Sequence Analysis

VINYL: Variant prIoritizatioN bY survivaL analysis

Matteo Chiara[1,2*], Pietro Mandreoli[1], Marco Antonio Tangaro[2], Anna Maria D'Erchia[2,3], Sandro Sorrentino[4], Cinzia Forleo[4], David S. Horner[1,2], Federico Zambelli[1,2] and Graziano Pesole[2,3]

[1]Department of Biosciences, University of Milan, Milan, Italy
[2]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy
[3]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari "Aldo Moro", Bari, Italy
[4]Cardiology Unit, Department of Emergency and Organ Transplantation, University of Bari "Aldo Moro", Bari, Italy
*To whom correspondence should be addressed.
Associate Editor: XXXXXXX
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract:
Motivation: Clinical applications of genome re-sequencing technologies typically generate large amounts of data that need to be carefully annotated and interpreted to identify genetic variants potentially associated with pathological conditions. In this context, accurate and reproducible methods for the functional annotation and prioritization of genetic variants are of fundamental importance. **Results:** In this paper, we present VINYL, a flexible and fully automated system for the functional annotation and prioritization of genetic variants. Extensive analyses of both real and simulated datasets suggest that VINYL can identify clinically relevant genetic variants in a more accurate manner compared to equivalent state of the art methods, allowing a more rapid and effective prioritization of genetic variants in different experimental settings. As such we believe that VINYL can establish itself as a valuable tool to assist healthcare operators and researchers in clinical genomics investigations.
Availability: VINYL is available at http://beaconlab.it/VINYL and https://github.com/matteo14c/VINYL.
Contact: matteo.chiara@unimi.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Applications of modern high throughput genome sequencing technologies to healthcare and clinical practice are driving a major breakthrough in medical science (Saudi Mendeliome Group 2015, UK10K Consortium 2015, Kowalski et al 2019). The unprecedented ability to interrogate the (more than) 3 billion pairs of nucleotides that compose our genome in a systematic and reliable manner, provides a formidable tool for the characterization and functional annotation of the human variome, the complete set of genetic variants in the human population (Gurdasani et al 2015, Kowalski et al 2019, Nagasaki et al 2015). The capacity to link genetic variants with phenotypic traits, pathological conditions, and/or positive or adverse reactions to therapies and medications is of instrumental importance for the development of informed approaches to medical science, such as precision medicine (Lu et al, 2014), that is the ability to treat patients based on their genetic background, or predictive medicine (Kotze et al, 2015) where risk factors for various diseases can be accounted beforehand and suitable measures instituted to prevent a future condition or mitigate its severity. Accordingly, numerous countries and institutions worldwide are already undertaking or are planning to launch large-scale projects aiming to sequence an increasing proportion of their population. These include, among the others, the UK10K project in the United Kingdom (UK10K Consortium et al, 2015), the All of Us research program by the NIH (All of Us Research program investigators, 2019), the French Plan for Genomic medicine funded by the French Ministry of Health (Lethimonnier and Levy, 2018), and the European '1+ Million Genomes' initiative promoted by the European Community (Saunders et al 2019).

While this push to sequence an unprecedented number of human genomes is driving a new revolution in medical science, the need to handle, analyze and interpret large collections of "big" genomic data is posing major challenges which at present remain unresolved (Alyass et al 2015, Klein et al 2017, Horowitz et al, 2019, Stark et al 2019). The limitations are both technical, due to the need to develop dedicated infrastructures for the handling, sharing and processing of sensitive human data (Saunders et al, 2019); and methodological, due to the need to integrate multiple bioinformatics tools into complex analytical workflows, which require a substantial effort for their set up and optimization (Canzoneri et al 2019, Ginsburg and Phillips 2018, Servant et al 2014).

A typical Next Generation Sequencing assay can detect in the order of tens of thousands or even millions of genetic variants, all of which need to be carefully annotated and evaluated to identify genetic traits potentially associated with a pathological condition (Elbeck et al 2017). However, the large majority of these variants are likely to represent standing genetic variation and are not relevant from a clinical perspective (Pickrell et al, 2014, Wilson et al, 2014). Variant prioritization is a simple procedure,

commonly used in clinical studies, to reduce the number of genetic variants that need to be evaluated manually. Briefly, a series of filters and criteria are established based on the predicted functional effects of the variants, their overall prevalence in the human population and other relevant considerations, in order to retain only variants of potential clinical relevance (Eilbeck et al, 2017). Subsequently, these variants are subjected to careful manual evaluation by expert clinicians to identify candidate causative variants potentially involved in the molecular pathogenesis of the disease. Although conceptually simple, "variant prioritization", represents a delicate and fundamental step in the application of genomics in clinical settings (Frebourg et al 2014, Jalali et al 2017). Excessively stringent criteria might result in the exclusion of interesting candidate variants, conversely lenient criteria, can significantly impact turn-around times and subsequent analyses. Additionally, although rigorous expert designed guidelines for the interpretation and analysis of genetic variants in clinical settings are currently available, it is not uncommon for different operators to apply slightly different criteria and filters when performing variant prioritization, thus limiting the overall reproducibility of the results of this type of analysis (Pabinger et al 2014). For example, several studies apply approaches based on the calculation of composite scores that integrate different types of resources and information for the identification of variants of potential clinical relevance. However, the exact procedure used in the calculation of these scores, and the relative importance of their different components are not set in a clear, systematic and reproducible manner (Eilbeck et al, 2017). Preventing the re-application of these methods and/or their adaptation to a different case of study.

In this paper, we present VINYL, a novel system for the prioritization of genetic variants. As suggested by guidelines and recommendations derived from clinical practice (Richards et al, 2015), VINYL computes a variant prioritization score that aggregates different sources of evidence and annotations obtained from publicly available resources. Several studies (Cirulli et al 2015, Lee et al 2014, Moutsianas et al 2015, Guo et al 2016, Li et al 2008) have reported that cohorts of affected individuals harbor an excess of deleterious or slightly deleterious variants at disease-associated loci with respect to unaffected controls. VINYL applies this logic to identify high scoring variants that are more likely to be associated with a pathological condition. Different scoring systems are compared by evaluating genetic profiles of a population of affected individuals and a population of matched controls. The scoring system that maximizes the number of high scoring, potentially pathogenic, variants in affected individuals and that results in a reduced number of high scoring variants in the population of matched controls is selected. Finally an automatic procedure based on "survival analysis", an analytical procedure that evaluates the enrichment in high scoring variants associated with scores above a certain cut-off, is applied to derive an optimal score threshold for the identification of variants of potential clinical relevance. VINYL is completely flexible and allows the design of customized scoring systems based on different levels of functional annotations, that can be adapted/optimized to different use cases and scenarios.

Extensive simulations based on publicly available data show that VINYL is capable of identifying clinically relevant genetic variants in a more efficient manner with respect to equivalent state-of-the-art methods. Similarly, by applying our method to a cohort of 38 patients with a diagnosis of cardiomyopathy (Forleo et al,2017) and to a large collection of 200 exome trios, from a cohort of patients affected by developmental disorders (Deciphering Developmental Disorders Study, 2015), we show that our tool is capable of identifying the large majority of the variants that were previously classified as Pathogenic/Likely Pathogenic by careful expert manual curation on the same datasets, while prioritizing only a limited number of variants in populations of unaffected individuals.

We believe that by providing a rapid, systematic and reproducible approach for the prioritization of genetic variants, VINYL can represent a practical tool to assist clinicians in variant prioritization in large scale clinical studies. The tool is currently available at: http://beaconlab.it/VINYL. To facilitate its usage and to improve the reproducibility of the analyses VINYL is incorporated into a dedicated instance of the popular Galaxy workflow manager (Afgan et al, 2018), along with a highly curated collection of tools and resources for the annotation of genetic variants.

