A filtering strategy to reduce reference bias in measurements of allele-specific expression

Joshua Bradley
Advisor: Zia Khan

Abstract
Understanding how genetic variation leads to phenotypic variation is a fundamental goal in genetics. One of the challenges of this goal is to distinguish variation in a genome that has an effect on phenotype from variation that has no effect. One way to determine if a genetic variant is likely to have a phenotypic effect is to determine if that variant affects gene regulation. Presumably through their effect on gene regulation these variants affect organismal phenotypes. Examining heterozygous sites in a genome can identify the presence of one class of variants, called cis-regulatory variants. In the absence of a functional variant, sequencing based measurements of gene expression should show no bias toward one allele. The presence of allelic bias implies that functional genetic variation affects gene regulation. One key issue in identifying these allele-specific events is a systematic bias toward the reference allele, an artifact of read alignment that creates false positive allele-specific expression (ASE). While N masking is one popular method used to reduce reference bias, it does not eliminate all sources of reference bias. We present a novel metric, the *homology score*, to characterize Single Nucleotide Polymorphisms (SNPs) based on the homology of their surrounding genomic region. When combined with N-masking, we are able to identify and filter out sources of reference bias not accounted for by use of N-masking alone. In general, this metric can assist in removing SNPs likely to yield false positive ASE.

Introduction

Determining the relationship between genetic variation and phenotypic variation has been a long-standing goal of evolutionary and medical genetics. Genetic variation can lead to and increase susceptibility to a wide variety of genetic disorders [1-6]. Advancements in high throughput sequencing have made whole-genome association studies feasible generating detailed profiles of genetic variation. This has led to identifying functional variants underlying phenotypic differences in diverse populations [7, 8]. While many variants are uncovered by high throughput sequencing, distinguishing functional variants associated with loss of function or disease from the broader background of variants that have no effect remains a topic of interest.

Variation in gene expression can modulate phenotype. Past studies have shown that gene expression is heritable and can be mapped as a quantitative trait [9]. Furthermore, gene expression regulated by allele-specific effects, which can arise from events such as genomic imprinting and X chromosome inactivation, is quite common throughout the genome [10, 11]. One way to study phenotypic effects of genetic variation is to look at how heterozygous variants affect the expression of a gene. In the absence of a functional variant, sequencing based measurements will show the allele from each parent equally expressed at a 1:1 ratio. A functional

variant is identified when the allele from one parent is expressed at a significantly higher level than the other allele.

An important aspect of allele-specific analysis is the reliable quantification of ASE at variant sites. ASE is one metric used to differentiate regulatory variants but can entail many false positives due to biases. Appropriate strategies must be applied to account for the various sources of bias currently known to affect ASE analysis [12, 13]. One confounding factor is the presence of a systematic reference bias when aligning to a single reference genome [14]. Reference bias is a type of technical bias first introduced when mapping sequencing reads to a genome. Reads containing the reference allele are more likely to map correctly than reads containing the alternate allele. Reads containing the alternate allele at SNP loci have an automatic 1 bp mismatch, thereby, increasing the chance they will be discarded due to exceeding a fixed mismatch threshold by the alignment tool. Reference bias can be even more pronounced in genomic regions where multiple SNPs reside. Reads containing multiple SNP loci can result in multiple mismatches and consequently be discarded for exceeding the threshold.

Several strategies have been proposed to reduce reference bias [14-17]. When the haplotype of both parents are known, mapping reads to each genome and combining results is a superior approach [12]. Unfortunately, phased sequencing is only available for a small number of highly sequenced cell lines making this approach impractical in many situations. Therefore, strategies relying on mapping reads to a single genome must be considered. N-masking known SNP locations in the genome so reads from both alleles have an equal chance of mapping is one approach used however it does not eliminate all bias [14]. We present a metric to identify and filter SNPs in homologous regions, one source of reference bias that N masking alone does not resolve [18]. When applied as a post-processing step to N masking, our approach is able to detect regions of reference bias not resolved by N-masking.

Methods

Our simulation study was conducted using the human genome (GRCh37) and a set of highly confident genotype SNP calls for NA12878 from the National Institute of Standards and Technology (NIST).

