bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

Viola Ravasio[1], Marco Ritelli[1], Andrea Legati[2], Edoardo Giacopuzzi[1,*]

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

[2]Unit of Molecular Neurogenetics, Fondazione IRCCS Istituto Neurologico 'Carlo Besta', via Temolo 4, 20126 Milan, ITALY

*To whom correspondence should be addressed

email: edoardo.giacopuzzi@unibs.it

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 previous hard-filters, and calculates for each variant a score from 0.0 to 1.0, allowing application of different thresholds based on desired level of sensitivity and specificity. GARFIELD-NGS processes standard VCF file input using Perl and Java scripts and produces 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.

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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[1,13–15]. This issue often implies additional costs for variants validation by Sanger sequencing, at least in diagnostic settings[5,16]. False positive calls pose serious challenges in downstream data analysis, introducing erroneous missense and loss of function variants, like frameshift INDELs, that are targets of most analysis work-flows[17,18].

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[19] is now the most adopted filtering method. Besides its robust performances, VQSR gives optimal results when applied to large datasets, since it needs a large set of variants to train a machine learning algorithm[20]. 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 developing 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[19], while for ION data there are only few reported strategies[14].

Machine learning (ML) approaches have been proven effective in solving classification problems in complex systems[21] and are rapidly diffusing also in the genomic field[22]. 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 been successfully applied to the analysis of

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

genomic variants and several ML based models have been developed to predict impact of genomic variants on protein functionality[23,24] or regulatory region[25,26]. ML algorithms are also implemented in GATK VQSR strategy for false variant filtering on large datasets[20].

Here we propose a new tool, Genomic vARiants FIltering by dEep Learning moDels in NGS (GARFIELD-NGS), that relies on neural networks algorithm to effectively classify true and false variants. GARFIELD-NGS can be applied in single sample WES analysis and it is particularly effective on 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.

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

Prediction models

To develop a new tool for variant filtering based on neural network machine learning algorithm, we first collected 22 different WES experiments for the NA12878 sample (Supplementary Table S1), 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 comparison with the gold-standard calls provided by Genome in a Bottle Consortium (GIAB). Variants datasets were then randomly splitted in pre-training, training, validation and test sets as described in Material and Methods (Supplementary Table S2).

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 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 S3. AUC values of final models on training and validation sets were > 0.90 for all variants groups but Illumina SNPs, showing a slightly worst performance with AUC almost 0.80 (Supplementary Fig. S1). Predictions of the proposed models resulted stable over 1,000 simulations (see Methods), showing high rate of prediction label concordance and low standard error values (Supplementary Fig. S2).

Features importance

To better understand contribution of the single features, we computed features importances for each prediction models and compared results with features distributions in the corresponding variants dataset. Features importance computed by H2O are reported in Figure 1, while a summary description of all features is provided in Supplementary Table S4. Most features in Illumina INDELs / SNPs and ION INDELs / SNPs revealed low performances in distinguishing false and true calls, as suggested by low r2 values and AUC scores (Supplementary Fig. S3-S7). Only for Illumina INDELs, QD (a specific score developed to predict variants confidence in Illumina dataset) and QUAL (variant quality) features emerged as good single predictors and they are confirmed as the most important features in the corresponding model (Figure 1a). As expected, for most features we observed good concordance between ranking based on variable importance within each model and ranking based on AUC values.

Interestingly, coverage related and strand-bias metrics are usually within the top discriminating

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

variables in each model, such as SOR and FS for Illumina SNPs or SSSB and FDP for ION SNPs and INDELs, respectively (Figure 1). For ION models we observed high impact for platform specific features evaluating the flow-space data, such as MLLD (mean log-likelihood delta per read) and RBI (distance of bias parameters from zero). Other features related to known platform-specific issues also resulted with high importance, such as PBP (position of variants along the reads) for ION SNPs and HRUN (length of homopolymer run) for ION INDELs (Figure 1b,d). Analaysis of features scaled importance over 1,000 simulated models (see Methods) resulted in very low standard errors, suggesting features stability (Supplementary Table S5).

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 (CP) ranging from 0.0 to 1.0, with higher values associated to true variants. 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, while Illumina SNPs model showed slightly reduced performances with test set AUC 0.7998 (Figure 2). Accuracy is > 0.90 for all variants categories. CP values clearly distinguish true from false variants in test set for Illumina INDELs, ION INDELs, and ION SNPs (Figure 3). Differences are smaller between median values for Illumina SNPs: true calls 0.955, false calls 0.926.

Applying the maximum accuracy filtering thresholds, GARFIELD-NGS correctly classifies more than 95% of true variants and reduced false positive variants significantly as shown in Supplementary Fig. S8. Performances at different thresholds are reported in Supplementary Table S6. 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, as shown in Figure 2. 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 (Figure 2a, b).

Comparison between GARFIELD-NGS, hard-filters and VQSR

Variants in our 4 test sets were re-analysed using hard-filters for Illumina[19] and ION[14] data, as described in methods. In Illumina INDELs, ION INDELs and ION SNPs groups, GARFIELD-NGS

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

outperformed hard-filters, showing higher accuracy, while obtaining comparable performances on Illumina SNPs (Figure 4, and Supplementary Table S6).

