GARFIELD-NGS: Genomic vARiants FIltering by dEep Learning moDels in NGS

Viola Ravasio, Edoardo Giacopuzzi*

Department of Molecular and Translational Medicine, University of Brescia, Viale Europa 11, 25123 Brescia, ITALY

*To whom correspondence should be addressed

email: edoardo.giacopuzzi@unibs.it

Abstract

Exome sequencing approach is extensively used in research and diagnostic laboratories to discover pathological variants and study genetic architecture of human diseases. Even if present platforms produce high quality sequencing data, false positives variants remain an issue and can confound subsequent analysis and result interpretation.

Here, we propose a new tool named GARFIELD-NGS (Genomic vARiants FIltering by dEep Learning moDels in NGS), which uses deep learning algorithm to dissect false and true variants in exome sequencing experiments performed with Illumina or Ion platforms. GARFIELD-NGS consists of 4 distinct models tested on NA12878 gold-standard exome variants dataset (NIST v.3.3.2): Illumina INDELs, Illumina SNPs, ION INDELs, and ION SNPs. AUC values for each variant category are 0.9267, 0.7998, 0.9464, and 0.9757, respectively. GARFIELD-NGS is robust on low coverage data down to 30X and on Illumina two-colour data, as well.

Our tool outperformed previously proposed hard-filter, and calculates for each variant a score from 0 to 1, allowing application of different thresholds based on the desired level of sensitivity and specificity. GARFIELD-NGS process standard VCF file input using Perl and Java scripts and produce a regular VCF output. Thus, it can be easily integrated in existing analysis pipeline. GARFIELD-NGS is freely available at https://github.com/gedoardo83/GARFIELD-NGS.

Introduction

Whole exome sequencing (WES) is a powerful method ideally designed to rapidly investigate all the coding sequences in human genome at base resolution, allowing to detect a wide spectrum of genetic variations[1–3]. In the latest years great advances were taken in Next Generation Sequencing (NGS) field and WES experiments have become faster, cheaper and easier to perform. These improvements encouraged the diffusion of WES through research laboratories, and allowed its translation from basic research to clinical use[4,5]. Indeed, WES has rapidly become a popular approach to discover new disease genes in rare Mendelian disorders[6–8], as well as to evaluate risk alleles in complex disorders[9,10].

Even if WES is now easy and affordable to perform, data analysis remains a critical and difficult step due to the quantity and complexity of information obtained from each experiment[11,12]. Previous studies have shown that genetic variants identified by exome sequencing often carries a significant proportion of false positive calls, especially INDELs[1,13,14]. This issue often imply additional costs for variants validation by Sanger sequencing, at least in diagnostic settings[5,15]. False positive calls poses serious challenges in downstream data analysis, introducing erroneus missense and loss of function variants, like frameshift INDELs, that are targets of most analysis work-flows[16,17].

Effective bioinformatic approaches to filter out false positive calls have been developed for Illumina NGS data and Variant Quality Score Recalibration (VQSR) method from GATK best practises[18] is now the most adopted filtering method. Besides its robust performances, VQSR applies only to large datasets including at least tens of samples, since it needs a large set of variants to train a machine learning algorithm[19]. This limits its application on single sample data, that could often occur in rare disease research projects or in diagnostic settings. Moreover, few filtering methods are available for ION WES data, since the low spread of WES on this platform has led to low interest in development of specific bioinformatic tools. As results, variant filtering strategies for single samples or trio analysis are today usually limited to hard filtering of variants based on a combination of quality parameters. For Illumina sequencing data, GATK best practises are the most widely adopted hard-filters[18], while for ION data there are only few reported strategies[13].

Machine learning (ML) approaches have been proven effective in solving classification problems in complex systems[20] and are rapidly diffusing also in the genomic field[21]. Indeed, ML algorithms revealed especially useful when the state of an object can not be deduced by single features or their linear combination, since they can integrate different layer of information and reveal hidden patterns in input data. In this way, ML models are often able to compute a robust probability value useful in object state classification. This approach has successfully applied to the analysis of

genomic variants and several ML based models have been developed to predict impact of genomic variants on protein functionality[22,23] or regulatory region[24,25]. ML algorithms are also implemented in GATK VQSR strategy for false variant filtering on large datasets[19].

