FANCY: Fast Estimation of Privacy Risk in Functional Genomics Data

Gamze Gürsoy[1,2], Fabio C.P. Navarro[1,2] and Mark Gerstein*[1,2,3]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[3]Department of Computer Science, Yale University, New Haven, CT 06520, USA

September 18, 2019

*pi@gersteinlab.org; Corresponding Author

Abstract

Functional genomics sequence data is becoming clinically actionable, raising privacy concerns. However, quantifying the privacy leakage by genotyping is difficult due the heterogeneous nature of the sequencing techniques. Thus, we present FANCY, a tool that rapidly estimates number of leaking variants from raw RNA-Seq, ATAC-Seq and ChIP-Seq data, without explicit genotyping. FANCY employs supervised regression using overall sequencing statistics as features and provides an estimate of the overall privacy risk before data release.

With the surge of genomics data and the decreasing cost of sequencing [1], genome privacy has become an important topic of interest. Traditional DNA sequencing, functional genomics [2], and molecular phenotype [3, 4] datasets create a great number of quasi-identifiers, which, in turn, can be used to re-identify or characterize individuals without their consent. The surge in different types of widely available functional genomics data increases the correlations between phenotype and genotype datasets, which in turn increases the possibility of re-identifying individuals. Functional genomics data allows for a detailed characterization of disease states and susceptibility. Broad dissemination of this data can promote advances. Although functional genomics experiments are not performed for genotyping purposesrather, the aim is to study phenotypes and basic biologyexperimental procedures have led to the raw reads containing a substantial amount of patients' variants, which raises privacy concerns. In contrast to DNA sequencing data, few tools can currently assess the risk of potential privacy loss from functional genomics data. Rapidly quantifying the number of leaking variants from functional genomics data is essential before the release of the data. This is particularly important when multiple assays are performed on samples from the same individuals. That is, a single data type may not leak enough variants, but a combination of different functional genomics data can pose significant privacy risk as different assays target different regions in the genome with different coverage profiles [e.g., RNA sequencing (RNA-Seq) targets expressed exons, whereas H3K27ac chromatin immunoprecipitation sequencing (ChIP-Seq) targets the noncoding genome on the promoter and enhancer regions] and depth profiles (i.e., some assays have spread out peaks while others are more punctuate). Such quantification is possible by genotyping the raw sequences and overlapping them with the gold-standard genotypes from the individuals as many of the inferred genotypes could be false positives. The limitations of such an approach is the requirement of large resources for genotyping as well as the presence of a gold-standard genotype data belonging to the patient. While it is possible to genotype the raw reads with current genotyping tools, an average variant calling pipeline needs to be radically re-parametrized to better fit with different assays given that software is typically optimized for whole-genome sequencing.

Table 1: The maximum and minimum number of variants leaked in each experiments and the RMSE of our predictions in these test datasets.

Assay	number of test datapoints	max. number of variants	min. number of variants	RMSE
RNA-Seq	4740	12,928	379	422.76
ATAC-Seq	6753	238,481	65	4503.10
ChIP-Seq	1082	210,567	2665	5381.22

In this study, we developed a supervised learning method to infer the number of leaking single nucleotide variants (SNVs) from raw functional genomics data. Our primary goal was to quantify the SNV leakages in raw sequences without needing genotyping and a gold-standard genotype list. We built a Gaussian Process Regression model that takes the assay type, sequencing features such as depth ($\overline{d}$) and breadth (b) of the coverage, and the statistical properties of the depth distribution such as standard deviation(σ), skewness ($\pm s$), and kurtosis(k) as input. The model then predicts the cumulative number of leaking SNVs with an R^2 of 0.99 for training and 0.90 for independent test RNA-Seq, 0.99 for independent test Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq), and 0.99 for independent test ChIP-Seq datasets (Figure 1). FANCY can separately estimate the number of rare and common variants and outputs each estimated number with a predicted upper and lower bound in the 95% confidence level. FANCY also outputs a simple qualitative warning message for the risk of sharing (green: can be shared, yellow: attention needed, red: cannot be shared). We used mean squared error as our loss function in the regression model [see Table 1 for Root Mean Squared Error (RMSE)]. Our predictions are in great agreement with the true number of leaking variants in all the independent test datasets (Figure 1). To easily interpret the performance of our predictions, we calculated the deviation from the true values by calculating (predicted value - true value) / true value; the negative values indicate underprediction, i.e number of predicted SNVs are lower than the true number of leaking SNVs (Supplementary Figure 1). On average, we had 8% prediction error for all of the independent test sets.

If a functional genomics experiment is leaking more than 1,000 variants, its privacy risk for re-identification is at maximum regardless of the absolute value of the number of variants. However, the risk assessment is more valuable for experiments that are leaking a low number of variants as mis-predicting these values by only a few might result in the release of private data. Thus, we developed another regression model (see Online Methods) that aims to predict the leakage more precisely when the number of leaking variants is low. This second model had an RMSE of 75.64, 74.8, and 74.0 for independent test RNA-Seq, ATAC-Seq, and ChIP-Seq datasets, respectively, in which the maximum number of leaking variants is 1,000 (Supplementary Figure 2). We also calculated the number of under-predicted (predicted value is lower than true value) leaking variants and found that we have no under-predicted leaking variants that are smaller than 400 and only three under-predicted leaking variants between 400 and 500.

