Beyond DBGWAS: Exploring de Bruijn Graph in an efficient manner

Hector Roux de Bézieux

25/03/2019

Division of Biostatistics, University of California, Berkeley
LBBE/CNRS, Universit de Lyon

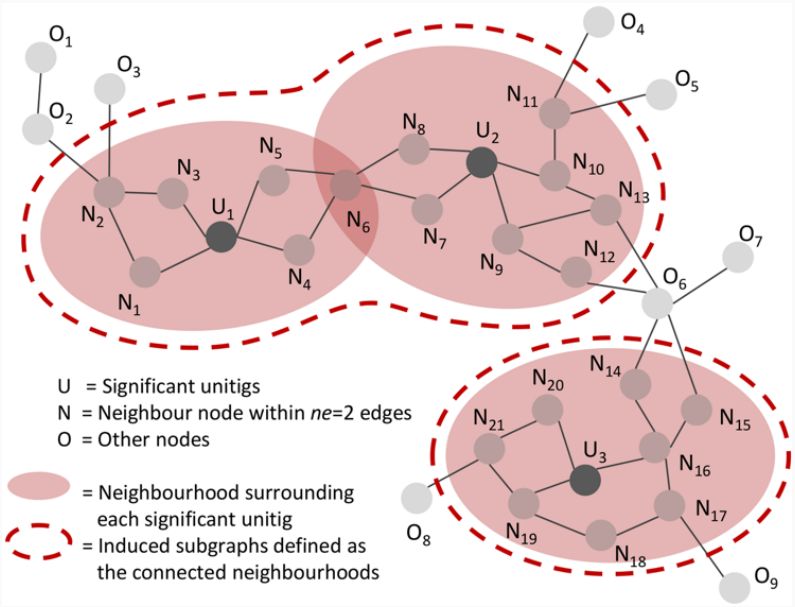
Table of contents

Introduction

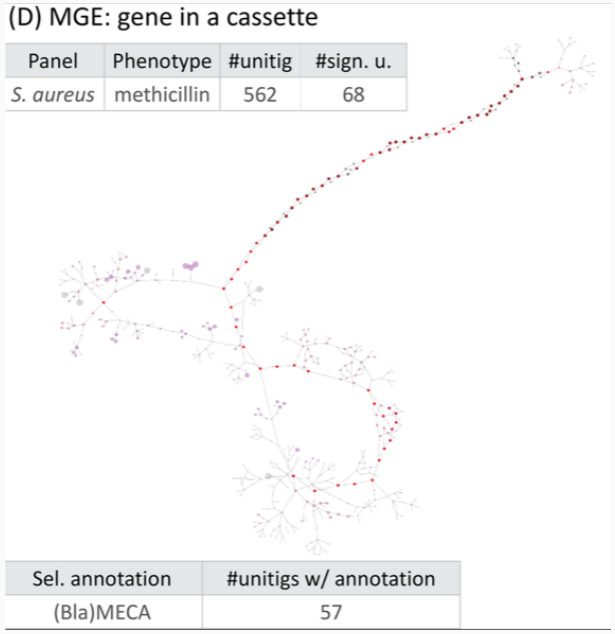
Motivation

Current format of DBGWAS has two limitations:

- Need to select an *nh* parameter to define the neighborhood (a).
- Low power to detect complex structures, as a gene cassette in (b).



(a)



(b)

Figure 1: [Jaillard et al., 2018]

Motivation (continued)

