part of the genome which represents less than 2% of the whole genome, between 20 000 and 50 000 variants are found (Gilissen *et al.* 2012). One of the first steps in reducing the number of possible variants is to exclude known variants through filtering them against e. g. a variant database. This reduces the number of possible variants with about 90%. Still, this leaves a lot of variants that need to be prioritized according to their relevance as causing of the patients disease (Gilissen *et al.* 2012). Here, bioinformatics tools can be of great help. Algorithms for prioritizing the variants through integrating more information about them beyond their potential deleteriousness and rarity can reduce the number of possible causal variants (Smedley *et al.* 2015).

To improve these tools and evaluate them, they need to be tested (benchmarked) on data with known diseases and mutations. Such data is typically a genome from a genetic database that is spiked with a known disease-causing mutation (Smedley *et al.* 2016). However, these are curated genomes and may be too "perfect" compared to exomes/genomes sequenced from patients. In an attempt to deal with this, a software called SIMdrom was developed for simulating genetic data to use for benchmarking, by the computational group at the Institute of Medical Genetics and Human Genetics at Charité in Berlin. This simulated genetic data was later recognized as a possible part in the Prioritization of Exome Data by Image Analysis (PEDIA) study, where there at the moment is a need for simulated exomes.

1.1 Purpose

The purpose of the project is to assess the reliability of the software SIMdrom, to find out how similar genomes sampled using SIMdrom are to real genomes (such as the 1000 Genomes Project (1KG) genomes, The 1000 Genomes Project Consortium 2015). Since the simulated genomes are being used in other projects (such as benchmarking other tools) it is important to know how realistic they are. Through this validation, improvement possibilities of SIMdrom might be recognized which in turn can aid in the improvement of the tools for finding disease-causing mutations. For this validation two goals were set:

1. Make a statistical validation of SIMdrom.
2. Compare different genetic training sets for the PEDIA study.

2 Background

Humans are diploid organisms, which means our genome exists in two copies. These two versions of our DNA are distributed over 22 autosomal chromosome pairs, and one pair of sex chromosomes, the gonosomes, XX (female) or XY (male) (Holmquist & Wienberg 2001). Each part on a chromosome is called a locus, and each chromosomal pair share the same locus for a specific gene. However, the same locus on a chromosome pair may carry different bases, or sequences of bases. The different locus variants are called alleles and can occur at different frequencies across populations (The 1000 Genomes Project Consortium 2015). The genotype defines what set up of alleles the chromosome pair carries at a locus, and can be either homozygous or heterozygous. The homozygous genotype means that the chromosomes carry the same allele, and the heterozygous genotype carries two different alleles (Lam & Mueller 2001). When the genotype is expressed it affects the phenotype, which are the observable traits in an individual. Different genotypes at a locus can be connected to different phenotypes, hence the composition of alleles at a locus may decide what phenotype the individual displays (Scriver 2001). A phenotype can be everything from eye colour to a disease syndrome and can be caused by just one specific allele or several genotypes in combination with environmental factors (White & Rabago-Smith 2011, Bamshad et al. 2011, Manolio et al. 2008).

To determine what variation of alleles and genotypes an individual carries, their sequenced exome or genome is compared to a reference sequence (Nielsen *et al.* 2011). The reference sequence is built from a consensus of several individual genomes (O'Leary *et al.* 2016). Alleles that differ from the reference allele at a locus are called variants. A typical human genome contains millions of variants, a few of these can be pathogenic and cause disease (The 1000 Genomes Project Consortium 2015, Bamshad et al. 2011). To predict if a variant is deleterious many aspects must be taken into consideration, e. g. the region on the chromosome, is it in a gene, will it cause a change in the amino acid sequence etc. Therefore, algorithms developed to incorporate this information in assessing variants deleteriousness are very useful (Kircher *et al.* 2014). To find the causative mutation in a patient is still not an easy task as each exome can contain up to roughly 50 000 variants (Gilissen *et al.* 2012), hence more algorithms can be of help in prioritizing the affected genes (Smedley *et al.* 2015).

2.1 Simulation of variants using SIMdrom

In this project, the Java based software SIMdrom will be used to simulate human exomes, or more precisely variants in the human exome. As input it takes variant call format (VCF) files, a file format for storing variable positions across the genome along with a reference. The files consist of three parts, the first is meta-information about the file, second a header describing the different columns, and last all the variable positions and information about them (Danecek *et al.* 2011).

An example can be seen in Figure 1 below, displaying the three fields. Here, the site 234 on chromosome 1 has one reference allele G and one alternate allele T. The given information concerns the alternate allele, in this example it has a phred quality score of 30 and has passed all filters. The INFO column is a flexible field which can contain many different attributes describing the samples. In Figure 1, the INFO column contains information about the allele frequency (AF) of the alternate allele which is 0.25, and its allele count (AC) of 25 in a total of 100 alleles (AN) of the called genotypes in a population. The attributes defined in the INFO column must be explained in the meta-information field to be valid (Figure 1). There is also one sample present in the file (sample1) which is heterozygous at this position, it carries one reference allele and one alternate allele (0/1) (Danecek *et al.* 2011). In this project, the information about allele frequencies, allele counts and counts of genotypes will be of interest. With SIMdrom it is possible to use this information to sample a new VCF-file containing an individual with a sampled set of genotypes.

```
##fileformat=VCFv4.2

##FORMAT=<ID=GT,Number=1,Type=String,Description=”Genotype”>

##INFO=<ID=AC,Number=A,Type=Integer,Description=”Allele count in genotypes, for each ALT allele, in the same order as listed”>

##INFO=<ID=AF,Number=A,Type=Float,Description=”Allele frequency, for each ALT allele, in the same order as listed”>

##INFO=<ID=AN,Number=1,Type=Integer,Description=”Total number of alleles in called genotypes”>

#CHROM    POS    ID        REF   ALT   QUAL   FILTER   INFO                     FORMAT   sample1

1         234    rs4567    G     T     30     PASS     AF=0.25;AC=25;AN=100     GT       0/1
```

Figure 1. An example of a VCF file. The meta-information field marked '##' describing the information in the file, the header marked '#' describes all the ten columns, the second row is a data line containing information about a variant position.

There are different ways to perform the simulation, depending on what output is desired and what input is available. In this project, three different variant databases are used, the 1KG (The 1000 Genomes Project Consortium 2015), Exome Aggregation Consortium (ExAC), and Genome Aggregation Database (gnomAD, Lek et al. 2016). All of them can be used as input for SIMdrom, with slightly different sampling techniques. The 1KG genomes are multiVCF-files which means that each file contains genotype information about multiple samples. From these files, one possibility is to use SIMdrom to randomly pick one sample. The ExAC and gnomAD in contrast, do not contain any genotype information about the samples, however they contain allele frequencies for all the variants, also within different populations. From the allele frequencies, it is possible to sample genotypes using SIMdrom.