We used RNA-Seq data from 432 individuals generated by the gEUVADIS project [5], H3K27ac ChIP-Seq data from 100 individuals generated by the PsychENCODE Consortium [6], ATAC-Seq data from 344 individuals generated by the BrainGVEX project[6], and ATAC-Seq data from 288 individuals generated by the PsychENCODE Consortium[6]. We then used the GATK Best practices from RNA-Seq and DNA data [7, 8] to call SNVs and small insertions and deletions (see Supplementary Figure 3 for statistics). We treated each chromosome separately, which resulted in 25,152 data points. We randomly divided the data in half to use as training and test sets. We also separately validated our model by using 308 data points from the RNA-Seq study by Kilpinen et al. [9].

Genotyping from DNA sequences is the process of comparing the DNA sequence of an individual to that of the reference human genome. To be able to successfully genotype, one needs a substantial depth of sequencing reads for each base pair. According to the Lander-Waterman statistics for DNA sequencing, when random chunks of DNA are sequenced repeatedly, the depth per base pair follows a Poisson distribution with a mean that can be estimated from the read length, number of reads, and the length of the genome [10]. For example, as RNA-Seq aim to sequence expressed genes, one would expect that sequencing per depth per base pair does not follow Poisson statistics. Thus, genotyping using reads from RNA-Seq experiments is biased towards variants that are in the exonic regions. Conversely, ChIP-Seq is biased against RNA-Seq, where it targets non-coding genome such as promoters and enhancers (see Figure 1).

We hypothesized that the statistical properties of the depth per base pair distribution are strong indicators of the number of variants that can be inferred from functional genomics data. We used a total of six sequencing features: 1) the average depth per base pair ($\overline{d}$); 2) the total fraction of the genome that is represented at least by one read (i.e., the breadth, $b = \sum_{i=1}^{N} \delta(d_i)$, such that $\delta(d_i) = 1$ if $d_i > 0$, $b = 0$ otherwise and N is the total number of nucleotides in the genome); 3) the standard deviation of the depth distribution; 4) skewness (i.e., whether the distribution is larger on the right or left side of the mean); 5) kurtosis (i.e., whether or not the depth distribution has big tails); and 6) the type of the experiment (i.e., RNA-Seq, ATAC-Seq, or ChIP-Seq).

We did a leave one feature out test and found that the mean depth ($\overline{d}$) and breadth (b) of the coverage had the greatest effect on the performance of the predictor (see Supplementary Figure 4). We then created predictors by using only (1) mean depth, (2) breadth, and (3) mean depth and breadth as the features. However, this predictor performed worse than the original model. These results show that although breadth is the highest contributing feature, all of our features contributed to the final model; indeed, the RMSE is the lowest when we use all of the features.

In addition to estimating privacy risk, FANCY can also be used to plan functional genomics experiments (i.e., for a target number of SNVs, one can back-calculate the required sequencing statistics). Moreover, we also developed a Random Forest classifier (Supplementary Methods) as a plug-in that predicts the type of the assay (RNA-Seq vs. ATAC-Seq vs. ChIP-Seq) by using the sequencing statistics as features, which can be useful to the community for samples with missing metadata. This classifier has an average accuracy of 96.8%, precision of 94.9%, recall of 90.2%, and F1 score of 93.3% (Supplementary Figure 5).

