fastSKAT:

Sequence Kernel Association Tests for large sets of markers

…and applications for analyzing LDL cholesterol in whole-genome sequencing data

Ken Rice

University of Washington

CHARGE Consortium

TOPMed Data Co-ordinating Center

This work has not been published previously

What is SKAT?

- SKAT (Wu, Lee et al, 2011) tests association between a trait and **multiple** variants; maintains power well across many possible ‘signals’

- For M variants, N subjects, takes MN x min(M,N) steps
- In large WGS work (TOPMed, CHARGE-S, etc) this **limits** SKAT analysis – too slow and/or insufficient CPU time, even with parallel processing

How to do SKAT tests faster?

SKAT compares statistic to reference – a sum of min(M,N) terms;

$$\lambda_1\chi_1^2+\lambda_2\chi_1^2+\lambda_3\chi_1^2+\lambda_4\chi_1^2+\lambda_5\chi_1^2+\lambda_6\chi_1^2+\lambda_7\chi_1^2+\ldots+\lambda_{\min(M,N)}\chi_1^2$$

Approximate this by;

from Stochastic SVD

Satterthwaite approximation

$$\lambda_1\chi_1^2+\lambda_2\chi_1^2+\lambda_3\chi_1^2+\lambda_4\chi_1^2+\ldots+\lambda_{100}\chi_1^2 + \text{remainder term}$$

Or even less, if genotype data sparse

Instead of MN x min(M,N) time, takes MN x 100 time: **fast**

Stochastic SVD?

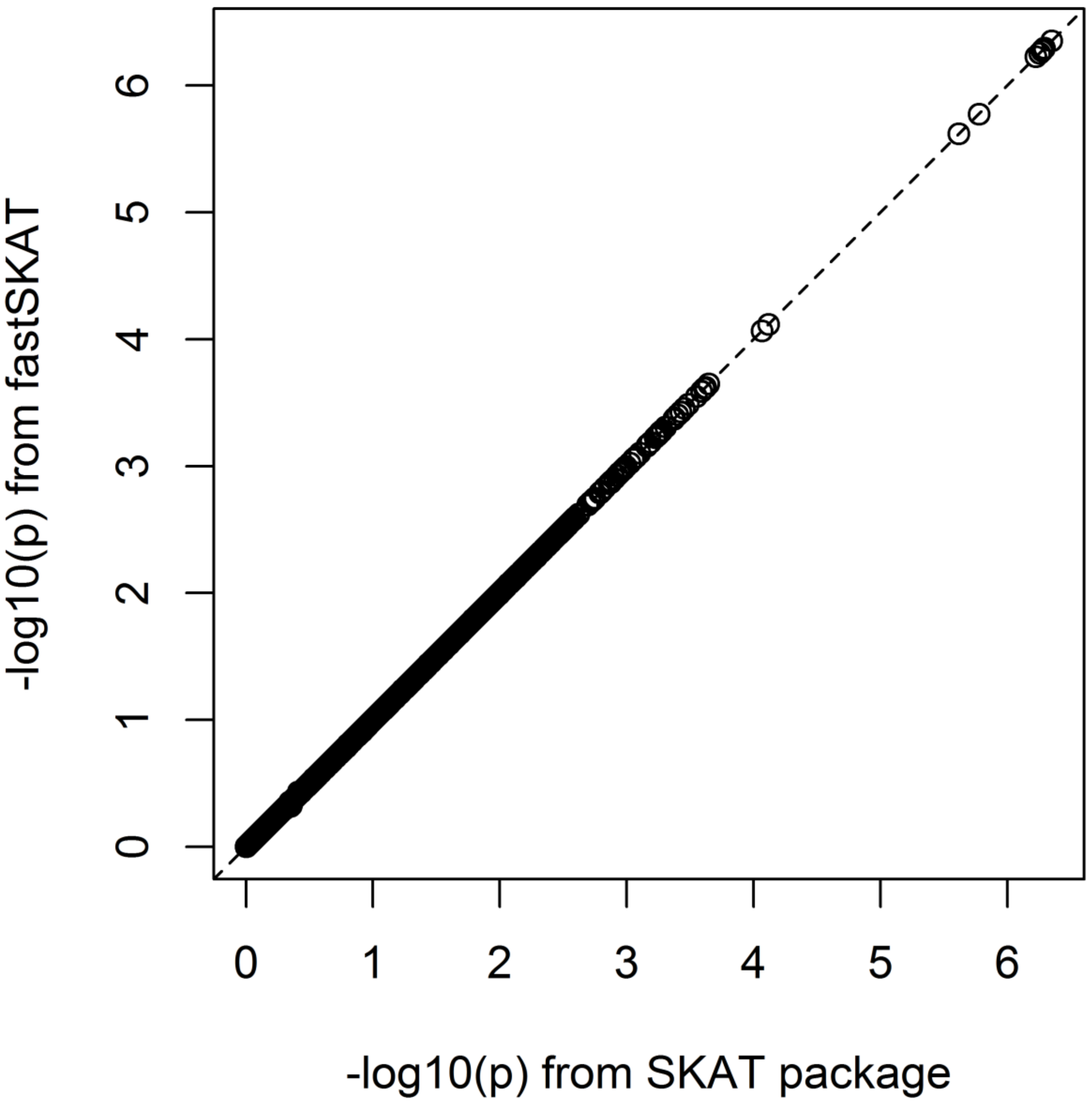
$\lambda_1, \lambda_2, \lambda_3, \dots , \lambda_{100}$

