Of Zeros and Ones

Rauschenberger, A.

2021

document version
Publisher's PDF, also known as Version of record

Acknowledgements

I would like to thank my PhD supervisors, Renée de Menezes, Marianne Jonker and Mark van de Wiel, for guiding me through this research and helping me to become a scientist. I should probably have listened to their advice more often. And I would like to thank my postdoc mentor, Enrico Glaab, for giving me the freedom to work on my ideas. I should maybe also listen to his advice more often.

Sincere thanks to all co-authors of the included papers, especially to Iuliana Ciocănea-Teodorescu for her patience with our numerous requests. I hope this is not the reason why she left for Sweden.

Many thanks to Carel Peeters for helpful discussions on Biostatistics in the Amsterdam Forest and to Léon-Charles Tranchevent for helpful discussions on Bioinformatics in the Belval Garden. I should have titled this thesis "Forest and Garden", but for once I listened to my supervisors. Also thanks to many other colleagues in Amsterdam and Luxembourg for our numerous conversations on science and other unserious topics.

I am grateful to the editors and anonymous reviewers of the included papers. Their time and effort helped me to improve my work. I am also grateful to the members of the reading committee for evaluating my thesis.

My appreciation goes to Nicolás Bellido Ortega for giving my favourite course at university. It was called "Statistics for the Social Sciences: Nonparametric and Robust Methods" (but longer), took place on Friday evenings from 17:45 to 21:00 (and beyond), and was attended by four students (or less). This is how statisticians are born.

Special thanks to my coursework supervisor, Neil Laws, and my dissertation supervisor, Tom Snijders, for guiding me through my studies. And special thanks to Gunther Tress for advising me during the first ten years of my academic career.

And I would like to thank my parents, Susanne and Michael, and my partner, Elodie, for their love and continuous support.

Abbreviations

AUC: **a**rea **u**nder the **c**urve

BMI: body mass index

ChIP: **ch**romatin **i**mmuno**p**recipitation

CNV: copy number variation

CpG: cytosine phosphate guanine

CRAN: Comprehensive R Archive Network

DNA: **d**eoxyribo**n**ucleic **a**cid

EM: expectation-maximisation

eQTL: expression quantitative trait locus

FDR: false discovery rate

GDC: Genomic Data Commons

GWA(S): genome-wide association (study)

indel: **in**sertion or **del**etion of bases

isomiR: microRNA isoform

LOH: loss of heterozygosity

meth-: methylation

mi-: micro

$\mathrm{N}(\cdot)$: Gaussian distribution

$\mathrm{NB}(\cdot)$: negative binomial distribution

$\mathrm{Pois}(\cdot)$: Poisson distribution

QTL: quantitative trait locus

RNA: **r**ibo**n**ucleic **a**cid

ROC: receiver operating characteristic

Seq: sequencing

SNP: single-nucleotide polymorphism

sQTL: splicing quantitative trait locus

TCGA: The Cancer Genome Atlas

Bibliography

Aben, N., Vis, D. J., Michaut, M., and Wessels, L. F. (2016). TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420. doi: 10.1093/bioinformatics/btw449.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750. doi: 10.1073/pnas.96.12.6745.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106. doi: 10.1186/gb-2010-11-10-r106.

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M., and Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, 8(1):1077. doi: 10.1038/s41467-017-01027-z.

Benjamini, Y. and Hochberg, Y. (1995). 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. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, 10(1):39. doi: 10.2202/1544-6115.1703.

Black, D. L. (2000). Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell*, 103(3):367–370. doi: 10.1016/s0092-8674(00)00128-8.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: Integrative L_1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017:7691937. doi: 10.1155/2017/7691937 (`ipflasso`).

Bøvelstad, H. M. and Borgan, Ø. (2011). Assessment of evaluation criteria for survival prediction from genomic data. *Biometrical Journal*, 53(2):202–216. doi: 10.1002/bimj.201000048.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140. doi: 10.1023/A:1018054314350.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nature Genetics*, 30(1):29–30. doi: 10.1038/ng803.

Bryan, M. S., Argos, M., Pierce, B., Tong, L., Rakibuz-Zaman, M., Ahmed, A., Rahman, M., Islam, T., Yunus, M., Parvez, F., et al. (2014). Genome-wide association studies and heritability estimates of body mass index related phenotypes in Bangladeshi adults. *PLoS One*, 9(8):e105062. doi: 10.1371/journal.pone.0105062.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242. doi: 10.3150/12-BEJSP11.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer, Berlin. doi: 10.1007/978-3-642-20192-9.