Figure

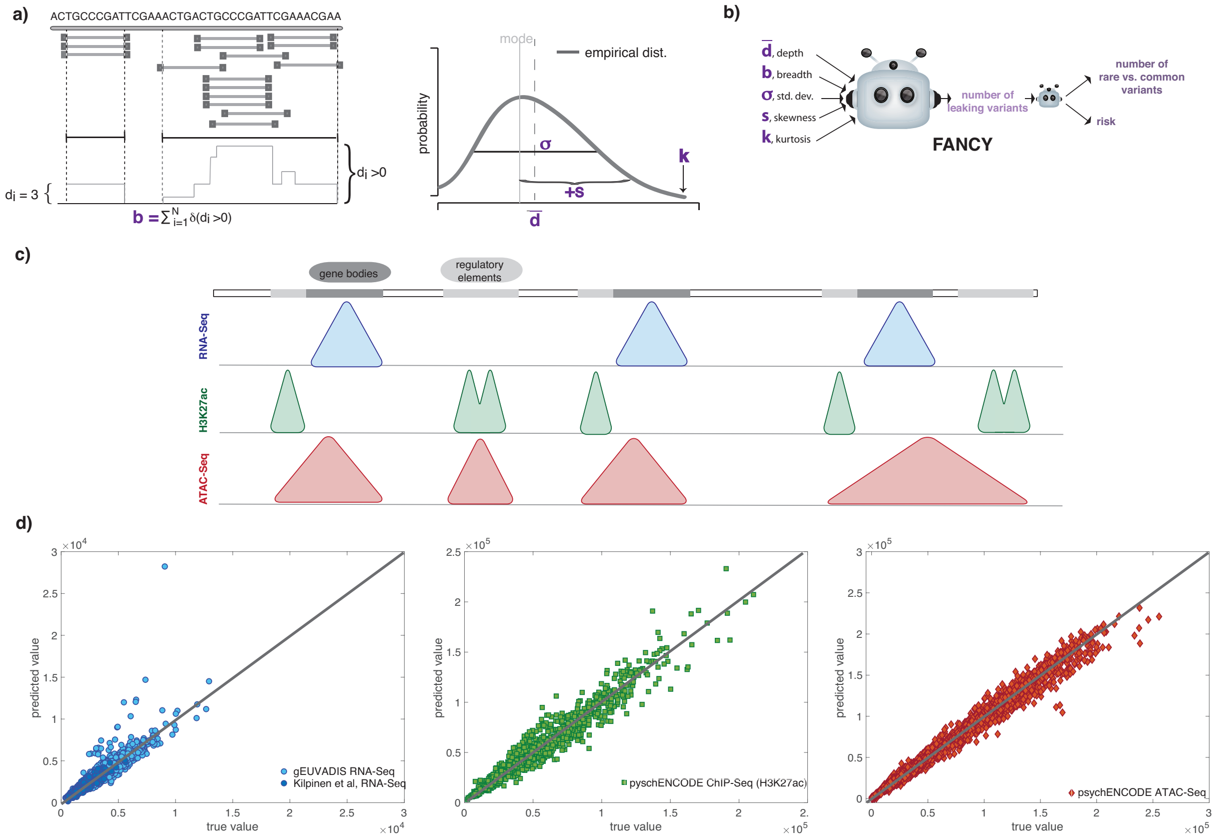


Figure 1: **Details of the FANCY and the results.** (**a**) The schematic of the features of FANCY: average depth; breadth, i.e number of nucleotides represented with at least one read; first, second and third moment of the depth distribution (standard deviation, skewness and kurtosis). If the distribution is skewed to the right hand side of the mode (mean is larger than the mode), then the skewness is positive, it is negative if the mean is smaller than the mode. (**b**) The schematic of inputs and outputs of FANCY. (**c**) The regions that are represented by each assay type. The reads of RNA-Seq are concentrated on the gene bodies, H3K27ac ChIP-Seq is concentrated on the non-coding genome (enhancers and promoters) and since ATAC-Seq covers the open chromatin, the reads are concentrated on both coding and non-coding rehions. (**d**) The performance of FANCY on the independent test datasets.

Methods

FANCY Details:

FANCY is a two-step method. The first step is a regression framework that uses a Gaussian Process Regression (GPR) model with a matern kernel [11]. We tried other regression models such as linear regression with and without interactions, different regression trees, and support vector machine (SVM) regression models. GPR outperformed other models in training (both in terms of RMSE and R^2; see Supplementary Table 1). We obtained the features as follows: We first aligned the raw functional genomics reads to the reference genome (bwa [12] is used for ChIP and ATAC-seq data; STAR [13] is used for RNA-Seq data). We then calculated the depth per base pair using samtools [14] and calculated the statistics. For the true number of SNVs, we used GATK [7, 8] (with appropriate parametrization for each assay type) to call the SNVs. After filtering low-quality SNVs as suggested by the best practice, we overlapped the remaining SNVs with the gold-standard SNVs generated from whole genome sequencing data to obtain the true number of SNVs. The second step is the estimation of rare vs. common variants. We divided the 1,000 Genomes data into rare and common categories based on minor allele frequency. For each individual in the 1,000 Genomes project, given an assay, we found the the rare variant density and used the mean density of all individuals with the predicted number of total leaking variants to estimate the number of rare vs. common variants. FANCY$_{\text{low}}$ is the version of FANCY that was trained on data with up to 1,000 SNV leakage.

Gaussian Process Regression:

GPR is a supervised learning method that is based on learning fitting functions to a given set of training data; in comparison, traditional regression models learn the parameters of a given function. GPR is a nonparametric method used to calculate the probability distribution over all functions that fit the data instead of calculating the probability distribution of a specific functions parameters. The advantage of using GPR is its ability to provide uncertainty estimations at a given confidence level. The disadvantage of this method is the computational complexity, which makes it unfeasible for large datasets. Since the sequencing statistics relate to the number of inferred variants differently in different regimes and for different assays, the relationship between features and the number of leaking variants cannot be modeled by general mathematical approaches such as generalized linear models (see Supplementary Figure 7). A Gaussian process can be defined by its mean and covariance functions as