The sampling of genotypes from allele frequencies utilizes the Hardy-Weinberg principle (eq. 1 and 2) to be able to calculate the probabilities for the sample being homozygous for the reference or the alternate, or heterozygous. Eq. 1 describes the allele frequencies in a diploid population at a locus with two different alleles, p and q. Eq. 2 can be derived from squaring eq. 1, which then describes the genotype frequencies for homozygous reference (p^2), heterozygous (*2pq*) and homozygous alternate (q^2). The Hardy-Weinberg principle describes genetic variation in a population at equilibrium, where the variation is constant between generations (Crow 1988). Therefore, to be able to use the principle the populations under investigation must be assumed to be at equilibrium. Under those assumptions, it is possible to calculate the frequencies of homozygous and heterozygous individuals in a population using eq. 1 and 2. If there are more than two alternate alleles at the same locus the equations are altered through adding the number of additional alleles to eq.1 before it is squared. For three possible alleles, there will be six possible genotypes, etc.

$$p + q = 1 \qquad \text{(eq. 1)}$$

$$p^2 + 2pq + q^2 = 1 \qquad \text{(eq. 2)}$$

In SIMdrom, each of the calculated genotype frequencies are given an interval between zero and one, equivalent to their size (an example can be seen in Figure 2). The genotype is then determined by sampling a number between zero and one, e. g. if the random number x is larger than p^2 but smaller than p^2+$2pq$ the resulting genotype is heterozygous (GT in Figure 2).

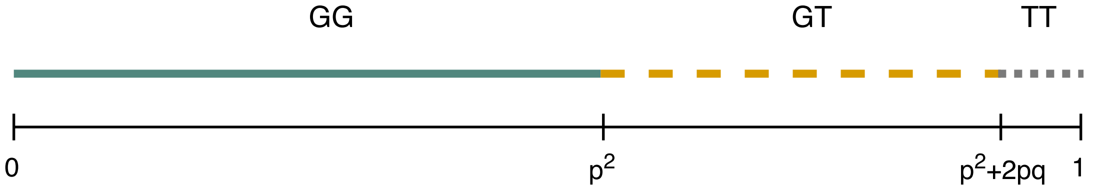


Figure 2. An example of the sampling of genotypes in SIMdrom. The sampling can be described as picking a random number between zero and one, where each genotype (GG, GT and TT) covers an interval corresponding their calculated or measured frequency in a population.

Another sampling alternative for ExAC is to use the measured number of homozygous and heterozygous alleles (AChom, AChet, and AChemi if gonosome, in the ExAC INFO column), to calculate the real genotype frequencies in the population. This alternative gives a more exact relationship between the different genotypes within the population, since they are measured and not predicted through Hardy-Weinberg. Hence, an individual with genotypes sampled through homozygous and heterozygous counts could come closer to resembling the reality. In gnomAD the allele counts are replaced with genotype counts (GC in the gnomAD INFO column) which can be used in a similar way in SIMdrom.

2.2 Validation of SIMdrom using PCA

To investigate how close SIMdrom sampled genomes are to real genomes, the idea was to use a method described by Novembre et al. (2008). In their paper, genetic variance from different populations within Europe were visualized using a principal components analysis (PCA) software called smartpca (Patterson *et al.* 2006). Their theory was that genetic data contains population specific differences even within Europe and that it can be visualized with PCA (Novembre *et al.* 2008).

In this project, the method is used to visualize the difference between sampled and real data. The smartpca software is specially developed for analysing the variance in genetic data. Each individual in the data gets represented by a row *i* in a matrix *M*, and each genetic variant by a column *j*. The genotypes as they are described in the VCF-files, are translated into the number of reference alleles at that position (0, 1 or 2). An example of such a matrix can be seen in Figure 3, displaying three individuals and four variable positions, where individual 1 have one reference allele (heterozygous) at variant 1. After a normalization step, the PCA is performed through singular value decomposition of the matrix *M* (Patterson *et al.* 2006). After the analysis, the eigenvectors that describe the most of the variance between the data sets can be visualized in a plot.

	variant 1	variant 2	variant 3	variant 4
individual 1	1	1	0	2
individual 2	1	2	0	0
individual 3	1	2	1	2

Figure 3. Genetic data turned into a matrix consisting of three individuals, four variable positions, and their genotypes 0, 1 or 2 representing the number of reference alleles at that position.

2.3 Simulating exomes for the PEDIA study

Another way of assessing the reliability of SIMdrom (and evaluate a possible area of application) is to see how the simulated data performs against real data in a project. The PEDIA study combines three different methods in monogenetic disease diagnosis, genotype, phenotype, and image analysis. The aim of the study is to make it easier to identify disease-causing mutations.

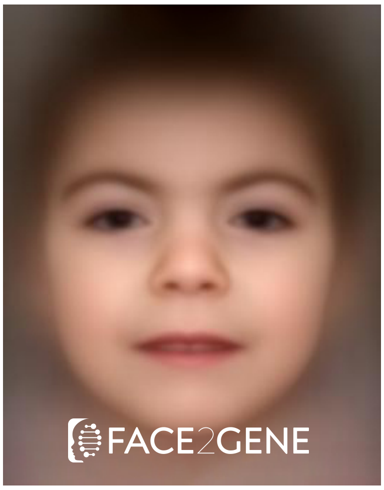


Figure 4. This is an example of a mask from Face2Gene representing a genetic syndrome. Displayed here is a mask for the Cornelia de Lange syndrome.

The image analysis is performed by an algorithm developed by Face2Gene (FDNA INC. 2017). The algorithm takes a photo of the patient's face as input and then compares biometrics from the patient's face to so-called "masks" (Figure 4). The masks represent different syndromes, constructed from patients with a known diagnose. Through the comparison to the masks, the patient can be classified to possible syndromes. The syndromes are scored by the algorithm, where the one that best fits the patient's facial measurements get the highest score.

The phenotype prioritization is performed using three different algorithms; Feature match also by Face2Gene, Phenomizer, and Bayesian Ontology Query Algorithm (BOQA) (FDNA INC. 2017, Köhler et al. 2009, Bauer et al. 2012). Using different ontology search methods, the algorithms prioritize diseases based on similarities to the described phenotypes of a patient. Given a set of Human Phenotype Onology (HPO) terms from a patient, the algorithms score possible diseases and their associated genes (Köhler et al. 2009, Bauer et al. 2012). The HPO is a controlled phenotype vocabulary for describing clinical abnormalities (Köhler *et al.* 2017). The molecular prioritization of the exome is performed using Combined Annotation Dependent Depletion (CADD) which generates a score from the deleteriousness of the variants in the patient's exome. The score represents how probable it is that a dysfunctionality of a gene is involved in a given phenotype (Kircher *et al.* 2014).

In the PEDIA study the three methods, phenotype prioritization, gene prioritization, and image analysis (Figure 7), get represented as scores per gene. These scores are then used to build a feature vector for a support vector machine (SVM), a supervised machine learning model. Through training on solved cases, the SVM learns to calculate a decision boundary for classifying affected and unaffected genes (Cortes & Vapnik 1995).

