Adjusted Likelihood-Ratio Test for Variants with Unknown Genotypes

Ronald J. Nowling and Scott J. Emrich
Computer Science & Engineering
University of Notre Dame
Notre Dame, IN 46656
(rnowling,semrich)@nd.edu

Abstract

Association tests performed with the Likelihood-Ratio Test (LR Test) can be an alternative to F_{ST}, which is often used in population genetics to find variants of interest. Because the LR Test has several properties that could make it preferable to F_{ST}, we propose a novel approach for modeling unknown genotypes in highly-similar species. To show the effectiveness of this LR Test approach, we apply it to single-nucleotide polymorphisms (SNPs) associated with the recent speciation of the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii* and compare to F_{ST}.

1 Introduction

Fixation index, or F_{ST}, has been used extensively in population genetics analyses (see [5, 10, 15] for insect-focused studies). F_{ST} is a score between 0 and 1 calculated from population frequencies of known alleles. To identify variants for further analysis, researchers often calculate F_{ST} for each single nucleotide polymorphism (SNP) individually, average individual F_{ST} scores over larger regions (windows), rank them using these scores, and then select interesting SNPs or regions based on an arbitrarily-chosen cutoff (e.g., top 500 or top 0.1%).

An alternative approach is performing Likelihood-Ratio Tests (LR Tests) using Logistic Regression (LogReg) models [2]. For each SNP, a LogReg model is trained, and then a LR Test is performed between the LogReg model and a null model based on the class probabilities [7]. LR Tests report p-values that can be used to identify statistically-significant variants relative to this null model. Note that LR Tests have been used extensively in human genome-wide association studies (GWAS) [1].

Population analysis of heterogeneous insect genomes often faces two challenges: small sample sizes and unknown genotypes. Because F_{ST} does not take samples sizes into account, the same F_{ST} score could be reported with 2 or 100 samples, as long as the observed frequencies of the alleles are identical. In contrast, LR Tests can account for sample sizes when determining the p-value of a SNP, which helps control type I errors (false positives).

Another concern is unknown genotypes that result from a variety of challenges, both biological (i.e., high levels of heterozygosity) and experimental (i.e., lower sampling coverage than expected). In humans and other organisms, unknown genotypes are often imputed using tools such as IMPUTE2 [8, 11] before performing single SNP association tests using tools such as SNPTEST [12]. Unknown genotypes in insect genomes, however, are rarely imputed because of the difficulty in doing so accurately with limited samples.

Rather than imputing unknown genotypes, we propose a framework that handles unknown genotypes directly. We make the conservative (uninformative) assumption that each unknown genotype has an equal probability of being each genotype. We then ensure that this assumption is reflected in the conditional class probabilities calculated by the LogReg models (Section 2.3). Then, in Section 3.2, we validate these resulting LogReg models by comparing predicted probabilities to analytically-calculated probabilities.

In Section 3.3, we compare the properties of F_{ST} and our LR Test approach using simulated data. We demonstrate that the p-values computed by the LR Test vary with the number of unknown genotypes and underlying sample sizes, while the F_{ST} scores do not.

As a specific example of a real-world application, we apply our LR Test framework to ≈1.7 million SNPs from the recently speciated malaria vectors *Anopheles gambiae* and *Anopheles coluzzii* from [5]. These data derive from a single chromosome arm (2L) containing relatively strong regions of differentiation [10, 15]. Identifying specific sequence-based differences is highly valuable for molecularly characterizing such closely-related species and ultimately to help understand speciation in these model systems [13]. Even though PCA analysis of samples from the two species has shown strong evidence for strong similarity within species

and clear differences between species [15], localizing key variants is ongoing work [14]. At a significance level of 1%, we find that as many as 522 positions on chromosome arm 2L are statistically significant after correcting for multiple comparisons. Of 1,633 positions with the highest possible F_{ST} score (1), only twenty overlap with this set of 522 significant positions.