$$f(x) \sim (\mu(x), \sum(x))$$

A Gaussian process assumes that the distribution of the values of functions $p(f(x_1), f(x_2), ..., f(x_N))$ at a set of points $(x_1, x_2, ..., x_N)$ is jointly Gaussian with a mean $\mu(x)$ and covariance $\sum(x)$, where $\sum_{ij} = k(x_i, x_j)$. K is a kernel function, which determines the similarity between data points x_i and x_j. If these points are deemed similar by the kernel, we expect the output of the functions at these points to be similar as well. For each x_i, yi in our training dataset, we can write a function $f(x_i)$ such that

$$y_i = f_i(x_i) + \varepsilon_i$$

,where $\varepsilon_i \sim N(0, \sigma^2)$. Therefore, for any input vector $(x_1, x_2, ..., x_N)$, $f(x)$ has a joint Gaussian distribution. The covariance (kernel) function k is generally taken as Gaussian (i.e., squared exponential kernel). However, in this application, we found that a Matern kernel performs better. We used fivefold cross-validation to avoid overfitting and a separate test dataset to validate our model.

Random Forest Details:

Random Forest classifiers combine several decision trees that use multiple subsets of data from a training sample to produce better predictive performance versus a single decision tree. The advantage of a Random Forest classifier is that it handles high dimensionality in data as well as missing values. It works via the following principles: assume we have an observation y_i and the feature associated with it is $x_{i,j}$. Here, $i = 1,..,N$, $j = 1,...,M$ and N and M are the number of observations and features, respectively. We first take a subset from N number of training data randomly with replacement. We then take a subset of M features randomly. We split the node iteratively by finding the feature associated with the best split. With this iteration, we grow the largest tree. We then repeat these steps to aggregate n number of trees to create our Random Forest. We generated 30 trees using a five-fold cross-validation and an independent test set to validate our model. We also compared the performance of the Random Forest Classifier to that of a single decision tree, logistic regression with linear and quadratic discriminants, SVM and k-nearest neighbor, and found that the Random Forest resulted in the best accuracy (Supplementary Table 2).

Dataset:

In total, we had 13,537 data points from ATAC-Seq, 9,456 data points from RNA-Seq, and 2,159 from ChIP-Seq. In order to understand whether an imbalance in the number of categories affected our model selection, we repeatedly sub-sampled 2,159 data points from the RNA-Seq and ATAC-Seq categories and trained multiple regression models. We found that GPR is the best regression model in each case (Supplementary Table 3).

Implementation and Data:

We used the MATLAB Statistical and Machine Learning Toolbox as well as pyGPs Python package [15] for Gaussian Process Learning. A python and MATLAB implementation of FANCY, FANCY$_{\text{low}}$, Random Forest Classifier, as well as the custom scripts to generate the features can be found at https://github.com/gersteinlab/FANCY. We also provide jupyter notebooks so that users can optimize the parameters in the regression model based on their own data.

References

[1] Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biology*, 2011;12(8):125.

[2] Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. *Nature Communications*, 2018

[3] Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature*, 2012;44(5):603-608.

[4] Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 2016;13(3):251-256.

[5] Lappalainen T et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* , 2013;501:506-511

[6] Wang D et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* , 2018;362:6420

[7] DePristo M et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 2011;43(5):491-498.

[8] Van der Auwera GA et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 2013;43:11.10.1-33.

[9] Kilpinen H et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* , 2013;342:744747

[10] Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 1988;2(3):231-239.

[11] Rasmussen CE, Williams CKI. Gaussian Processes for machine learning. *The MIT Press*, 2006

[12] Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 2009;25:1754-1760

[13] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013;29(1):15-21.

[14] Li H et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009;25(16):2078-2079.

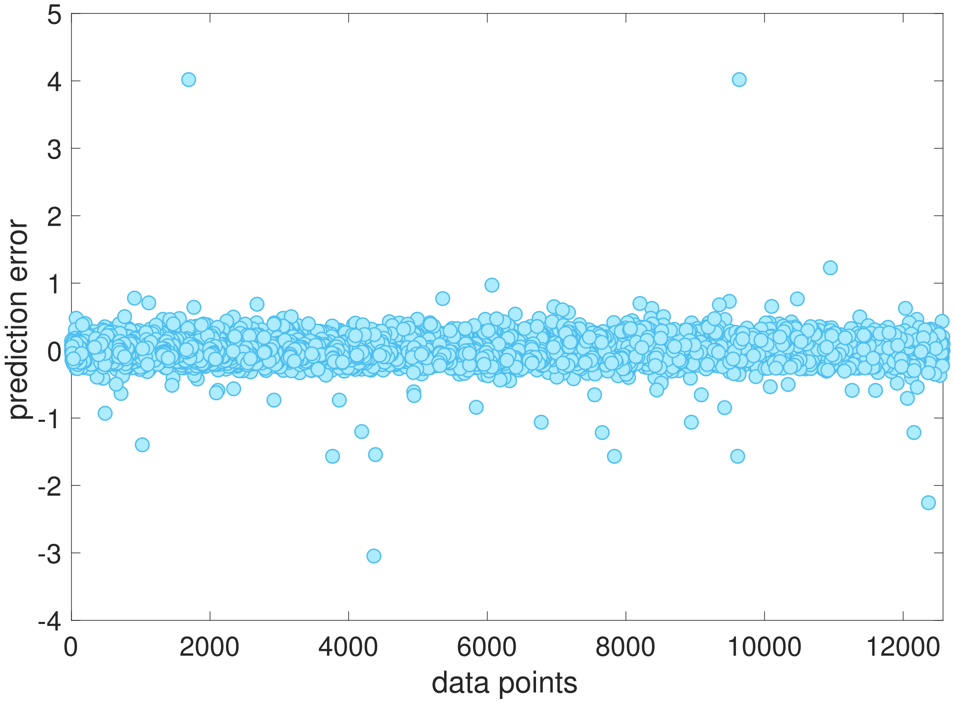
[15] Neumann M, Huang S, Marthaler DE and Kersting K pyGPs – A Python Library for Gaussian Process Regression and Classification. *Journal of Machine Learning Research*,2015;16:2611-2616.