Largest improvements are seen for INDELs. Accuracy of GARFIELD-NGS reached 0.93 and 0.91 for Illumina and ION INDELs, respectively, compared to 0.86 and 0.80 calculated for previous hard-filters. GARFIELD-NGS confirmed better performances also at 0.99 TPR threshold.

Furthermore, we compared GARFIELD-NGS performances with GATK VQSR filters on Illumina data. VQSR calculated a VQLOD value for each variant and then different filtering thresholds could be set based on desired level of sensitivity. GARFIELD-NGS outperformed GATK VQSR when applied on INDELs variants: VQLOD reached an AUC value of 0.6783, with 0.79 accuracy applying the 99.9 tranche filter, while GARFIELD-NGS reached 0.92 AUC and 0.93 accuracy when applying maximum accuracy threshold. We obtained comparable performances for SNPs: VQLOD AUC value was slightly higher and the two methods showed comparable accuracy, with GARFIELD-NGS recognizing more false positive variants. Comparisons with VQSR filters are reported in Figure 4, and detailed in Supplementary Table S6. VQLOD ROC curve is reported in Supplementary Fig. S9.

Characterization of false positive variants identified by GARFIELD-NGS

Comparing false variants filtered by each method, GARFIELD-NGS models identified most of false variants also reported by hard filters. Meanwhile, the proportion of false variant missed by our method and identified by other filters remains generally low (Supplementary Fig. S10). Overall, GARFIELD-NGS models variants seems less dependent on single variant quality metrics like QD, QUAL and GQ, being able to recognize also false variants with high values that are usually classified as true by hard-filters and VQSR. When dealing with ION variants, our models are less influenced by coverage metric (FDP) and distribution of alternate allele observation between forward and reverse (FSAF, FSAR, STB, STBP) reads, being able to filter also variants with high coverage and balanced representation, that are usually retained by hard-filters. Distributions of features values differentially filtered by GARFIELD-NGS and other filters are reported in Supplementary Fig. S11-16.

Assessment of GARFIELD-NGS on external replication datasets

To verify generalization of our models for practical use, we first applied GARFIELD-NGS on 35 Illumina and 32 ION WES experiments performed on distinct samples. The percentage of variants filtered as false positive by our models is significantly higher when applied to rare variants (ExAC MAF < 10^{-5}) compared to common variants for both Illumina and ION data (p-value 8.84×10^{-11} and

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$6.15x10^{-10}$, respectively), as shown in Supplementary Fig. S17.

To further verify GARFIELD-NGS performances and estimate optimal threshold for practical use, we tested our models on a set of variants validated by Sanger sequencing. We obtained 0.958 and 0.878 accuracy on Illumina INDELs and SNPs, respectively. Test on ION variants resulted in 0.804 and 0.955 accuracy for INDELs and SNPs, respectively. Results of external validation and suggested filtering thresholds are reported in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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[1,13–15] poses serious challenges for downstream data analysis and result interpretation. Even if numerous tools have been proposed to analyse Illumina and/or ION data[27,28], GATK remains the most adopted variant caller for Illumina data, while the Torrent Variant Caller (TVC) is almost the only one adopted for ION data. Both approaches produce a discrete percentage of false positive calls, as we observed in our datasets as well (Supplementary Table S2). Taken singularly, variants features calculated by variants callers showed poor performance in predicting false and true calls (Supplementary Fig. S3-7), suggesting that their integration in a prediction model could be a more effective strategy. Thus, we decided to develop GARFIELD-NGS, a filtering method based on neural networks that integrates variant features reported by GATK or TVC (Supplementary Table S4) and can be applied directly to variant callers output to improve performances of current WES analysis pipelines. Deep neural networks have been widely applied in genomic studies[29] and provided effective solutions to generate predictions from complex data, such as splicing prediction from RNA-Seq[30] or identification of binding domains in DNA or RNA sequences[31,32]. The multilayer-perceptron alghoritm used here is especially effective to extrapolate useful classification when a large number of labeled data area available, as in our training datasets. To developed GARFIELD-NGS, we used WES data obtained by sequencing the NA12878 reference sample (Supplementary Table S1) and determined false and true variants by comparison with the gold-standard calls provided by Genome In A Bottle consortium (GIAB) (Supplementary Table S2). In 2013, GIAB has distributed the first set of gold standard calls for NA12878 sample based on integration of 13 different datasets obtained using different NGS technologies[33]. This constantly updated set of variants is now broadly accepted as a standard for variant identification benchmarking. Given a standard VCF4.2 file, GARFIELD-NGS calculates for each variant a score ranging from 0.0 to 1.0, with higher values associated to true calls (confidence probability, CP). The tool is composed of 4 models, specifically developed on INDELs or SNPs variants coming from Illumina or ION experiments (Supplementary Table S3).