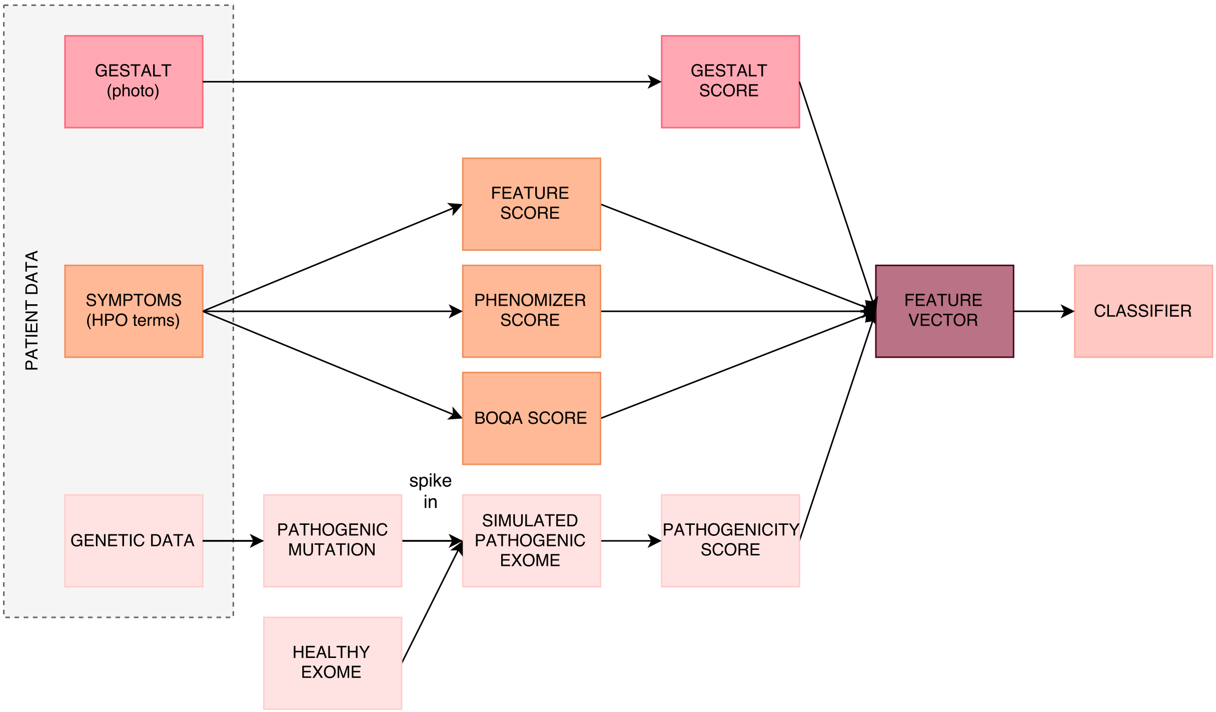


Figure 5. The different parts in the PEDIA study, used to build the feature vector for the classifier.

A current problem in the PEDIA study is the lack of exomes with known diagnosis, since most of the genetic data from patients consist of single-gene tests or cases from literature. Exomes with known diagnosis are needed for training and testing the classifier. Therefore, to generate more genetic data, 1KG genomes have been used as background to spike-in patients' causal variants into (Figure 5). This data has then been used in the training set (together with the patient's image and described phenotype), and the existing patients' exomes in the test set.

In this project, SIMdrom simulated exomes will be used in one of three training sets (Figure 7) for the PEDIA classifier, as an evaluation of the training sets and to see how the simulated exomes perform against real data.

3 Materials and method

3.1 Goal 1

In the first goal, the software smartpca was used to visualize differences between SIMdrom simulated exomes/genomes and real genomes from the 1000 Genomes Project. To document the method and to be able to run everything simultaneously the workflow management system Snakemake was used (Köster & Rahmann 2012). A complete documentation of the workflow for goal one can be found at https://github.com/sofiaanas/smartflow.

3.1.1 Preparation

In the preparation step, the ExAC exome database, the gnomAD genome database and 1KG genomes were downloaded from their respective databases (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/ 2017-02-14, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ 2017-04-06, https://storage.googleapis.com/gnomad-public/release-170228/vcf/genomes/ 2017-04-27). The analysis was divided into two parts, simulation of exomes using ExAC and simulation of exomes using gnomAD, both compared to real 1KG genomes but cut to different regions. For the ExAC analysis, the simulated exomes and real 1KG samples were filtered to regions with a selected coverage (regions with a depth of at least 20 in 80% of samples) in ExAC (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/coverage/ 2017-04-12). In the gnomAD analysis, the simulated and real genome samples were filtered to regions covered in RefSeq (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/ARCHIVE/ANNOTATION_RELEASE.105/GFF/ 2017-04-26). The real samples were randomly hand-picked from the 1KG samples-list (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/20130606_sample_info.xlsx 2017-04-06) to match the populations simulated from ExAC and gnomAD, (Appendix A 1).

3.1.2 Simulation

The SIMdrom simulation was performed in two different ways for each analysis (ExAC and gnomAD) to visualize differences between the simulation options. The ExAC exomes were simulated using the allele frequency (AF) for five different populations African/African American (AFR), Latin American (AMR), East Asian (EAS), Finnish (FIN), Non-Finnish European, (NFE), creating 50 exomes, ten from each population. The same populations were also simulated using the allele counts (AC) of homozygotes and heterozygotes. The gnomAD exomes were simulated using allele frequency too, for the same five populations resulting in 50 exomes (ten of each population). In addition to that, another 50 exomes were simulated using the genotype counts (GC). In total 100 exomes were simulated from ExAC and 100 from gnomAD.

3.1.3 Analysis and visualization

To be able to compare the differences between the simulations and the real samples, the simulated files were merged with the real samples to perform the smartpca. The different simulation options (AF, AC, GC) within each analysis (ExAC, gnomAD) were merged first, with each other, second, with the real samples (Table 1), and third, with each other and the real samples (Table 2). This resulted in eight different analysis files which can be seen in Table 1 and Table 2. The files were then modified somewhat to be able to be used in later steps. New ID's were assigned to the variants because some did not have a previous identification. As a result from the file-merging, genotypes homozygous for the reference appear encoded as './.', these were corrected to '0/0'. Regions considered to have linkage disequilibrium (LD) were pruned using PLINK (Purcell & Chang 2017, Chang et al. 2015), removing variants displaying correlations between genotype allele counts (using same settings as in Novembre et al. 2008, removing variants within every 50 single nucleotide polymorphism (SNP) displaying pairwise squared correlations greater than 80%). Otherwise LD can induce correlations in the nearby columns in the matrix, since nearby variants will display the same pattern (Patterson *et al.* 2006).

Table 1. The files were merged before the smartpca analysis to be able to compare the different simulation options and their similarity to the real samples, therefore the merges seen in the table below were created.

	ExAC AF	**gnomAD AF**	**real 1KG (R)**
ExAC AC	E:AF+AC	X	E:AC+R
gnomAD GC	X	G:AF+GC	G:GC+R
real 1KG (R)	E:AF+R	G:AF+R	X

Table 2. In addition to the merges above the most important merges were the ones where both the different simulation options and the real samples were present. The merges in the left column can be found in Table 1

	real 1KG (R)
E:AF+AC	E:AF+AC+R
G:AF+GC	G:AF+GC+R

Smartpca takes a parameter file as input. This file contains the path to three files, the genotype file, the SNP file and the individuals file, and some parameters for the analysis. The genotype file contains all the genotypes encoded as in Figure 3, but each row represents a variant and each column an individual. The SNP file contains all the variants, their positions and their ID's. Last, the individuals file contains all the individuals, their sex (in this project all displayed as unknown, U) and population labels. Examples of these files can be seen in Figure 6 below.

a

```
2 2 1
0 2 2
1 1 1
```

b

```
1   234    rs305
1   789    rs860
1   1123   rs1194
```

c

```
AF_AFR_1   U   AF_AFR
AC_AMR_1   U   AC_AMR
HG00140    U   R_NFE
```

Figure 6. Displays examples of the genotype file (a), SNP file (b), and individuals file (c).