SNP Selection

As part of the Genome In A Bottle (GIAB) Project, NIST has compiled a set of 1,671,942 highly accurate genotype variant calls for NA12878, the pilot genome for the GIAB Consortium [19]. After downloading this reference material (v2.19), the following filtering steps were applied:

1. Filter out SNPs that lie within 10bp of an indel
2. Filter for variants labeled as SNPs, discard all others
3. Filter for only heterozygous SNPs
4. Filter for only bi-allelic SNPs

After filtering, 1,510,938 (~90% of the original set) SNPs were left over. This filtered set represents the collection of SNPs was used throughout the simulation study unless noted otherwise.

Read Simulation

Mason was used to simulate RNA-seq reads at each SNP loci with a 1:1 ratio of the reference allele and alternate allele [20]. To achieve a 1:1 ratio of the reference allele and alternate allele at every SNP loci, two versions of the human genome were created: 1) a genome with all reference alleles at SNP loci and 2) a genome with all alternate alleles at SNP loci. Then 200,000,000 single-end 50 bp reads were simulated from each genome based on an Illumina sequencing error profile and a per-base mismatch probability of 0.01. Reads from each genome were combined and only reads overlapping SNP loci (10,780,849 reads) were kept for testing.

Read Mapping

After N masking all SNP loci in the reference genome, simulated reads were mapped using STAR, allowing up to 2 mismatches and only accepting reads that mapped uniquely to the reference genome [21]. Of the original 10,780,849 reads simulated for testing, 9,906,349 (91.89%) mapped successfully under these alignment conditions.

To calculate the homology score for each SNP, all reads were mapped to the reference genome with standalone BLAT using parameters recommended by the University of California Santa Cruz to replicate web-based BLAT results [22].

To reproduce similar BLAT results, use the following parameters when running BLAT

```
blat -stepSize=5 -maxDnaHits=2 -repMatch=2253 -minScore=0 -minIdentity=0
```

As part of this study, a real-world RNA-seq dataset of GM12878 was used. Reads were first trimmed with ngsShort using parameters recommended by the author [23]. Reads were then mapped using STAR, allowing up to 2 mismatches and only accepting reads that mapped uniquely to the reference genome [21]. Reads were mapped to the human genome without masking and then mapped again against a genome with SNP loci N-masked.

Homology Score

We define the homology score as follows:

Let *s* denote a SNP that has *t* overlapping reads and we assume that each read has been processed by BLAT. The read homology score m_r for a given read *r* is defined as

$$b_{r_1} = \text{maximum BLAT score for read } r$$

$$b_{r_2} = \text{second top BLAT score for read } r$$

$$\delta_r = \frac{b_{r_1} - b_{r_2}}{b_{r_1}}$$

$$m_r = 1 - \delta_r$$

Reads that have only one BLAT hit (i.e. a unique match was found) are assigned a homology score of 0. The SNP homology score is computed by the average over all t reads

$$F_s = \frac{1}{t}\sum_{c=1}^{t} m_c$$

A SNP with a score of 0 will represent a SNP in a unique region of the genome. By definition, a SNP with a score of 1 will most certainly be in a homologous region of the genome as it would be made up entirely of reads that had exact matches elsewhere in the genome. While studies involving short reads may require a different alignment tool, it should not affect the calculation of the SNP homology score as the conceptual idea remains: to quantify the homology of each SNP region based on the reads supporting it.

Results & Discussion

N-masking the reference genome before mapping reads is a popular strategy used to eliminate allelic bias in allele-specific expression analysis. By masking known SNP loci, reads mapping to the SNP loci that contain the alternate allele are not at a disadvantage of exceeding sequence mismatch thresholds. To evaluate how well N-masking affects allelic bias, reads were simulated across SNP loci at a 1:1 ratio of the reference allele and alternate allele. A deviation from this ratio would be the result of bias. MASON was used to generate RNA-seq reads because it accurately models properties of RNA-seq regularly found in Illumina sequencing platforms [20]. We confirmed that the distribution of reads matched expectation and SNPs had a varied range of coverage. After simulating reads at all SNP loci, reads were mapped to the reference genome where 730,118 of 1,591,769 (46%) SNPs (excluding indels) showed some allele imbalance (i.e. sites without an exact ratio of 0.5). After N-masking the reference genome at SNP loci and remapping all reads, 206,833 of 1,556,921 (13%) SNPs (excluding indels) showed allele imbalance. This confirms prior results that N-masking can eliminate overall reference bias (Fig 1).