2 Methods

2.1 Implementation of VINYL

VINYL is implemented as a Laniakea (Tangaro et al, 2018) Galaxy (Afgan et al, 2018) instance based on Galaxy release 18.05. Annotation of VCF files is performed by the Annovar software (Wang et al, 2010), using a collection of "standard" resources maintained by the Annovar developers along with a selection of custom annotation tracks. These include the OregAnno database (Griffith et al 2008), the Ensembl regulatory build annotation (Zerbino et al 2016), the NHGRI-EBI GWAS catalog (Buniello et al, 2019) and the ncER score, which provide fine-grained annotations of non-coding and regulatory genomic elements (Wells et al 2019). A complete list of the annotations that are currently supported by VINYL along with a brief description is reported in Supplementary Table S1.

VINYL is implemented as a collection of Perl and R scripts and is composed of 3 main modules:

- *the optimizer*, which derives the best scoring system
- *the threshold optimizer*, that calculates the ideal score threshold for the prioritization of variants
- and the *score calculator*, the main tool which computes the scores by integrating different types of annotations.

VINYL is currently available from http://beaconlab.it/VINYL. The source code is available at https://github.com/matteo14c/VINYL. A detailed manual to the usage of VINYL is provided at http://beaconlab.it/VINYL/manual.

2.1.1 Computation of the pathogenicity score

VINYL computes its score directly from annotated VCF files. Annotations that should be considered for the computation of the score can be specified by a plain text configuration file. This can include both annotations that are provided by Annovar, or equivalent tools for the annotation of genetic variants, but also custom annotations provided by the user (See Supplementary Materials and Methods). Currently VINYL can discriminate between 12 different types of functional annotations, including -among the others- databases of human genetic variation (RV), the predicted functional effects of the variants (FE) and/or their presence/absence in databases of clinically relevant genetic variants (DB). A complete list is reported in Supplementary Table S2 (and in the online manual). The score itself is computed as a linear aggregation of the different types of functional annotations by the following formula:

$$\textit{Pat Score} = w^{db}\textit{DB} + w^{rv}\textit{RV} + w^{fe}\textit{FE} + w^{ns}\textit{NS} + w^{or}\textit{OR} + w^{eq}\textit{eQ} + w^{ad}\textit{AD} + w^{mi}\textit{mi} + w^{Reg}\textit{Reg} + w^{Tf}\textit{TF} + w^{GW}\textit{GW} + w^{Sp}\textit{Sp}.$$

Where w^{db}, w^{rv}, w^{fe}, w^{ns}, w^{or} , w^{eq}, w^{ad}, w^{mi}, w^{Reg}, w^{Tf}, w^{Gw} and w^{Sp}, represent the weights (the relative importance) of each component of the score. Single components of the score are computed according to the following rules:

- **DBs of pathogenic variants (DB)**: the score is incremented for variants reported to be Pathogenic or Likely Pathogenic in publicly available resources of clinically relevant variants. The score is decreased for variants that are reported as "Benign" or "Likely Benign". Users can provide a description of the disease and its symptoms using a simple configuration file. Only entries that match these keywords are considered for the computation of the score. In the current implementation of VINYL the Clinvar (Landrum et al, 2014) database is used as the main source for the annotation of disease-associated genetic variants
- **Rare Variants (RV):** the score is increased if a genetic variant shows a Minor Allele Frequency (MAF) lower than a user-defined cutoff -typically the prevalence of the disease- in public databases of human genetic variation
- **Functional effect of the variant (FE)**: the score is increased if the variant is predicted to have a deleterious functional effect (i.e. splicing variants, stop-gain, frameshift variants).
- **Disruptive non-synonymous (NS):** the score is incremented for NS variants that are predicted to have a disruptive effect. Tools to be considered for the evaluation of the effect of NS variants can be specified at runtime. Predictions are derived from the dbNFSP database (Liu et al, 2016) version 3.5a.
- **Overrepresentation (OR):** if a genetic variant with MAF ≤0.01 (the frequency cut-off that is normally considered for the definition of "common" SNPs) is found in N or more affected individuals the score is incremented. The value of N is

specified at runtime. Default value is set to 10% of the cohort size (which corresponds to a odds ratio of 10 or more).

- **eQTLs (eQ):** the score is incremented by this value when a variant is associated with an eQTL according to the GTEx study (GTEx Consortium, 2013). A list of relevant tissues for the annotation of eQTLs can be provided by users in the form of a simple text file.
- **Disease-associated genes (AD):** the score is incremented if a genetic variant is associated with genes previously implicated in the disease or in similar pathological conditions. Users can provide a list of disease-related genes by a simple text file
- **miRNA binding site (mi):** the score is increased if the variant is associated with a known miRNA binding site.
- **Regulatory element (Reg):** the score is incremented if the variant is part of a genomic regulatory element (promoter, enhancer, silencer), according to the OregAnno database (Griffith et al 2008) or the Ensembl regulatory build annotation (Zerbino et al 2016)
- **TF binding site (TF):** the score is increased if the variant is associated with a transcription factor binding site, according to the OregAnno database (Griffith et al 2008) or the Ensembl regulatory build annotation (Zerbino et al 2016)
- **GWAS (GW):** the score is incremented if the variant is associated with a phenotypic trait relevant to the pathological condition according to one or more GWAS studies. Similar to the DB score, only entries matching a user-specified list of keywords are considered for the computation of this score
- **Splicing variants (Sp):** the score is incremented if the variant is reported to have a deleterious effect on a splice site according to the dbscsnv11 (Jian et al, 2014) database.

Users can configure the behavior of VINYL by additional parameters and configuration files that specify a disease model (Autosomal Dominant, Autosomal Recessive or X-linked), a list of symptoms associated with the disease, define a set of genes implicated with the pathological condition of interest and/or list tissues to be considered for the evaluation of expression quantitative trait loci (eQTLs). When a disease model is specified, scores associated with genotypes that are not compatible with the model of inheritance of the disease get a linear penalization of 40%. See Supplementary Figure 1 a detailed description of conceptual workflow implemented by VINYL.

2.1.2 Optimization of the pathogenicity score

Genetic algorithms, as implemented in the genalg (Willighagen and Ballings, 2015) R library, are used to identify optimal weights for the components of the pathogenicity score, by performing a search on the parameter space. As illustrated in Supplementary Figure S1, score distributions are computed for a population of affected individuals (A) and a population of healthy controls (C). The optimal scoring system and the corresponding threshold for the identification of potentially pathogenic variants are established by comparing scores distributions by means of an iterative survival analysis based on the Wang Allison method (Wang et al, 2004). Cut-off values spanning from the maximum score to the minimum score, with an interval of 0.5, are evaluated for every scoring system, and the number of potentially clinically relevant variants as identified in the A and C populations are recorded. A Fisher's exact test is subsequently used to evaluate the over-representation of clinically relevant variants in A with respect to C.

The scoring system (and the corresponding threshold value) that maximizes the difference between the number of potentially clinically relevant variants in A and that-at the same time- minimizes the number of potentially pathogenic variants identified in the C is selected. The following equation is used to define the optimality criterion:

$$Optimal\ Score = argmax < 0.2 * -log10(Fpv) + 0.6 * log2(Ffc) - 0.2 * PC >$$

Fpv= p-value for the over-representation of likely pathogenic variants in A according to the Fisher's exact Test. Ffc= ratio between the proportion of likely pathogenic variants identified in A and C respectively. PC= number of potentially pathogenic variants identified in C. The coefficients of the equation have been derived empirically, to obtain a reasonable balance between the maximization of the number of potentially pathogenic variants identified in A, and the minimization of the, likely false positive, pathogenic variants identified in C.

2.1.3 Utilities for the post-processing of VINYL's output

All the utilities for the post-processing of VINYL's output files are implemented in the form of standalone R scripts. Principal Component Analysis is performed by means of the prcomp R function from the stats package (R Core Team 2018). Graphical representation of the results using the R ggplot2 package (Wickham 2016).

2.2 Simulated dataset

Disease-causing genetic variants were simulated by means of the Hapgen2 (Su et al 2011) program, using the haplotype files of the TSI (Toscani in Italia) population from the 1000G study.