The three files were created (for each analysis) through cutting and pasting columns together from the VCF-files. Finally, smartpca was run on the different analysis sets and the results were visualized using R.

3.2 Goal 2

In the second goal, SIMdrom simulated exomes were used to create the genetic base in a training set for the PEDIA classifier. The simulated exomes were a part in one of three training sets consisting of 203 exomes each (Figure 7). The other two sets were real genomes from the 1KG and real exomes from an in-house data set of healthy individuals of Iranian heritage (IRN). Later in the project more data was available both for the train and test sets. Then 320 exomes could be used as the genetic part in the training set (still consisting of simulated exomes from ExAC, 1KG genomes and IRN exomes, spiked-in with patients' disease-causing variants) and 19 patient exomes were used in the test set.

After training, the classifier was tested on the real patient data. The difference in performance of the classifier was analysed to see if any of the three training sets was a better fit to the real data.

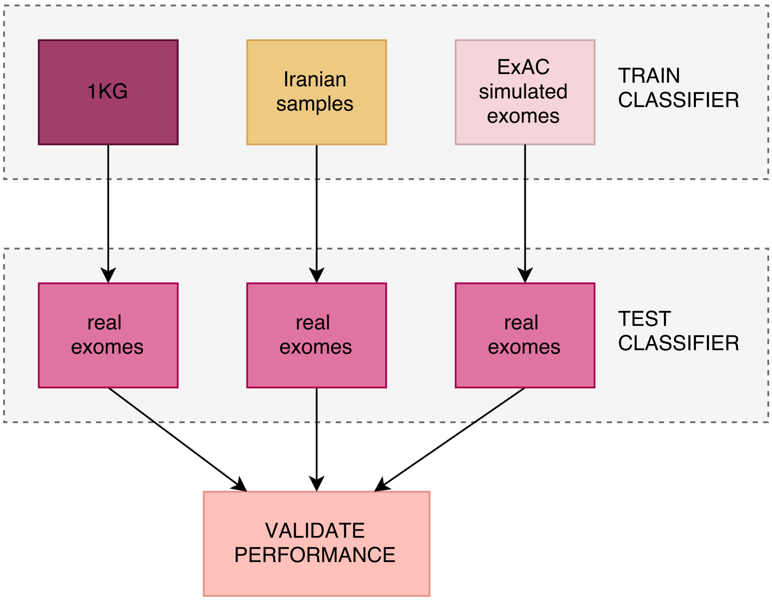


Figure 7. Simplified workflow for the use and validation of three different genetic data sets in the training part of the PEDIA classifier.

3.2.1 Preparing and scoring the genetic data

A workflow for using 1KG as the genetic base in a training set was already in place in the study, hence the use of ExAC and Iranian samples were implemented in that workflow as well. The download and preparation of the ExAC was performed as in goal one, except here filtered to RefSeq regions, so were also the Iranian samples. Additionally, the sets were filtered down to only contain rare variants, using a cut-off at allele frequency of maximum 0.01.

The reason for this is that only rare diseases are being studied, these are not caused by common variants, hence these can be discarded. Both sets were annotated using Jannovar, (Jäger *et al.* 2014). With SIMdrom first 203 and later 320 exomes were sampled from the ExAC NFE population's homozygous and heterozygous allele counts (since this simulation option seemed to work better than using allele frequency, in goal one). Further, to be able to use the data sets in the training, known disease-causing mutations from patients were spiked-into the otherwise healthy exomes. The sets were then filtered down to only include genes with clinical relevance using the Online Mendelian Inheritance in Man (OMIM) database (McKusick-Nathans Institute of Genetic Medicine 2017). This resulted in an average of around 300 genes per exome in the data sets. The deleteriousness of the variants was assessed by the CADD score. Since the classification is performed per gene, the maximum CADD score was used when more than one variant was present in a gene.

3.2.2 Training, testing and cross-validation

The genetic data with the CADD scores were then combined with the feature vector of the phenotype and image analysis scores. The feature vector for each training set then consists of all the five scores from the different algorithms (Figure 5), ordered per gene.

A random forest classifier with a forest size of 100, was trained on the three different training sets, and tested on the set containing real patients' exome data. Since there were few real patient exomes available in the PEDIA study at the moment (first only 18, later only 19), a cross-validation was implemented in addition to the training and testing. A stratified k-fold cross-validation was performed on the three sets containing 203 exomes each. The data was divided into 76 partitions, since there are 76 unique pathogenic disease-causing variants in the total data set. Every partition was validated exactly once, as a test set. The performance of the classifier on the different data sets was visualized and evaluated using receiver operating characteristics (ROC) curves and precision-recall (PRC) curves.

4 Results

4.1 Goal 1

The results from smartpca were visualized in R, results from the analysis of the two sets, E:AF+AC+R (ExAC) and G:AF+GC+R (gnomAD), can be seen in Figure 8, Figure 9 and Figure 10 below. The first two principal components (PC1 and PC2) contain most of the variance in the data sets. Out of 138 eigenvalues in the ExAC analysis PC1 and PC2 represent around 4 and 2 percent of the variance (in their eigenvalues: 1 = 5.3 and 2 = 2.5). In the gnomAD analysis the two first PCs also represent 4 and 2 percent of the variance (in their eigenvalues 1 = 5.0 and 2 = 2.6, out of 118 eigenvalues in total). These PCs were therefore picked to visualize the data sets. Figure 8 displays the exomes simulated using the allele frequencies (AF). First, the five simulated populations by themselves (a) and then together with the real samples from the same populations (b).