Campbell, F. and Allen, G. I. (2017). Within group variable selection through the Exclusive Lasso. *Electronic Journal of Statistics*, 11(2):4220–4257. doi: 10.1214/17-EJS1317.

Chaturvedi, N., de Menezes, R. X., and Goeman, J. J. (2017). A global×global test for testing associations between two large sets of variables. *Biometrical Journal*, 59(1):145–158. doi: 10.1002/bimj.201500106.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Annals of Statistics*, 37(5A):2523–2542. doi: 10.1214/08-AOS651.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71. doi: 10.1093/nar/gkv1507 (`TCGAbiolinks`).

Cortes, C. and Mohri, M. (2004). AUC optimization vs. error rate minimization. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 313–320. MIT Press, Cambridge.

Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J. S. (2000). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genetics*, 24(4):340–341. doi: 10.1038/74153.

Cule, E., Vineis, P., and De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372. doi: 10.1186/1471-2105-12-372.

Culverhouse, R., Suarez, B. K., Lin, J., and Reich, T. (2002). A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471. doi: 10.1086/338759.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Dey, K. K. and Stephens, M. (2018). CorShrink : Empirical Bayes shrinkage estimation of correlations, with applications. *bioRxiv*. doi: 10.1101/368316.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911. doi: 10.1111/j.1467-9868.2008.00674.x.

Frazee, A. C., Langmead, B., and Leek, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(1):449. doi: 10.1186/1471-2105-12-449.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1). doi: 10.18637/jss.v033.i01 (`glmnet`).

Gade, S., Porzelius, C., Fälth, M., Brase, J. C., Wuttig, D., Kuner, R., Binder, H., Sültmann, H., and Beißbarth, T. (2011). Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer. *BMC Bioinformatics*, 12(1):488. doi: 10.1186/1471-2105-12-488.

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99. doi: 10.1093/bioinformatics/btg382 (`globaltest`).

Goeman, J. J., van de Geer, S. A., and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493. doi: 10.1111/j.1467-9868.2006.00551.x (`globaltest`).

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537. doi: 10.1126/science.286.5439.531.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2):100–107. doi: 10.1016/S0168-9525(00)02176-4.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448. doi: 10.1080/01621459.2014.993077.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618. jstor: 24308572.

Huang, X., Stern, D. F., and Zhao, H. (2016). Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival–evidence from TCGA pan-cancer data. *Scientific Reports*, 6:20567. doi: 10.1038/srep20567.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(Suppl. 1):S96–S104. doi: 10.1093/bioinformatics/18.suppl_1.S96.

Huisman, M., Poppelaars, J., van der Horst, M., Beekman, A. T., Brug, J., van Tilburg, T. G., and Deeg, D. J. (2011). Cohort profile: the longitudinal aging study Amsterdam. *International Journal of Epidemiology*, 40(4):868–876. doi: 10.1093/ije/dyq219.

Hulse, A. M. and Cai, J. J. (2013). Genetic variants contribute to gene expression variability in humans. *Genetics*, 193(1):95–108. doi: 10.1534/genetics.112.146779.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679. doi: 10.1038/89044.

Kooperberg, C. and LeBlanc, M. (2008). Increasing the power of identifying gene×gene interactions in genome-wide association studies. *Genetic Epidemiology*, 32(3):255–263. doi: 10.1002/gepi.20300.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511. doi: 10.1038/nature12531.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29. doi: 10.1186/gb-2014-15-2-r29 (`limma-voom`).

le Cessie, S. and van Houwelingen, H. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, 51(2):600–614. doi: 10.2307/2532948.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323. doi: 10.1186/1471-2105-12-323.

Lo, A., Chernoff, H., Zheng, T., and Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, 112(45):13892–13897. doi: 10.1073/pnas.1518285112.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42(2):413–468. doi: 10.1214/13-AOS1175.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297. doi: 10.1093/nar/gks042.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models. 2nd ed.* Chapman and Hall, London.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278. doi: 10.1093/biomet/asn007.

Menezes, R. X., Mohammadi, L., Goeman, J. J., and Boer, J. M. (2016). Analysing multiple types of molecular profiles simultaneously: connecting the needles in the haystack. *BMC Bioinformatics*, 17(1):77. doi: 10.1186/s12859-016-0926-8 (`SIM`).

