`poolr`: An Extensive Set of Methods for Gene-Based Testing

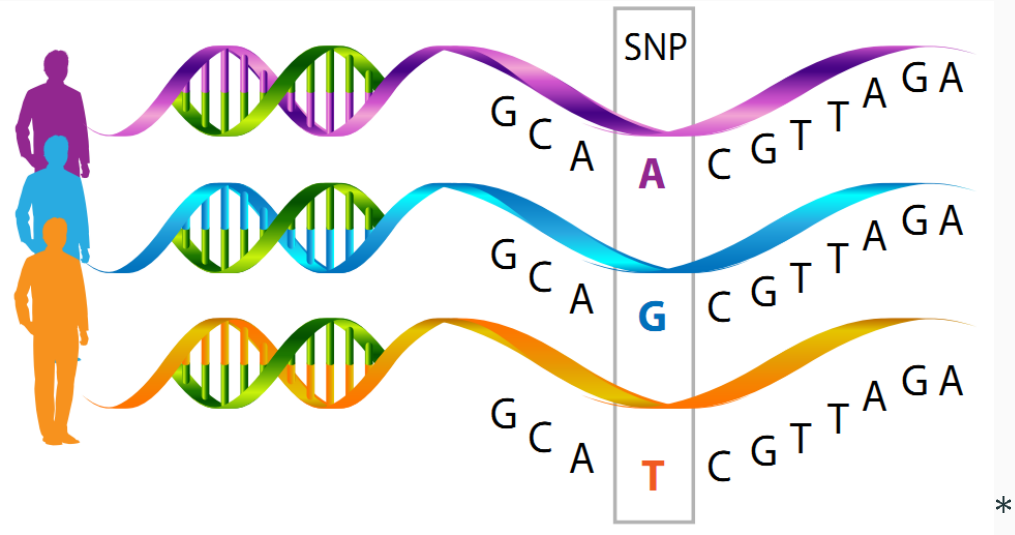
Ozan Çınar[1] Wolfgang Viechtbauer[1]

European Bioconductor Meeting 2019, Brussels, Belgium
09.12.2019

[1]Maastricht University

Genome-Wide Association Studies (GWAS)

- GWAS: Examininig the associations between single-nucleotide polymorphisms (SNPs) and a phenotype[1]



*

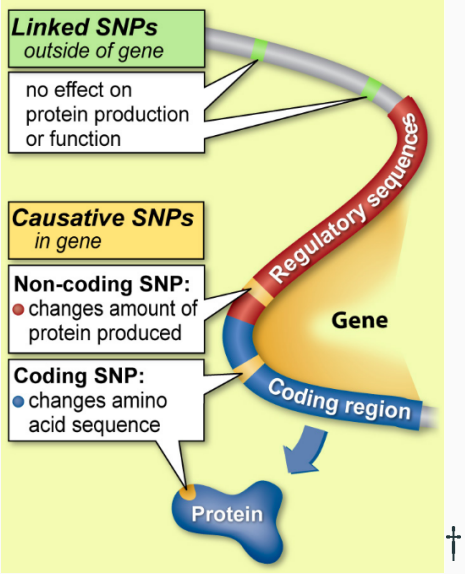
- Nowadays testing more than a million SNPs simultaneously[2]
  - $\mathrm{E}(\#(\mathrm{FP})) = 0.05 \times 10^6 = 50000$
  - Severe multiple testing corrections, e.g., 5×10^{-8} with the Bonferroni

---

* https://neuroendoimmune.wordpress.com/2014/03/27/dna-rna-snp-alphabet-soup-or-an-introduction-to-genetics/

Gene-Based Testing and Independence Assumption

- Combining the p-values of SNPs that belong to a gene
  - Accounts for polygenic effects[3]
  - #(Genes) << #(SNPs) $\rightarrow$ May improve power[4]
- Several methods for combining p-values: Fisher[5], Stouffer[6], Binomial Tests[7], Bonferroni[8], Tippett[9]



†

- Independence assumption $\rightarrow$ linkage disequilibrium (LD) is ignored[10,11]
- Common adjustment techniques: Effective number of tests[12–15], permutation tests[16], deriving the test statistic under dependence[17]

† https://learn.genetics.utah.edu/content/precision/snips/

Available R Packages and Missing Points

- Available packages via CRAN and Bioconductor
  - **Independent Tests**: `metap`[18], `survcomp`[19], `aggregation`[20], `gap`[21]
  - **Dependent Tests**: `CombinePValue`[22], `EmpiricalBrownsMethod`[23], `TFisher`[24], `harmonicmeanp`[25]
- Points still to be addressed
  - Identicality assumption between the LD and correlation/covariance matrices (effective number of tests and Stouffer under dependence)
  - Need for raw data and high computation time (permutation tests)
  - Applicable only to one-sided tests (under dependence)
  - Imprecise approximations to the covariance matrix (under dependence)

The `poolr` package - Base Functions

- `fisher()`, `stouffer()`, `invchisq()`, `binotest()`, `bonferroni()`, `tippett()`

  ```
  > args(fisher)

  function (p, adjust = "none", R, m, size = 10000, threshold,
      side = 2, batchsize, ...)
  NULL
  ```

- The vector of *p*-values (`p`) and the LD matrix (`R`) are sufficient
- Adjustment techniques for dependence (`adjust`)
  - Effective number of tests (`c("nyholt", "liji", "gao", "galwey")`)
  - Empirically-derived null distributions
  - Test statistic under dependence (for both one- and two-sided tests)