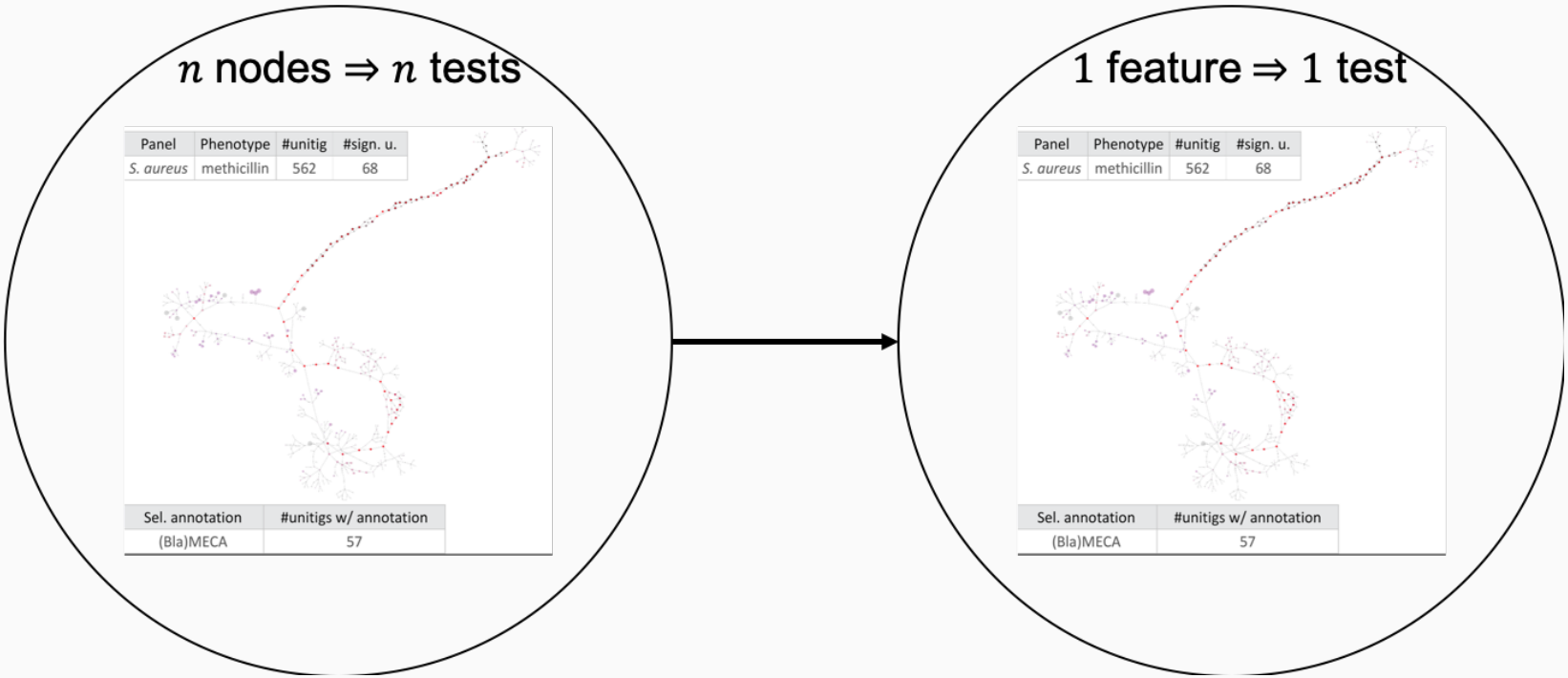
We also want to preserve strong features of DBGWAS

- Correcting for population structure.
- Remain reference-free as long as possible.
- Discover significant SNPs, gene cassettes or even species.
- Good interpretation and visualization tools.

Approach and notation

Approach

Instead of testing at the node level and trying to combine in a heuristic manner, test all possible subgraphs.



Notation

We have a set of n observations $\mathcal{D} = \{\mathcal{G}_i, y_i\}_{i=1}^n$, where

- $\mathcal{G}_i$ is a graph (part of the full k-mer graph)
- y_i is a binary phenotype.

We denote by $\mathcal{G} = U_{i=1}^n \mathcal{G}_i$, the full k-mer graph. For every subgraph $\mathcal{H} \in \mathcal{G}$, we note $z_{i,\mathcal{H}} = (\mathcal{H} \cap \mathcal{G}_i \neq \emptyset)$ and $z_{\mathcal{H}} = (z_{1,\mathcal{H}}, \ldots, z_{n,\mathcal{H}})$.

For all $\mathcal{H} \in \mathcal{G}$, we want to test $z_{\mathcal{H}} \perp Y$.

Tarone's trick

Why use Tarone's trick

Testing all subgraphs in a naive manner is not possible. The number of tests to run is much too large

1. to be computationally tractable.
2. to give reasonable power to any test.