After simulating reads at all SNP loci and mapping reads to a N-masked genome, many biased SNPs still remain. 80,872 out of 1,511,294 (5%) SNPs that exhibited allele imbalance showed expression of just one allele. The remaining SNPs showed severe to modest allele imbalance as illustrated in Figure 2. Our results point to the limitations of using N-masking to remove allelic bias. We hypothesized that a significant portion of the allelic bias that remains after N-masking occurs in homologous regions. While reads containing either allele will have an equal chance to map correctly to a N-masked genome, differentiating the correct mapping of reads between homologous regions becomes ambiguous. When SNPs occur in homologous regions of the genome, the allele at the SNP loci can help distinguish these ambiguous mappings. Reads containing the reference allele will map to one homologous SNP region while reads containing the alternate allele will map to a different homologous SNP region. This phenomenon confounds allele-specific studies where it is presumed that N-masking will not affect the mapping of reads significantly. Of the total 10,780,849 simulated reads, 9,906,349 (91.9%) reads mapped uniquely to the genome with no masking while 9,735,608 (90.3%) reads

mapped uniquely to the N-masked genome. Since reads may map ambiguously in highly homologous regions, SNP loci in these regions are not amenable to ASE analysis. Therefore we developed a *homology score* that identifies homologous regions near SNPs *de novo* based on how unique the read alignment around SNPs is. BLAT hit scores provide an excellent opportunity to exploit when quantifying unique alignments.

The homology score was significantly higher (p-value < 2.2e-16, Wilcoxon rank sum) for SNPs that were one-allele biased compared to those that showed a mixed bias of both alleles (Fig 3). We confirmed that this pattern is not dependent on read coverage. Overall, our results indicate that after N-masking 5% of biased SNPs are located in highly homologous regions. We can use appropriate homology score thresholds to remove biased SNPs (true positives) at the risk of removing SNPs that show no bias (false positives). We determined the receiver operating characteristic (ROC) curve and precision-recall (PR) curve along a range of thresholds for the homology score. Score thresholds that remove extremely biased SNPs but do not remove any unbiased SNPs are ideal (Fig 4). Weighing both sensitivity and specificity equally, our results indicated the optimal threshold for the homology score to be 0.6. While in practice it would be best to use a more stringent threshold, the PR curve shows a greater loss in recall as precision rises steadily.

We applied our homology score to a RNA-seq data set from the lymphoblastoid cell line GM12878 (Coriell). Mapping reads to the genome with no N-masking resulted in the discovery of 44,853 SNPs. 41,329 (92%) of the discovered SNPs showed some degree of allele imbalance. After N-masking the genome and remapping reads, 43,930 SNPs were identified. 40,456 (92%) of these SNPs showed some degree of allele imbalance as well. Surprisingly 767 SNPs were identified when mapping reads to the N-masked genome that were not present when mapping reads to the genome with no N-masking. This supports arguments from previous work that N-masking negatively impacts the ability of mapping tools to align reads. Applying our homology score approach (using the optimal threshold based on simulation results) removed a total of 8,303 SNPs. 7,728 of these SNPs showed significant bias while 575 SNPs did not.

Our results show a significant portion of SNPs that escape N-masking are extremely biased (expressing only one allele) and exist in homologous regions. Applying a homology score filter as a post-processing step to N-masking helps eliminate these SNPs. Although the effects of allelic bias diminish as read length increases, our strategy will be pertinent in experiments when longer reads are considered. Extremely biased SNPs display effects of genomic imprinting that are often the most interesting. In the future, we will continue to study other sources of bias not accounted for by the N-masking approach or homology score. While the homology score seeks to identify extremely biased SNPs, other unknown sources responsible for modestly-biased SNPs still exist (Fig 2).