Galinsky et al (2016, AJHG) use it for fastPCA; **fastSKAT** does **inference**

Does it work?

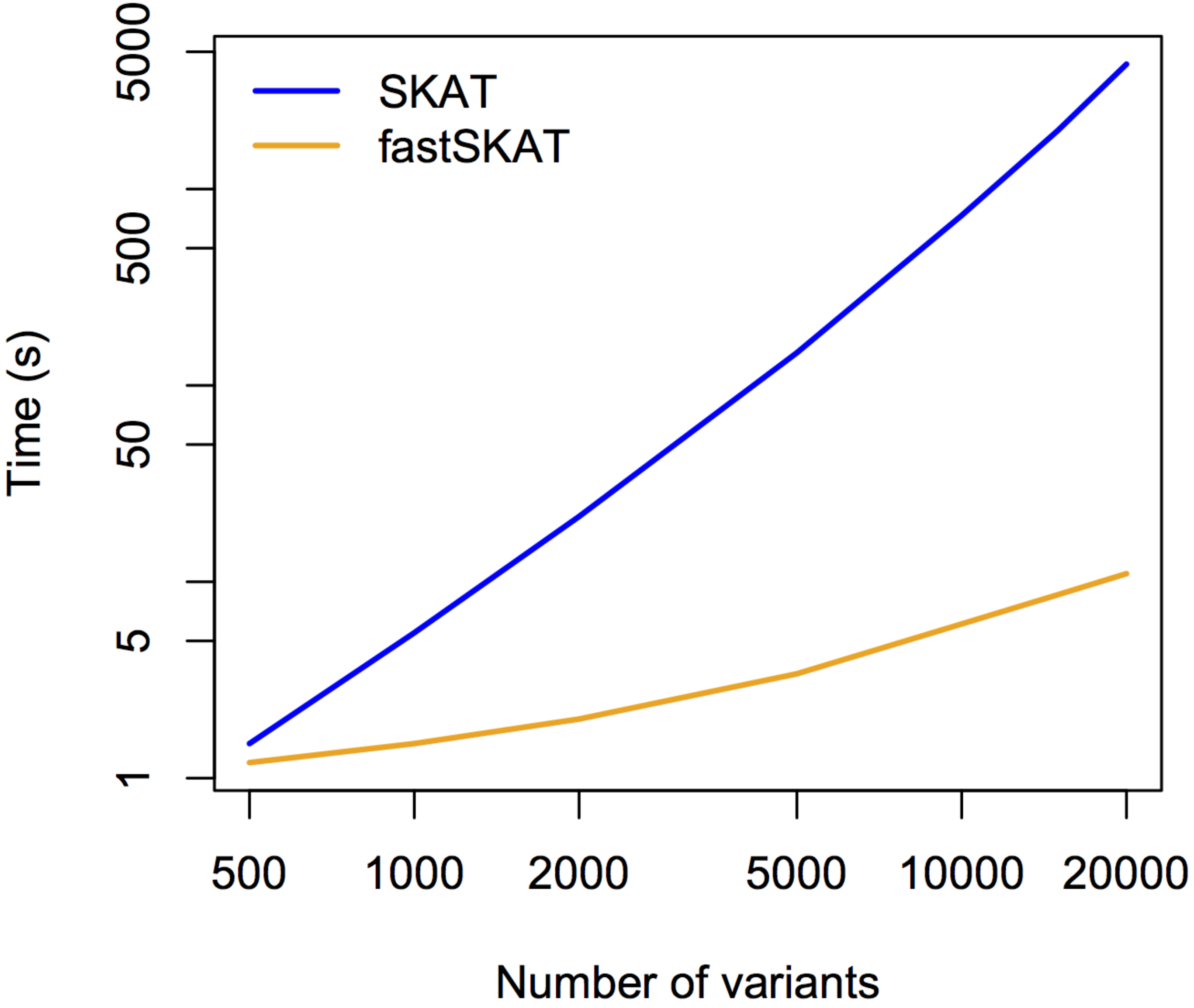
- Yes, as well as SKAT does;

LDL-C; 17259 gene regions with 1k-7k variants within ± 50 kb



How much faster?