Our method 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 (Figure 2). While hard-filters only perform a boolean classification of

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

variants in true or false categories, GARFIELD-NGS calculates a prediction values ranging from 0.0 to 1.0, with distinct distributions between false and true variants (Figure 3). This allows tuning of variant filtering threshold depending on the desired accuracy and specificity or even integration of CP value as prioritization score rather than variant filter. The maximum accuracy thresholds retain > 95 % of true calls while reducing false calls by 36-80 %, depending on variant category (Supplementary Fig. S8). Even when applying a threshold corresponding to 0.99 TPR, GARFIELD-NGS maintains > 0.86 accuracy (Figure 4 and Supplementary Table S6). 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[34,35]. 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. This hypothesis is supported by the analysis of sequencing context around the identified SNPs as shown in Supplementary Fig. S18.

Nowadays, the most applied strategy for false positive variants filtering on Illumina are the GATK hard-filters and VQSR method[19,20]. Alternative pipelines have been proposed such as GotCloud[36], VarScan[37] and SNPSVM[38], which combine variant calling and variant filtering, or VariantMetaCaller[39] and BAYSIC[40], which integrate results of different variant callers to increase sensitivity and specificity. However, only few tools are available to refine SNPs and INDELs called using the widely adopted GATK. Moreover, these filtering tools are usually developed for specific experimental settings, like tarSVM for microfluidic based sequencing[41], or require additional information or pedigree data to perform variant filtering, such as VarBin[42] and LR[43]. Concerning ION data, widely adopted strategies for variant filtering are lacking, and only few filtering methods are reported[14]. GARFIELD-NGS predictions outperformed hard-filters for Illumina[19] and ION[14] data in 3 variants categories, while results are comparable on Illumina SNPs (Figure 4 and Supplementary Table S6). A strong improvement in variant filtering was observed for INDELs on both Illumina (maximum accuracy 0.9355, TPR 0.9779, FDR 0.06) and ION data (maximum accuracy 0.9117, TPR 0.9542, FDR 0.0707). Even if Illumina SNPs AUC value is lower than those obtained from other models, GARFIELD-NGS performs as well as GATK hard-filters, showing a maximum accuracy of 0.9435, 0.9949 TPR and 0.0535 FDR. Compared to GATK VQSR, GARFIELD-NGS confirmed better performance on INDELs and comparable results on SNPs, with a slight increase in false positive detection (Figure 4, Supplementary Fig. S9).

We compared false positive variants filtered by GARFIELD-NGS and other strategies demonstrating that our models recognize most false calls identified by other filters on both ION and Illumina data (Supplementary Fig. S10). Features distributions in false positive variants recognized

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

specifically by GARFIELD-NGS suggested that our tool can identify also challenging false calls with high coverage and quality, which escape hard-filters and VQSR. Overall, GARFIELD-NGS models seem less dependent on single variant quality metrics like QD, QUAL and GQ and ION models are less influenced by coverage metric (FDP) and distribution of alternate allele observation between forward and reverse reads (FSAF, FSAR, STB, STBP) (Supplementary Fig. S11-16).
When applied on replication WES datasets, the percentage of variants filtered as false positive by our models is significantly higher when applied to rare variants (ExAC MAF < 10^{-5}) compared to common variants for both Illumina and ION data (p-value $8.84x10^{-11}$ and $6.15x10^{-10}$), as shown in Supplementary Fig. S17. This supports the general applicability of our methods, since sequencing errors are expected to generate stochastic alterations presenting as rare or private variants, instead of common polymorphisms. When applied to Sanger validated variants GARFIELD-NGS confirmed to be effective in variants filtering, reaching 0.804 – 0.958 accuracy (Table 1).
Overall, 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[17,18]. GARFELD-NGS can be successfully applied to SNPs filtering as well, with performances comparable to hard-filters or VQSR.
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.

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 (Supplementary Table S1 and Supplementary Fig. S19). 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[19,20]. 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 4.2 files, variants were decomposed, normalized and left aligned using vt tool[44]. Focusing on exome regions, we considered for further analysis only variants

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

located in RefSeq coding exons plus 5bp flanking regions and overlapping high confident regions defined in NIST v.3.3.2 data (ftp://ftp-trace.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[33]. This set of gold standard calls is based on integration of 13 different datasets obtained using different NGS technologies and represents a broadly accepted standard for variant identification benchmarking. Detailed description of variants identified for each experiment is given in Supplementary Table S1.

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. The defined variant datasets consider only variants identified in WES data sources, so that true negative calls are represented only by called variants not present in NIST reference dataset and correctly identified as false by the model. Similarly, false positive variants are represented by calls not present in the NIST reference dataset and not filtered by the model.