The latest version of Hapgen2 was obtained from https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html, while haplotype files from the 1000G project were obtained from https://mathgen.stats.ox.ac.uk/impute/impute_v1.html#Using_IMPUTE_with_the_HapMap_Data.

Three different distributions of Odds Risk Ratios, with an average of 3, 10 and 20 respectively, were simulated using the rnorm function in R. Standard deviation was set to 10% of the average.

To simulate different sequencing strategies, cohorts of different size (25, 50 and 100 affected individuals and matched number of controls) and each including a variable number of polymorphic positions (1000, 5000 and 10000) were simulated, for a total of 9 (3x3) distinct datasets. Disease-associated variants have been simulated by randomly selecting a fixed number of 75 rare variants (Minor Allele Frequency $\leq$0.001), with equivalent proportion of variants associated with different predicted functional effects, including: splice site variants, variants in promoter regions, frameshift variants, variants in miRNA target regions, stop-gain/stop-loss variants. To test the ability of the variant prioritization methods applied in this study to identify both "novel" clinically relevant variants (i.e variants not already associated with a pathological condition), and variants already associated with a pathological condition at every iteration, a small number (between 5 and 10) of variants already implicated in a known pathological condition according to the ClinVar database were selected. Lists of genes associated with the "simulated" pathological condition and related phenotypic terms were derived accordingly.

2.3 Real datasets

Genetic profiles of the 38 unrelated cardiomyopathy patients included in Forleo et al (Forleo et al, 2017) and of 1133 family trios of individuals affected by developmental disorders, as recruited by the DDD (Deciphering Developmental Disorders Study, 2015) study were retrieved from the EGA database under the EGAS00001002506 and EGAD00001001413 accession respectively. Candidate likely pathogenic variants for the 38 cardiomyopathy patients were obtained directly from Table 1 of Forleo et al. (2017). For the DDD study pathogenic and/or likely pathogenic variants were obtained from the DECIPHER (Firth et al, 2009) database. In this case, only 200 subjects affected by autosomal dominant developmental disorders and for which one or more pathogenic or likely pathogenic variants were already reported in DECIPHER were considered in our analyses.

2.4 Execution of VINYL

The VINYL pipeline was executed with default parameters both for the analysis of real and simulated data. In the analysis of the Forleo et al (2017) dataset the TSI (Toscani in Italia) population from the 1000G (The 1000 Genomes Project Consortium, 2015) study was used as the "control" population. In the analysis of the 200 subjects from the DDD cohort, parents were used as the control population. Text files with the description of the symptoms associated with the pathological conditions under study were obtained simply by combining the most recurrent words (5 occurrences or more) as reported in the original study, i.e Table S1, as available from Forleo et al for the cardiomyopathy dataset, and the EGAF00000883031.txt file as available from the EGA database for the DDD cohort.

For the DDD dataset, a modified keyword configuration file was created to instruct VINYL to incorporate annotations of allele frequency (DDD_AF) and of *de-novo* mutations (TEAM29_FILTER) according to Denovogear (Ramu et al, 2013) as provided by the DDD. The cut-off frequency for rare alleles was consistently set to 10e-4.

2.5 Execution of Privar and KGGseq

The latest versions of KGGseq (Li et al 2012) and Privar (Zhang et al 2013) were obtained from http://grass.cgs.hku.hk/limx/kggseq/ and http://paed.hku.hk/genome/software.html,respectively. Privar was executed using the "Literature-based strategy" with default parameters. A custom list of disease-associated genes (identical to that used for VINYL) was provided by means of the "-customlist" parameter. KGGseq was applied using the strategy illustrated in the e manual for the prioritization of genetic variants associated with rare Mendelian diseases; "the "--candi-list" and "--phenotype-term" parameters were used to provide a list of

disease-associated genes and a list of symptoms of the disease under study, respectively. In the analysis of the DDD data, the "--genotype-filter" was applied to identify *de novo* mutations. Both lists were completely identical to the lists used to provide equivalent information to VINYL. Consistent with the parameters used in VINYL, the cut-off frequency for rare alleles was set to 10e-4.

3 Results

3.1 VINYL: an automated tool for variant prioritization

VINYL provides a fully automated system for variant prioritization, which -according to the guidelines used in clinical practice- is based on the integration of different types of functional annotations and resources. By leveraging the Galaxy workflow manager, VINYL is made available through a powerful and user-friendly web-based graphical interface and allows collaborative and highly reproducible analysis of large amounts of data. Encrypted data volumes are used to ensure data protection. Users can upload their data to VINYL in the form of plain VCF files. Variants annotation is performed by the Annovar software (Wang et al, 2010), which is available in VINYL along with an extensive collection of resources for the annotation of genetic variants (see Table S1). Additional information used for the computation of the pathogenicity score, including for example the symptoms and prevalence of the pathological condition under study, the model of inheritance of the disease, the type of predicted functional effects that should be considered deleterious, a list of tools used for the prediction of the functional effects of genetic variants, and custom annotation tracks (see below) can be specified by users at run-time using simple configuration files in plain text format (see Material and Methods, and Supplementary Materials).

The main output of VINYL consists of a tabular file, where variants are ranked according to their score. A score threshold for the prioritization of variants that are more likely to be clinically relevant is derived automatically. Additional utilities (see below) can be used to perform more fine-grained analyses for the identification of genes that display a significant over-representation of high scoring variants (burden analysis), the identification of variants with similar functional annotations, or for the stratification of patients in groups by dimensionality reduction techniques. Along with a carefully curated collection of tools and resources for the functional annotation of genetic variants, the VINYL Galaxy instance incorporates also a collection of reference data, including VCF files of 26 distinct human geographic populations from the 1000 Genomes study (The 1000 Genomes Project Consortium, 2015), which can provide a suitable background control population for most clinical studies. The features contained in VINYL and the rationale used in the implementation of the tool are briefly outlined in Figure 1.

3.2 Evaluation of VINYL on simulated data

To evaluate the ability of VINYL to identify genetic variants of clinical relevance, we performed extensive simulations of disease-associated variants derived from real human haplotypes. Different scenarios were simulated to evaluate the impact of cohorts of different size (25, 50 and 100 individuals), the strength of the association of variants with pathological conditions (odd risk ratios of 3, 10 and 20), different functional effects (both on protein coding genes and on functional regulatory elements), and the total number of variants included in the call-set (1000, 5000 and 10000): a proxy for the simulation of different sequencing strategies (from targeted resequencing of a limited number of genes to exome sequencing). Results obtained by VINYL on these datasets were compared with those attained by two other popular methods for the prioritization of genetic variants: *Privar* (Zhang et al, 2013) and *KGGseq* (Li et al, 2012). As outlined in Supplementary Table S3 and Figure 2A, VINYL demonstrates an improved sensitivity in the prioritization of genetic variants potentially associated with a disease compared to both KGGseq and Privar, resulting in a significantly increase in AUC (area under the curve) in all the simulations performed in this study. These results suggest that the approach adopted by our method can outperform currently available state of the art methods in the prioritization of disease-associated genetic variants. As expected, since these methods are not devised for the prioritization of genomic variants in non coding genomic elements, VINYL is more effective than than either of Privar or KGGseq in the identification of variants associated with this type of genomic elements (Supplementary Figure S2A). However, a marginal, but significant, improvement in the correct prioritization of variants associated with protein coding genes (Supplementary Figure S2A) is also observed. More importantly, (Supplementary Figure S2B) we observe that while recovering a larger proportion of the variants associated with a pathological condition in our simulations, the lists of variants prioritized by VINYL is in general more compact if compared to those derived by KGGseq and Privar. Suggesting that the approach adopted by VINYL can be highly effective in decreasing the number of clinically relevant variants that should be subjected to manual evaluation. As expected the performances of VINYL are strongly influenced by the composition and the size of the input dataset, as we observe an increase in sensitivity when large cohorts of patients are analyzed (Table S3). Moreover (see Supplementary Table S4 and Supplementary materials), since unlike other similar methods, VINYL requires a population of matched controls to optimize its scoring system, we observe that the choice/availability of the "correct" background population has major implications on the overall accuracy. In fact, (Supplementary Table 4, and Supplementary Materials) the use of a "mismatched" control population, can often result in the incorrect prioritization of a consistent number of population biased alleles. To mitigate this issue for studies where genetic profiles of a matched control population are not available, VCF files filtered from population biased alleles have been incorporated in the main Galaxy VINYL instance to serve as an alternative reference.