Here we propose a new tool, Genomic vARiants FIltering by dEep Learning moDels in NGS (GARFIELD-NGS), that rely on neural networks algorithm to effectively classify true and false variants. GARFIELD-NGS can be applied in single sample WES analysis and it is effective on SNPs and INDELs variants derived from both Illumina or ION platform. It is robust on medium and low coverage dataset and can be applied to experiments based on the recent 2-colour Illumina chemistry, as well.

Results

Prediction models

We developed 4 distinct models addressing INDELs and SNPs for both Illumina and ION platforms. After optimization of hyper-parameters and model refinement, we generated 4 prediction models with distinct architectures optimized for each class of variants. All 4 models present 5 hidden layers, using Tanh or Rectifier activation functions for SNPs and INDELs models, respectively. Different specific values of rho, epsilon, l1, and l2 were obtained for each model as shown in Supplementary Table S1. Features importances for each model are reported in Supplementary Fig. S1. No single feature emerged as strong predictor in all Illumina SNPs / INDELs and ION SNPs / INDELs, but coverage related and strand-bias metrics are usually in the top 5 variables. AUC values of final models on training and validation sets were > 0.9 for all variants groups but Illumina SNPs, showing a slightly worst performance with AUC almost 0.8 (see Supplementary Fig. S2).

Prediction models performances on test sets

GARFIELD-NGS contains 4 models specifically optimized for Illumina INDELs, Illumina SNPs, ION INDELs, and ION SNPs datasets. Based on each model, our tool calculates for each variant in VCF file a confidence probability (P true) ranging from 0 to 1. Actual performances of our models were evaluated using independent test sets of ~ 80,000 SNPs and ~ 2,000 INDELs.

AUC values > 0.90 were obtained for Illumina INDELs, ION INDELs and ION SNPs: 0.9269, 0.9464, and 0.9757, respectively. Otherwise, Illumina SNPs model showed slightly reduced performances with test set AUC 0.7998 (see Figure 1).

P true value clearly distinguish true from false variants in test set for Illumina INDELs, ION INDELs, and ION SNPs (see Figure *2)*. Calculated median values are: Illumina INDELs true calls 0.964, false calls 0.229; ION INDELs true calls 0.947, false calls 0.096; ION SNPs true calls 0.968, false calls 0.081. Differences are smaller between median values for Illumina SNPs: true calls 0.955, false calls 0.926. Notably, a distinct distribution of true and false variants can not be observed evaluating single features (see Supplementary Fig. S3-S6).

Maximum accuracy is > 0.90 for all variants categories: Illumina INDELs 0.9355, Illumina SNPs 0.9435, ION INDELs 0.9117, ION SNPs 0.9919. Applying filtering threshold corresponding to maximum accuracy, we obtained the following TPR and FDR values: Illumina INDELs 0.9779 and 0.0604, Illumina SNPs 0.9949 and 0.0536, ION INDELs 0.9542 and 0.0707, and ION SNPs 0.9974 and 0.0056, respectively. Comparing with proportion of false calls present in test sets (see

Supplementary Table S2), GARFIELD-NGS score allowed significant reduction of false positives: Illumina INDELs 75%, Illumina SNPs 35%, ION INDELs 76%, ION SNPs 68%. Thresholds for variant filtering according to maximum accuracy or 0.99 TPR are reported in Table 1 and additional thresholds in Supplementary Table S3.

Moreover, GARFIELD-NGS was tested on medium and low coverage experiments, using variants sets obtained from sequence data downsamped to 60X and 30X mean coverage. AUC values calculated on downsampled sets (60X / 30X) are similar to those obtained with full data: Illumina INDELs 0.9042 / 0.8933, SNPs 0.6609 / 0.6307; ION INDELs 0.8663 / 0.8174, SNPs 0.9522 / 0.9221 (see Figure 1).

Finally, we tested our Illumina models on variants generated by the recent two-colour Illumina chemistry, using data from HiSeqX experiments. GARFIELD-NGS predictions achieved AUC values of 0.9676 in INDELs and 0.8584 in SNPs from HiSeqX variant sets (see Figure 1a, b).

Comparison between GARFIELD-NGS and previous hard-filters

Variants in our 4 test sets were re-analysed using previously proposed hard-filters for Illumina (Van der Auwera et al. 2013) and ION (Damiati et al., 2016) data, as described in methods. In all 4 variants groups, GARFIELD-NGS outperform previous filters, showing higher accuracy and comparable or higher TPR (see Table 1).