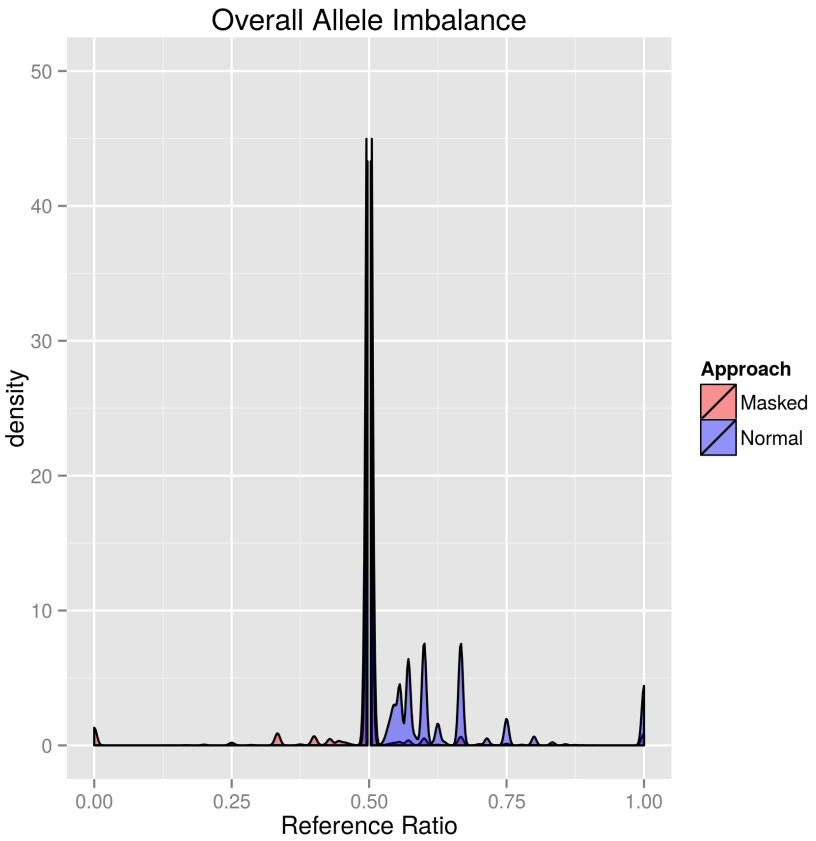

Figure 1 – Elimination of reference bias by N-masking: Reads were simulated at SNP loci across the entire genome at a 1:1 ratio. After mapping reads to an unmodified genome and an N-masked version of the genome, allele imbalance was calculated for each SNP. Mapping reads to an unmodified genome led to 46% of all SNPs showing allele imbalance. By N-masking the genome first, only 13% of all SNPs show allele imbalance. The distribution of reference bias shows that N-masking the reference genome before mapping reads eliminates overall reference bias.