Model description

H2O's deep learning method is based on multi-layer perceptron algorithm implemented in a neural network model with feedforward multilayer architecture, trained with stochastic gradient descent using back-propagation. In this model, the weighted combination $\alpha = \sum^{n}_{i=1} w_i x_i + b$ of input signals is aggregated, and then an output signal $f(\alpha)$ is transmitted by the connected neuron. Here, x_i and w_i

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

represent the firing neuron's input values and their weights, respectively. The function *f* represents the nonlinear activation function used throughout the network and the bias *b* represents the neuron's activation threshold. The network contains multiple hidden layers of nonlinearity consisting of numerous interconnected neuron units with Tanh, Rectifier, or Maxout activation functions *f*.
The l1 and l2 regularization parameters are implemented to prevent overfitting. They act modifying loss function to minimize loss: $L'(W, B \mid j) = L(W, B \mid j) + \lambda_1 R_1(W, B \mid j) + \lambda_2 R_2(W, B \mid j)$.
In l1 regularization $R_1(W, B \mid j)$ represents the sum of all l1 norms of the weights and biases in the network. l2 regularization via $R_2(W, B \mid j)$ represents the sum of squares of all the weights and biases in the network. Learning process occurs by tuning weighs to minimize the errors in labeled training data.

Evaluation of features importance

We used variant features reported in VCF file version 4.2 produced by GATK and TVC variant callers to train deep learning algorithms predicting true out of false variants. Detailed description of features is reported in Supplementary Table S4. To estimate contribution of each feature, we analysed their distributions across all variants using logistic regression model to estimate their ability to distinguish false and true calls (Supplementary Fig. S3-S6). Classification capability of each feature on different class of variants was also evaluated using ROC curve (Supplementary Fig. S7). Features importance within each prediction model was calculated using the Gedeon method[45] as implemented in H2O package, reporting scaled variable importances. The scaled importance represents the relative importance across all variables, scaled to 1.

Development of prediction models

We used variant features reported in VCF4.2 file output by GATK and TVC variant callers to train deep learning algorithms predicting true out of false variants. We included 18 features for ION variants and 10 for Illumina variants (Supplementary Table S4). 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-03 stopping tolerance), generating at least 10,000 different models. Models were ranked according to cross-validation AUC and the best five hyper-parameters

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 (Supplementary Fig. S20). 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 S3. Finally, GARFIELD-NGS prediction performance for each variants category was evaluated on test sets using the corresponding model.

Features and models stability

Stability were assessed by 1,000 simulations for each of the four models. In each simulations we removed a random 1% of the original training set and then performed model training with the same parameters used for the original model. To evaluate models stability, we analysed concordance of predictions on test sets across the 1,000 simulated models, measuring concordance of prediction label and standard error of the output value. Features stability were assessed measuring the standard error of scaled importance for each feature across the simulated models.

Comparison with hard-filters and VQSR

Variants in our 4 test sets were re-analysed using hard-filters for Illumina, as described in GATK best practises[19],and ION[14] 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.

Variant Quality Score Recalibration (VQSR) was applied separately to VCF files from each sample according to parameters described in GATK best practises for WES experiments[19], to reproduce filtering on single samples. This method reports for each variant a VQLOD value and it can be applied as hard filter, choosing the desired filtering tranche. We compared performances of VQSR and GARFIELD-NGS on the test sets variants, considering both VQLOD value distribution or the 4 suggested hard-filtering thresholds (100, 99.9, 99, 90 tranches). VQSR could not be applied effectively to ION data, since the VCF file produced by TVC lack several features evaluated by this filtering method.

GARFIELD-NGS validation on external dataset

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To test GARFIELD-NGS on completely independent data, we assessed how our models filter variants from WES data not processed by our pipeline. We obtained VCF files for 35 Illumina (mean coverage >60 X) and 32 ION (mean coverage > 90 X) WES experiments. Illumina data were generated on either HiSeq 2000 or HiSeq 1000, using Agilent SureSelect All Exon v4 or v5 kits for exome capture. Variants were identified using GATK v.3.3 or v.3.4 following GATK best practices. ION data were generated on Ion Proton, using Hi-Q chemistry and AmpliSeq Exome RDY kit for exome capture. Variants were identified using TVC v.5.0 or v.5.2. Variants were decomposed, left aligned and normalized using vt tool and then annotated with MAF in human population from ExAC v0.3.1. We analysed the percentage of rare (MAF < 10^{-5}) and common variants that were filtered by our models using the max accuracy thresholds in each sample.

To validate GARFIELD-NGS and estimate optimized thresholds for practical usage, we applied our models to a set of external variants that were previously validated by Sanger sequencing. We collected 65 (41 SNPs, 24 INDELs) and 101 (67 SNPs, 34 INDELs) variants from 95 and 46 different samples for Illumina and ION based sequencing, respectively (Table 1). Illumina variants are derived from gene panel target resequencing (86 samples) and WES experiments (9 samples). Gene panel data were generated on MiSeq platform using TruSeq custom amplicon assay (64 samples) or Nextera rapid capture assay (22 samples) for target region capture. Samples resulted in mean coverage between 70 and 650X. WES data were generated on HiSeq 2000 platform using Agilent SureSelect All Exon v4 / 5 for target region capture. Samples resulted in mean coverage between 70 and 650X.Variants were aligned to hg19 reference genome using BWA and variants were identified using GATK UnifiedGenotyper. ION data were generated on Ion Proton platform, using Hi-Q chemistry, PI v3 chip and AmpliSeq Exome RDY kit for exome capture. Samples resulted in mean coverage between 80 and 120X. Sequencing reads were aligned to hg19 and variants were identified using TVC v.5.2. Based on results from Sanger sequencing, we assessed the performance of our models to distinguish false and true variants and estimated optimal thresholds for variant filtering.

