Bayesian Inference of Posterior Error Probabilities for Disease Mutation Association

Guy Karlebach[1]

[1]Department of Computer Science, Fitchburg State University, Pearl Street, Fitchburg, 01420, MA, USA.

Contributing authors: gkarleba@fitchburgstate.edu;

Abstract

Associations between single-nucleotide polymorphisms (SNPs) and phenotype are an important research paradigm in genomics and have been extensively studied, especially for the purpose of bettering our understanding of he emergence and development of disease. Various methodologies accommodating different data types, data qualities and genomic properties have been developed. In this work, we focus on a specific aspect of the analysis, namely the computation of an association score using a statistical procedure. We argue that Bayesian inference using a mixture prior that is learned from the data leads to better predictions than existing approaches, and suggest its incorporation into existing tools. To demonstrate its power, we implement the new procedure and compare its performance to popular tools on simulated data. We also detect associations between disease and SNPs using a real, modest-size RNA-Sequencing dataset, showing that the method can produce useful insights and has broad applicability.

Keywords: Bayesian Inference, Region of Practical Equivalence, Single-Nucleotide Polymorphism

1 Introduction

Association between single-nucleotide polymorphisms, whether those that correspond to change in DNA or those resulting from somatic mutations, and phenotype have been widely used in genomics for the purpose of understanding conditions for which a mechanistic understanding is lacking. DNA- and RNA-Sequencing have generated a large number of high-throughput datasets, which can be utilized for this purpose, both at the single-cell and at whole-sample resolution [1–3]. Extensive work has been

done on the topics of data analysis and variant identification, both at the bulk and single-cell level [4–6]. Association studies may be conducted genome-wide [7], over the transcriptome [8, 9] or combine insights derived at multiple levels of resolution [10]. Typically, a statistical procedure is employed in order to separate real associations from spurious ones, in particular in computational tools that are intended for use by the research community [11, 12]. The latter aspect of the analysis is the focus of this paper. When associations are ranked using Null-Hypothesis Significance Testing, multiple testing correction procedures are applied to the resulting p-values in order to discard spurious associations [13–15]. This is essential due to the large number of tests that are required with high-throughout data. As an alternative, the Bayes Factor has been suggested for evaluating Bayesian models of association [16]. In this work we suggest a Bayesian approach that accounts for multiple comparisons, and at the same time produces probabilities of association. It is based on the Region of Practical Equivalence(ROPE) [17], which has been suggested as a summary of the posterior, namely for producing posterior error probabilities (PEP). It does so by summing the posterior density in a predefined region of no effect. We harness an application of the ROPE for multiple comparisons that has been used for the study of differential gene expression and splicing [18] and adapt it for finding associations between SNPs and phenotypes. By using a mixture prior with components for significant and spurious associations, where spurious association value fall within the ROPE, we are able to incorporate the multiple comparison adjustment of posterior probabilities into the data modeling. The PEPs are then obtained by using the ROPE summary of the posterior. We argue that the new approach is more accurate than existing ones, and furthermore that PEPs are simpler to integrate over different data types, thereby improving the quality of studies that combine differential gene expression, alternative splicing and genomic elements that influence them. In the following section (Methods) we describe this procedure in more detail. We then demonstrate its effectiveness in the Results section.

2 Methods

For modeling allele occurrence in cases and controls, we construct a Bayesian model that is one component of the hierarchical Bayesian model described in Karlebach et al. [18, 19]. In Karlebach et al., changes in gene isoform proportions between cases and control samples are modeled using the Aitchison perturbation [20]. We use the same parameter for changes in allele probabilities, and the observed data, that is the alleles at each sample, are modeled as categorical variables. Denoting the vector of probabilities of the different alleles in control samples as $\vec{\phi}$, the prior distribution of $\vec{\phi}$ is:

$$\vec{\phi} \sim Dir(\vec{1}) \tag{1}$$

As in Karlebach et al., denote the vector that represents changes in allele probabilities as $\vec{\alpha}$. Then the prior of $\vec{\alpha}$ is a mixture of two Dirichlet distributions: one uniform over all points in its support (as the prior of $\vec{\phi}$), and one concentrated within the Region

of Practical Equivalence(ROPE) [17]:

$$\vec{\alpha} \sim \theta \cdot Dir(\vec{1}) + (1-\theta) \cdot Dir(\vec{K}) \tag{2}$$