3.3 Evaluation of VINYL on real data

3.3.1 Cardiomyopathy dataset

VINYL was applied to a dataset composed of 38 Italian patients affected by different types of cardiomyopathies, which were previously subjected to genotyping by targeted resequencing of a panel of 115 genes. As described in Forleo et al (Forleo et al, 2017) expert manual curation identified a total of 27 likely pathogenic variants in 26 out of 38 patients. VINYL prioritized a total of 50 variants (4.02%) on this dataset, notably, all the 27 variants selected by manual curation were recovered (Figure 2B). Only 1 out of 3739 genetic variants in the control population was prioritized by VINYL (Figure 2B). Notably although the number of variants prioritized Privar and KGGseq on the same dataset was consistently higher: 84 and 83 (compared to 50) respectively neither Privar nor KGGseq were able to recover the complete collection of the 27 variants identified by manual curation (19 and 21 for Privar and KGGseq, respectively). Conversely, if compared to VINYL, both KGGseq and Privar identified an increased number of potentially clinically relevant variants: 21 and 22 respectively, in the population of unaffected controls (Figure 2B). Taken together these observations suggest that- at least on this dataset- VINYL can provide a more accurate prioritization of clinically relevant variants compared with equivalent state-of-the-art methods. Interestingly, similar to our observations on simulated data, genetic variants prioritized by VINYL, but not identified as potentially relevant by KGGseq and Privar, are substantially enriched for functional annotations related to regulatory elements, and/or non protein coding genomic elements (Supplementary figure S3A). For example, we observe that 10 (43%) out the 23 variants prioritized by VINYL, but not included in the list of likely pathogenic variants identified by Forleo et al, are associated with mutations in regulatory sequences and or/miRNA target regions. Since annotation of non coding genetic elements was not included in the criteria for the prioritization of genetic variants in the original work, it is unsurprising that those variants were not considered by Forleo et al. Further interpretation of the clinical relevance of these variants, would require careful examination by expert teams of clinicians, and lies outside the scope of the current work. On the other hand, it is interesting to notice that variants prioritized by Privar and KGGseq but not by VINYL are highly enriched in rare, missense variants. Importantly, (Supplementary Figure S3B) we observe that only a limited ($\leq$2 out of 18) number of the methods for the evaluation of the effects of missense variants included in the dbFNSP database predict a potentially deleterious effect for these variants. Additionally, (Supplementary Figure S3C) highly variable

estimates of Minor Allele Frequency (MAF) are observed when different databases of human genetic variation (ExAC, gnomad, 1000G) are considered. With the majority of the variants prioritized exclusively by KGGseq and Privar showing a difference of more than 5 fold between the smallest and largest value of MAF as reported in public databases.

3.3.2 Developmental Disorders dataset

The DDD (Deciphering Developmental Disorders) study (Deciphering Developmental Disorders Study, 2015) is one of the most complete and highly curated resource of genetic profiles of individuals affected by a pathological condition, matched with phenotypic data and annotations of pathogenic and likely pathogenic variants, as reviewed by expert teams of clinicians. Along with the genotypes of 13,462 individuals affected by developmental disorders, DDD also incorporates genome and exome sequences of their parents (Trio sequencing), and fine grained annotations of phenotypic traits of all the subjects included in the study according to the HPO ontology (Robinson et al, 2008). To evaluate the performance of our method, we applied VINYL to a collection of exome sequencing data of 200 patients affected by autosomal dominant disorders as described in Deciphering Developmental Disorders Study (EGAD00001001413) and for which at least one pathogenic or likely pathogenic variant was already reported in the DECIPHER database (Firth et al, 2009). Genetic profiles of the parents were used to provide a matched control population.

VINYL assumes that individuals affected by a pathological condition should display an excess of potentially pathogenic variants at disease associated loci with respect to a population of healthy controls. To corroborate this hypothesis we compared distributions of the top 3 highest scoring variants for every individual, as obtained by applying VINYL to the 200 trios described above, on:

- S1: 94 genes displaying a significant enrichment in disruptive *de-novo* mutations in individuals affected by developmental disorders (Deciphering Developmental Disorders Study, 2015);
- S2: 983 genes associated with autosomal dominant developmental disorders according to the DDG2P database (Wright et al, 2015);
- S3: all the (16454) human genes that are not reported to be associated with developmental disorders according to the DDG2P.

As outlined in Figure 3A, a significant enrichment of highly scoring variants (Wilcoxon p-value 1e-46 and 1e-24, for S1 and S2 respectively) is observed in affected individuals at disease related genes, while no enrichment is observed at genes that should not be associated with the pathological conditions under study. The number of genetic variants selected for prioritization by each of VINYL, KGGseq and Privar on this dataset is reported in Figure 3B. Interestingly, and consistent with our previous findings, we observe that, while being more compact, the list of variants selected by VINYL includes a larger fraction (96% compared to 89% and 91% for KGGseq and Privar respectively) of the variants that were identified as potentially pathogenic by expert manual curation on the same dataset. Similar to our previous observations we notice (Figure 3C) that missense variants prioritized by KGGseq and Privar, but not by VINYL, show contrasting estimates of allele frequency in different databases of human genetic variation and are enriched for non synonymous substitutions that are predicted to be deleterious only by a limited number of the tools incorporated in the dbFNSP database. Conversely, we observe that genotypes prioritized by VINYL include a larger proportion of variants associated with an increased risk of defective developmental phenotypes according to GWAS studies (Supplementary Figure S4A). Unsurprisingly, considering that exome sequencing does not provide a systematic representation of non coding functional genomic elements, in this case we do not observe a strong enrichment of annotations associated with regulatory elements. Notably (Supplementary Figure S4B) all the 3 methods identify a common set of 46 non-synonymous variants that are not reported to be associated with development disorders according to the DDD; 95.6% (44) of these variants are reported as variants of unknown clinical significance in the DECIPHER database.

3.3.3 Post-processing of the results

VINYL incorporates helper applications and utilities to facilitate the post-processing of the data and the interpretation of the results. These include a dimensionality reduction analysis tool, based on Principal Component Analysis (PCA), which can be used to identify groups of patients with similar/related disease-associated mutations; a "burden analysis" utility which can assist in the identification of genes showing a significant increase of pathogenic or likely pathogenic variants; as well as helper methods for the graphical representation of the relative importance of the different components of the scoring systems derived by VINYL (barplot utility) and of their contribution to the final score assigned to single variants (heatmap utility). These utilities produce explicative graphical outputs and accept tabular files generated by VINYL as their main input. An example of the application of PCA and burden test analysis to the Forleo et al dataset is depicted in Supplementary Figure S1. The PCA analysis displayed in Supplementary Figure S1A clearly separates controls from affected individuals. Interestingly 2 distinct groups of patients are observed: group 1 is formed exclusively by patients affected by DCM (dilated cardiomyopathy) while group 2 incorporates patients affected by ARVC (arrhythmogenic right ventricular cardiomyopathy) and HCM (hypertrophic cardiomyopathy). As depicted in Supplementary Figure S1B, the output of VINYL's burden test analysis consists of a panel where, for every gene the distribution of VINYL pathogenicity scores observed in the cohort of affected individuals is compared to the corresponding distribution in the control population. A Mann Whitney Wilcoxon test is used to identify genes showing a significant increase in the score. Only genes with a p-values ≤ 0.15 are reported. To facilitate a rapid comparison, score distributions are represented in the form of boxplots. Dotted lines are used to indicate the "pathogenicity" cut-off value as provided by the user.