Data availability and implementation

The Illumina and ION datasets analysed in the present study are available from the SRA archive repository or Thermo Fisher Cloud as described in Supplementary Table S1. Releases are freely available at: https://github.com/gedoardo83/GARFIELD-NGS

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. 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

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

15. Jiang, Y., Turinsky, A. L. & Brudno, M. The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. *Nucleic Acids Res.* **43,** 7217–28 (2015).

16. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genet. Med. **15,** 733–47 (2013).

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

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

19. 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).

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

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

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

23. 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).

24. 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).

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

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

27. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5,** 17875 (2015).

28. Cornish, A. & Guda, C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *Biomed Res. Int.* **2015,** 456479 (2015).

29. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12,** 878 (2016).

30. Leung, M. K. K., Xiong, H. Y., Lee, L. J. & Frey, B. J. Deep learning of the tissue-regulated splicing code. *Bioinformatics* **30,** i121-9 (2014).

31. Liu, F. *et al.* De novo identification of replication-timing domains in the human genome by deep learning. *Bioinformatics* **32,** 641–9 (2016).

32. Zhang, S. *et al.* A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* **44,** e32 (2016).

33. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–51 (2014).

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

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

36. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25,** 918–25 (2015).

37. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–5 (2009).

38. O'Fallon, B. D., Wooderchak-Donahue, W. & Crockett, D. K. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics* **29,** 1361–6 (2013).

39. Gézsi, A. *et al.* VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics* **16,** 875 (2015).

40. Cantarel, B. L. *et al.* BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics* **15,** 104 (2014).

41. Gillies, C. E. *et al.* tarSVM: Improving the accuracy of variant calls derived from microfluidic PCR-based targeted next generation sequencing using a support vector machine. *BMC Bioinformatics* **17,** 233 (2016).

42. Durtschi, J., Margraf, R. L., Coonrod, E. M., Mallempati, K. C. & Voelkerding, K. V. VarBin, a novel method for classifying true and false positive variants in NGS data. *BMC Bioinformatics* **14 Suppl 13,** S2 (2013).

43. Hwang, K.-B. *et al.* Reducing false-positive incidental findings with ensemble genotyping and logistic regression based variant filtering methods. *Hum. Mutat.* **35,** 936–44 (2014).

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

45. Gedeon, T. D. Data mining of inputs: analysing magnitude and functional measures. *Int. J. Neural Syst.* **8,** 209–18 (1997).

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. We acknowledge dr. Valeria Cinquina for technical support in Sanger sequencing validations.

Author Contributions

VR performed ROC curve analysis. EG conceived the study, performed NGS data analysis, variant calling and statistical analysis. AL provided Sanger validated variants for Illumina models validation. MR provided Sanger validated variants for ION models validation. VR and EG developed prediction models. VR and EG wrote the manuscript and prepared figures.

Competing financial interests

The author(s) declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure Legends

Figure 1. Features importances in GARFIELD-NGS models

Importance of each feature in prediction models is reported as scaled importance in Illumina INDELs (a), ION INDELs (b), Illumina SNPs (c), and ION SNPs (d) models. The scaled importance represents the relative importance across all variables, scaled to 1.

Figure 2. 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 3. Distributions of GARFIELD-NGS score for true and false variants

GARFIELD-NGS models assign a score from 0.0 to 1.0 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. True and false distribution are significantly different in all groups (t-test p values < 2.2e-16).

Figure 4. Comparison between GARFIELD-NGS and hard-filters

Performances of GARFIELD-NGS, hard-filters and VQSR were compared for Illumina (a) and ION (b) datasets, reporting accuracy, true positive rate (TPR) and specificity (left to right in each panel).