This result suggests that the adjusted LR Test may be more specific than averaging SNP F_{ST} values across larger windows as performed by [10] and can better address unknown and heterogeneous genotypes than F_{ST} alone. We provide a reference implementation using scikit-learn in Asaph, a variant analysis toolkit. Note that since this framework uses common methods, it can also be easily implemented using alternative programming language/libraries if needed.

2 Methodology

2.1 Data sets

Details on the sequencing and variant calling (including filtering) for the 16 mosquito samples from Cameroon studied here are given in [5, 10, 15].

As part of the assessment of our method vs. F_{ST}, we simulated a single variant. We used fifty individuals per population for the sweep over unknown genotypes, and for each combination, we converted the appropriate number of samples' genotypes to unknown genotypes before computing the two metrics. For the sweep over population sizes, we increased population sizes in multiples of two.

2.2 Analytical Equations for Probabilities

In diploid organisms, SNPs for individual samples can be thought of as multi-sets over the nucleotides A, T, C, and G. For example, the homozygous A, homozygous T, and heterozygous genotypes would be represented as the following multi-sets, respectively: $\{A, A\}$, $\{T, T\}$, and $\{A, T\}$.

We can calculate the probability that an individual belongs to population one of two conditioned on its genotype as follows:

$$\begin{aligned} P(y=1|gt) &= \frac{P(gt|y=1)P(y=1)}{P(gt)} \\ &= \frac{\frac{N_{gt,1}}{N_1} \cdot \frac{N_1}{N}}{\frac{N_{gt}}{N}} \\ &= \frac{N_{gt,1}}{N_{gt}} \end{aligned} \tag{1}$$

For unknown genotypes, we make the uninformative assumption that the unknown genotype could be any of the possible genotypes with equal probability. In particular, we do not want to assume that we can accurately infer the true genotype of an unknown genotype from the known genotypes among sampled individuals. Additionally, we do not want to infer the class probability based on the distribution of the unknown genotypes across the classes. Note that this is a significant difference between this method traditional human GWAS analysis, because in the latter imputation is often required prior to running LR Tests.

Mathematically, we can define the conditional class probability for the unknown genotype as the union of of the conditional class probabilities for each of the known genotypes. Note that the known genotypes are mutually exclusive.

$$\begin{aligned} P(y=1|gt) &= \frac{P(gt|y=1)P(y=1)}{P(gt)} \\ &= \frac{N_{gt,1} + \frac{1}{3}N_{\{\},1}}{N_{gt} + \frac{1}{3}N_{\{\}}} \end{aligned} \tag{2}$$

$$\begin{aligned} P(y=1|\{\}) &= \frac{P(\{\}|y=1)P(y=1)}{P(\{\})} \\ &= \frac{N_1}{N} \end{aligned} \tag{3}$$

2.3 Logistic Regression Model

Assume that we have N samples with V biallelic positions. Each position has a reference allele and an alternative allele, and at each position, each sample has one of three genotypes (homozygous reference, homozygous alternate, or heterogyzous).

For each position, we encode the variants as a feature matrix $\mathbf{X}$ with dimensions $N \times 3$. We represent each genotype for each position as one of three categorical variables. If sample i has the homozygous reference genotype at position k, then we set $\mathbf{X}_{i,1} = 1$. If sample i has the homozygous alternate genotype at position k, then we set $\mathbf{X}_{i,2} = 1$. If sample i has the heterozygous genotype at position k, then we set $\mathbf{X}_{i,3} = 1$. If the genotype of sample i is unknown at position k, then the row contains zeros in every column.

From the samples' population labels, we define a N-length vector $\mathbf{y}$ of class labels. We then fit the parameters of a Logistic Regression model with the form [7]:

$$P(\mathbf{y}_i = 1|\mathbf{X}_i) = \frac{1}{1 + \exp(-\beta \cdot \mathbf{X}_i + \beta_0)} \tag{4}$$

where $\mathbf{y}_i$ is the class label and $\mathbf{X}_i$ is the feature vector for a single sample i and β is the P-length weight vector

and β_0 is the intercept. We trained the model using Stochastic Gradient Descent (SGD) and an L_2 penalty. (For the experiments in this paper, we performed 10,000 epochs of training for each model.)