Using Tarone's trick Tarone [1990], we can solve both issues

Tarone's trick in an example

Fisher's exact test for a two-by-two table:

Variable	Favors soccer	Favors rugby	Rows totals
Comes from the south	6	5	10
Comes from the north	8	1	10
Cols Totals	14	6	20

Conditional on the marginals, we have a hyper-geometric distribution and an associated p-value of ≈ 0.16

Minimal p-value and testable hypothesis

Before looking at the data, we can compute the minimal possible p-value. Because we have integer counts, it is not zero. The minimal p-value is obtained with this distribution of the data.

Variable	Favors soccer	Favors rugby	Rows totals
Comes from the south	4	6	10
Comes from the north	10	0	10
Cols Totals	14	6	20

The minimal p-value is ≈ 0.11

Tarone's trick

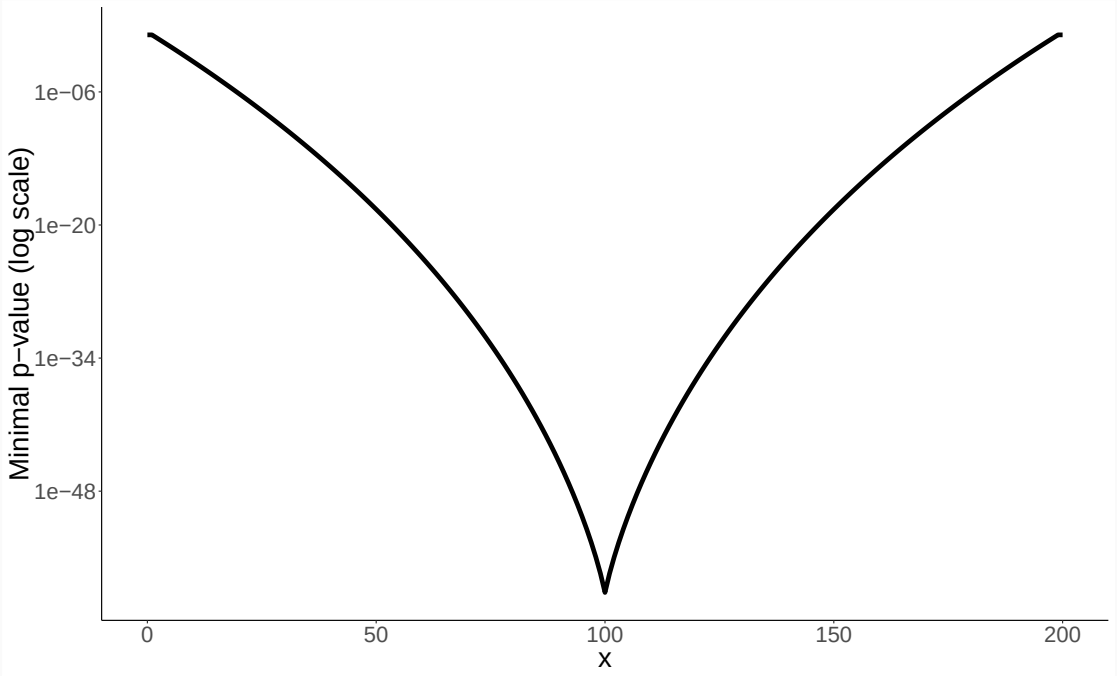
We want to test N hypotheses $h \in \mathbf{H}$. Tarone's trick relies on computing, for various values of $k \in [1, \ldots, N]$, $m(k) = \left|\{h \in \mathbf{H} | p^*(h) \leq \frac{\alpha}{k}\}\right|$. Then we identify $k_0 = \min_k \{k \in [1, \ldots, N] | m(k) \leq k\}$.

We can then define $\mathcal{R} = \{h \in \mathbf{H} | p^*(h) \leq \frac{\alpha}{k_0}\}$ and we only test the hypothesis in $\mathcal{R}$. We can then control the family-wise error rate (FWER) at a level α by rejecting each test $h \in \mathcal{R}$ iif $p(h) \leq \frac{\alpha}{k_0}$.

This has been used for regular GWAS by Llinares-López et al. [2015].