- For N=5000;
- Exploits sparse genotypes, here
- 3 orders of magnitude faster, for large M

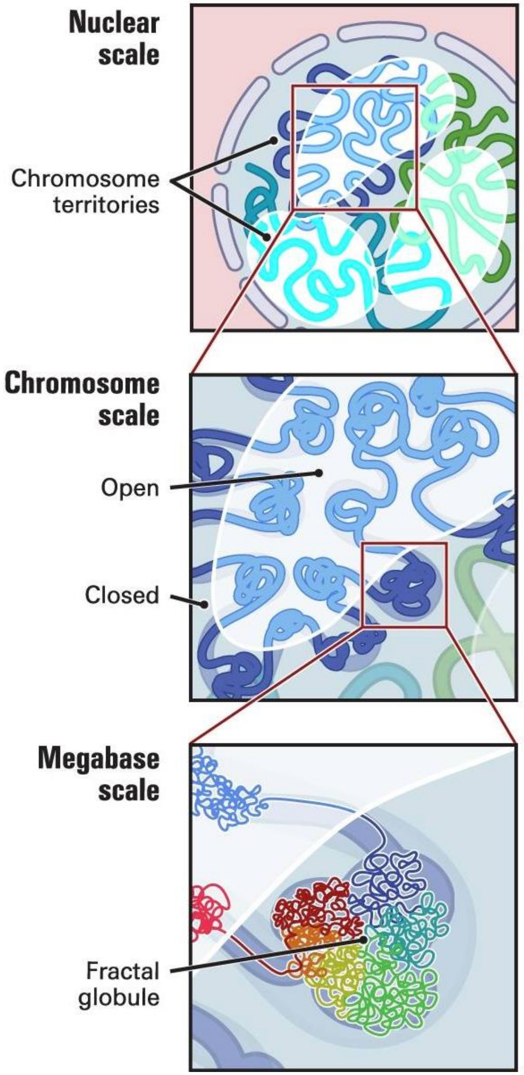


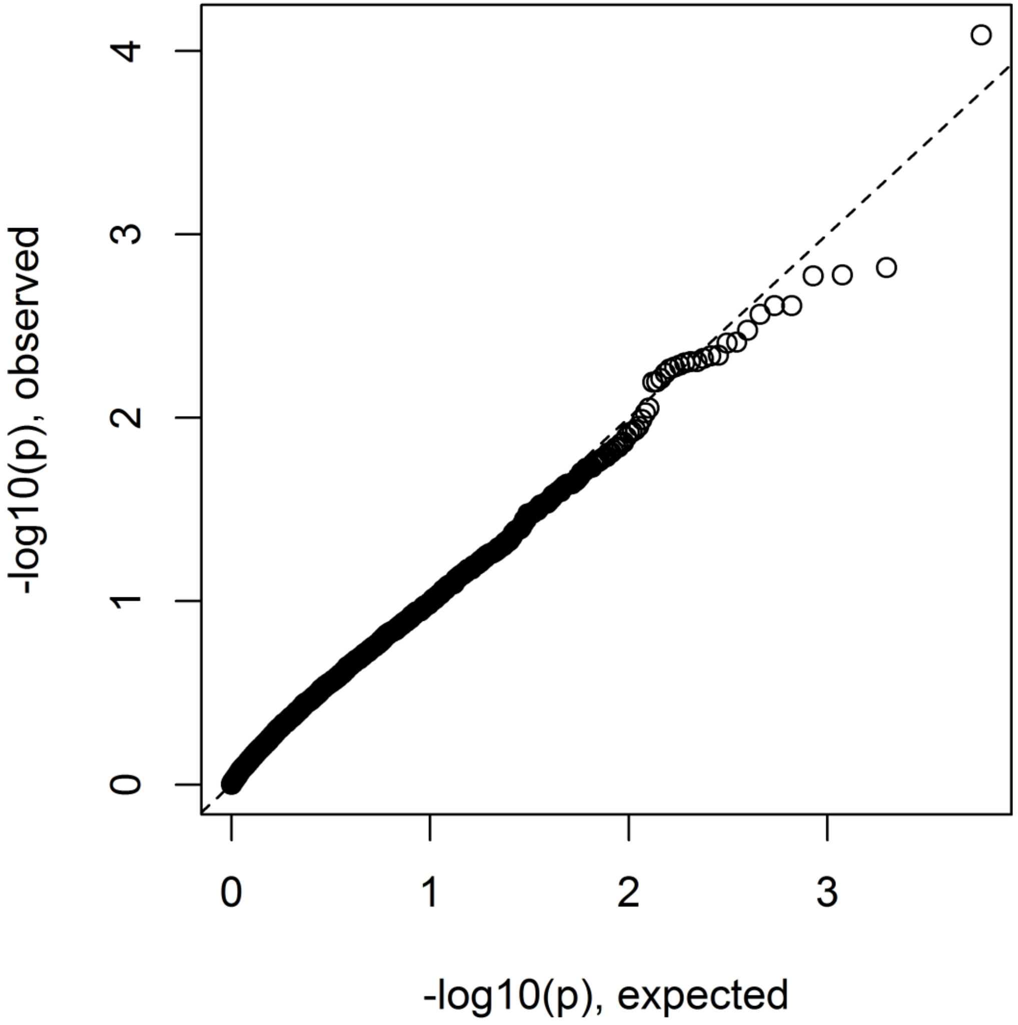
What new stuff can it do?