In the "standard case", we fit a LogReg model on the feature matrix $\mathbf{X}$ for each position and vector $\mathbf{y}$ of class labels described above.

To adjust the conditional class probabilities, we employ the following revised training procedure. We form a new $3N \times 3$ feature matrix $\tilde{\mathbf{X}}$ and a new $3N$ vector $\tilde{\mathbf{y}}$ of class labels by duplicating each data point three times (since there are three possible genotypes). For unknown genotypes, we set each copy to one of the three known genotypes. Thus, the conditional class probabilities for the known genotypes will incorporate a key assumption of our method: that each unknown genotype has an equal probability of being one of the known genotypes (i.e., "uninformative prior."). We also set the LogReg model intercept to the fraction of samples in class one versus all of the samples and fix the intercept so it is not altered during the SGD training. This ensures that the conditional class probabilities for the unknown genotypes are determined by the ratio of class one samples to all samples. Lastly, we train the weights of the LogReg model using SGD.

Note that for predicting the conditional class probabilities, we utilize the original feature matrix $\mathbf{X}$ and class labels $\mathbf{y}$, regardless of training method.

2.4 Likelihood-Ratio Test

The log likelihood for the Logistic Regression model is given by [7]:

$$\log L(\beta, \beta_0 | \mathbf{X}, y) = \prod_{i=1}^{N} \log y_i P(\mathbf{y}_i = 1 | \mathbf{X}_i) + (1 - \mathbf{y}_i) \log(1 - P(\mathbf{y}_i = 1 | \mathbf{X}_i)) \quad (5)$$

To perform the Likelihood-Ratio Test, two LogReg models are trained. The first model (the alternative), trained as described in Section 2.3, contains additional independent variables (features) not in the null model. (In our case, the null model only contains the intercept and thus, predicts the conditional class probabilities using the ratio of class one samples to all samples.) The weights (β_1, β_0) from the two models are used to compute the log likelihoods. The difference G between the two is calculated by:

$$G = 2(\log L(\beta^1, \beta_0^1 | \mathbf{X}^1, \mathbf{y}) - \log L(\beta^0, \beta_0^0 | \mathbf{X}^0, \mathbf{y})) \quad (6)$$

The p-value for the difference in log likelihoods is calculated using the χ^2 distribution:

$$p = P[\chi^2(df) > G] \quad (7)$$

where df is the difference in the number of degrees of freedom (weights) between the two models.

2.5 Corrected Significance Level

We used a significance level of $\alpha = 0.01$ (1%). Following the method of [6], we performed a PCA analysis of the *Anopheles* SNPs and found that 15 principal components were needed to explain 99.9% of the variance. Using their modified version of Bonferroni correction, we used $0.01/15 = 6.66 \times 10^{-4}$ as the cutoff.

2.6 Ranking SNPs with $\mathbf{F}_{ST}$

To rank the SNPs, we first calculated the the F_{ST} score for each position using VCFTools [3]. Scores which were invalid (nan) or negative were to set to zero. Then, we sorted the SNP positions in descending order by their F_{ST} scores.

2.7 Asaph

Our method was evaluated using Asaph, our toolkit for variant analysis. Asaph was implemented in Python using Numpy / Scipy [18], Matplotlib [9], and Scikit Learn [16] and is available at `https://github.com/rnowling/asaph` under the Apache Public License v2.