The minimal p-value is strictly increasing (after some point)

If we name $x = \sum_{i=1}^{n} 1\{Y_i = 1\}$, we have



In general, for $x > x' \geq \max(n_1, n_2), p^*(x) > p^*(x')$. So, if a subgraph is not testable, any subgraph that contains it is not testable either. We can use Frequent Subgraph Mining (FSM) algorithms to explore the De Bruijn graph.

Accounting for population structure

K-means on node matrix

We use the node absence / presence matrix $\begin{bmatrix} 0 & 1 & ... \\ 1 & 1 & ... \\ ... & ... & ... \end{bmatrix}$ of n samples by m nodes to run the k-mean algorithm. We can then obtain a categorical variable $c_i \in \{1, \ldots, K\}$ for each sample.

CMH tests with Tarone's trick

Updating previous notation, we now have $\mathcal{D} = \{\mathcal{G}_i, y_i, c_i\}_{i=1}^n$ and we want to test $z_{\mathcal{H}} \perp Y | C$. Tarone's trick works for any test that relies on the discreteness of the data, including the CMH test [Cochran, 1954, Mantel and Haenszel, 1959].

However, we loose the increasing property of the initial p-value.

Defining an envelope

We define $\tilde{p}^*(\mathcal{H}) \equiv \min_{\mathcal{H}' \supseteq \mathcal{H}} p^*(\mathcal{H}')$. Then we recover the wanted property. Papaxanthos et al. [2016] proved that the envelope can be computed in $O(k \log k)$.

We need to modify the FSM algorithm to prune the graph based on the envelope, instead of the frequency (*work in progress*).

References

Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex van Belkum, Vincent Lacroix, and Laurent Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS genetics*, 14(11):e1007758, 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007758. URL `http://www.ncbi.nlm.nih.gov/pubmed/30419019`.

RE Tarone. A modified bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.

Felipe Llinares-López, Dominik G. Grimm, Dean A. Bodenham, Udo Gieraths, Mahito Sugiyama, Beth Rowan, and Karsten Borgwardt. Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics*, 31 (12):i240–i249, 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btv263.

William G Cochran. Some Methods for Strengthening the Common $\chi 2$. *Biometrics*, 10(4):417–451, 1954.

Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4): 719–748, 1959.

Laetitia Papaxanthos, Felipe Llinares-Lopez, Dean Bodenham, and Karsten Borgwardt. Finding significant combinations of features in the presence of categorical covariates. *Nips*, (Nips):2271–2279, 2016. ISSN 10495258. doi: 10.1007/s00464-011-2087-1.

Peter B. Gilbert. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(1):143–158, 2005. ISSN 00359254. doi: 10.1111/j.1467-9876.2005.00475.x.

Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jørgensen, Nicole Shapiro, Philip D. Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z. Demaere, Rayan Chikhi, Niranjan Nagarajan, Christopher Quince, Fernando Meyer, Monika Balvočiut, Lars Hestbjerg Hansen, Søren J. Sørensen, Burton K.H. Chia, Bertrand Denis, Jeff L. Froula, Zhong Wang, Robert Egan, Dongwan Don Kang, Jeffrey J. Cook, Charles Deltel, Michael Beckstette, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu Wei Wu, Steven W. Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter Meinicke, Michael D. Barton, Thomas Lingner, Hsin Hung Lin, Yu Chieh Liao, Genivaldo Gueiros Z. Silva, Daniel A. Cuevas, Robert A. Edwards, Surya Saha, Vitor C. Piro, Bernhard Y. Renard, Mihai Pop, Hans Peter Klenk, Markus Göker, Nikos C. Kyrpides, Tanja

Woyke, Julia A. Vorholt, Paul Schulze-Lefert, Edward M. Rubin, Aaron E. Darling, Thomas Rattei, and Alice C. McHardy. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nature Methods*, 2017. ISSN 15487105. doi: 10.1038/nmeth.4458.

Fredrik H. Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013. ISSN 00280836. doi: 10.1038/nature12198.

Jun Wang, Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle Lechatelier, Pierre Renault, Nicolas Pons, Jean Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, S. Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jian Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012. ISSN 00280836. doi: 10.1038/nature11450.