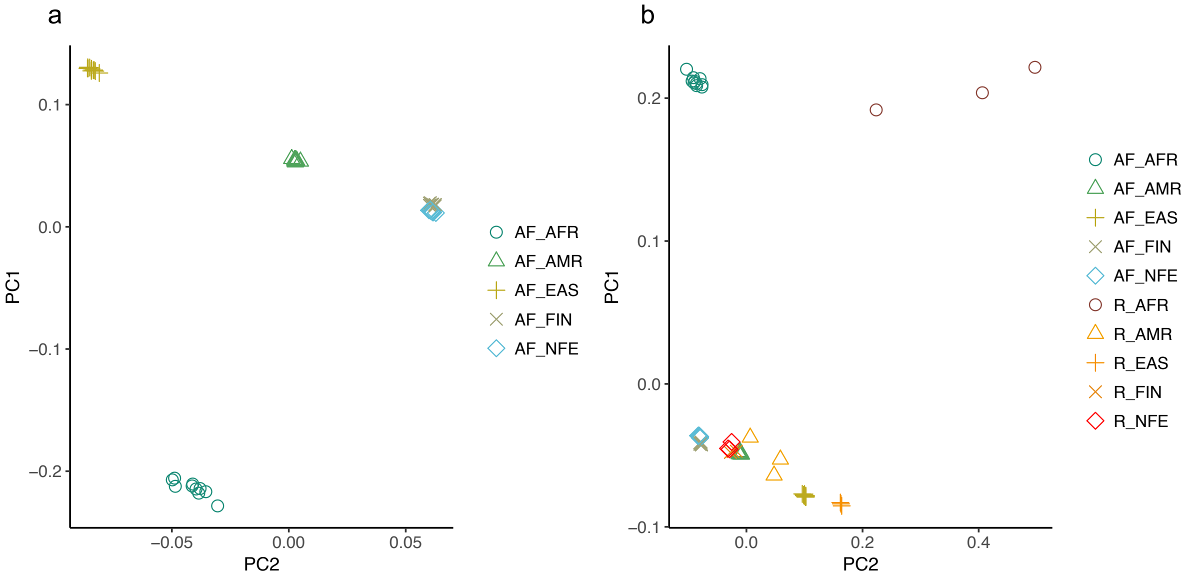


Figure 8. The results from the ExAC analysis (set E:AF+AC+R) visualized using the two first principal components (PC1 and PC2). The simulated exomes (using allele frequencies, AF) are first visualized by themselves (a) and then together with their real counterpart (b), for the five different populations.

In the results from the first goal, the exomes simulated using allele frequency from ExAC, form four well separated population clusters (Figure 8 a). In the first principal component (PC1) the variance separates the AFR population from the four other populations, with the EAS population furthest away. This is expected since the AFR populations typically contain more variants than other populations (The 1000 Genomes Project Consortium 2015). This is also the case with the exomes simulated using allele frequencies from ExAC where the AFR exomes contain ~35500 variants, which is around 5500 more than the other populations. The FIN and NFE creates one cluster, the cause for this might be their similar number of variants (~29700 and ~29400) and the fact that these two populations also are geographically closest to each other.

Another possible reason for the patterns is the out of Africa (OoA) model, where the modern-day populations are thought to have emerged from Africa and migrated across the world. In addition, the event is thought to have been accompanied by a population bottleneck which means the modern-day populations have evolved from one small gene-pool (Campbell & Tishkoff 2008). This means that the similarities seen between the non-African populations (AMR, EAS, NFE and FIN) could be explained by their probable genetic similarities, as opposed to the AFR population.

When the real samples are added to the figure (Figure 8 b) the pattern between the populations is preserved but minimized, and they cluster in the vicinity of their simulated counterpart. This points toward similarity between the simulated and the real genetic data. The only exception is the AFR samples. An explanation for this could be that they are from three different sub-populations within the AFR population (Appendix A 1). Hence, one possible reason is that there are differences in the number of variants also within the AFR population. This is also displayed by the real AMR samples, although not in the same ratio (The 1000 Genomes Project Consortium 2015). Another reason is that not only the number of variants differ between AFR sub-populations but that they also display genetic diversity due to population structure (Campbell & Tishkoff 2008).

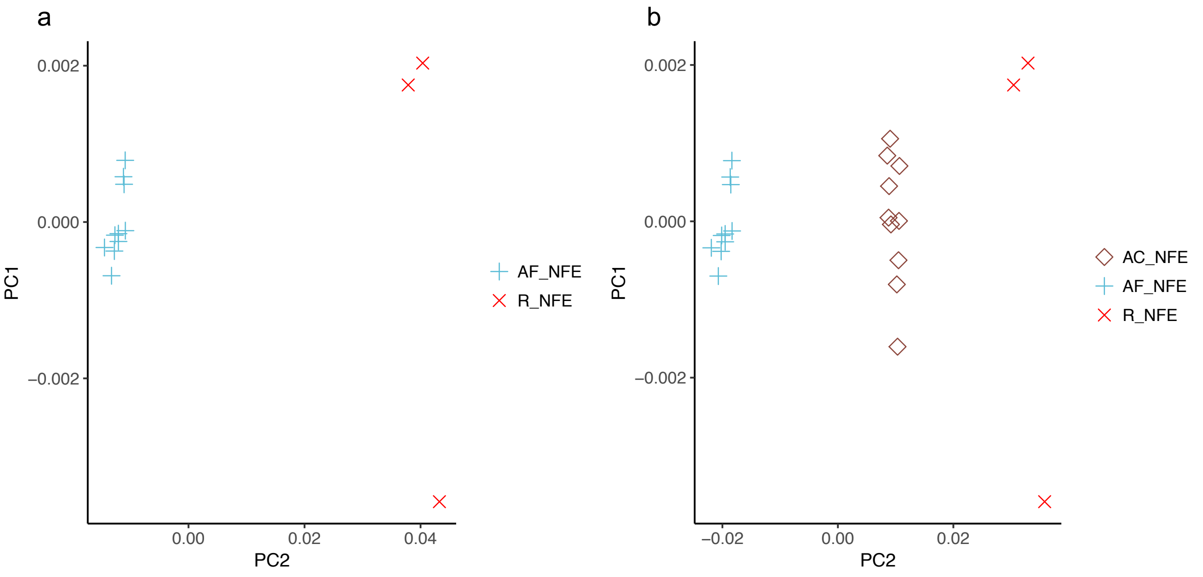


Figure 9. Results from the same analysis as in Figure 8. The two sets of NFE samples, simulated and real are displayed in a. In the b, NFE exomes simulated using homogenous and heterogeneous allele counts (AC) were added to the plot.

To be able to get a closer look at how one population varies between real and sampled sets, the two NFE populations were visualized on their own. In Figure 9 a, they can be seen as two distinct clusters. Exomes simulated using the homozygous and heterozygous allele counts (AC) from ExAC were added to the plot to be able to see differences between the two simulation options with respect to the real samples. As can be seen in Figure 9 b, they form a cluster closer to the real samples than the exomes simulated using allele frequency.

It was expected that the simulation using the allele count might perform better, since using the actual counts of homozygous, heterozygous and hemizygous alleles might come closer to resembling the reality, than predicting them using the observed allele frequencies. This is because of the assumption that the population under investigation is in Hardy-Weinberg equilibrium. The assumptions state that the population must be large, of constant size, under no selective pressure, the mating must be random and that there are no new mutations. Deviations from Hardy-Weinberg could suggest that some of these assumptions are false (Crow 1988). One assumption that could be false is random-mating, which means that the populations could have internal population structure due to e. g. geographical separation, which in-turn could alter the genotype frequencies (Crow 1988) This can be seen in the paper by Novembre et al. 2008, where populations within Europe form clusters close to their geographical origin.