Menezes, R. X., Rauschenberger, A., 't Hoen, P. A., and Jonker, M. A. (2020). A powerful test for spliceQTL effects. *In preparation.*

Mironov, A. A., Fickett, J. W., and Gelfand, M. S. (1999). Frequent alternative splicing of human genes. *Genome Research*, 9(12):1288–1293. doi: 10.1101/gr.9.12.1288.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–19. doi: 10.1038/ng0102-13.

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, 5:4698. doi: 10.1038/ncomms5698.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777. doi: 10.1038/nature08903.

Münch, M. M., Peeters, C. F., van der Vaart, A. W., and van de Wiel, M. A. (2019). Adaptive group-regularized logistic elastic net regression. *Biostatistics*. doi: 10.1093/biostatistics/kxz062 (`gren`).

Pecanka, J., Jonker, M. A., International Parkinson's Disease Genomics Consortium, Bochdanovits, Z., and van der Vaart, A. W. (2017). A powerful and efficient two-stage method for detecting gene-to-gene interactions in GWAS. *Biostatistics*, 18(3):477–494. doi: 10.1093/biostatistics/kxw060.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772. doi: 10.1038/nature08872.

Rauschenberger, A., Ciocănea-Teodorescu, I., Jonker, M. A., Menezes, R. X., and van de Wiel, M. A. (2019). Sparse classification with paired covariates. *Advances in Data Analysis and Classification*. doi: 10.1007/s11634-019-00375-6 (`palasso`).

Rauschenberger, A., Glaab, E., and van de Wiel, M. A. (2020a). Predictive and interpretable models via the stacked elastic net. *Bioinformatics*. doi: 10.1093/bioinformatics/btaa535 (`starnet`).

Rauschenberger, A., Jonker, M. A., van de Wiel, M. A., and Menezes, R. X. (2016). Testing for association between RNA-Seq and high-dimensional data. *BMC Bioinformatics*, 17(1):118. doi: 10.1186/s12859-016-0961-5 (`globalSeq`).

Rauschenberger, A., Menezes, R. X., van de Wiel, M. A., van Schoor, N. M., and Jonker, M. A. (2020b). Semi-supervised mixture test for

detecting markers associated with a quantitative trait. *In preparation*. (`semisup`).

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 13(4):163. doi: 10.1016/s0168-9525(97)01103-7.

Reid, S. and Tibshirani, R. (2016). Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, 17(2):364–376. doi: 10.1093/biostatistics/kxv049.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140. doi: 10.1093/bioinformatics/btp616.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25. doi: 10.1186/gb-2010-11-3-r25 (`edgeR`).

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887. doi: 10.1093/bioinformatics/btm453.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332. doi: 10.1093/biostatistics/kxm030.

Robinson, M. R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J. E., Vinkhuyzen, A., Berndt, S. I., Gustafsson, S., et al. (2015). Population genetic differentiation of height and body mass index across Europe. *Nature Genetics*, 47(11):1357–1362. doi: 10.1038/ng.3401.

Rodríguez-Girondo, M., Kakourou, A., Salo, P., Perola, M., Mesker, W. E., Tollenaar, R. A., Houwing-Duistermaat, J., and Mertens, B. J. (2017). On the combination of omics data for prediction of binary outcomes. In Datta, S. and Mertens, B. J., editors, *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*, pages 259–275. Springer, Cham. doi: 10.1007/978-3-319-45809-0_14.

Roehle, A., Hoefig, K. P., Repsilber, D., Thorns, C., Ziepert, M., Wesche, K. O., Thiere, M., Loeffler, M., Klapper, W., Pfreundschuh, M., et al. (2008). MicroRNA signatures characterize diffuse large B-cell lymphomas and follicular lymphomas. *British Journal of Haematology*, 142(5):732–744. doi: 10.1111/j.1365-2141.2008.07237.x.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517. doi: 10.1093/bioinformatics/btm344.

Sanchez-Carbayo, M., Socci, N. D., Lozano, J., Saint, F., and Cordon-Cardo, C. (2006). Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays. *Journal of Clinical Oncology*, 24(5):778–789. doi: 10.1200/JCO.2005.03.2375.

Schoenmaker, M., de Craen, A. J., de Meijer, P. H., Beekman, M., Blauw, G. J., Slagboom, P. E., and Westendorp, R. G. (2006). Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European Journal of Human Genetics*, 14:79–84. doi: 10.1038/sj.ejhg.5201508.

Senchaudhuri, P., Mehta, C. R., and Patel, N. R. (1995). Estimating exact p values by the method of control variates or Monte Carlo rescue. *Journal of the American Statistical Association*, 90(430):640–648. doi: 10.1080/01621459.1995.10476558.