1 Suplementary Information

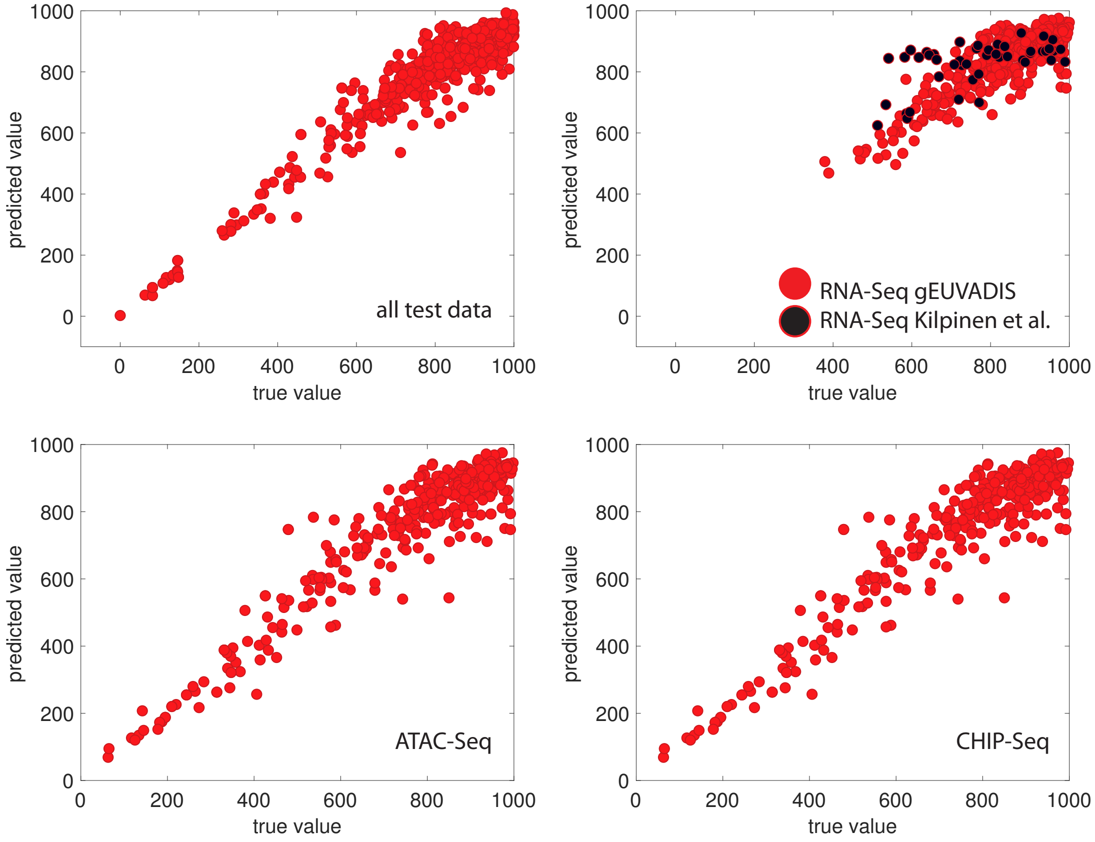
Contents

List of Figures

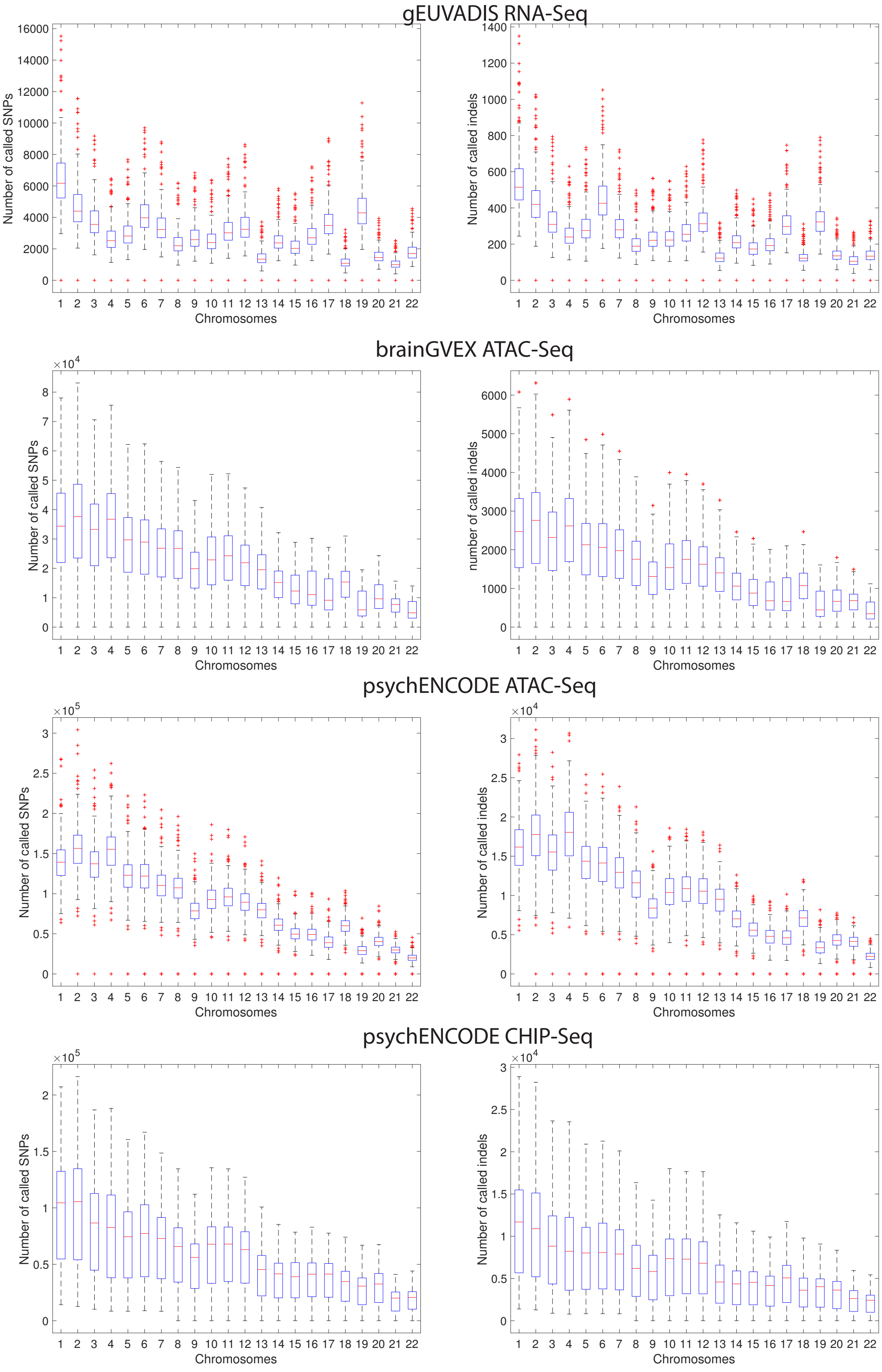
List of Tables



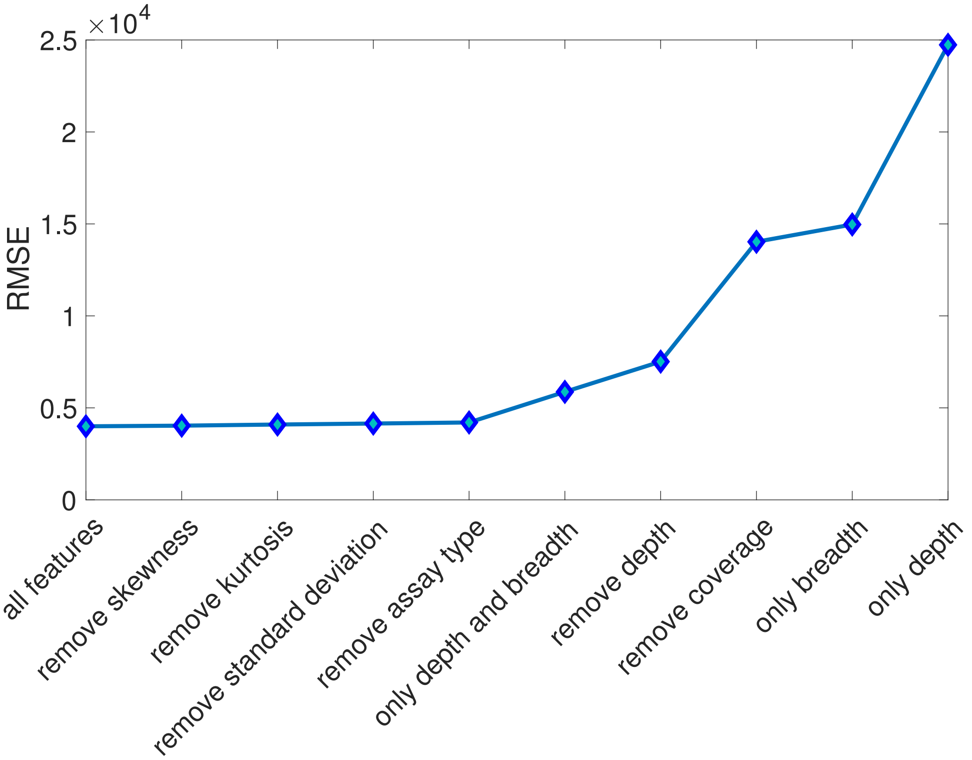
Supplementary Figure 1: **The ratio of the error in the predictions with respect to true values. Negative values indicate that the predicted values are lower than the true values, positive values indicate the predicted values are greater than the true values. Zero indocates perfect prediction.**