Relevant improvements are seen for INDELs. In the best scenario, GATK hard-filters applied on Illumina INDELs dataset reached 0.8665 accuracy, 0.9934 TPR and 0.1456 FDR, while GARFIELD-NGS had a maximum accuracy of 0.9355, with 0.9779 TPR and 0.06 FDR. Even at 0.99 TPR threshold, GARFIELD-NGS showed better performances with 0.9326 accuracy and 0.0736 FDR.

Considering ION INDELs, the maximum accuracy for previous filters correspond to low setting and resulted in 0.8033 accuracy, 0.9659 TPR and 0.1920 FDR. GARFIELD-NGS had a maximum accuracy of 0.9117, with 0.9542 TPR and 0.0707 FDR. At 0.99 TPR threshold, GARFIELD-NGS confirmed better performances with 0.8607 accuracy and 0.1517 FDR.

GARFIELD-NGS demonstrated best performances on ION SNPs, where it achieved 0.9919 maximum accuracy, with 0.9974 TPR and 0.0056 FDR.

Implementation and availability

Prediction models are compiled in Java and implemented in GARFIELD-NGS perl script to perform automated variant scoring on VCF files. Source code is freely available at: https://github.com/gedoardo83/GARFIELD-NGS

Discussion

Filtering out false variants from WES results is a long standing challenge in data analysis. Indeed, the high proportion of false calls, especially INDELs, generated by both Illumina and ION platforms[1,13,14] poses serious challenges for downstream data analysis and result interpretation. To develop a new method for variant filtering, we first collected 22 different WES experiments for the NA12878 sample (see Supplementary Table S4), generating a dataset of 178,450 Illumina variants (173,116 SNPs / 5,334 INDELs) and 181,479 ION variants (177,362 SNPs / 4,117 INDELs). True and false calls were determined by comparing to the gold-standard calls provided by Genome in a Bottle Consortium (GIAB). In 2013 Genome in a Bottle Consortium (GIAB), part of the National Institute of Standards and Technology (NIST), has distributed the first set of gold standard calls based on integration of 13 different datasets of this sample obtained using different NGS technologies[26]. This constantly updated set of variants is now broadly accepted as a standard for variant identification benchmarking.

As expected, we observed high proportion of false calls, especially INDELs, in our unfiltered datasets of WES variants from both Illumina and ION platforms (see Supplementary Table S2). Nowadays, the most applied strategy for false positive variants filtering on Illumina is the GATK VQSR method[18,19], which has been proven effective, but applies only to large datasets including at least tens of samples. Concerning ION data, widely adopted strategies are lacking. In this scenario, variant filtering strategies for single samples or trio analysis are today usually limited to hard filtering of variants based on a combination of quality parameters. For Illumina sequencing data, GATK best practises are the most adopted hard-filters[18], while for ION data there are only few reported strategies[13]. However, taken singularly, variants features calculated by variants callers do not clearly distinguish false and true calls (see Supplementary Fig. S3-S6), suggesting that their integration in a prediction model could be a more effective strategy.

Following this approach, we developed GARFIELD-NGS tool, that relies on deep learning models to discriminate between true and false variants in WES experiments integrating variant features reported by GATK or TVC variant callers (see Supplementary Table S5). Given a standard VCF file, it calculates for each variant a score ranging from 0 to 1, reflecting probability of being a true call (P true). The tool is composed of 4 models, specifically developed on INDELs or SNPs variants coming from Illumina or ION experiments (see Supplementary Table S1).

GARFIELD-NGS revealed robust performances on all 4 variants categories, showing high AUC values: 0.9041 for Illumina INDELs, 0.7998 for Illumina SNPs, 0.9464 for ION INDELs, and 0.9757 for ION SNPs. GARFIELD-NGS predictions maintain robust performances when applied to

results from medium (60 X) or low (30 X) mean coverage data or to data from the recently introduced Illumina 2-colour chemistry (see Figure 1).

While previous hard-filters only perform a boolean classification of variants in true or false categories, GARFIELD-NGS calculates a prediction values ranging from 0 to 1, with distinct distributions between false and true variants (see Figure 2). This allows tuning of variant filtering threshold depending on the desired accuracy and specificity or even integration of P true value as prioritization score rather than variant filter. The maximum accuracy thresholds retain > 95 % of true calls while reducing false calls by 35-76 %, depending on variant category. Even when applying a threshold corresponding to 0.99 TPR, GARFIELD-NGS maintains > 0.86 accuracy and reduces false calls by 37-80 % (see Table 1).