Shmulevich, I. and Zhang, W. (2002). Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, 18(4):555–565. doi: 10.1093/bioinformatics/18.4.555.

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1. doi: 10.18637/jss.v039.i05.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., Foekens, J. A., and Martens, J. W. (2008). Subtypes of breast cancer show preferential site of relapse. *Cancer Research*, 68(9):3108–3114. doi: 10.1158/0008-5472.CAN-07-5644.

Song, L., Langfelder, P., and Horvath, S. (2013). Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics*, 14(1):5. doi: 10.1186/1471-2105-14-5 (`randomGLM`).

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42(11):937–948. doi: 10.1038/ng.686.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960. doi: 10.1126/science.1160342.

Taylor, M. B. and Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, 31(1):34–40. doi: 10.1016/j.tig.2014.09.001.

Telonis, A. G., Magee, R., Loher, P., Chervoneva, I., Londin, E., and Rigoutsos, I. (2017). Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Research*, 45(6):2973–2985. doi: 10.1093/nar/gkx082.

Ternès, N., Rotolo, F., Heinze, G., and Michiels, S. (2017). Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biometrical Journal*, 59(4):685–701. doi: 10.1002/bimj.201500234.

The Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Mills Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120. doi: 10.1038/ng.2764.

The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968):789–796. doi: 10.1038/nature02168.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), 67(1):91–108. doi: 10.1111/j.1467-9868.2005.00490.x.

van de Wiel, M. A., Berkhof, J., and van Wieringen, W. N. (2009). Testing the prediction error difference between 2 predictors. *Biostatistics*, 10(3):550–560. doi: 10.1093/biostatistics/kxp011.

van de Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., van der Vaart, A. W., and van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128. doi: 10.1093/biostatistics/kxs031 (`ShrinkBayes`).

van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N., and Wilting, S. M. (2016). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, 35(3):368–381. doi: 10.1002/sim.6732 (`GRridge`).

van de Wiel, M. A., te Beest, D. E., and Münch, M. M. (2019). Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, 46(1):2–25. doi: 10.1111/sjos.12335.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):25. doi: 10.2202/1544-6115.1309 (`SuperLearner`).

van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J., and Wessels, L. F. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, 25(18):3201–3216. doi: 10.1002/sim.2353.

van Iterson, M., Boer, J. M., and Menezes, R. X. (2010). Filtering, FDR and power. *BMC Bioinformatics*, 11(1):450. doi: 10.1186/1471-2105-11-450.

van Steen, K. (2011). Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, 13(1):1–19. doi: 10.1093/bib/bbr012.

van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L. (2009). Survival prediction using gene expression data: a review and comparison. *Computational Statistics and Data Analysis*, 53(5):1590–1603. doi: 10.1016/j.csda.2008.05.021.

van Wieringen, W. N., van de Wiel, M. A., and van der Vaart, A. W. (2008). A test for partial differential expression. *Journal of the American Statistical Association*, 103(483):1039–1049. doi: 10.1198/016214507000001319 (`PDGEtest`).

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59(2):254–262. doi: 10.1111/1541-0420.00032.

Waegeman, W., Dembczyński, K., and Hüllermeier, E. (2019). Multi-target prediction: a unifying view on problems and methods. *Data Mining and Knowledge Discovery*, 33(2):293–324. doi: 10.1007/s10618-018-0595-5.

Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C., and Jurisica, I. (2011). Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27(24):3399–3406. doi: 10.1093/bioinformatics/btr591.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178. doi: 10.1093/nar/gkq622.

Westfall, P. H. (2005). Combining P values. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley, Hoboken. doi: 10.1002/0470011815.b2a15181.

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1):29–51. doi: 10.1177/0962280209105024.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259. doi: 10.1016/S0893-6080(05)80023-1.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114–1120. doi: 10.1038/ng.3390.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society:*

Series B (Statistical Methodology), 68(1):49–67. doi: 10.1111/j.1467-9868.2005.00532.x.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. Morgan and Claypool. doi: 10.2200/S00196ED1V01Y200906AIM006.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429. doi: 10.1198/016214506000000735.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320. doi: 10.1111/j.1467-9868.2005.00503.x.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198. doi: 10.1073/pnas.1119675109.

Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One*, 9(1):e85150. doi: 10.1371/journal.pone.0085150.

About the author

Armin Rauschenberger studied first Economics (BSc) at the University of St Andrews, the University of Geneva and the Autonomous University of Madrid and then Applied Statistics (MSc) at the University of Oxford. He currently works at the University of Luxembourg.