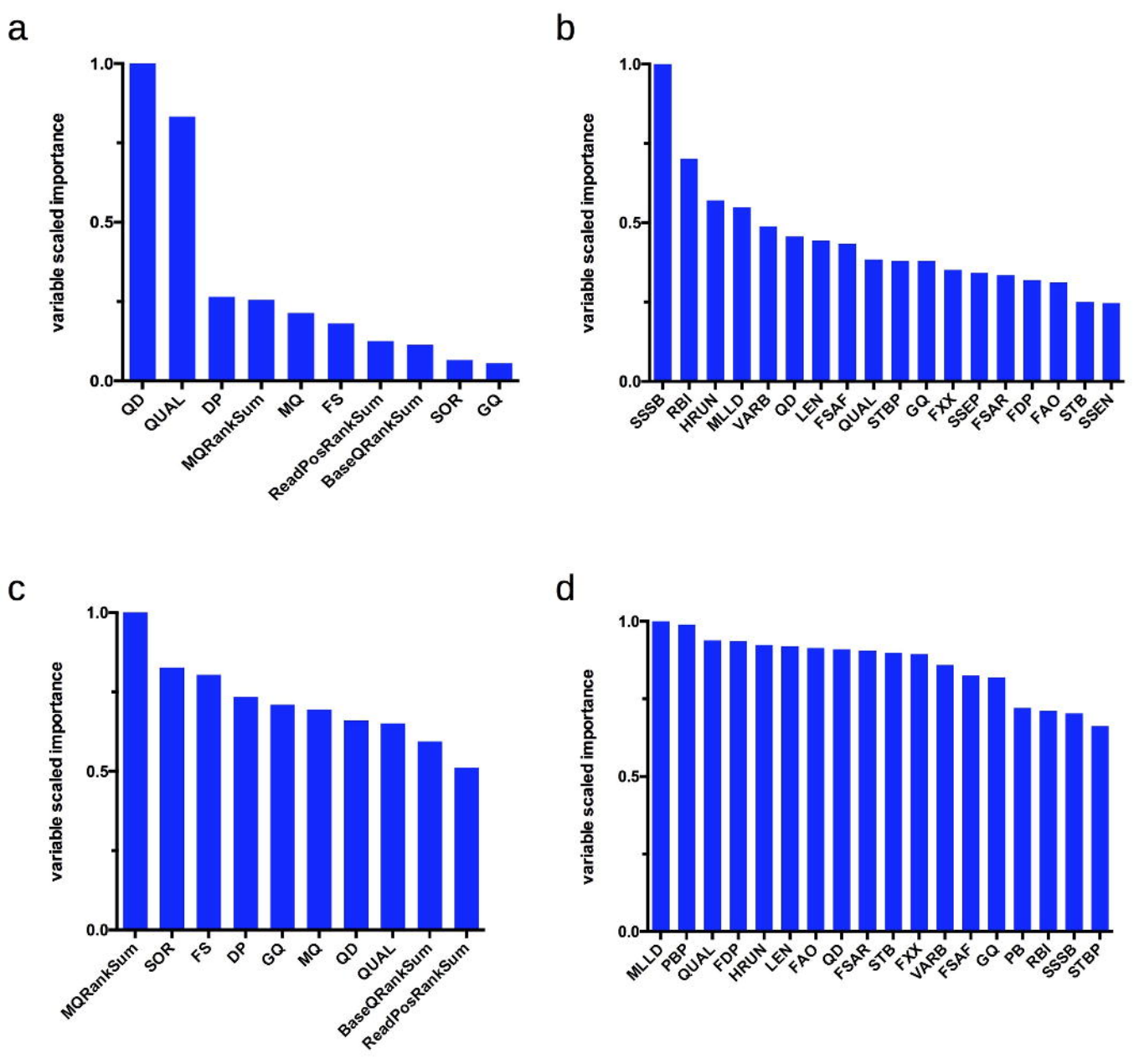
bioRxiv preprint doi: https://doi.org/10.1101/149146; this version posted September 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