In our implementation, we followed [18] and set $K = 50$. With a large number of tests, it is impractical to compute the posterior distribution of the mixture weight parameter for all the tests jointly, but the mode can be efficiently obtained through numerical optimization. The mode can then be used as a prior for each test/SNP separately. When the data does not contain any significant associations, the prior will restrict all the posteriors into the ROPE. With larger fractions of significant tests, the weight of the prior's first component increases, and more associations will be detected. This approach has the attractive property that performing more tests only reduces the number of findings if the added tests contain a smaller fraction of true associations. Now the i^{th} allele in the control samples can be modeled as:

$$allele\_control_i \sim Cat(\vec{\phi}) \tag{3}$$

And the j^th allele in the case samples as:

$$alelle\_case_j \sim Cat(\frac{\vec{\phi}_j \cdot \vec{\alpha}_i}{\sum_{1 \leq k \leq T} \vec{\phi}_k \cdot \vec{\alpha}_k}) \tag{4}$$

The Categorical variable can represent a number of variants, but can also represent heterozygote and homozygote genotypes, and even include a category for samples where no call was made, in order to test for technical issues with the caller or the data. Both the search for the mode of the mixture weight θ and the sampling from the posterior are implemented using the R interface to Stan, rstan [21].

Once a sample from the posterior is obtained, the posterior error probability is the fraction that falls within the ROPE. Two slightly different summaries are possible here, as described in [19]. One produces the probability of change for each individual proportion, and the other the probability that a change occurred at the genomic site, without assigning it to a specific allele/proportion. Whichever summary is chosen, we can then set a probability threshold for true associations, or we can choose to include the largest group of associations such that their expected PEP is less than a threshold such as 0.05.

The PEP can be used as-is in calculations of the joint probability of different events, such as the occurrence of a somatic mutation and a change in gene expression. If PEP_{DE} is the PEP for differential expression of a gene and PEP_{SNP} is the PEP for a specific SNP occurring in that gene, the probability of both events occurring is $(1 - PEP_{DE}) \cdot (1 - PEP_{SNP})$. PEPs for both differential gene expression and alternative splicing are computed by the HBA-DEALS tool [18], although at the moment it is only applicable to genes with more than a single expressed isoform. However, it is rather straightforward to generalize the HBA-DEALS model for genes with a single expressed isoform.

While using MCMC is computationally demanding, the bottleneck of the calculation

is the computations of the posteriors of the individual alleles, which can be distributed over multiple cores. With extensive computational resources being dedicated to next generation sequencing pipelines, investing resources for improvements in analysis results is quite reasonable. All the calculations in this work were done using a multicore personal computer.