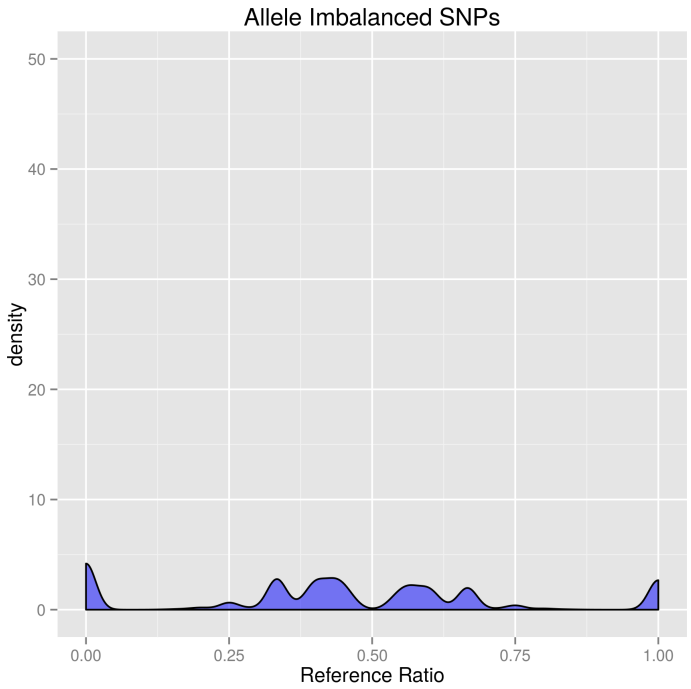

Figure 2 – Density plot of the reference ratio in allele-imbalanced SNPs after N-masking: After mapping reads to an N-masked genome, 13% of all SNPs maintain an allele imbalance that deviates from the simulated 1:1 ratio. To study the source of allelic bias after mapping reads to a N-masked genome we looked at the distribution of allelic imbalanced SNPs. The density plot illustrates that allelic imbalance in 5% of these SNPs is due to the expression of a single allele while the other allele-imbalanced SNPs show modest bias.

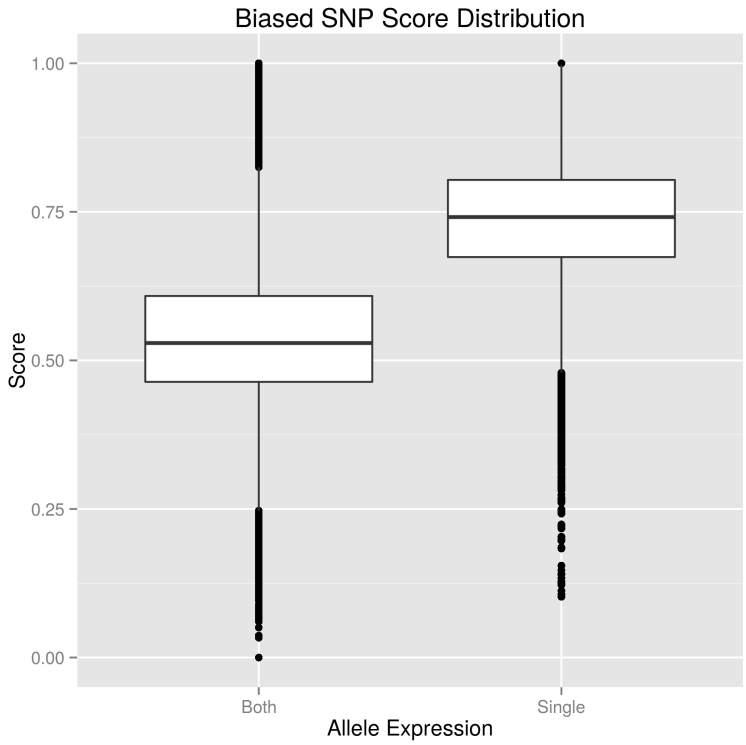

Figure 3 – Score distribution of highly biased SNP: SNPs where one allele was expressed tended to have a higher homology score than SNPs where both alleles were expressed. A Wilcoxon rank sum test confirmed the distribution of the homology scores for SNPs where a single allele is expressed is significantly higher than SNPs where both alleles are expressed (p-value < 2.2e-16). This indicates that extremely biased SNPs expressing a single allele are located in regions of the genome identified as highly homologous.