The barplot utility included in the Galaxy VINYL implementation (Supplementary Figure S6A) provides a quick and intuitive manner to evaluate the relative importance of different components of the score in the scoring system derived by VINYL. For example, Supplementary Figure S6A illustrates a direct comparison of the scoring systems obtained by VINYL for the Forleo et al and the DDD dataset. The 2 scoring systems show important differences. In particular we notice that the components of the score associated with regulatory/non coding features have higher importance in the Forleo et al dataset. Conversely, the components of the score associated with GWAS studies seems to be more relevant for the prioritization of genetic variants in the DDD dataset.

Among its main outputs, VINYL provides a detailed breakdown of score calculation for every single variant. As illustrated in Supplementary Figure S6B, heatmap representation of individual components of the score computed by VINYL can be used as a quick and effective tool to outline the main features of the variants that were prioritized by tool. For example, in Supplementary Figure S6B the e red square highlights a "cluster" of variants associated with regulatory genomic elements in the Forleo et al dataset.

4 Discussion

The application of genome sequencing technologies to clinical practice is promising a major advance in clinical sciences. However, the systematic integration of genomics in clinical applications poses several challenges, most of which remain unresolved at present. The availability of rapid and effective methods for the accurate "prioritization" of clinically relevant genetic variants is certainly one of the most critical issues in this respect.

Here we introduce VINYL, a fully automated, highly reproducible and customizable system for the annotation and prioritization of genetic variants. The strategy adopted by VINYL is based on well established guidelines and best practices that are currently applied in large scale studies. VINYL calculates a composite score which combines different types of functional annotations. An optimization procedure based on the comparisons of the genetic profiles of a population of affected individuals with a matched control population, is then applied to derive the best scoring system and the ideal threshold for the identification of potentially clinically relevant variants The procedure is completely automatic and the scoring system implemented by VINYL can be easily adapted to different use cases and scenarios.

By performing extensive simulations of disease-associated variants and analysing real data derived from different disease models, we show how VINYL can be effectively applied for the identification of clinically relevant genetic variants in different experimental settings. Importantly, all the comparisons performed in this study demonstrate that the lists of variants prioritized by our novel method, although being more compact, incorporate a larger proportion of the genetic variants that were selected by expert manual curation on the same datasets, with respect to other two state-of-the-art methods. Moreover VINYL is designed to allow high levels of flexibility and can be adapted to different experimental settings. For example, VINYL allows the incorporation of customized annotations a feature that to our knowledge is not incorporated in other similar methods. Additionally, different types of functional evidence can be included or removed from the computation of the score simply by editing/modifying plain configuration text files (see Supplementary Materials). Finally, users can effectively limit or increase the importance of different types of genomic features in the calculation of the final score, allowing a complete customization of the scoring system.

For example, the importance of components of the score associated with GWAS studies or eQTLs can be limited or reduced to 0 in clinical/diagnostic studies aiming to identify genetic variants associated with strictly monogenic disorders or, conversely, increased if the aim of the study is to identify/characterize relatively common genetic variants that could be associated with an increased risk to develop a disorder.

In this respect, we observe that the helper applications and tools incorporated in the Galaxy instance of VINYL (see Supplementary Figures S5 and S6 and Supplementary Materials) can be of great help for comparing different scoring systems, and for understanding which different types of functional genomic elements are associated with variants prioritized by our method. Importantly, this approach is bound to increase its performance over time as it will greatly benefit from the growing number of publicly available data that are being deposited in dedicated databases of genotype-phenotype association such as dbGAP (Mailman et al 2007) and EGA (Lappalainen et al 2015). The availability of more data will help in the construction of more accurate scoring systems for specific diseases, which in turn could become applicable also to the analysis of single samples.

By building on the popular Galaxy workflow manager, VINYL is accessible through a simple yet powerful web interface, which enables collaborative work and facilitates the reproducibility of bioinformatics analyses. A crucial consideration for a more effective analysis of large scale datasets, and for their integration. Taken together, we believe that, in the light of the results presented in the current study, VINYL will represent a valuable resource to assist in the annotation and prioritization of genetic variants in clinical studies.

Acknowledgments

We thank ELIXIR-IIB and ReCaS-Bari for providing computational facilities and Annarita Armenise for technical assistance.

Funding

This work was supported by the Italian Ministero dell'Istruzione, Università e Ricerca (MIUR): PRIN 2017 and CNRbiomics (PIR01_00017); H2020 Projects ELIXIR-EXCELERATE, EOSC-Life, and EOSC-Pillar, and Elixir-IIB.

Conflict of Interest: none declared.

Figures and tables legends

Table1: Sensitivity and specificity on simulated data. Levels of sensitivity and specificity of VINYL, Privar and KGGseq on simulated data. A) Dataset with 1000 polymorphic sites. B) Dataset with 5000 polymorphic sites. C) Dataset with 10000 polymorphic sites. Sizes of the simulated cohorts (25,50 or 100 individuals) are reported in the first column. Tools are indicated in the second column. Corresponding levels of sensitivity and specificity attained by each tool, are reported in the subsequent columns. Columns 3 to 4, 5 to 6 and 7 to 8, report the values for the simulation of pathogenic variants with an odd Risk Ratio of 3, 10 and 20 respectively.

Figure 1: Outline of the variant prioritization strategy adopted by VINYL. Genetic variants identified from a cohort of affected individuals (orange) and a cohort of healthy controls (purple) are subjected to variant annotation. A scoring algorithm is subsequently used to compute a pathogenicity score based on the predicted functional effect of the variants. Different scoring schemes are evaluated and distributions of pathogenicity scores are compared between the 2 cohorts (affected and controls). The scoring system that maximizes the difference of the score distribution between the two populations is selected. The corresponding cut-off score for the identification of potentially pathogenic variants is identified by selecting the threshold that maximizes the number of potentially pathogenic variants in the cohort of affected individuals, while at the same time minimizing the number of potentially pathogenic variants in the control population.

Figure 2: Sensitivity and specificity of VINYL on simulated and read data. A) Distribution of AUC (Area Under the Curve) ROC values for KGGseq, VINYL and Privar in the detection of simulated pathogenic variants. Distributions of AUC are represented in the form of boxplots. Top, middle and lower panels indicate simulations with odd Risk Ratio values of 3, 10 and 20 respectively. B) Comparison of VINYL, Privar and KGGseq on real data. Left: proportion of variants in the population of affected individuals prioritized by each tool. Middle: Proportion of variants prioritized by each tool in the control population. These are likely to represent false positive calls. Right: Proportion of manually curated pathogenic variants according to Forleo et al 2017 recovered by each tool. Orange=VINYL, Blue=KGGseq, Green=Privar.

Figure 3: Comparison of variant prioritization method on the DDD dataset. A) Boxplot of normalized VINYL scores distribution in affected and non-affected individuals in the DDD dataset on S1: 94 gene highly enriched in de-novo mutations in patients affected by developmental disorders, S2: 983 genes associated with monogenic, autosomal dominant developmental disorders according to DDG2P, S3: genes not associated with developmental disorders. B) Comparison of VINYL, Privar and KGGseq on DDD data. For every tool, a bar plot is used to represent the total number of variants prioritized in this dataset (Tot), the total number of variants reported in DECIPHER recovered by each tool (Decip) and the total number of variants prioritized by each tool which are not reported as pathogenic or potentially in DECIPHER (notDecip). A dotted line is used to indicate the number (200) of pathogenic and/or potentially pathogenic variants reported in the DECIPHER database for this cohort. C) Evaluation of missense variants prioritized by VINYL, KGGseq and Privar, but not reported in DECIPHER. Top: histogram of the total number of tools as incorporated in the dbFNSP database, that support a deleterious effect for the variants. Bottom, proportion of variants associated with a 5 fold difference in the Minor Allele Frequency reported by different resources of human genetic variation (1000 genomes, Exac, Gnomad and Topmed). 5Fold: proportion of variants that show a 5fold difference or greater between MAF estimates provided by different databases of human genetic variation. No5fold: proportion of variants that do not show a 5 fold difference in MAF estimates. Orange=VINYL, Blue=KGGseq, Green=Privar.