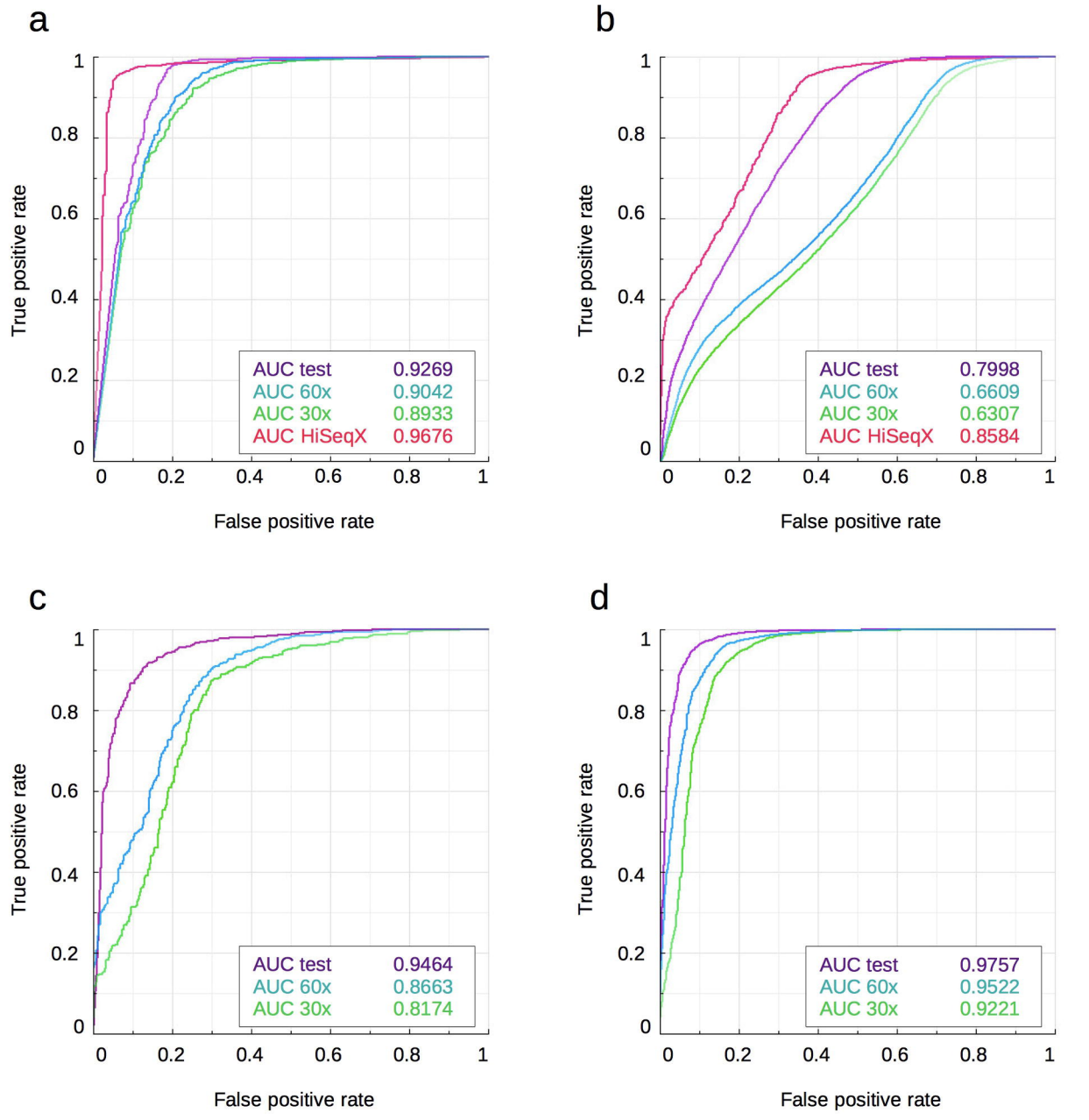
Tables

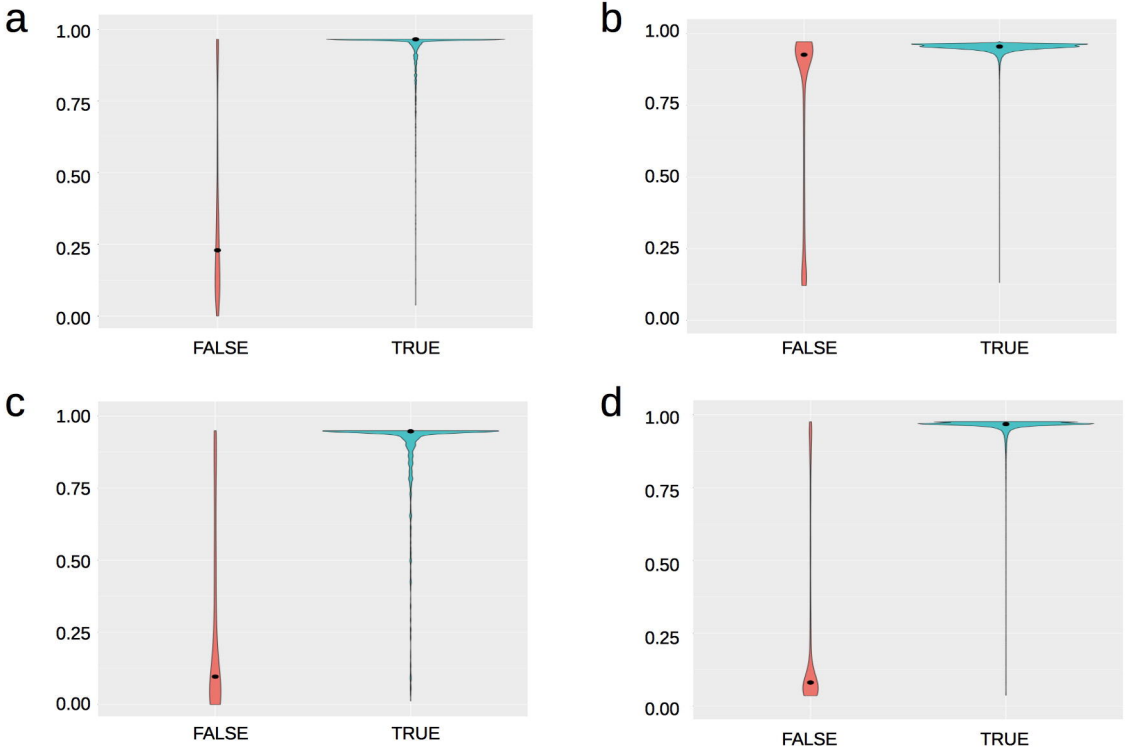
Table 1. GARFIELD-NGS performance on independent replication data.