Overall, lower performances emerged for Illumina SNPs model. This may be explained by the peculiar nature of Illumina false SNPs, which are often systematic errors induced by specific sequence context[27,28]. This kind of information are not captured by variant annotations generated by GATK and evaluated by GARFIELD-NGS models, making our approach less effective on Illumina SNPs.

GARFIELD-NGS predictions outperformed previously proposed hard-filter for Illumina (Van der Auwera et al. 2013) and ION (Damiati et al., 2016) data in all 4 variants categories (see Table 1). GARFIELD-NGS score showed a strong improvement on INDELs variants for both Illumina (maximum accuracy 0.9355, TPR 0.9779, FDR 0.06) and ION data (maximum accuracy 0.9117, TPR 0.9542 TPR, FDR 0.0707). Thus, our tool effectively reduces false INDEL calls and could be useful to improve WES results interpretation considering that many work-flows search for variants that potentially alter gene function, especially loss of function variants like frameshift INDELs[16,17]. Even if Illumina SNPs AUC value is lower than those of other models, GARFIELD-NGS still perform better than GATK hard-filters showing a max accuracy of 0.9435, with 0.9949 TPR and 0.0535 (see Table 1).

Overall, these results define GARFIELD-NGS as a robust tool for all type of Illumina and ION exome data, with particular focus on single or small multi-sample experiments. GARFIELD-NGS script performs automated variant scoring on VCF files and returns a standard VCF output with prediction score added as INFO tags. Thus, it can be easily integrated in already established analysis pipelines.

Materials and Methods

Data sources

Data used in model training, validation and test were based on 19 high-coverage exome sequencing experiments on the NA12878 reference sample, produced by either Illumina or Ion Torrent platforms (see Supplementary Table S4 and Supplementary Fig. S7).

Illumina dataset contains 9 exome sequencing experiments from Sequence Read Archive (SRA), produced on Illumina HiSeq 2000 / 2500 platforms. Mean coverage ranges from 77X to 164X, with > 85% of bases covered at least 20X.

ION dataset includes 10 exome sequencing experiments produced on ION Proton platform: 6 obtained as aligned reads from Ion Community, and 2 as in-house exome experiments. For in house sequencing, NA12878 gDNA was obtained from Coriell Cell Repository and exome libraries were prepared from 100ng gDNA using ION AmpliSeq Exome RDY kit. Hi-Q PI OT2 200 kit was used for ISP template preparation using 8 µl of 100pM exome library and products were sequenced using Hi-Q PI Sequencing 200 kit and PI v3 chips on Ion Proton platform. The mean coverage ranges from 120X to 270X, with > 92% of bases covered at least 20X.

To generate medium and low coverage datasets for models validation, BAM file of Illumina and ION experiments were downsampled to 30X and 60X mean coverage by random sampling using samtools.

Additionally, we included an HiSeqX dataset consisting of 3 genome sequencing experiments produced on Illumina HiSeqX platform. Mean coverage ranges from 27X to 52X, with > 76% of bases covered at least 20X.

Variant calling

Illumina data were analysed following GATK best practices[18,19]. Briefly, sequencing reads were aligned to hg19 reference genome using BWA-mem v.0.7.1, followed by duplicate marking with Picard v.1.119 and BAM file realignment using GATK 3.6. Variants were then identified using GATK Haplotype Caller 3.6 with stand_emit_conf and stand_call_conf set to 10 and 30, respectively. Ion Torrent data were processed using Torrent Suite v.5.0.2 and Torrent Variant Caller (TVC) v.5.0.2. Briefly, sequencing reads were aligned to hg19 reference genome using TMAP, followed by BAM file realignment and variant identification with TVC v.5.0.2, using standard parameters provided by manufacturer for AmpliSeq Exome protocol. The same pipelines were used to identify variants in 30X / 60X downsampled experiments. GATK and TVC were selected as the most widely adopted variant callers for Illumina and Ion Torrent data. To provide comparable

representation of alleles across VCF files, variants were decomposed, normalized and left aligned using vt tool[29]. Focusing on exome regions, we considered for further analysis only variants located in RefSeq coding exons plus 5bp flanking regions and overlapping high confident regions defined in NIST v.3.3.2 data (ftp://ftptrace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/).
True and false variants in these regions were determined based on comparison with NA12878 gold-standard calls from NIST v.3.3.2[26].
Detailed description of variants identified for each experiment is given in Supplementary Table S4.

