

Bacterial and metagenome GWAS

Hector Roux de Bézieux

JBIMS Data Sciences Workshop, February 18th 2020

Division of Biostatistics, University of California Berkeley

Table of contents

1. Motivation
2. Existing GWAS methods
3. DBGWAS
4. Improvements to DBGWAS

Motivation

Computational pan-genomics

Genome as a single string

```
CAATAAGGCTTGGAAATTEACCCGCTCCTGCCCGCGTCTGGAGTTCACCCGCTCCTGCCCGCGTATTATATTCAACTCTCTG
CAATAAG : CTTGGAAATTEACCCGCTCCTGCCCGCGTCTGGAGTTCATTATATTCAACTCTCTG
CAATAAGGCTTGGAAATTEACCCGCTCCTGCCCGCGTCTGGAGTTCATTATATTCAACTCTCTG

(a) Unaligned sequences
```

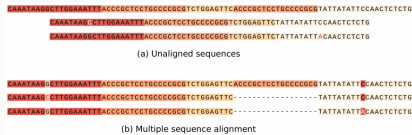


```
CAATAAG : CTTGGAAATTEACCCGCTCCTGCCCGCGTCTGGAGTTCACCCGCTCCTGCCCGCGTATTATATTCAACTCTCTG
CAATAAG : CTTGGAAATTEACCCGCTCCTGCCCGCGTCTGGAGTTC-----TATTATATTCAACTCTCTG
CAATAAG : CTTGGAAATTEACCCGCTCCTGCCCGCGTCTGGAGTTC-----TATTATATTCAACTCTCTG

(b) Multiple sequence alignment
```

(from *The Computational Pan-Genomics Consortium*, 2016)

Genome as a single string



(from *The Computational Pan-Genomics Consortium*, 2016)

Ill-suited approximation for current sequencing data:



- Discarding accessory genes, rearrangements and repeated regions.
- Problem for: microbes, viruses, metagenomes, human diseases, anything hard to assemble.
- Was really always a problem, even for simpler situations.

Existing GWAS methods

Method overview in human

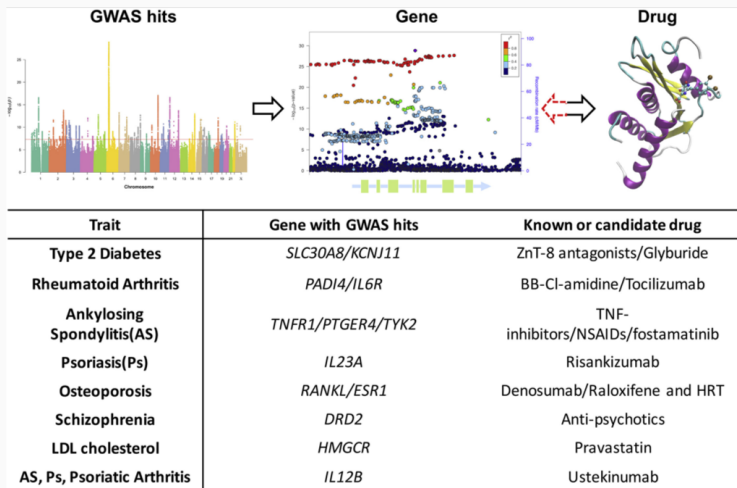


Figure 1: [Visscher et al., 2017]

In bacterial genomes and metagenomes



TTCG**C**TCGTA



TTCG**A**TCGTAT

k-mers are easy to analyse but hard to interpret




TTCGCTCGTA
TTCG
TCGC
CGCT
GCTC
CTCG
TCGT
CGTA
GTAT




TTCGATCGTAT
TTCG
TCGA
CGAT
GATC
ATCG
TCGT
CGTA
GTAT

Mapping to a reference is too dependent on its quality



TTCGCTCGTA



TTCGATCGTAT

- Easy to interpret
- Good for validation
- Dependent on good reference genomes
- Hard to analyze SNPs, genes, species at once.