References

Afgan E, Baker D, Batut B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537–W544. doi:10.1093/nar/gky379

All of Us Research Program Investigators, et alThe "All of Us" Research Program. N Engl J Med. 2019 Aug 15;381(7):668-676. doi: 10.1056/NEJMsr1809937.

Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics. 2015 Jun 27;8:33. doi: 10.1186/s12920-015-0108-y. PMID: 26112054; PMCID: PMC4482045.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 2019, Vol. 47 (Database issue): D1005-D1012.

Canzoneri R, Lacunza E, Abba MC. Genomics and bioinformatics as pillars of precision medicine in oncology. Medicina (B Aires). 2019;79(Spec 6/1):587-592.

Cirulli E.T, Lasseigne B.N, Petrovski S, Sapp P.C, et al Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science. 2015;347:1436–1441.

Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519(7542):223-228. doi:10.1038/nature14135
Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18(10):599–612. doi:10.1038/nrg.2017.52
Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84(4):524-533. doi:10.1016/j.ajhg.2009.03.010
Forleo C, D'Erchia AM, Sorrentino S, et al. Targeted next-generation sequencing detects novel gene-phenotype associations and expands the mutational spectrum in cardiomyopathies. PLoS One. 2017;12(7):e0181842.. doi:10.1371/journal.pone.0181842
Frebourg T. The challenge for the next generation of medical geneticists. Hum Mutat. 2014 Aug;35(8):909-11. doi: 10.1002/humu.22592.
Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. Health Aff (Millwood). 2018;37(5):694–701. doi:10.1377/hlthaff.2017.1624
GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–585. doi:10.1038/ng.2653
Griffith OL, Montgomery SB, Bernier B, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 2008;36(Database issue):D107–D113. doi:10.1093/nar/gkm967
Guo M.H, Dauber A, Lippincott M.F, Chan Y.-M, et al. Determinants of power in gene-based burden testing for monogenic disorders. Am. J. Hum. Genet. 2016;99:527–539.
Gurdasani, D, Carstensen, T, Tekola-Ayele, F, Pagani, L, et al. The african genome variation project shapes medical genetics in Africa. Nature. 2015 517, 327–332. doi: 10.1038/nature13997
Horowitz CR, Orlando LA, Slavotinek AM, et al. The Genomic Medicine Integrative Research Framework: A Conceptual Framework for Conducting Genomic Medicine Research. Am J Hum Genet. 2019;104(6):1088–1096. doi:10.1016/j.ajhg.2019.04.006
Jalali Sefid Dashti M, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques. 2017 Jan 1;62(1)
Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534–13544. doi:10.1093/nar/gku1206
Klein ME, Parvez MM, Shin JG. Clinical Implementation of Pharmacogenomics for Personalized Precision Medicine: Barriers and Solutions. J Pharm Sci. 2017 Sep;106(9):2368-2379. doi: 10.1016/j.xphs.2017.04.051.
Kowalski MH, Qian H, Hou Z et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. 2019 23;15(12):e1008500. doi: 10.1371/journal.pgen.1008500.
Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015;47(7):692–695. doi:10.1038/ng.3312
Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–D985. doi:10.1093/nar/gkt1113
Lee S, Abecasis G.R, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 2014;95:5–23.
Lethimonnier F, Levy Y. Genomic medicine France 2025. Ann Oncol. 2018 1;29(4):783-784. doi: 10.1093/annonc/mdy027.
Li B, Leal S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 2008;83:311–321.
Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 2012;40(7):e53. doi:10.1093/nar/gkr1257
Liu X, Wu C, Li C and Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. Human Mutation. 37:235-241.
Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007;39(10):1181–1186. doi:10.1038/ng1007-1181
Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 2015;11:e1005165.
Nagasaki, M, Yasuda, J, Katsuoka, F, Nariai, N, Kojima, K, Kawai, Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 japanese individuals. Nat. Commun. 2015 6;8018. doi: 10.1038/ncomms90181038/ncomms9018
Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014;15(2):256–278. doi:10.1093/bib/bbs086
Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet. 2014;30(9):377-389. doi:10.1016/j.tig.2014.07.007
Ramu A, Noordam MJ, Schwartz RS, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods. 2013;10(10):985-987. doi:10.1038/nmeth.2611
Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–424. doi:10.1038/gim.2015.30
R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2018
Saudi Mendeliome Group. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. Genome Biol. 2015;16(1):134. doi:10.1186/s13059-015-0693-2
Saunders G, Baudis M, Becker R et al. Leveraging European infrastructures to access 1 million human genomes by 2022. Nat Rev Genet. 2019 20(11):693-701. doi: 10.1038/s41576-019-0156-9.
Stark Z, Dolman L, Manolio TA, et al. Integrating Genomics into Healthcare: A Global Responsibility. Am J Hum Genet. 2019;104(1):13–20. doi:10.1016/j.ajhg.2018.11.014
Servant N, Roméjon J, Gestraud P, et al. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. Front Genet. 2014;5:152. 30. doi:10.3389/fgene.2014.00152
Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011;27(16):2304–2305. doi:10.1093/bioinformatics/btr341
The 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. Nature. 2015 526, 68–74. doi: 10.1038/nature15393
Tangaro MA, Donvito G, Antonacci et al. Laniakea: an open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures. bioRxiv 472464; doi: https://doi.org/10.1101/472464
UK10K Consortium, Walter K, Min JL, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82–90. doi:10.1038/nature14962
Wang C, Li Q, Redden DT, Weindruch R, Allison DB. Statistical methods for testing effects on "maximum lifespan". Mech Ageing Dev. 2004 Sep;125(9):629-32. Erratum in: Mech Ageing Dev. 2006 Jul;127(7):652.
Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. 2010 Nucleic Acids Research, 38:e164,
Wells A, Heckerman D, Torkamani A, et al. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nat Commun. 2019;10(1):5241. Published 2019 Nov 20. doi:10.1038/s41467-019-13212-3
Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag NewYork, 2016
Willighagen E and Ballings M genalg: R Based Genetic Algorithm. R package version 0.2.0. https://CRAN.R-project.org/package=genalg. 2015
Wilson BA, Petrov DA, Messer PW. Soft selective sweeps in complex demographic scenarios. Genetics. 2014;198(2):669-684. doi:10.1534/genetics.114.165571
Zerbino DR, Johnson N, Juetteman T, et al. Ensembl regulation resources. Database (Oxford). 2016;2016:bav119. doi:10.1093/database/bav119
Zhang L, Zhang J, Yang et al PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. Bioinformatics. 2013 Jan 1;29(1):124-5. doi: 10.1093/bioinformatics/bt

TABLES

Table1: Sensitivity and specificity on simulated data. Levels of sensitivity and specificity of VINYL, Privar and KGGseq on simulated data. A) Dataset with 1000 polymorphic sites. B) Dataset with 5000 polymorphic sites. C) Dataset with 10000 polymorphic sites. Sizes of the simulated cohorts (25,50 or 100 individuals) are reported in the first column. Tools are indicated in the second column. Corresponding levels of sensitivity and specificity attained by each tool, are reported in the subsequent columns. Columns 3 to 4, 5 to 6 and 7 to 8, report the values for the simulation of pathogenic variants with an odd Risk Ratio of 3, 10 and 20 respectively.