3 Results

3.1 Simulation

In order to test the suggested approach on data where the ground truth is known, we generated 10 simulated SNP datasets of cases and controls, where the genotype can be homozygous for the reference allele, heterozygous or homozygous for a single alternative allele. This setting has the advantage that it is compatible with popular tools and is also focused on the core task of scoring association between genotype and phenotype. Since the changes in allele proportions due to somatic mutations are not restricted to a specific type of mapping, we assign equal probability to each possible change. Each sample has 10,000 sites, including 100 in which allele proportions change between cases and controls. RNA-Seq experiments often include a modest number of samples, and therefore we create 50 control samples and 50 case samples in each dataset. For comparison to the new procedure, we run the association tests that are implemented in PLINK [11] and SNPTEST [12]. The arguments for PLINK were *--assoc - --adjust*, the arguments for the frequentist test of SNPTEST were *-frequentist 1 -method newml* , the arguments for the Bayesian test of SNPTEST were *-bayesian 1 -method score* and the arguments for the script implementing the new method were $mcmc.warmup = 20000, mcmc.iter = 200000, allele.level = F$. The first two procedures produce BH-corrected p-values, the third produces Bayes Factors and the latter PEPs. We set an FDR threshold of 0.05 for the frequentist scores, and for compatibility select the largest set of events such that their mean PEP is at most 0.05. For the SNPTEST Bayesian score, we set a threshold of a fold-change of 2. The fraction of true proportion-changes that were detected for each of the 10 datasets are shown in Figure 1, as well as the fraction of false-positives out of the total number of predictions. As can be seen in the figure, the new method (ROPE in the figure) detects on average more true associations (21.5%) with a smaller fraction of false-positives (0.5%) compared to the next best method (SNPTEST) for which these values and 19.4% and 6%, respectively. The highest rate of false-positives on any of the datasets is 5% for ROPE vs. 14% for SNPTEST. The highest recall rate for ROPE is 32% vs. 24% for SNPTEST. We therefore conclude that the new method is more sensitive and accurate than existing procedures.

3.2 Application to RNA-Sequencing Data

To demonstrate the method on a real dataset, we use the RNA-seq dataset of Hooks et al. [22], including 30 hepatoblastoma tumor vs. 30 normal samples, and excluding the cell line samples. Fastp [23] was used for sample quality control , the STAR aligner [24] to align reads to the human genome, and cellSNP [25] in order to call variants in

each sample. As a list of variants for cellSNP, we chose variants that are predicted to destabilize protein structure [26]. The three proportions to compare between cases and controls were a genotype of 0/0 (homozygous for the reference, or non-destabilizing allele), a genotype with at least one affected allele (0/1 or 1/1), and no variant call at the site. Significant change in the latter proportion is used as an indicator that calling at that site was not reliable and should be discarded. Since the mutations are predicted to destabilize protein structure, we aggregate all the mutations that fall within the same gene. Hence if a gene does not have a destabilizing mutations it would be homozygous for all the reference alleles that fall within that gene, and if a destabilizing variant occurs at any position the protein product will be destabilized. After variant predictions, as an additional quality control measure, we removed genomic sites where cellSNP did not call any allele for more than 25% of the samples.
After computing the PEPs for each proportion of each gene, we selected the largest set of changing proportions such that their mean PEP is at most 0.05. We performed the analysis at the single-proportion level and not at the gene level in order to be able to detect significant differences in variant-call rates between cases and controls. The genes in which such sites were detected were removed from the results, as their analysis could have been affected by confounding factors. Seven significantly mutated genes passed all filters (Table 1). Figure 2 displays the ratio between the number of case samples and control samples in which each gene was mutated. As the figure shows, all the genes were mutated in more case samples than control samples, as should be expected, since control samples are from normal tissues which typically will not include protein-destabilizing mutations. Figure 3 shows a histogram of the library-size normalized log-read-counts for the gene PIK3R2, in cases and control samples. PIK3R2 is significantly affected by destabilizing mutations in case samples compared to controls, and it is also down-regulated in case samples, possibly as a cellular response to its defective protein product. In fact, all of the detected genes were down-regulated in case samples. Their mean expression log-fold-changes are displayed in Figure 4. As can be seen in the figure, all the genes had at least a twofold reduction (log-fold change of -1) in mean expression in controls compared to cases. This finding suggests an interesting possibility that protein-destabilizing mutations in hepatoblastoma are systematically followed by up-regulation of expression, potentially contributing to pathology.