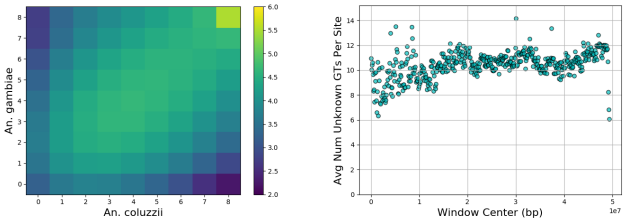
3 Experimental Results

3.1 Genotypes for Many *Anopheles* Variants are Unknown

To motivate our work, we analyzed the prevalence of unknown genotypes among the ≈1.7 million positions described in Section 2.1. For each site, we counted the number of unknown genotypes per species, which is given as a 2D histogram (with log counts) in Figure 1a. The unknown genotypes seemed to occur equally in both species. Fewer than 3% of all positions have known genotypes for each sample, while for as many as 25% of the positions, none of the genotypes are known for any of the samples in at least one population (data not shown).

We also analyzed the presence of unknown genotypes across the 2L chromosome arm. We counted the number of unknown genotypes per site and computed averages over non-overlapping 100 Kbp windows (see Figure 1b. While, the number of unknown genotypes was highest from the beginning of the inversion region (at 25 Mbp) to the end of the arm, on average more than half of the genotypes per site are unknown. Thus, unknown genotypes are highly common for this data set, which makes downstream analysis challenging.

Figure 1: Analysis of Sites on 2L with Unknown Genotypes. (a) Histogram ($\log_{10}$) of Unknown Genotypes Per Site By Species (b) Average Number of Unknown Genotypes Per Site in non-overlapping 100 Kbp Windows



(a) (b)

3.2 Mean Absolute Error of Proposed Training Method

We also evaluated the agreement of the conditional class probabilities computed by Logistic Regression (LogReg) models. For each of 800 SNPs with between zero and all-but-one unknown genotypes sampled from the *Anopheles* data set, we trained models with the standard approach and with our proposed approach described in Section 2.3. We calculated the probability for each of the four possible genotypes using each of the models. Lastly, we calculated the mean absolute error (MAE), broken down by genotype, between the probabilities from the LogReg models and the analytical probabilities.

The MAEs are reported in Table 1. With the standard training method, the LogReg model achieves a MAE as large as 0.23. With the new training approach, the largest MAE is as low as 0.0081. For the case of the unknown genotype, the error is reduced to 0, as expected.

Table 1: Mean Absolute Errors (MAE) of Analytical vs Logistic Regression-Estimated Probabilities

	Standard	**Corrected**
Homo. 1	$1.3 \times 10^{-}1$	$1.5 \times 10^{-}4$
Homo. 2	$1.3 \times 10^{-}2$	$8.1 \times 10^{-}3$
Het.	$1.7 \times 10^{-}2$	$8.1 \times 10^{-}3$
Unknown	$2.3 \times 10^{-}1$	0.

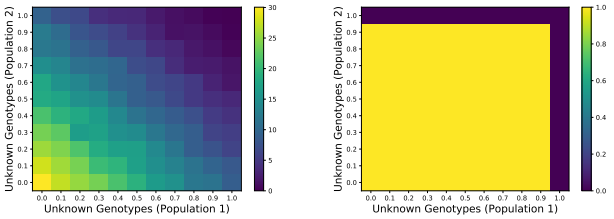
3.3 Varying of the Number of Samples and Unknown Genotypes

The Likelihood-Ratio Test (LR Test) differs from F_{ST} in two significant ways: its p-value incorporates the number of the samples and, because of our proposed training method, the percentage of unknown genotypes is also factored in. We illustrate these differences in comparisons on simulated data (see Section 2.1).

First, we considered a fixed difference where samples in one class have one homogeneous genotype and samples in the second class have the other homogeneous genotype. We swept over different combinations of percentages of samples with unknown genotypes from each population. Except for cases where all of the samples in a single class have unknown genotypes, the F_{ST} scores for all combinations are one. In contrast, the LR Test p-values increase as the percentage of unknown genotypes increase, as desired (see Figure 2).