DBGWAS

Constructing a De Bruijn Graph

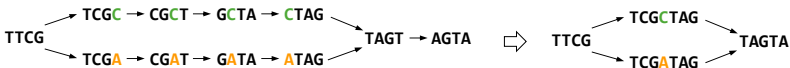
A) Fork pattern



B) Bubble pattern

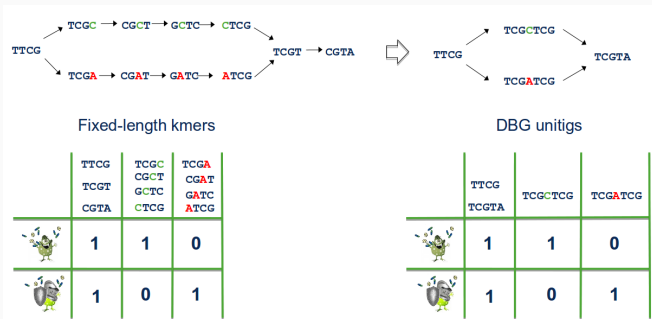


C) Compacted graph



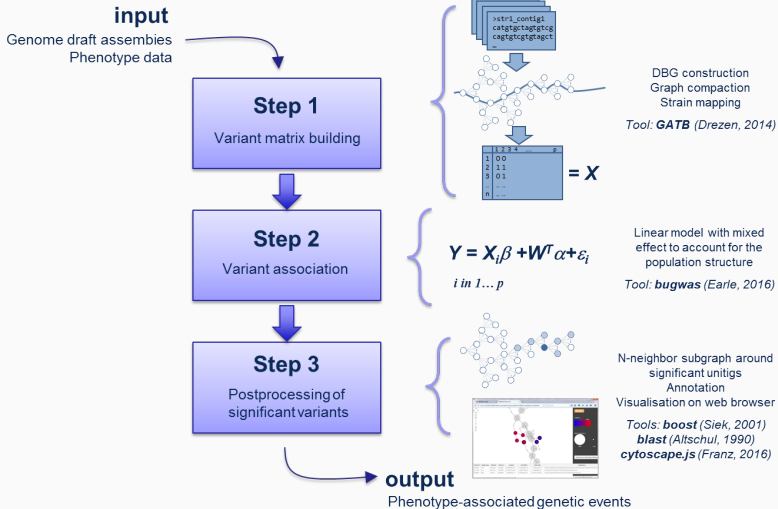
- Widely used in assembly and variant calling methods.
- A node is called an unitig

De Bruijn Graphs eliminates redundancy



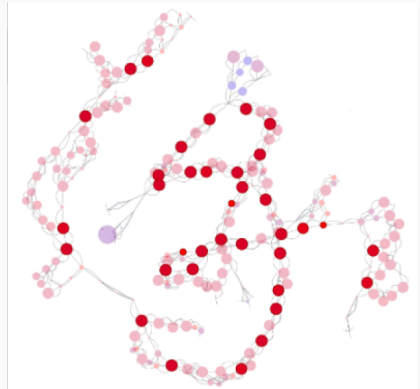
- No change in information: set of unique presence/absence profiles is the same.
- Easier to interpret: Compaction eliminates local redundancy: fewer, longer sequences. of each unitig.

Full workflow



Example: whole plasmid inclusion for *P. aeruginosa* amikacin resistance

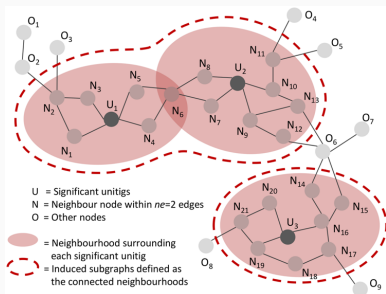
- Linear subgraph with mostly red nodes: presence of the entire sequence is associated with resistance.
- Neighborhoods connect top kmers separated by less significant ones.
- Maps to pHS87b plasmid recently described as being involved in resistance.