Acknowledgments

This work has been done in collaboration with Laurent Jacob at LBBE/CNRS, Universit de Lyon. Inputs were provided by Fanny Perraudeau, Joe McMurdie and Christian Sieber at Whole Biome, and Sandrine Dudoit at UC Berkeley.

Appendix

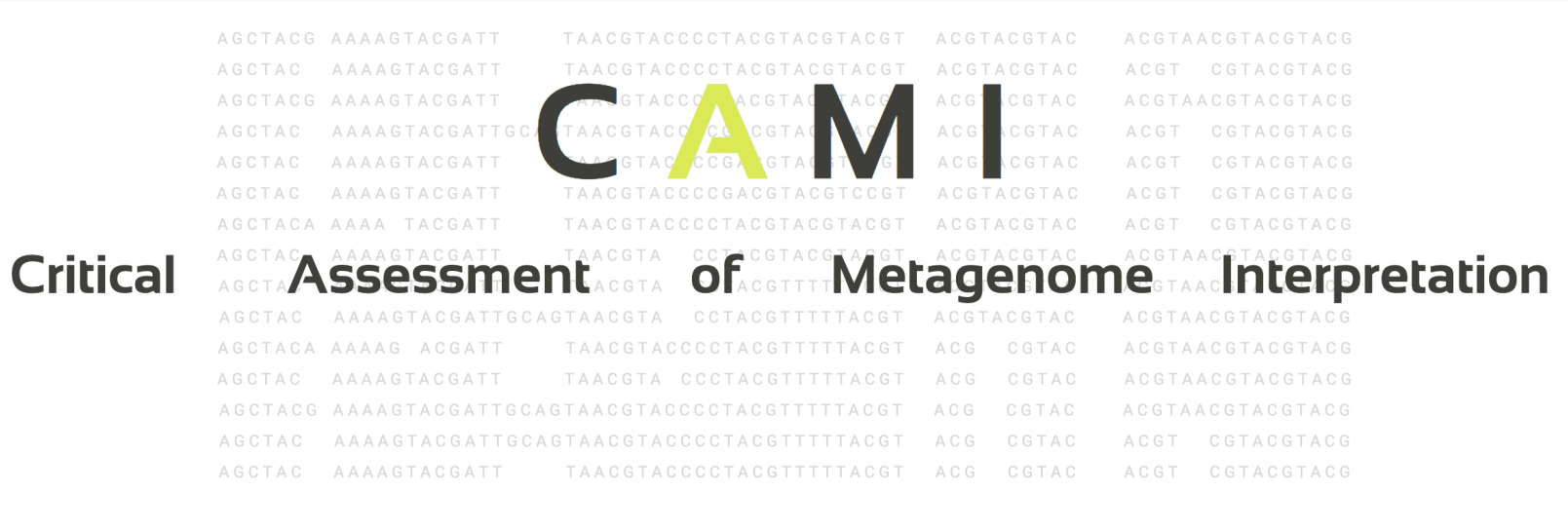
Musing on multiple testings

We note $\mathcal{R} = \{h \in \mathbf{H} | p^*(h) \leq \frac{\alpha}{k_0}\}$

- We proved that, if $|\mathcal{R}| \leq \sqrt{n}$, then Tarone's trick with FWER is less conservative than the FDR, with the same α level.
- As Gilbert [2005] pointed out, controlling the FDR on $\mathcal{R}$ controls the FDR on $\mathbf{H}$.
- Actually, there are no reason to use the same test to define $\mathcal{R}$ and then to test on $\mathcal{R}$. We still control the FDR (or the FWER) on $\mathbf{H}$.

Datasets

- Single species from the original publication: 282 bacterial genomes of Pseudomonas aeruginosa along with their drug (amikacin) resistance/sensitivity phenotype. Many results have been validated in the lab.
- Simulated metagenomics data from CAMI [Sczyrba et al., 2017]. The contigs are known, the genes will be revealed at some point.



Datasets (Continued)

- Simulate data from real datasets. Randomly add contigs from a hold-hover sample with probability $\pi = \pi(y_i)$.
- Real diabetes datasets from Karlsson et al. [2013], Wang et al. [2012]. Many results are known.