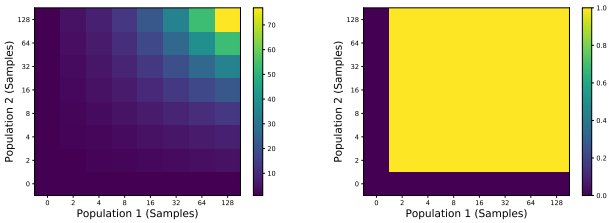
In the second comparison, we re-considered the fixed difference, but with different combinations of sample sizes in each class. We calculated the LR Test p-value and F_{ST} score for each combination (see Figure 3). As before, the F_{ST} scores for each combination were one, except when one of the populations had zero samples. The LR Test p-values decreased as the number of samples increased.

Figure 2: Adjusted Likelihood-Ratio Test p-Values ($-\log_{10}$) and F_{ST} Scores for Different Percentages of Unknown Genotypes for a Fixed Difference



(a) LR Test p-Values (log) (b) F_{ST} Scores

Figure 3: Adjusted Likelihood-Ratio Test p-Values ($-\log_{10}$) and F_{ST} Scores for Different Combinations of Sample Sizes for a Fixed Difference



(a) LR Test p-Values (log) (b) F_{ST} Scores

3.4 Analysis of SNPs from the *Anopheles* Data set

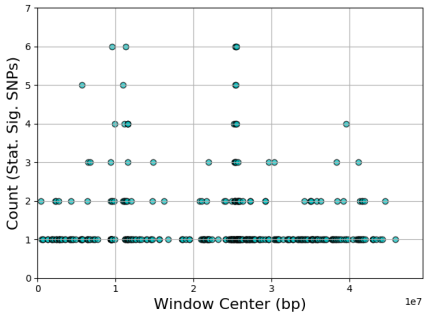
We applied the adjusted Likelihood-Ratio Test (LR Test) to perform single SNP association tests on two data sets of SNPs from the *Anopheles gambiae* and *Anopheles coluzzii* species. We first calculated q-values, a measure of significance in terms of the false discovery rate (FDR) [17, 19]. None of the SNPs, however, satisfied a q-value threshold of 0.01 (FDR of 1%).

Next, we then used the PCA-based method of [6] to determine a less conservative significance threshold (see Section 2.5). Our chosen significance level of $\alpha = 0.01$ (1%) was corrected to $0.01/15 = 6.66 \times 10^{-4}$. At that level, 522 SNPs passed the revised threshold.

For initial validation, we "binned" these 522 SNPs across the 2L chromosome in non-overlapping 10 Kbp windows—combining our method with that of [10]—and found three interesting regions: 10 Mbp, 25 Mbp, and 40 Mbp. Significantly, the 25 Mbp region and 40 Mbp region corresponds to the 2La inversion boundaries, the frequencies of which are known to differ between these samples [10, 15]. The high concentration in the 10 Mbp region is a novel result, and has been provided to our biological collaborators.

We also briefly analyzed the top 20 (as ranked by their p-values) statistically-significant SNPs individually. The first- (position 25,396,564), third- (position 21,707,904), and fifth-ranked (position 25,403,885) SNPs are located within the resistance to dieldrin (Rdl) gene, which has been previously associated with insecticide resistance in *A. gambiae* and other insects [4, 10].

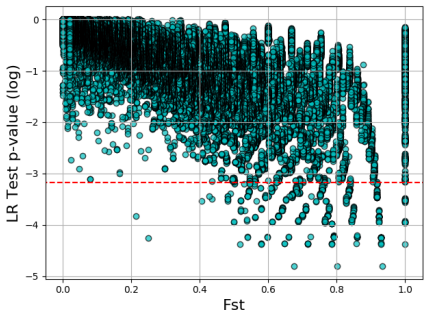
Figure 4: Counts of 522 Statistically-Significant SNPs Appearing in 10 Kbp Windows Across 2L Chromosome



We compared the adjusted LR Test p-values to the F_{ST} scores for the SNPs (see Figure 5). Notably, 1,633 SNPs have F_{ST} scores of 1, but only 20 were found in the set of 522 statistically-significant SNPs. The small number of statistically-significant SNPs with $F_{ST} = 1$ was most likely due to unknown genotypes.