The `poolr` package - Multivariate Theory

- `mvnconv()`: Covariances among the (transformed) p-values[17]

  ```
  > args(mvnconv)

  function (R, side = 2, target, cov2cor = FALSE)
  NULL
  ```

- `target` is set to
  - `"m2lp"` for `fisher()`
  - `"z"` for `stouffer()`
  - `"chisq1"` for `invchisq()`
  - `"p"` for effective number of tests

An Example Data

```
> round(grid2ip.p[1:4], 3)  # p-values in the gene GRID2IP

[1] 0.524 0.032 0.039 0.923

> length(grid2ip.p)  # Number of SNPs in the gene

[1] 23

> round(grid2ip.ld[1:4, 1:4], 3)  # LD matrix

             rs10267908 rs112305062 rs117541653 rs11761490
rs10267908        1.000       0.199      -0.185     -0.143
rs112305062       0.199       1.000       0.144     -0.004
rs117541653      -0.185       0.144       1.000     -0.098
rs11761490       -0.143      -0.004      -0.098      1.000
```

Applying `poolr` on the Example Data

```
> fisher(p = grid2ip.p, adjust = "empirical", R = grid2ip.ld)

number of p-values combined (k): 23
combined p-value: 0.0024 (95% CI: 0.00154, 0.00357)
test statistic: 118.292 ~ chi-square(46)
adjustment: empirical

> # Stepwise algorithm
> fisher(p = grid2ip.p, adjust = "empirical", R = grid2ip.ld,
+        size = c(1000, 10000, 100000), threshold = c(.5, .05, 0))

> # Using batches to avoid memory allocation problems when
> # generating a large empirical distribution
> fisher(p = grid2ip.p, adjust = "empirical", R = grid2ip.ld,
+        size = 1000000, batchsize = 1000)
```

Applying `poolr` on the Example Data

```
> fisher(p = grid2ip.p, adjust = "generalized",
+        R = mvnconv(R = grid2ip.ld, side = 2))

number of p-values combined (k): 23
combined p-value: 0.000765
test statistic: 38.338 ~ chi-square(14.908)
adjustment: Brown's method
```

Getting `poolr` and Future Works

- Available at: `https://github.com/ozancinar/poolr`

```
> require(devtools)
> install_github("ozancinar/poolr")
```

- Adding `poolr` to CRAN
- Papers to be published
  - Presentation of the package
  - Comparison of the methods with a simulation
- Adding methods to estimate the covariances from the p-values alone (assuming compound symmetry)

Thanks for the Listening

ozan.cinar@maastrichtuniversity.nl

References

[1] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95, 2005.

[2] R. C. Johnson, G. W. Nelson, J. L. Troyer, J. A. Lautenberger, B. D. Kessing, C. A. Winkler, and S. J. O'Brien. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC genomics*, 11(1):724, 2010.

[3] Jaeyoon Chung, Gyungah R Jun, Josée Dupuis, and Lindsay A Farrer. Comparison of methods for multivariate gene-based association tests for complex diseases using common variants. *European Journal of Human Genetics*, page 1, 2019.

[4] B. Lehne, C. M. Lewis, and T. Schlitt. From snps to genes: Disease association at the gene level. *PloS one*, 6(6):e20133, 2011.

[5] R. A. Fisher. *Statistical Methods for Researchers (4th. ed.)*. Edinburgh: Oliver and Boyd, 1932.

[6] S. A. Stouffer, E. A. Suchman, L. C. Devinney, Shirley A. Star, and Robin M. Williams Jr. *The American Soldier: Adjustment During Army Life (Studies in Social Psychology in World War II*, volume 1. Princeton: Princeton University Press, 1949.

[7] B. Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156 – 158, 1951.

[8] J. M. Bland and D. G. Altman. Multiple significance tests: The bonferroni method. *British Medical Journal*, 310(6973):170, 1995.

[9] L. H. C. Tippett. *The Methods of Statistics*. London: Williams & Norgate, 1931.

[10] M. Slatkin. Linkage disequilibrium: Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477 – 485, 2008.

[11] J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946 – 1978, 2014.

[12] D. R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765 – 769, 2004.

[13] J. Li and L. Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221 – 227, 2005.

[14] 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.

[15] N. W. Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7):559 – 568, 2009.

References

[16] V. Moskvina, K. M. Schmidt, A. Vedernikov, M. J. Owen, N. Craddock, P. Holmans, and M. C. O'Donovan. Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. *Eur J Hum Genet*, 20(8):890–6, 2012.

[17] M. B. Brown. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987 – 992, 1975.

[18] Michael Dewey. *metap: Meta-Analysis of Significance Values*, 2017. R package version 0.8.

[19] M S Schroeder, A C Culhane, J Quackenbush, and B Haibe-Kains. survcomp: An r/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011.

[20] Lynn Yi and Lior Pachter. *aggregation: p-Value Aggregation Methods*, 2018. R package version 1.0.1.

[21] J H Zhao. gap: Genetic analysis package. *Journal of Statistical Software*, 23(8):1–18, 2007.

[22] Hongying Dai. *CombinePValue: Combine a Vector of Correlated P-Values*, 2014. R package version 1.0.

[23] William Poole. *EmpiricalBrownsMethod: Uses Brown's Method to Combine P-Values from Dependent Tests*, 2017. R package version 1.5.0.

[24] H. Zhang, T. Tong, J. E. Landers, and Z. Wu. Tfisher tests: Optimal and adaptive thresholding for combining p-values. *arXiv*, 1801.04309, 2018.

[25] Daniel J Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.