Definition of variant datasets for model development

For both Illumina and ION platforms we merged variants from all experiments resulting in 178,450 Illumina variants (173,116 SNPs / 5,334 INDELs) and 181,479 ION variants (177,362 SNPs / 4,117 INDELs). SNP and INDEL variants were considered separately in subsequent analysis, generating four groups: Illumina INDELs, Illumina SNPs, ION INDELs, and ION SNPs. Variants in each group were then splitted randomly in 4 independent datasets to be used in models development: pre-training, training and validation sets were used to develop and refine prediction models; test sets contained ~ 50% of overall variants and were used to assess prediction performances. Since both Illumina and ION platforms have high accuracy on SNP calls, SNPs sets contained a strongly unbalanced proportion of true calls. To avoid overfitting on true calls, pre-training and training sets were balanced by randomly removing true calls so that they contain at least 20 % of false variants. Additionally, we assembled a 60X and a 30X test sets merging variants derived from downsampled experiments (see data sources) and randomly selecting ~ 50% of overall variants. HiSeqX test set was obtained merging variants from 3 HiSeqX experiments (see data sources).
Detailed description of the final datasets used in this study is reported in Supplementary Table S2.

Development of prediction models

We used variant features reported in VCF file output by GATK and TVC variant callers to train deep learning algorithms predicting true out of false variants. Features with constant values were not considered.
For ION SNP variants we included 18 features: FAO, FDP, FSAF, FSAR, FXX, GQ, HRUN, LEN, MLLD, QUAL, QD, RBI, SSEN, SSEP, SSSB, STB, STBP, and VARB.
For ION INDEL variants we included 18 features: FAO, FDP, FSAF, FSAR, FXX, GQ, HRUN, LEN, MLLD, PB, PBP, QUAL, QD, RBI, STB, STBP, SSSB and VARB.
For Illumina SNP and INDEL variants we included 10 features: BaseQRankSum, DP, FS, GQ, MQ,

MQRankSum, QD, QUAL, ReadPosRankSum, SOR.

Detailed description of selected features is reported in Supplementary Table S5, while distributions of each feature values in all variants analysed in this study are reported in Supplementary Fig. S3-S6.

INDELs and SNPs were treated separately for each platform, generating 4 distinct prediction models: Illumina INDELs, Illumina SNPs, ION INDELs, and ION SNPs. Deep learning models development was performed using H2O 3.10.4.5 (http://www.h2o.ai).

First, hyper-parameters were optimised for each model using corresponding training sets and 10 fold cross-validation. We used random search to explore space of 6 hyper-parameters: l1, l2, rho, epsilon, hidden layers and activation function. Search was conducted with early stopping based on log-loss (5 stopping rounds with 10E-3 stopping tolerance), generating at least 10,000 different models. Models were ranked according to cross-validation AUC and the best five hyper-parameters combinations were used for further model refinement. For each combination we first performed unsupervised pre-training with autoencoder on pre-training sets using 1,000 epochs and early stopping based on log-loss (10 stopping rounds with 10E-5 stopping tolerance). Prediction models were than initiated with the corresponding pre-training model and refined on training and validation sets using 1,000 epochs and early stopping as above (see Supplementary Fig. S8). For each group of variants, a final prediction model was selected based on AUC value on validation set. The architecture of each model is reported in Supplementary Table S1.

Finally, GARFIELD-NGS prediction performance for each variants category was evaluated on test sets using the corresponding model.

Comparison with previous hard-filters

Variants in our 4 test sets were re-analysed using previously developed hard-filters for Illumina, as described in GATK best practises[18] and ION[13] data. For Illumina data we created 2 sets of filtered variants using quality based metrics and then adding genotype quality (GQ) filter after GQ refinement, as described in GATK protocols. Instead, for ION data we created 3 sets of filtered variants applying hard, medium and low stringency filters proposed in the original paper.

Data availability

The Illumina datasets analysed during the current study are available in the SRA archive repository, https://www.ncbi.nlm.nih.gov/sra. ION datasets are available from Thermo Fisher Cloud, https://ion-torrent.s3.amazonaws.com/datasets/HiQ. Accession codes are given in Supplementary Table S4.

References

1. Zhang, G. *et al.* Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. *BMC Genomics* **16,** 581 (2015).

2. Petersen, B.-S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D. & Franke, A. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet.* **18,** 14 (2017).

3. Kadalayil, L. *et al.* Exome sequence read depth methods for identifying copy number changes. *Brief. Bioinform.* **16,** 380–392 (2015).

4. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312,** 1880–7 (2014).

5. Bowdin, S. *et al.* Recommendations for the integration of genomics into clinical practice. *Genet. Med.* **18,** 1075–1084 (2016).