Table 1 Significantly Mutated Genes

Entrez Gene ID	Ensembl Gene ID	Gene Symbol
5296	ENSG00000105647	PIK3R2
145173	ENSG00000187676	B3GLCT
3339	ENSG00000142798	HSPG2
51199	ENSG00000100503	NIN
63895	ENSG00000154864	PIEZO2
64283	ENSG00000214944	ARHGEF28
9820	ENSG00000044090	CUL7

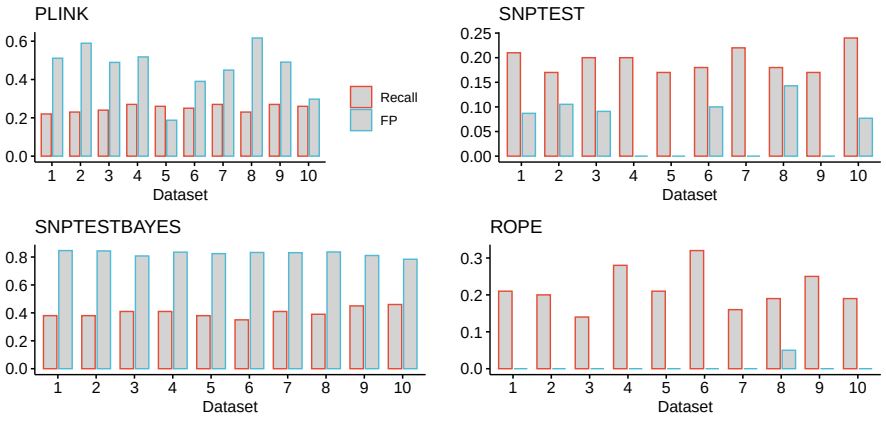

Fig. 1 Recall and false-positives fractions (y-axis, red- and blue- outlined bars, respectively) on each of 10 simulated datasets (x-axis)

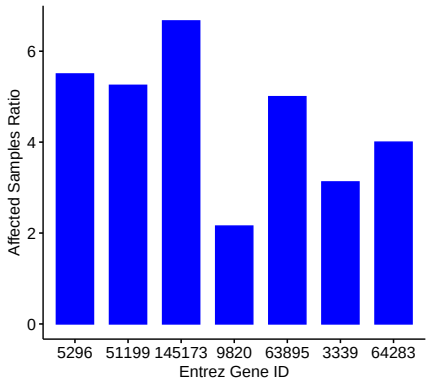

Fig. 2 Ratio of the number of samples affected by a destabilizing mutation in case samples vs that number in control samples(y-axis), shown for seven genes in which this difference is significant(x-axis).

4 Conclusion

In this work we presented a novel procedure for identifying associations between SNPs and phenotypes. Our procedure improves the accuracy of significance scores and produces a posterior error probability that can be easily interpreted and used for downstream analysis. Using a modest-size RNA-Seq dataset, we were able to detect mutations that occur significantly more frequently in hepatoblastoma samples vs. control samples, and postulated a possible link between these mutations and disease pathology. In the future, we plan to apply the new procedure to a broad range of datasets and integrate gene expression, alternative splicing and SNP information using posterior error probabilities, in order to elucidate the interplay between the three processes. Better recall and fewer false positives can help us establish a clearer understanding of the roles of SNPs, and their induced expression and splicing changes, in disease.

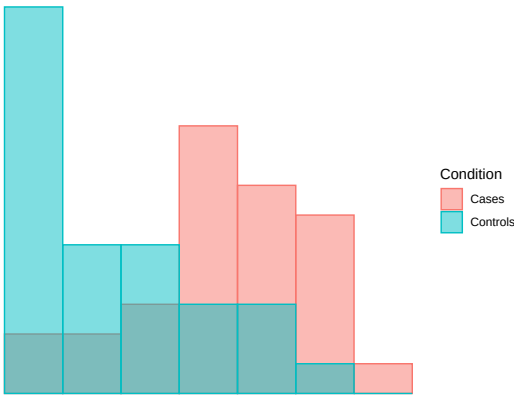

Fig. 3 Read-count in after sample depth normalization and log-transformation, in controls(red) and cases(blue), for the gene PIK3R2, which is also significantly mutated in case samples.