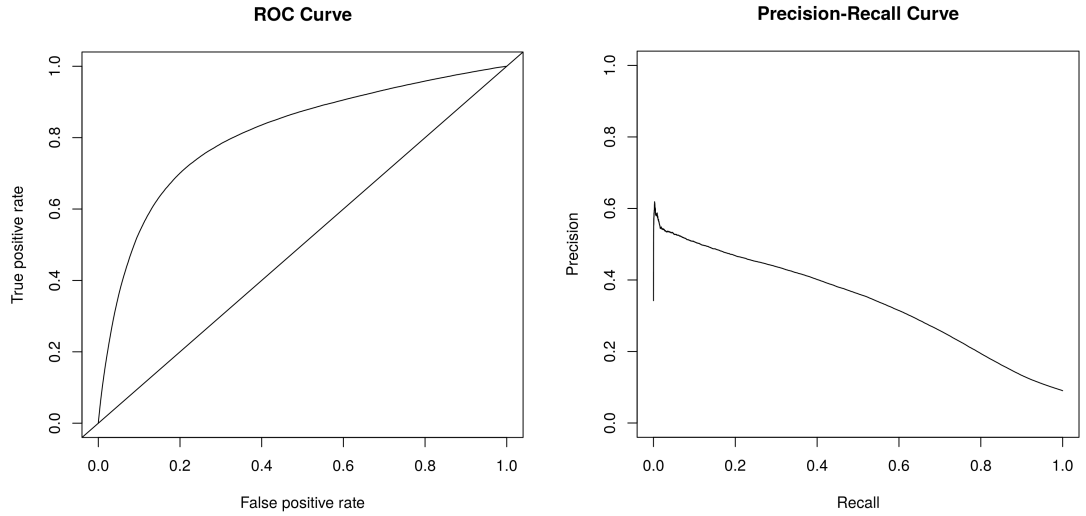

Figure 4 – ROC curve and PR curve: The ROC curve has an area under the curve (AUC) of 0.81. Overall, there are gains in the true positive rate (TPR) up to the optimal threshold, (> 73%)), trading off a false positive rate (FPR) up until about 23% FPR. After an FPR of 23%, we do not see significant gains in TPR for a tradeoff of increased FPR. [24]

References

1. Steinthorsdottir, V., et al., *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nat Genet, 2007. **39**(6): p. 770-5.
2. Steinthorsdottir, V., et al., *Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes.* Nat Genet, 2014. **46**(3): p. 294-8.
3. Harismendy, O., et al., *9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response.* Nature, 2011. **470**(7333): p. 264-8.
4. Raber, J., Y. Huang, and J.W. Ashford, *ApoE genotype accounts for the vast majority of AD risk and AD pathology.* Neurobiol Aging, 2004. **25**(5): p. 641-50.
5. Chen, J.M., C. Ferec, and D.N. Cooper, *A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview.* Hum Genet, 2006. **120**(1): p. 1-21.
6. Hollams, E.M., et al., *MRNA stability and the control of gene expression: implications for human disease.* Neurochem Res, 2002. **27**(10): p. 957-80.
7. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
8. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
9. Yang, S., et al., *Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals.* BMC Genomics, 2014. **15**: p. 13.
10. Buckland, P.R., *Allele-specific gene expression differences in humans.* Hum Mol Genet, 2004. **13 Spec No 2**: p. R255-60.
11. Lo, H.S., et al., *Allelic Variation in Gene Expression Is Common in the Human Genome.* Genome Research, 2003. **13**(8): p. 1855-1862.
12. Stevenson, K.R., J.D. Coolon, and P.J. Wittkopp, *Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome.* BMC Genomics, 2013. **14**: p. 536.
13. Wood, D.L., et al., *Recommendations for Accurate Resolution of Gene and Isoform Allele-Specific Expression in RNA-Seq Data.* PLoS One, 2015. **10**(5): p. e0126911.
14. Degner, J.F., et al., *Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.* Bioinformatics, 2009. **25**(24): p. 3207-12.
15. Castel, S.E., et al., *Tools and best practices for allelic expression analysis.* bioRxiv, 2015.
16. van de Geijn, B., et al., *WASP: allele-specific software for robust discovery of molecular quantitative trait loci.* bioRxiv, 2014: p. 011221.
17. Satya, R.V., N. Zavaljevski, and J. Reifman, *A new strategy to reduce allelic bias in RNA-Seq readmapping.* Nucleic Acids Res, 2012. **40**(16): p. e127.

18. Castel, S.E., et al., *Tools and best practices for data processing in allelic expression analysis.* Genome Biol, 2015. **16**(1): p. 195.
19. Zook, J.M., et al., *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.* Nat Biotechnol, 2014. **32**(3): p. 246-51.
20. Holtgrewe, M., *Mason – A Read Simulator for Second Generation Sequencing Data*, in *Technical Report TR-B-10-06*. 2010, Freie Universität Berlin. p. 18.
21. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
22. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
23. Chen, C., et al., *Software for pre-processing Illumina next-generation sequencing short read sequences.* Source Code Biol Med, 2014. **9**: p. 8.
24. Sing, T., et al., *ROCR: visualizing classifier performance in R.* Bioinformatics, 2005. **21**(20): p. 3940-1.