6. Brown, T. L. & Meloche, T. M. Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* **108,** 109–114 (2016).

7. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

8. Wang, Z., Liu, X., Yang, B.-Z. & Gelernter, J. The Role and Challenges of Exome Sequencing in Studies of Human Diseases. *Front. Genet.* **4,** 160 (2013).

9. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44,** 623–30 (2012).

10. Kosmicki, J. A., Churchhouse, C. L., Rivas, M. A. & Neale, B. M. Discovery of rare variants for complex phenotypes. *Hum. Genet.* **135,** 625–634 (2016).

11. Lelieveld, S. H., Veltman, J. A. & Gilissen, C. Novel bioinformatic developments for exome sequencing. *Hum. Genet.* **135,** 603–614 (2016).

12. Bao, R. *et al.* Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.* **13,** 67–82 (2014).

13. Damiati, E., Borsani, G. & Giacopuzzi, E. Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. *Hum. Genet.* **135,** 499–511 (2016).

14. Boland, J. F. *et al.* The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum. Genet.* (2013). doi:10.1007/s00439-013-1321-4

15. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15,** 733–47 (2013).

16. Wang, S. & Xing, J. A Primer for Disease Gene Prioritization Using Next-Generation Sequencing Data. *Genomics Inform.* **11,** 191–199 (2013).

17. Isakov, O., Perrone, M. & Shomron, N. in *Methods in molecular biology* (ed. Shomron, N.) **1038,** 137–158 (Springer Science, 2013).

18. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).

19. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

20. de Ridder, D., de Ridder, J. & Reinders, M. J. T. Pattern recognition in bioinformatics. *Brief. Bioinform.* **14,** 633–647 (2013).

21. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16,** 321–332 (2015).

22. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24,** 2125–37 (2014).

23. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48,** 1581–1586 (2016).

24. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12,** 931–4 (2015).

25. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47,** 955–61 (2015).

26. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–51 (2014).

27. Allhoff, M. *et al.* Discovering motifs that induce sequencing errors. *BMC Bioinformatics* **14 Suppl 5,** S1 (2013).

28. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43,** e37 (2015).

29. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31,** 2202–4 (2015).

Acknowledgements

EG has been supported by “Fondazione Cariplo” and “Regione Lombardia” under the project: “La salute della persona: lo sviluppo e la valorizzazione della conoscenza per la prevenzione, la diagnosi precoce e le terapie personalizzate”, Grant Emblematici Maggiori 2015-1080.

Author Contributions

VR performed ROC curve analysis. EG conceived the study and performed NGS data analysis and variant calling. VR and EG developed prediction models. VR and EG wrote and reviewed the manuscript text and prepared figures.

Competing financial interests

The author(s) declare no competing financial interests.

Figure Legends

Figure 1. ROC curves of GARFIELD-NGS final models on test datasets

Performance of prediction models were assessed using ROC curves on test sets, 60X and 30X downsampled sets, and HiSeqX sets. Performances were evaluated separately on Illumina data (INDELs in a, SNPs in b) and Ion data (INDELs in c, SNPs in d). Values of area under the curve (AUC) are indicated in the graphical plots.