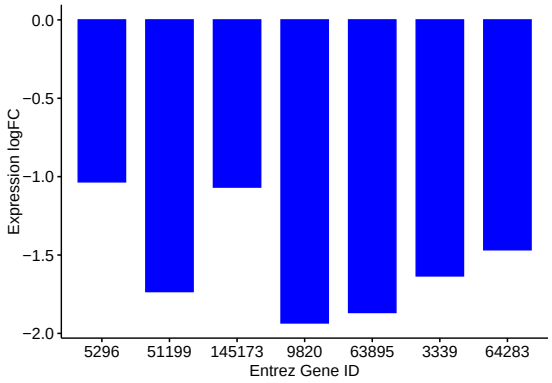

Fig. 4 Log-fold change of mean expression values in controls vs. cases, for genes that are significantly mutated in cases. All seven genes are down-regulated in controls.

Declarations

4.1 Competing Interests

The authors declare that they have no competing of interests.

4.2 Data Availability

An implementation of the method described in this work can be found at https://github.com/karleg/SNPROPE.
The data is used in this work is available publicly at the Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/

4.3 Ethics Declaration

Not applicable.

4.4 Consent to Publish Declaration

Not applicable.

4.5 Consent to Participate Declaration

Not applicable.

4.6 Author Contributions

G.K. performed all the work described in this manuscript.

4.7 Funding

There is no funding to declare. ==========

References

[1] Behjati, S., Tarpey, P.S.: What is next generation sequencing? Archives of disease in childhood - Education; practice edition **98**(6), 236–238 (2013) https://doi.org/10.1136/archdischild-2013-304340

[2] Shapiro, E., Biezuner, T., Linnarsson, S.: Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics **14**(9), 618–630 (2013) https://doi.org/10.1038/nrg3542

[3] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10**(1), 57–63 (2009) https://doi.org/10.1038/nrg2484

[4] Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S.: Genotype and snp calling from next-generation sequencing data. Nature Reviews Genetics **12**(6), 443–451 (2011) https://doi.org/10.1038/nrg2986

[5] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z.: A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics **15**(2), 256–278 (2013) https://doi.org/10.1093/bib/bbs086

[6] Muyas, F., Sauer, C.M., Valle-Inclán, J.E., Li, R., Rahbari, R., Mitchell, T.J., Hormoz, S., Cortés-Ciriano, I.: De novo detection of somatic mutations in high-throughput single-cell profiling data sets. Nature Biotechnology **42**(5), 758–767 (2023) https://doi.org/10.1038/s41587-023-01863-z

[7] Bush, W.S., Moore, J.H.: Chapter 11: Genome-wide association studies. PLoS Computational Biology **8**(12), 1002822 (2012) https://doi.org/10.1371/journal.pcbi.1002822

[8] Seo, J.-S., Ju, Y.S., Lee, W.-C., Shin, J.-Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.-O., Shin, J.-Y., Yu, S.-B., Kim, J., Lee, E.-R., Kang, C.-H., Park, I.-K., Rhee, H., Lee, S.-H., Kim, J.-I., Kang, J.-H., Kim, Y.T.: The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Research **22**(11), 2109–2119 (2012) https://doi.org/10.1101/gr.145144.112

[9] Tang, G., Liu, X., Cho, M., Li, Y., Tran, D.-H., Wang, X.: Pan-cancer discovery of somatic mutations from rna sequencing data. Communications Biology **7**(1) (2024) https://doi.org/10.1038/s42003-024-06326-y

[10] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., Geus, E.J.C., Boomsma, D.I., Wright, F.A., Sullivan, P.F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A.J., Lehtimaki, T., Raitoharju, E., Kahonen, M., Seppaa, I., Raitakari, O.T., Kuusisto, J., Laakso, M., Price, A.L., Pajukanta, P., Pasaniuc, B.: Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics **48**(3), 245–252 (2016) https://doi.org/10.1038/ng.3506