Additionally, the F_{ST} scores of some of the 522 statistically-significant SNPs were as low as 0.2. We attribute this result to our categorical encoding scheme, which considers genotypes, not alleles. In fact, the uncovered 2La inversion breakpoints are only fixed in one species and by definition have non-ideal F_{ST} scores.

Figure 5: Likelihood-Ratio Test p-Values vs F_{ST} Scores. Red dashed line indicates significance threshold.



4 Discussion and Conclusion

The Likelihood-Ratio Test (LR Test) has several properties that make it desirable for population genetics analysis. In particular, unlike the more commonly used F_{ST} metric, the LR Test provides p-value that can be used to identify statistically-significant variants relative to a null model based purely on class probabilities.

Challenges in the sequencing and assembly of insect genomes results in a high propensity for unknown genotypes, as illustrated in Section 3.1. Significantly, we demonstrated in Section 3.3 that our LR Test framework can adjust the calculated p-value in line with the percentage of unknown genotypes and smaller sample sizes to address unknown values without requiring highly difficult and often impossible genotype imputation these species.

Using the adjusted LR Test, 522 *Anopheles* SNPs were found to be statistically significant. Since F_{ST} only uses population frequencies and ignores unknown genotypes in their calculation, only 20 of the 1,633 SNPs with $F_{ST} = 1$ were among the 522 significant SNPs. Significantly, treating the heterozygous genotype separately may help uncover important non-fixed differences such as the ecologically important 2La inversion [10] rediscovered here.

When used in place of F_{ST}, the adjusted LR Test has the potential to substantially reduce false positives without requiring combining multiple loci together, as is often down with window analysis (see [10]). As such, the adjusted LR Test could significantly impact

population genetics by ranking specific sequence-based differences, which will be essential to quickly characterizing and ultimately helping understand speciation in highly similar species.

5 Acknowledgments

The authors would like to thank Nora Besansky, Michael Fontaine, Becca Love, and Aaron Steele for thoughtful discussions that provided the motivation for this effort.

References

[1] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.

[2] D. J. Balding. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–791, Oct 2006.

[3] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156, 2011.

[4] W. Du, T. Awolola, P. Howell, L. Koekemoer, B. Brooke, M. Benedict, M. Coetzee, and L. Zheng. Independent mutations in the Rdl locus confer dieldrin resistance to *Anopheles gambiae* and *An. arabiensis*. *Insect Mol Biol*, 14(2):179–183, 2005.

[5] M. C. Fontaine, J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217), 2015.

[6] X. Gao, J. Starmer, and E. R. Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32(4):361–369, 2008.

[7] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. Wiley, 3 edition, 2013.

[8] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, 5(6):1–15, 06 2009.

[9] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[10] M. K. N. Lawniczak, S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson, B. White, S. Redmond, L. Fulton, E. Appelbaum, J. Godfrey, et al. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, 330(6003):512–514, 2010.

[11] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. 11:499 EP –, Jun 2010. Review Article.

[12] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913, Jul 2007.

[13] A. P. Michel, W. M. Guelbeogo, O. Grushko, B. J. Schemerhorn, M. Kern, M. B. Willard, N' F. Sagnon, C. Costantini, and N. J. Besansky. Molecular differentiation between chromosomally defined incipient species of *Anopheles funestus*. *Insect Molecular Biology*, 14(4):375–87, 2005.

[14] A. Miles, N. J Harding, G. Botta, C. Clarkson, T. Antao, K. Kozak, D. Schrider, A. Kern, S. Redmond, I. Sharakhov, et al. Natural diversity of the malaria vector anopheles gambiae. 2016.

[15] D. E. Neafsey, M. K. N. Lawniczak, and D. J. Park. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science*, 2984, 2010.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[18] S. v. d. Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[19] J.D. Storey with contributions from A. J. Bass, A. Dabney, and D. Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2017. R package version 2.8.1.