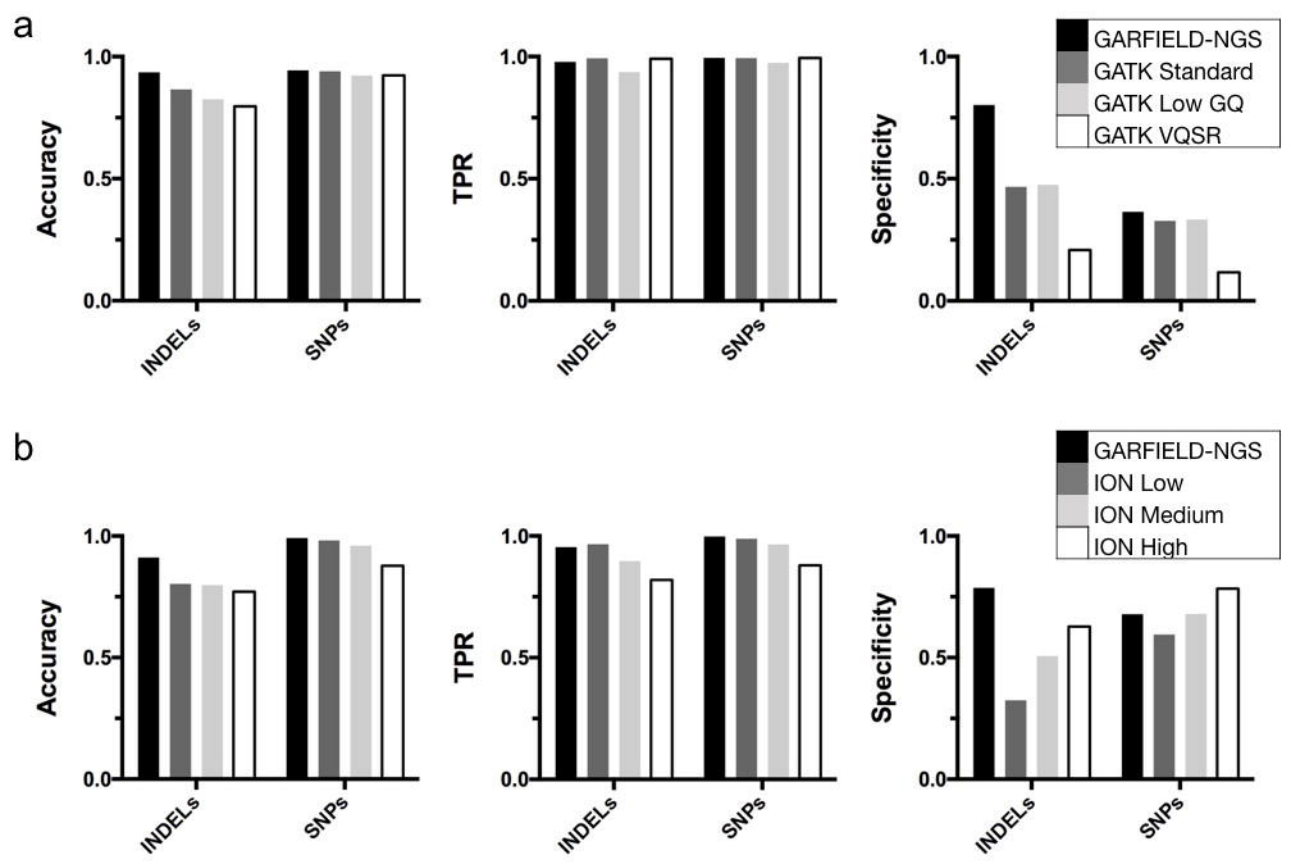
Performance of GARFIELD-NGS prediction when applied to external variants validated by Sanger sequencing. The number of samples and the number of true / false calls in each category is reported, together with the optimal threshold for filtering. Variants with CP < threshold are classified as false. TPR: true positive rate, TNR: true negative rate.

Platform	Samples	Category	TRUE	FALSE	Threshold	Accuracy	TPR	TNR
Illumina	95	SNPs	17	24	0.025	0.878	0.941	0.833
		INDELs	13	11	0.630	0.958	0.923	1.000
ION	46	SNPs	37	30	0.139	0.955	0.972	0.933
		INDELs	12	22	0.320	0.804	0.916	0.727


a
variable scaled importance
1.0
0.5
0.0
QD
QUAL
DP
MQRankSum
MQ
FS
ReadPosRankSum
BaseQRankSum
SOR
GQ
b
variable scaled importance
1.0
0.5
0.0
SSSB
RBI
HRUN
MLLD
VARB
QD
LEN
FSAF
QUAL
STBP
GQ
FXX
SSEP
FSAR
FDP
FAO
STB
SSEN
c
variable scaled importance
1.0
0.5
0.0
MQRankSum
SOR
FS
DP
GQ
MQ
QD
QUAL
BaseQRankSum
ReadPosRankSum
d
variable scaled importance
1.0
0.5
0.0
MLLD
PBP
QUAL
FDP
HRUN
LEN
FAO
QD
FSAR
STB
FXX
VARB
FSAF
GQ
PB
RBI
SSSB
STBP


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


a
Accuracy
TPR
Specificity
1.0
0.5
0.0
INDELs
SNPs
GARFIELD-NGS
GATK Standard
GATK Low GQ
GATK VQSR
b
Accuracy
TPR
Specificity
1.0
0.5
0.0
INDELs
SNPs
GARFIELD-NGS
ION Low
ION Medium
ION High