[11] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., Bakker, P.I.W., Daly, M.J., Sham, P.C.: Plink: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics **81**(3), 559–575 (2007) https://doi.org/10.1086/519795

[12] Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., Todd, J.A., Donnelly, P., Barrett, J.C., Burton, P.R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., Marchini, J.L., Morris, A.P., Spencer, C.C.A., Tobin, M.D., Cardon, L.R., Clayton, D.G., Attwood, A.P., Boorman, J.P., Cant, B., Everson, U., Hussey, J.M., Jolley, J.D., Knight, A.S., Koch, K., Meech, E., Nutland, S., Prowse, C.V., Stevens, H.E., Taylor, N.C., Walters, G.R., Walker, N.M., Watkins, N.A., Winzer, T., Todd, J.A., Ouwehand, W.H., Jones, R.W., McArdle, W.L., Ring, S.M., Strachan, D.P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E.K., Grozeva, D., Hamshere, M.L., Holmans, P.A., Jones, I.R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M.C., Owen, M.J., Craddock, N., Collier, D.A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A.H., Ferrier, I.N., Ball, S.G., Balmforth, A.J., Barrett, J.H., Bishop, D.T., Iles, M.M., Maqbool, A., Yuldasheva, N., Hall, A.S., Braund, P.S., Burton, P.R., Dixon, R.J., Mangino, M., Stevens, S., Tobin, M.D., Thompson, J.R., Samani, N.J., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C.W., Nimmo, E.R., Satsangi, J., Fisher, S.A., Forbes, A., Lewis, C.M., Onnie, C.M., Prescott, N.J., Sanderson, J., Mathew, C.G., Barbour, J., Mohiuddin, M.K., Todhunter, C.E., Mansfield, J.C., Ahmad, T.,

Cummings, F.R., Jewell, D.P., Webster, J., Brown, M.J., Clayton, D.G., Lathrop, G.M., Connell, J., Dominiczak, A., Samani, N.J., Marcano, C.A.B., Burke, B., Dobson, R., Gungadoo, J., Lee, K.L., Munroe, P.B., Newhouse, S.J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., , Genomics, T.B.i.R.G., Bruce, I.N., Donovan, H., Eyre, S., Gilbert, P.D., Hider, S.L., Hinks, A.M., John, S.L., Potter, C., Silman, A.J., Symmons, D.P.M., Thomson, W., Worthington, J., Clayton, D.G., Dunger, D.B., Nutland, S., Stevens, H.E., Walker, N.M., Widmer, B., Todd, J.A., Frayling, T.M., Freathy, R.M., Lango, H., Perry, J.R.B., Shields, B.M., Weedon, M.N., Hattersley, A.T., Hitman, G.A., Walker, M., Elliott, K.S., Groves, C.J., Lindgren, C.M., Rayner, N.W., Timpson, N.J., Zeggini, E., McCarthy, M.I., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A.V.S., Bradbury, L.A., Farrar, C., Pointon, J.J., Wordsworth, P., Brown, M.A., Franklyn, J.A., Heward, J.M., Simmonds, M.J., Gough, S.C.L., Seal, S., Susceptibility Collaboration, B.C., Stratton, M.R., Rahman, N., Ban, M., Goris, A., Sawcer, S.J., Compston, A., Conway, D., Jallow, M., Newport, M., Sirugo, G., Rockett, K.A., Kwiatkowski, D.P., Bumpstead, S.J., Chaney, A., Downes, K., Ghori, M.J.R., Gwilliam, R., Hunt, S.E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Deloukas, P., Leung, H.-T., Nutland, S., Stevens, H.E., Walker, N.M., Todd, J.A., Easton, D., Clayton, D.G., Burton, P.R., Tobin, M.D., Barrett, J.C., Evans, D., Morris, A.P., Cardon, L.R., Cardin, N.J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I.B., Howie, B.N., Marchini, J.L., Spencer, C.C.A., Su, Z., Teo, Y.Y., Vukcevic, D., Donnelly, P., Bentley, D., Brown, M.A., Cardon, L.R., Caulfield, M., Clayton, D.G., Compston, A., Craddock, N., Deloukas, P., Donnelly, P., Farrall, M., Gough, S.C.L., Hall, A.S., Hattersley, A.T., Hill, A.V.S., Kwiatkowski, D.P., Mathew, C.G., McCarthy, M.I., Ouwehand, W.H., Parkes, M., Pembrey, M., Rahman, N., Samani, N.J., Stratton, M.R., Todd, J.A., Worthington, J.: Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature **447**(7145), 661–678 (2007) https://doi.org/10.1038/nature05911