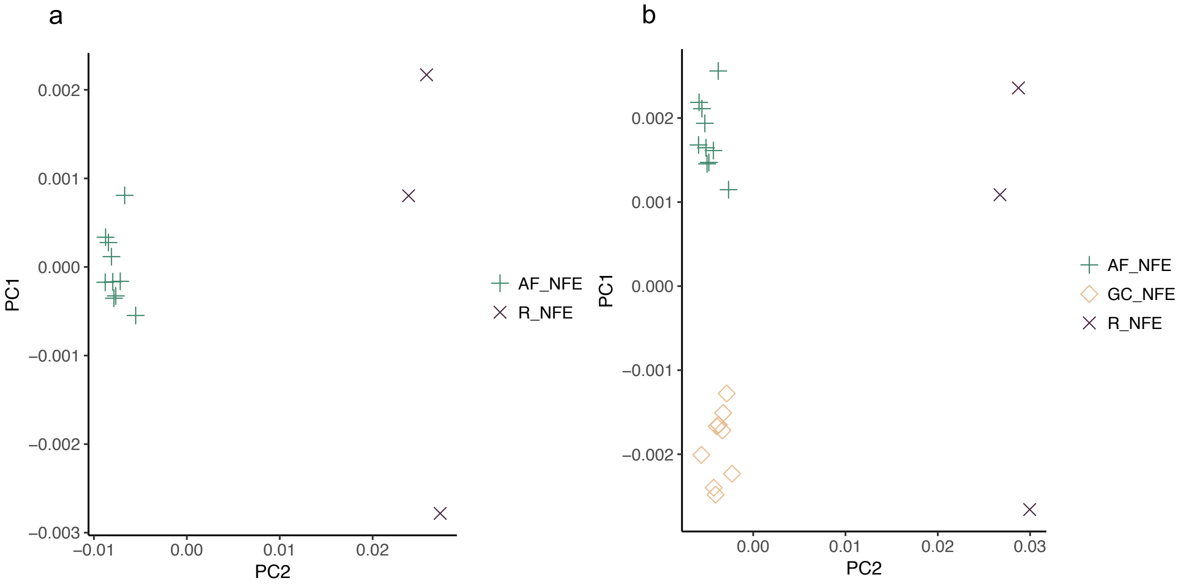


Figure 10. The results from the gnomAD analysis visualized using the first two principal components (PC1 and PC2). In the left figure (a), the exomes simulated using AF are displayed together with the real genomes. In b, exomes simulated using genotype counts (GC) are added to the plot.

The NFE populations simulated using gnomAD allele frequencies (AF) were also visualized together with the real NFE samples (Figure 10 a). In Figure 10 b, the exomes simulated using the genotype counts (GC) were added to the plot for the same reasons as with the ExAC analysis above. The results were similar (Figure 10 a), however the exomes simulated using the genotype count from gnomAD did not perform much better than the ones simulated using allele frequency (Figure 10 b).

4.2 Goal 2

As in goal one, the results were visualized in R. The results from the cross-validation of the three different datasets, containing 203 exomes each, can be seen in the two figures below (Figure 11 a and b). The first figure (a) describes the receiver operating characteristics (ROC). This is a measure of how well the classifier discriminates between the two classes. The true positive rate (y axis) is the number of correctly called positives (true positives), in this case the affected genes, against the actual number of affected genes in the data set (true positives + false negatives). The false positive rate (x axis) on the other hand is the number of wrongly classified unaffected genes (false positives) against all the unaffected genes (true negatives + false positives) (Flach 2011). The ExAC data set has a slightly better area under the ROC curve (AUROC, 0.947) than the two other data sets 1KG and IRN (Figure 11 a). A value of 1 is the highest, which means perfect discrimination between classes and 0.5 means that the classifier is guessing randomly (Flach 2011). Hence, the classifier trained and tested on the ExAC data set, performs slightly better in discriminating between affected and unaffected genes.

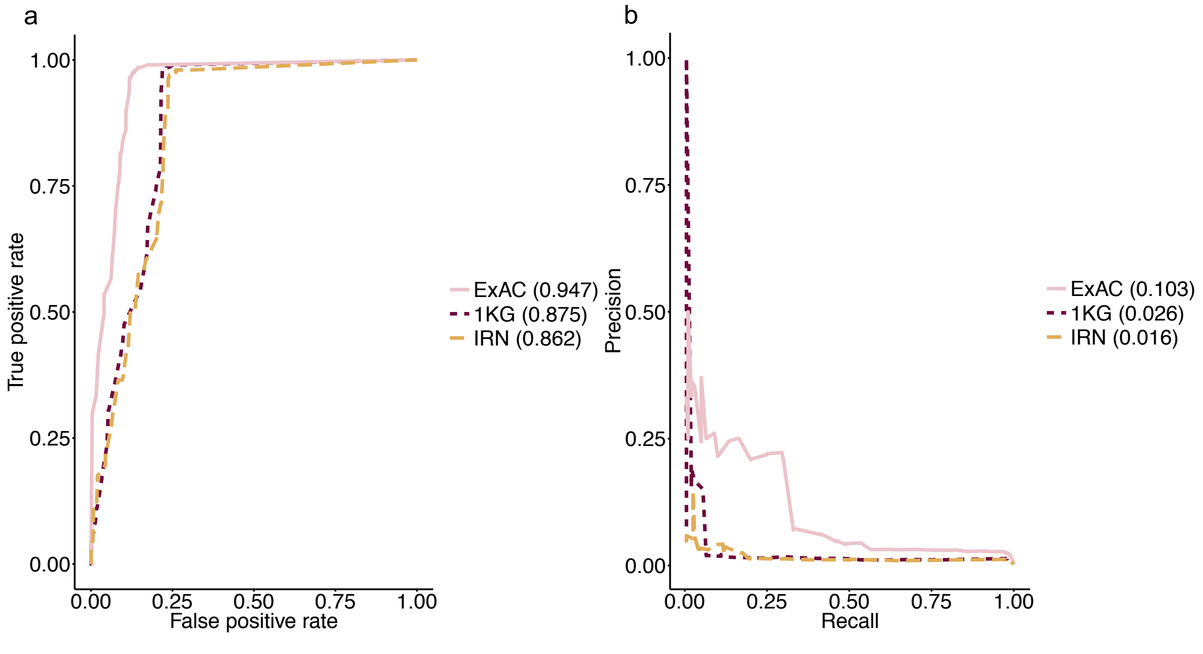


Figure 11. The two plots, a and b, displays two different performance evaluations of the PEDIA classifier when performing a cross-validation on the three different data sets (ExAC, 1KG and IRN). The left figure (a) displays their receiving operator characteristics (ROC) curves and the right figure (b) displays their precision-recall (PRC) curves. In brackets are the area under the curve for the three data sets.

The second figure (Figure 11 b) displays precision versus recall, a visualization of how precise the classifier is. Precision is calculated as the number of correctly classified affected genes against all the genes that are classified as affected (true positive + false positive). Recall measures how many of all the affected genes the classifier recognizes, the fraction of true positives against true positive and false negative. The precision-recall curve is a good evaluation when there is an imbalance between the classes in the data set (Tax *et al.* 2009). This is the case in the PEDIA study where the positive class (affected genes) is much smaller than the negative (unaffected genes).