Figure 2. Distributions of GARFIELD-NGS score for true and false variants

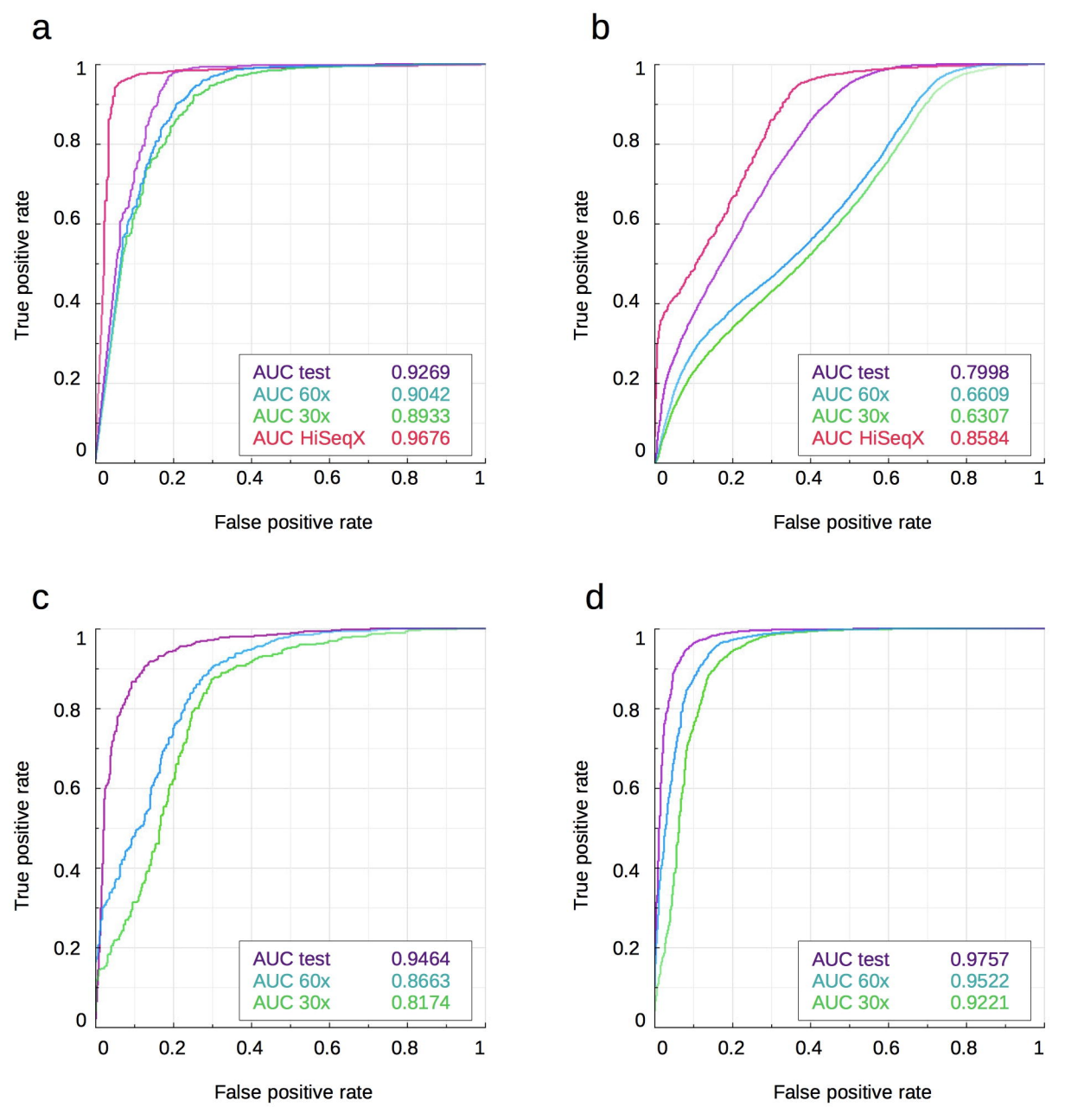
GARFIELD-NGS models assign a score from 0 to 1 to each variant. Distributions of GARFIELD-NGS score for true and false variants are clearly separated for Illumina INDELs (a), ION INDELs (c), and ION SNPs (d) test sets. Smaller difference is observed for Illumina SNPs (b). Black dots indicate median values.