1000							
A		*RiskRatio~3*		*RiskRatio~10*		*RiskRatio~20*	
		Sens	*Spec*	*Sens*	*Spec*	*Sens*	*Spec*
25	*VINYL*	73.01	99.66	78.89	99.29	89.95	99.69
	Privar	54.97	93.44	59.79	91.64	61.58	93.57
	KGGSeq	62.29	95,55	64.03	96.77	59.93	96.96
50	*VINYL*	78.64	99.52	86.14	99.36	94.35	99.72
	Privar	56.66	94.01	58.60	90.04	59.95	90.79
	KGGSeq	63.92	95.05	66.37	96.60	65.90	94.56
100	*VINYL*	82.69	99.89	90.47	99.04	96.27	99.82
	Privar	55.80	89.87	58.06	91.34	59.63	91.77
	KGGSeq	66.34	96.14	64.47	96.59	66.91	96.58
5000							
B		*RiskRatio~3*		*RiskRatio~10*		*RiskRatio~20*	
		Sens	*Spec*	*Sens*	*Spec*	*Sens*	*Spec*
25	*VINYL*	76.95	99.32	82.90	99.33	91.50	99.96
	Privar	58.27	92.47	61.67	91.70	63.30	92.99
	KGGSeq	63.21	95.55	64.52	95.93	60.60	96.13
50	*VINYL*	80.75	99.41	87.12	99.54	96.03	99.26
	Privar	56.84	93.54	58.84	89.80	62.40	90.36
	KGGSeq	63.68	94.42	69.78	95,94	67.35	94.10
100	*VINYL*	84.53	99.25	93,83	99,17	99.86	99.63
	Privar	58.47	89.89	62.42	90,59	60.78	91.06
	KGGSeq	66.86	95.47	66.53	96.27	67.71	96.36
10000							
C		*RiskRatio~3*		*RiskRatio~10*		*RiskRatio~20*	
		Sens	*Spec*	*Sens*	*Spec*	*Sens*	*Spec*
25	*VINYL*	78.41	99.48	84.01	99.40	92.23	99.66
	Privar	62.70	92.60	64.17	91.87	64.53	93.11
	KGGSeq	69.85	95.75	67.57	96.08	61.25	96.30
50	*VINYL*	82.95	99.55	88.44	99.63	98.29	99.45
	Privar	59.84	93.69	62.06	89.87	65.02	90.52
	KGGSeq	65.59	94.59	73.40	96.07	68.64	94.27
100	*VINYL*	88.38	99.42	97.32	99.39	99.29	99.85
	Privar	61.53	89.91	65.48	90.77	64.08	91.12
	KGGSeq	68.82	95.49	70.21	96.28	72.02	96.51

Figure 1

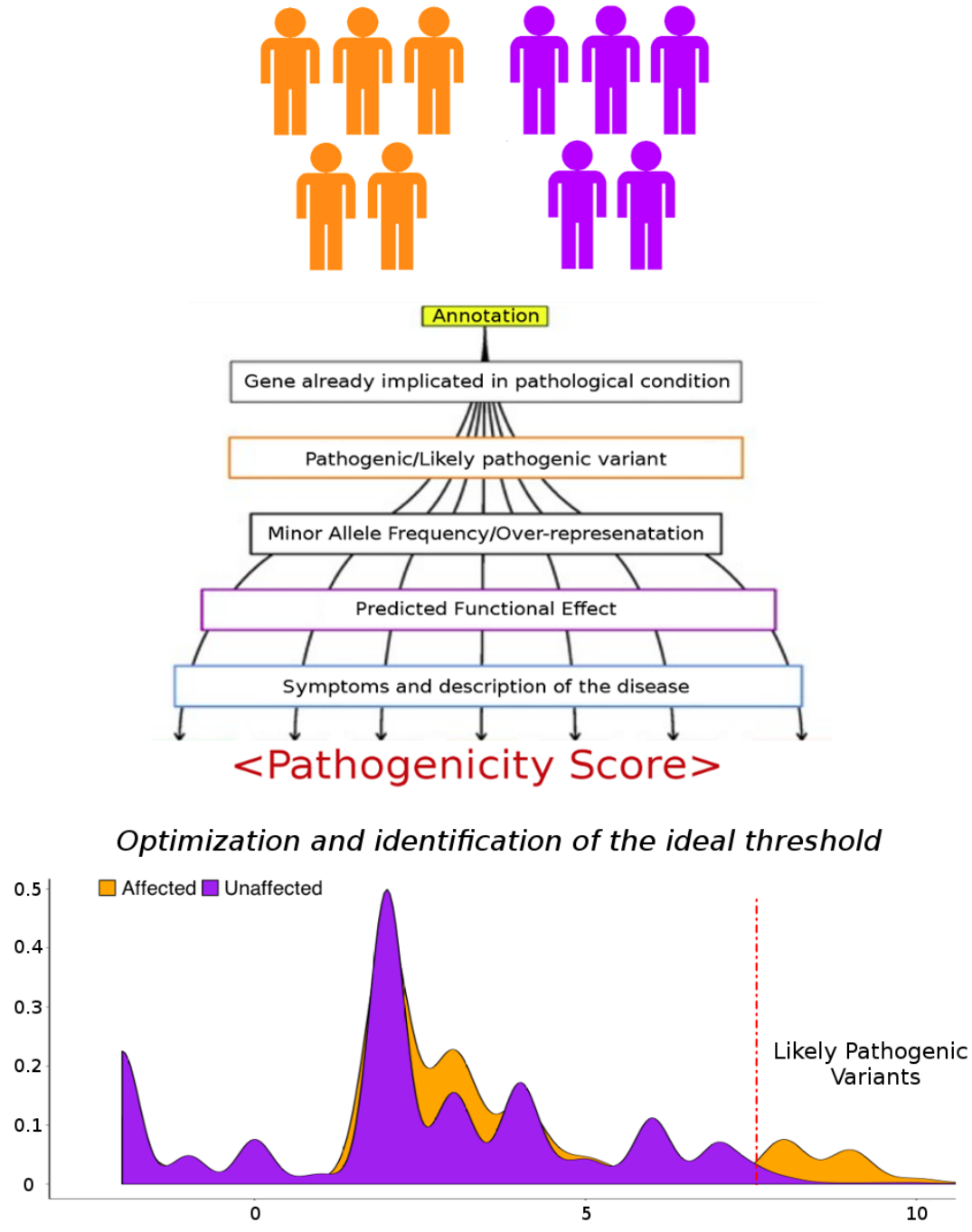

Figure 1: Outline of the variant prioritization strategy adopted by VINYL. Genetic variants identified from a cohort of affected individuals (orange) and a cohort of healthy controls (purple) are subjected to variant annotation. A scoring algorithm is subsequently used to compute a pathogenicity score based on the predicted functional effect of the variants. Different scoring schemes are evaluated and distributions of pathogenicity scores are compared between the 2 cohorts (affected and controls). The scoring system that maximizes the difference of the score distribution between the populations is selected. The corresponding cut-off score for the identification of potentially pathogenic variants is identified as the threshold that maximizes the number of potentially pathogenic variants in the cohort of affected individuals, while at the same time minimizing the number of potentially pathogenic variants in the control population.