Improvements to DBGWAS

Current limitations

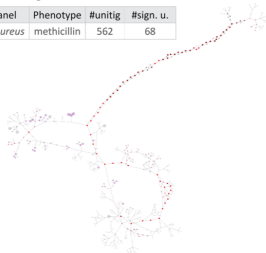
- Need to select a parameter to define the neighborhood (a).
- Low power to detect complex structures, as in (b).



(a)

(D) MGE: gene in a cassette

Panel	Phenotype	#unitig	#sign. u.
S. aureus	methicillin	562	68



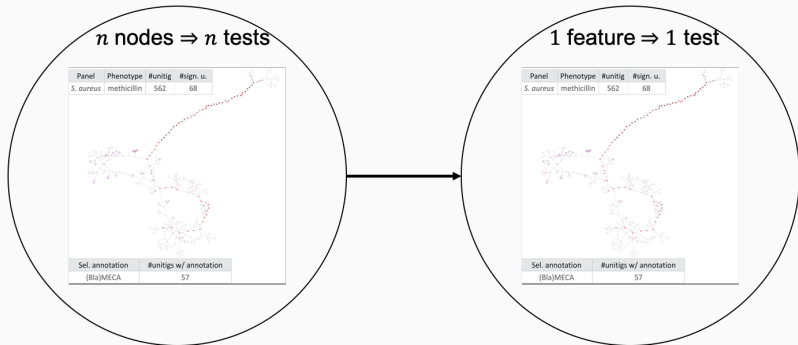
Sel. annotation	#unitigs w/ annotation
(Bla)MECA	57

(b)

[Jaillard et al., 2018]

Approach

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



Using 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

Greg and Susie



Search ID: Jkn384

"It's a little chilly in here. Throw another batch of resumes we have on file in the fire."

Greg is a recruiter. Greg throws away half of the CVs without looking at them. Greg is a bad recruiter.

Greg and Susie



Search ID: 1100384
"It's a little chilly in here. Throw another batch of resumes we have on file in the fire."

Greg is a recruiter. Greg throws away half of the CVs without looking at them. Greg is a bad recruiter.

Susie is a statistician. Susie throws away half of the hypotheses without looking at them. Is Susie a bad statistician?

Greg and Susie



"It's a little chilly in here. Throw another batch of resumes we have on file in the fire."

Greg

is a recruiter. Greg throws away half of the CVs without looking at them. Greg is a bad recruiter.

Susie is a statistician.

Susie throws away half of the hypotheses without looking at them. Is Susie a bad statistician?

Not

if you consider FWER and FDR. Of course, we do loose power.

Tarone's trick increases the rejection threshold

For discrete tests, the smallest possible p-value, or minimal p-value is not zero. So you can discard some hypotheses without testing them. This has been used for regular GWAS by Llinares-López et al. [2015], which proposed this FAIS algorithm.

- We build a common DBG from the k-mer decomposition.
- We define the features as the nodes of the graph.
- We tests them using a mixed-effect model.
- Improvements: define more complex features as subgraphs of the DBG.

Acknowledgments

Magali Jaillard and Leandro Ishi developed the DBGWAS algorithm.

This work has been done in collaboration with Laurent Jacob at LBBE/CNRS, Université de Lyon. Inputs were provided by Fanny Perraudou, Joe McMurdie, Christian Sieber and Benedicte Colnet at Pendulum, Sandrine Dudoit at UC Berkeley and Arnaud Mary at Université de Lyon.

Thank you for listening

Questions

References

- Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I Mccarthy, Matthew A Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. 2017. doi: 10.1016/j.ajhg.2017.06.005. URL <http://dx.doi.org/10.1016/j.ajhg.2017.06.005>.
- 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.

Examples