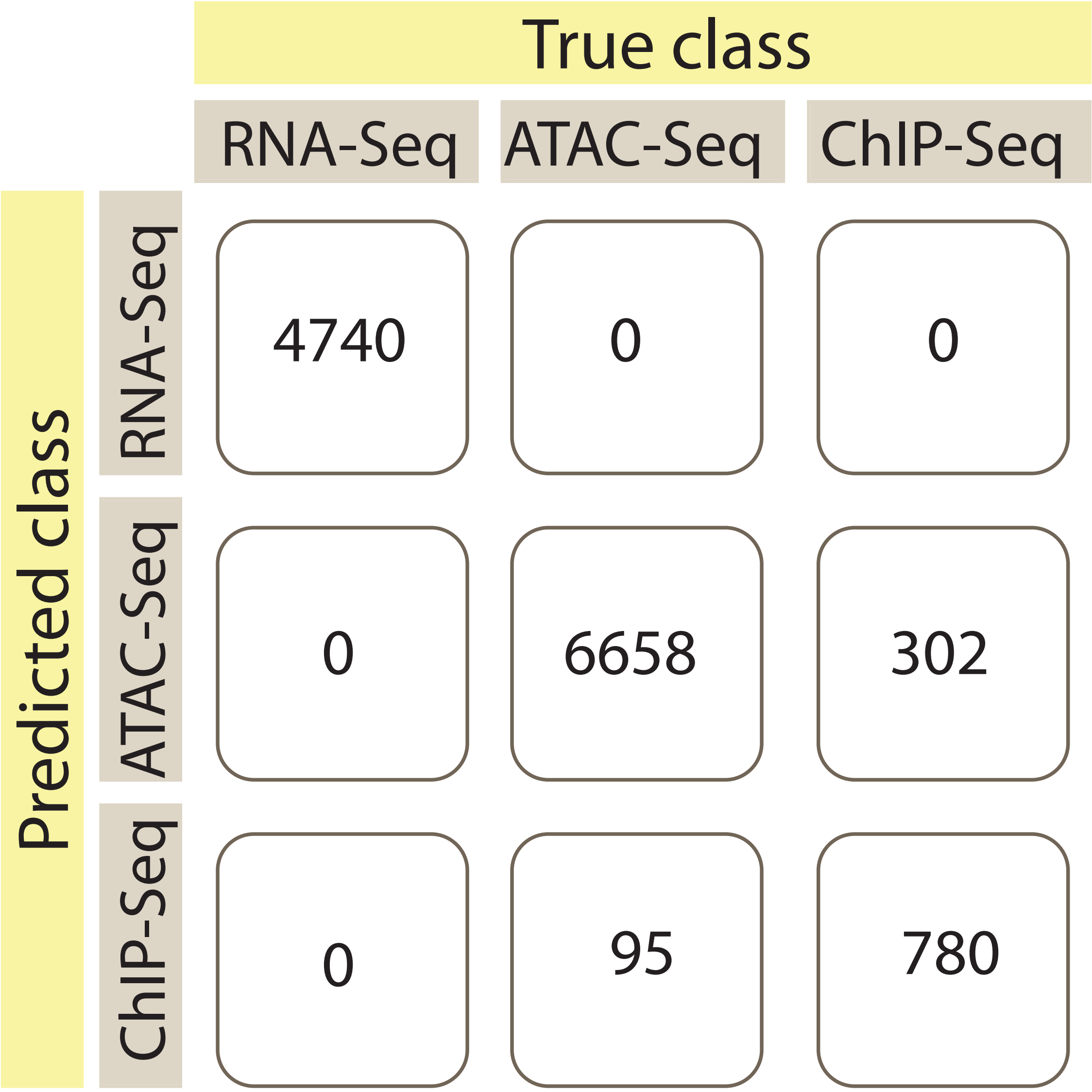
Supplementary Figure 2: **Performance of** FANCY$_{\text{low}}$ **for number of variants less than 1000. Performance for all the test dataset and also seperately for RNA-Seq, ATAC-Seq and CHIP-Seq are shown.**



Supplementary Figure 3: **Distribution of number of called SNPs and indels from functional genomics data.**