[13] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Statistical Methodology **57**(1), 289–300 (1995) https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[14] Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics **29**(4) (2001) https://doi.org/10.1214/aos/1013699998

[15] Shaffer, J.P.: Multiple hypothesis testing. Annual Review of Psychology **46**(1), 561–584 (1995) https://doi.org/10.1146/annurev.ps.46.020195.003021

[16] Stephens, M., Balding, D.J.: Bayesian statistical methods for genetic association studies. Nature Reviews Genetics **10**(10), 681–690 (2009) https://doi.org/10.1038/nrg2615

[17] Kruschke, J.K.: Rejecting or accepting parameter values in bayesian estimation. Advances in Methods and Practices in Psychological Science **1**(2), 270–280 (2018) https://doi.org/10.1177/2515245918771304

[18] Karlebach, G., Aronow, B., Baylin, S.B., Butler, D., Foox, J., Levy, S., Meydan, C., Mozsary, C., Saravia-Butler, A.M., Taylor, D.M., Wurtele, E., Mason, C.E., Beheshti, A., Robinson, P.N.: Betacoronavirus-specific alternate splicing. Genomics **114**(2), 110270 (2022) https://doi.org/10.1016/j.ygeno.2022.110270

[19] Karlebach, G., Hansen, P., Veiga, D.F., Steinhaus, R., Danis, D., Li, S., Anczukow, O., Robinson, P.N.: Hba-deals: accurate and simultaneous identification of differential expression and splicing using hierarchical bayesian analysis. Genome Biology **21**(1) (2020) https://doi.org/10.1186/s13059-020-02072-6

[20] Aitchison, J.: The Statistical Analysis of Compositional Data. Springer, ??? (1986). https://doi.org/10.1007/978-94-009-4109-0 . http://dx.doi.org/10.1007/978-94-009-4109-0

[21] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**(1) (2017) https://doi.org/10.18637/jss.v076.i01

[22] Hooks, K.B., Audoux, J., Fazli, H., Lesjean, S., Ernault, T., Dugot-Senant, N., Leste-Lasserre, T., Hagedorn, M., Rousseau, B., Danet, C., Branchereau, S., Brugières, L., Taque, S., Guettier, C., Fabre, M., Rullier, A., Buendia, M., Commes, T., Grosset, C.F., Raymond, A.: New insights into diagnosis and therapeutic options for proliferative hepatoblastoma. Hepatology **68**(1), 89–102 (2018) https://doi.org/10.1002/hep.29672

[23] Chen, S., Zhou, Y., Chen, Y., Gu, J.: fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics **34**(17), 884–890 (2018) https://doi.org/10.1093/bioinformatics/bty560

[24] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2012) https://doi.org/10.1093/bioinformatics/bts635

[25] Huang, X., Huang, Y.: Cellsnp-lite: an efficient tool for genotyping single cells. Bioinformatics **37**(23), 4569–4571 (2021) https://doi.org/10.1093/bioinformatics/btab358

[26] Janssen, K., Duran-Romaña, R., Bottu, G., Guharoy, M., Botzki, A., Rousseau, F., Schymkowitz, J.: Snpeffect 5.0: large-scale structural phenotyping of protein coding variants extracted from next-generation sequencing data using alphafold models. BMC Bioinformatics **24**(1) (2023) https://doi.org/10.1186/s12859-023-05407-9