In Figure 11 b, ExAC has a clearly better AUPRC (0.103) compared to the other two sets (1KG=0.026 and IRN=0.016). A better area here means that the classifier has a better balance between picking only affected genes and picking all the affected genes (Tax *et al.* 2009). The classifiers trained and tested on the IRN and 1KG data sets get really low AUPRC in this evaluation. In the case of the IRN, this was partly expected since the Iranian samples are real exomes. Therefore, more noise can exist in the data because of false calls. The 1KG samples on the other hand, are from a well annotated database and was thought to perform well. One reason for the 1KG to be unbalanced is that the 203 samples are randomly picked, which means that they most likely contain several different populations. This could be a difficulty in the cross-validation if the classifier is constantly trained and tested on different populations, making it difficult to find the affected genes due to population structure (Novembre *et al.* 2008). The ExAC exomes are simulated from only one population, the NFE, which could make it easier for the classifier. What clinicians are interested in is which classifier correctly classifies the affected gene in the most cases, which means ranking the correct gene at first place. Here, only the 1KG has a peak towards the highest precision meaning it correctly ranked an affected gene at the first position (Figure 11 b).

The same visualizations were performed for the results from the training on the simulated data using 203 exomes and testing on real data, and can be seen in Appendix B 2 and Appendix B 1. The results from the training and testing of the classifier on the three different sets (ExAC, 1KG and IRN) all display perfect curves. This is probably due to the small testing set (18 patients) of real exomes. Since the test set is small (~8% of the total data set, 18+203), the results may be due to chance. The training set could contain the same affected genes or causative variants as the test set, which means the classifier has seen the data before. Hence, it is biased towards classifying those genes correctly. This could be adjusted through removing such data from the training set, to make sure that the test set consists of unseen data. Since the curves are identical, it is not possible to say if one of the training sets is a better fit to the real test data.

Later in the project, more data was available making it possible to perform training on 320 exomes spiked with patients' causative mutations and training on 19 patient exomes. Also, the training set was controlled not to contain any of the genes in the test set. These results were visualized in the same way and can be seen in the figures. Here, it is possible to distinguish between the classifier trained on the three different sets. In the ROC evaluation, the IRN dataset perform better than the two other data sets, and has a slightly better AUROC (0.835). In the PRC evaluation, the ExAC correctly predicts the affected gene at the first rank in around 25 percent of the cases. The IRN performs even better in predicting the correct affected gene in around 40 percent of the cases, it also has the best AUPRC (0.519).

A reason why the IRN is performing better than both the other data sets might be that, as previously stated, the real Iranian exomes probably contains more noise. This might also be true for the real patient exomes, since when sequencing exomes for clinical purpose it is more important to keep possible disease-causing variants than to filter out false calls.

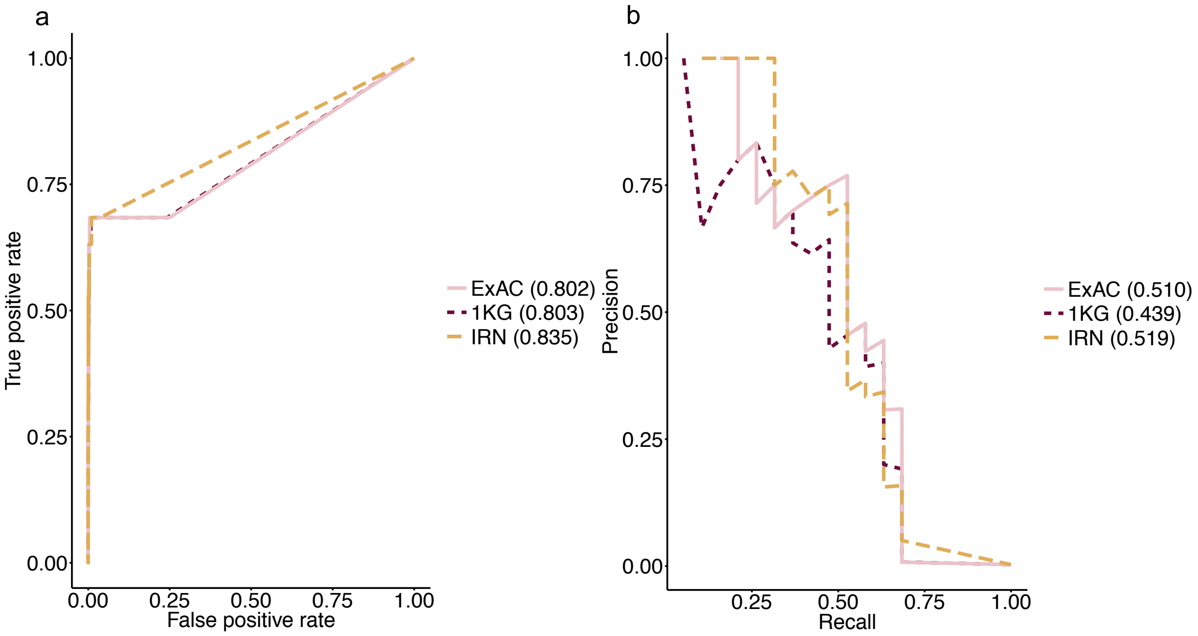


Figure 12. The two plots, a and b, displays two different performance evaluations of the PEDIA classifier when using the three different data sets (ExAC, 1KG and IRN) containing 320 exomes each as training sets and 19 real patient exomes as the test set. The left figure (a) displays their ROC curves and the right figure (b) their PRC curves. In brackets are the area under the curve for the three data sets.

5 Conclusions

The purpose of this project was to assess the reliability of SIMdrom simulated genetic data, to find out how similar it is to real genetic data. The aim was to make sure that the simulated genetic data is valid to use in other projects, such as the PEDIA study and for benchmarking molecular prioritization tools.

To further analyse the simulated exomes using smartpca, more real samples should be added since they seem to contain much internal variance. The other principal components (e. g. PC3 and PC4) could be investigated more closely. To have more confidence in the results from the second goal, the cross-validation could be repeated multiple times until the results converge. From the results in goal two, the IRN data set seems to be the best fit to the real patient data set. The reason for this could be that even though it seemed to be valuable to train and test on the same population (as with the ExAC NFE population in the cross-validation), training and testing on noise could be even more important. From the results in this project the best training set for the PEDIA classifier seems to be a data set containing noise, from the same population as the test set (the patient exomes). Still the test set is small (~8% of total data set), so to be sure that the results are not random, the classifier should be tested on a larger test set.

The SIMdrom simulated exomes performs well in both parts of this project. Even though the results from the first goal indicates that they are not perfectly similar to the real 1KG genomes, the results in the second goal indicates that they could be similar enough. More options could be developed to also be able to spike-in noise into the simulated exomes, which might make it a better fit for the PEDIA study. Future tests of their validity should include to test them against more real patient exomes in the PEDIA study. Before they are tested against a large enough set of real patient exomes, it is not possible to say if they are a good fit as the simulated genetic part in training sets for the PEDIA study.

6 Acknowledgements

I want to start by expressing my sincere gratitude towards my master thesis supervisor Max Schubach who was always there for me with support and inspiration, and who gave me the opportunity to make this project my own. I would also like to thank the universe (and Marie herself) for steering Marie Coutelier into the neighbouring office at the Institute of Medical Genetics and Human Genetics at Charité in Berlin, who encouraged me to keep on going even when the problem seemed too difficult (but also to take a break to think and discuss over a beer). I would also like to thank everybody else in the computational biology group, for nice coffee moments and for trying to teach me German. Finally, I would like to take the opportunity to thank Torsten Günther for being my subject reader.

References

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics 12: 745–755.

Bauer S, Köhler S, Schulz MH, Robinson PN. 2012. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics 28: 2502–2508.