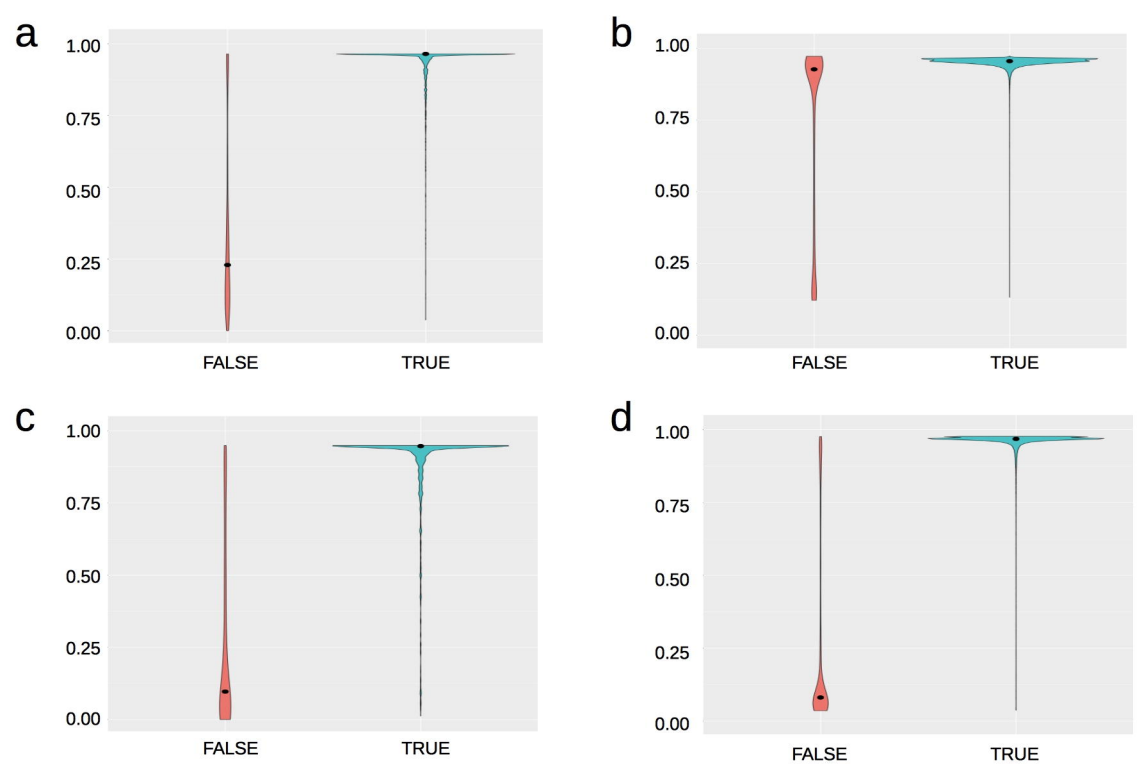
Tables

Table 1. GARFIELD-NGS performance and comparison with previous hard-filters

Performance of GARFIELD-NGS prediction applying threshold corresponding to maximum accuracy or 0.99 TPR and comparison with previous GATK[18] and ION[13] hard-filters. Variants with P true value < threshold are classified as false. TPR: true positive rate, FDR: false positive rate.

GARFIELD-NGS	**criterion**	**accuracy**	**TPR**	**FDR**	**specificity**	**threshold**
Illumina INDELs	0.99 TPR	0.9326	0.9901	0.0738	0.7517	0.4703
	max accuracy	0.9355	0.9779	0.0604	0.8021	0.6301
Illumina SNPs	0.99 TPR	0.9411	0.9900	0.0518	0.3901	0.8777
	max accuracy	0.9435	0.9949	0.0536	0.3644	0.8369
ION INDELs	0.99 TPR	0.8607	0.9906	0.1517	0.4786	0.0876
	max accuracy	0.9117	0.9542	0.0707	0.7863	0.4948
ION SNPs	0.99 TPR	0.9868	0.9900	0.0035	0.8047	0.5261
	max accuracy	0.9919	0.9974	0.0056	0.6786	0.1389
GATK hard-filters	**criterion**	**accuracy**	**TPR**	**FDR**	**specificity**	**threshold**
Illumina INDELs	Standard	0.8665	0.9934	0.1456	0.4670	-
	Low GQ	0.8254	0.9366	0.1510	0.4757	-
Illumina SNPs	Standard	0.9400	0.9943	0.0566	0.3275	-
	Low GQ	0.9223	0.9745	0.0572	0.3332	-
ION hard-filters	**criterion**	**accuracy**	**TPR**	**FDR**	**specificity**	**threshold**
ION INDELs	High	0.7707	0.8192	0.1336	0.6282	-
	Medium	0.7978	0.8969	0.1576	0.5064	-
	Low	0.8033	0.9659	0.1920	0.3248	-
ION SNPs	High	0.8779	0.8795	0.0043	0.7835	-
	Medium	0.9601	0.9650	0.0058	0.6800	-
	Low	0.9817	0.9885	0.0072	0.5948	-


a
True positive rate
False positive rate
AUC test 0.9269
AUC 60x 0.9042
AUC 30x 0.8933
AUC HiSeqX 0.9676
b
True positive rate
False positive rate
AUC test 0.7998
AUC 60x 0.6609
AUC 30x 0.6307
AUC HiSeqX 0.8584
c
True positive rate
False positive rate
AUC test 0.9464
AUC 60x 0.8663
AUC 30x 0.8174
d
True positive rate
False positive rate
AUC test 0.9757
AUC 60x 0.9522
AUC 30x 0.9221


a
1.00
0.75
0.50
0.25
0.00
FALSE
TRUE
b
1.00
0.75
0.50
0.25
0.00
FALSE
TRUE
c
1.00
0.75
0.50
0.25
0.00
FALSE
TRUE
d
1.00
0.75
0.50
0.25
0.00
FALSE
TRUE