Figure 2

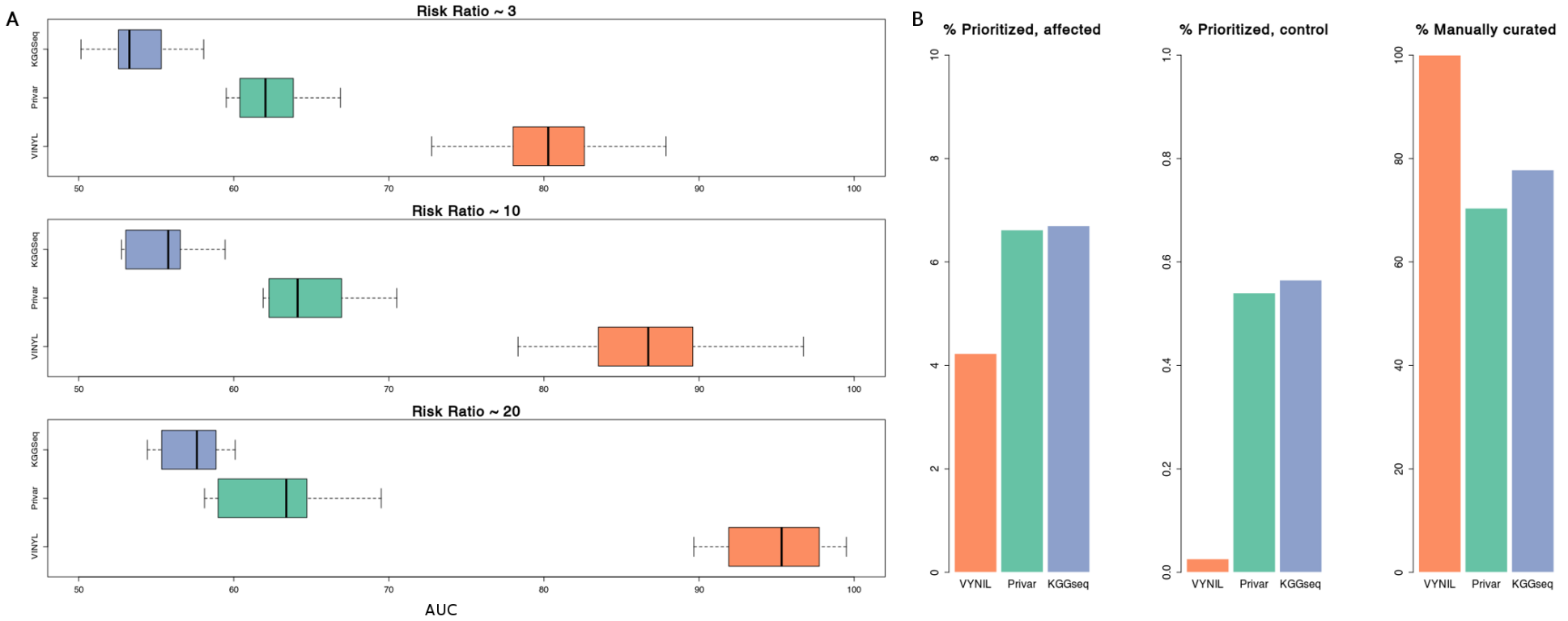

Figure 2: Comparison of variant prioritization methods on simulated and data and on the cardiomyopathy dataset. A) Distribution of AUC (Area Under the Curve) ROC values for KGGseq, VINYL and Privar in the detection of simulated pathogenic variants. Distributions of AUC are represented in the form of boxplots. Top, middle and lower panels indicate simulations with odd Risk Ratio values of 3, 10 and 20 respectively. B) Comparison of VINYL, Privar and KGGseq on cardiomyopathy patients data. Left: proportion of variants in the population of affected individuals prioritized by each tool. Middle: Proportion of variants prioritized by each tool in the control population. Right: Proportion of manually curated pathogenic variants according to Forleo et al 2017 recovered by each tool. Orange=VINYL, Blue=KGGseq, Green=Privar.

Figure 3

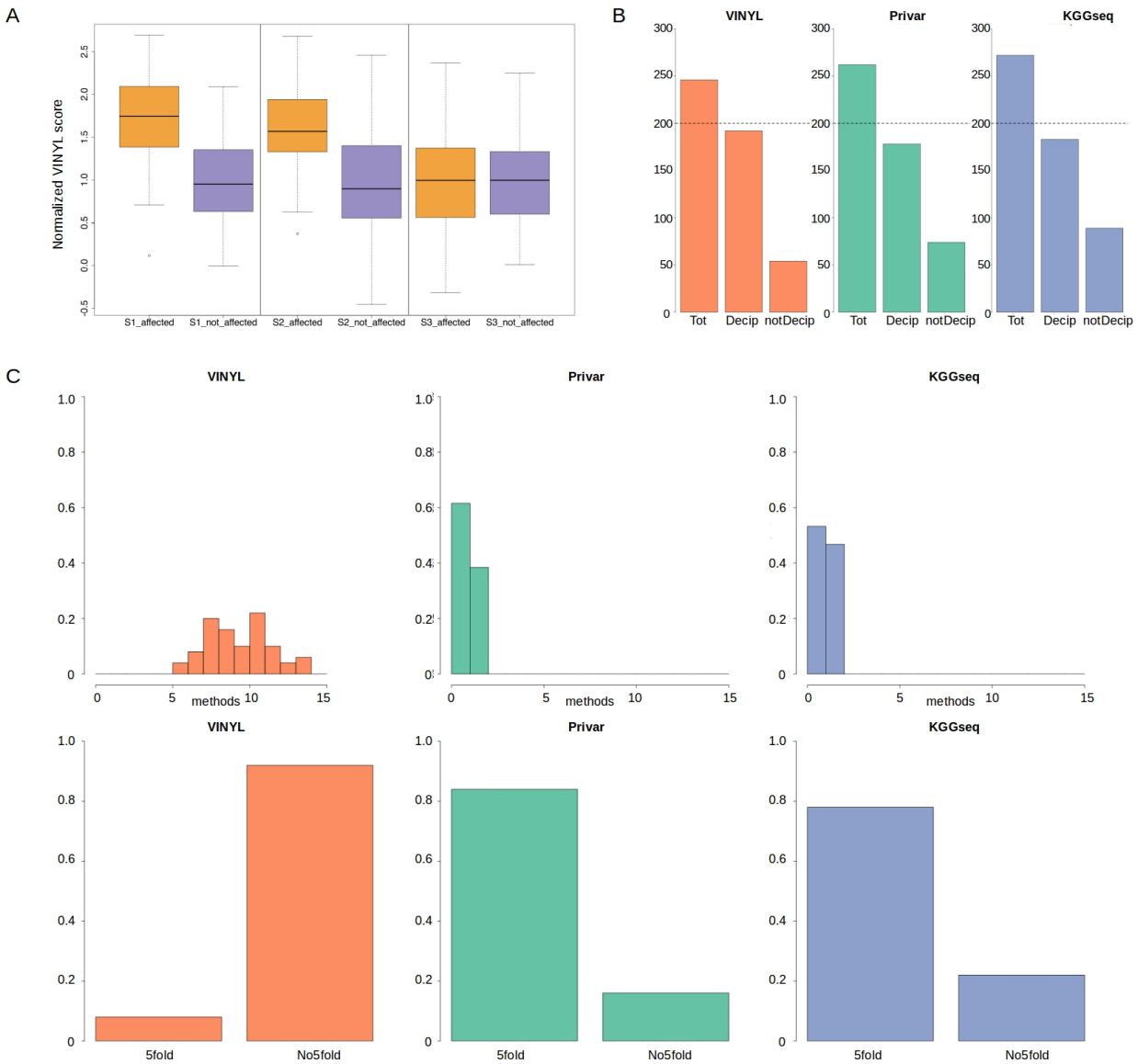

Figure 3: Comparison of variant prioritization method on the DDD dataset. A) Boxplot of normalized VINYL scores distribution in affected and non-affected individuals in the DDD dataset on S1: 94 gene highly enriched in de-novo mutations in patients affected by developmental disorders, S2: 983 genes associated with monogenic, autosomic dominant developmental disorders according to the DDG2P, S3: genes not associated with developmental disorders. B) Comparison of VINYL, Privar and KGGseq on DDD data. For every tool, a bar plot is used to represent the total number of variants prioritized in this dataset (Tot), the total number of variants reported in DECIPHER recovered by each tool (Decip) and the total number of variants prioritized by each tool which are not reported as pathogenic or potentially in DECIPHER (notDecip). A dotted line is used to indicate the number (200) of pathogenic and/or potentially pathogenic variants reported in the DECIPHER database for this cohort. C) Evaluation of missense variants prioritized by VINYL, KGGseq and Privar, but not reported in DECIPHER. Top: histogram of the total number of tools as incorporated in the dbFNSP database, that support a deleterious effect for the variants. Bottom, proportion of variants associated with a 5 fold difference in the estimated minor allele frequency, according to different resources of human genetic variation (1000 genomes, Exac, Gnomad and Topmed). Orange=VINYL, Blue=KGGseq, Green=Privar.