Campbell MC, Tishkoff SA. 2008. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annual review of genomics and human genetics 9: 403–433.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4: 7.

Cortes C, Vapnik V. 1995. Support-vector networks. Machine Learning 20: 273–297.

Crow JF. 1988. Eighty Years Ago: The Beginnings of Population Genetics. Genetics 119: 473–476.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

FDNA INC. 2017. Face2Gene: Phenotyping apps that facilitate comprehensive and precise genetic evaluations. FDNA, Boston, USA.URL: https://suite.face2gene.com/

Flach PA. 2011. ROC Analysis. 869–875.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. European Journal of Human Genetics 20: 490–497.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17: 333–351.

Holmquist GP, Wienberg J. 2001. Human Chromosome Evolution. eLS, doi 10.1002/9780470015902.a0001447.pub2.

Jäger M, Wang K, Bauer S, Smedley D, Krawitz P, Robinson PN. 2014. Jannovar: a java library for exome annotation. Human mutation 35: 548–55.

Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics 46: 310–315.

Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. 2009. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. The American Journal of Human Genetics 85: 457–464.

Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJS, DeMare LE, Devereau AD, de Vries BBA, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MWM, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JOB, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. 2017. The Human Phenotype Ontology in 2017. Nucleic acids research 45: D865–D876.

Köster J, Rahmann S. 2012. Snakemake - a scalable bioinformatics workflow engine. Bioinformatics 28: 2520–2522.

Lam WW, Mueller RF. 2001. Human Genetics: Principles. eLS, doi 10.1038/npg.els.0001873.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285–291.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. The Journal of Clinical Investigation 118: 1590–1605.

McKusick-Nathans Institute of Genetic Medicine. 2017. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, USA.URL: https://omim.org/

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443–451.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456: 98–101.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44: D733-745.

Patterson N, Price AL, Reich D. 2006. Population Structure and Eigenanalysis. PLoS Genetics, doi 10.1371/journal.pgen.0020190.

Purcell SM, Chang CC. 2017. PLINK 1.9. URL: http://www.cog-genomics.org/plink/1.9/.

Scriver CR. 2001. Allelic and Locus Heterogeneity. eLS, doi 10.1038/npg.els.0005481.

Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E, Zemojtel T, Buske OJ, Washington NL, Bone WP, Haendel MA, Robinson PN. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature protocols 10: 2004–2015.

Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, Jäger M, Hochheiser H, Washington NL, McMurry JA, Haendel MA, Mungall CJ, Lewis SE, Groza T, Valentini G, Robinson PN. 2016. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. The American Journal of Human Genetics 99: 595–606**Improved exome prioritization of disease.

Tax DMJ, Loog M, Duin RPW. 2009. Optimal Mean-Precision Classifier. SpringerLink, pp. 72–81. Springer, Berlin, Heidelberg,

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526: 68–74.

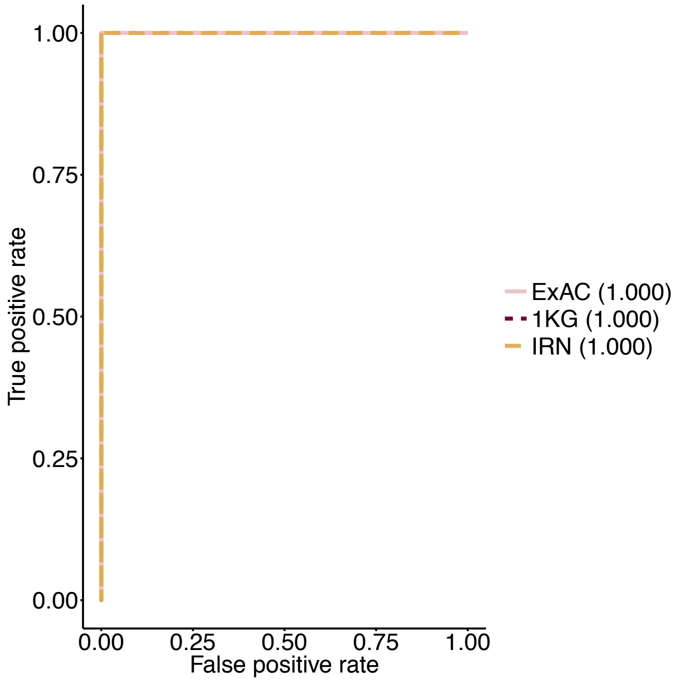
White D, Rabago-Smith M. 2011. Genotype-phenotype associations and human eye color. Journal of Human Genetics 56: 5–7.

Appendix A

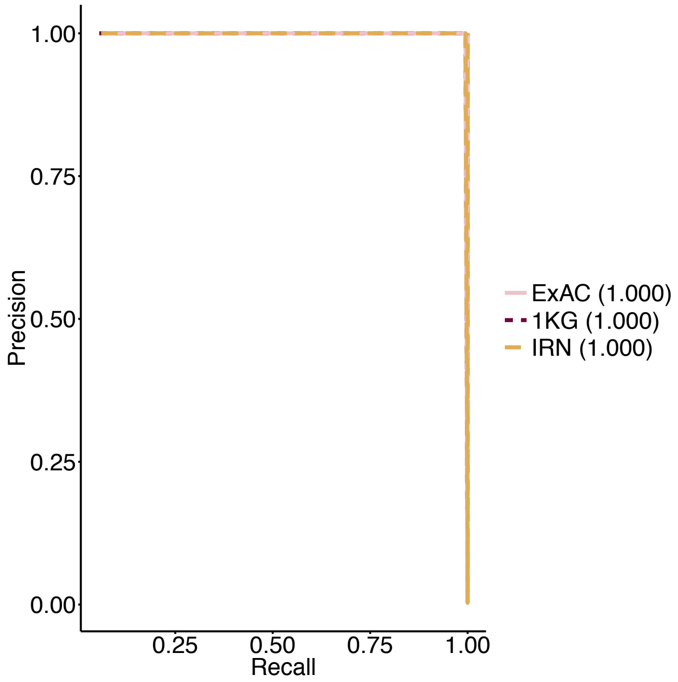
Appendix A 1. The table describes the populations used from ExAC, gnomAD and the real samples picked from 1KG.

ExAC	**gnomAD**	**1KG, real samples**
African/African American (AFR)	AFR	NA1920 (Kenya) NA19700 (African American, USA) NA18510 (Nigeria)
Latin American (AMR)	AMR	HG01550 (Colombia) HG01970 (Peru) NA19720 (Mexican, USA)
East Asian (EAS)	EAS	NA18620 (China) NA18940 (Japan) HG01600 (Vietnam)
Finnish (FIN)	FIN	HG00310 (Finland) HG00190 (Finland) HG00280 (Finland)
Non-Finnish European (NFE)	NFE	HG00140 (England and Scotland, UK) NA20510 (Italy) NA07000 (European, USA)

Appendix B



Appendix B 2. The results from the ROC evaluation of the training (using the three data sets ExAC, 1KG and IRN) and testing (using real patient data) of the PEDIA classifier.



Appendix B 1. The results from the PRC evaluation of the training and testing of the PEDIA classifier, for the same data sets as in Appendix B 2.