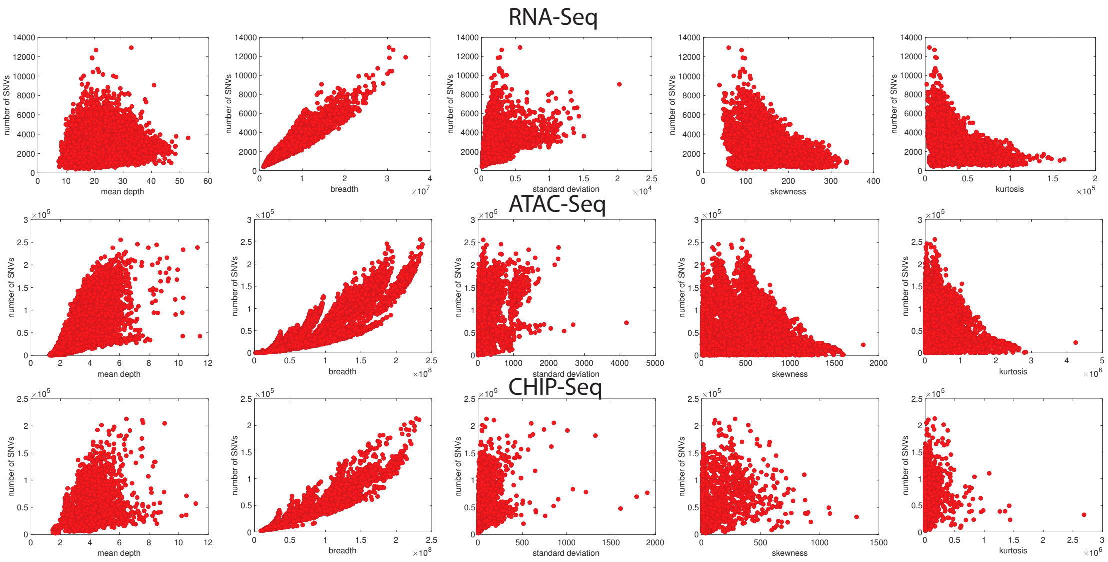


Supplementary Figure 4: **Sorted contribution of different features to the overall predictions.** Lower RMSE corresponds to better predictions. The difference between the first five points on the figure is smaller compared to other data points, however, lowest RMSE is still observed when using al the features in the model.

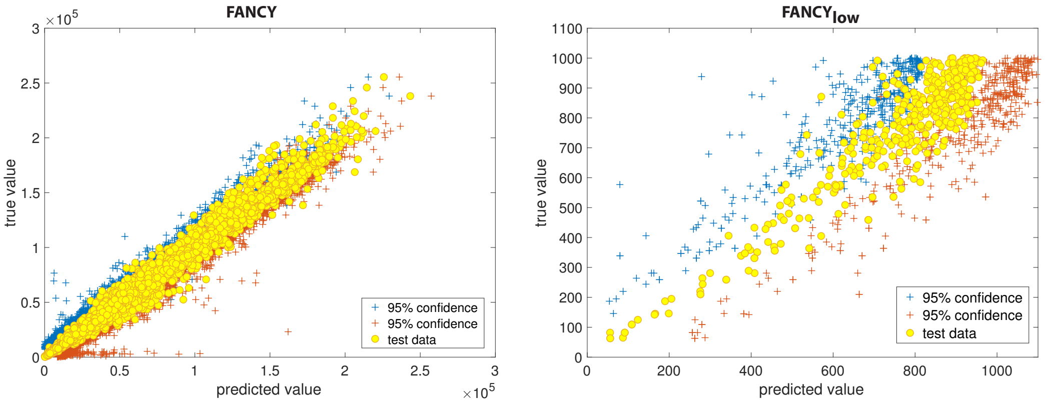


		True class		
		RNA-Seq	ATAC-Seq	ChIP-Seq
Predicted class	RNA-Seq	4740	0	0
	ATAC-Seq	0	6658	302
	ChIP-Seq	0	95	780

Supplementary Figure 5: **Random Forest classifier performance.** A Random Forest Classifier is developed to predict the type of the experimental assay by using sequencing statistics as features.



Supplementary Figure 6: **Relationship between sequencing statistics and number of leaking variants.**



Supplementary Figure 7: **Predicted vs. true values within 95% confidence level.**

Supplementary Table 1: Comparion of different regression model performances.

Model	RMSE	R^2
Linear Regression	15831	0.85
Decision Tree	5616	P0.98
SVM	6696	0.97
Random forest	4990	0.99

Supplementary Table 2: Comparion of different classification model performances.

Model	RMSE	R^2
Linear Regression	15831	0.85
Decision Tree	5616	P0.98
SVM	6696	0.97
Random forest	4990	0.99

Supplementary Table 3: Comparion of different regression model performances when the data is sub-sampled.

Model	RMSE	R^2
Linear Regression	11003	0.88
Decision Tree	7898	P0.94
SVM	11722	0.86
Random forest	6932	0.95
GPR	5544.3	0.97