Investigate large variant sets (10k-100k) defined by structural or functional criteria

- Topologically Associating Domains
- Histone marks

Topologically Associating Domains



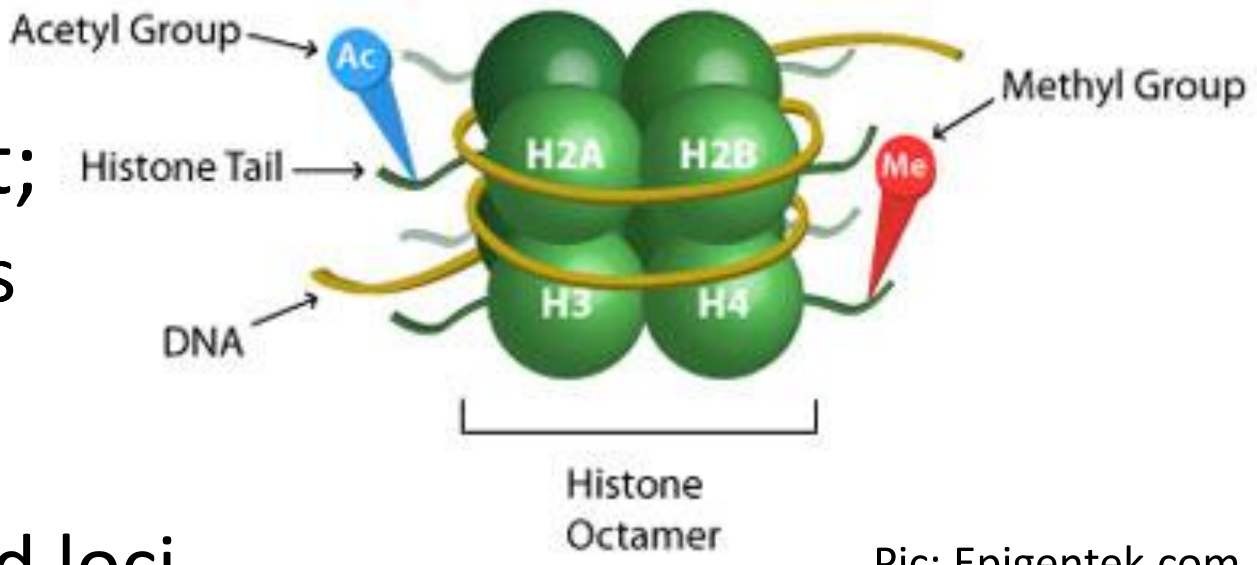


Lieberman-Aiden et al (2005, Science)

- average 1Mb, 10k-20k variants
- top hit contains APOE, not quite significant

- **fastSKAT** is 2400 times faster

Histone marks



Pic: Epigentek.com

- Analyzed rare variants that;
  - fall within regulatory marks of six different histones annotated in adult liver
  - within 500Kb of known lipid loci
  - **aggregated over a whole chromosome** (up to M=100k)
- Control: random variants in same regions
- Two signals ($p=10^{-5}$) on chromosome 19 (likely APOE)

Can **fastSKAT** handle…

- Binary data? Yes
- Survival data? Not yet
- Parallel processing? Will be straightforward
- Family data from pedigrees? Yes with mixed models (GMMAT)
- Empirical kinship matrices? Not yet

- Software: github.com/tslumley/bigQF
- Manuscript: read it on the plane home!
- Underlying math: Halko et al (ArXiv) *Finding structure with randomness*

Any questions?

Thanks to:

- Thomas Lumley, Jen Brody, Gina Peloso
- CHARGE Lipids Working Group
- TOPMed Analysis group and Data Coordinating Center
- Analysis Commons on DNAnexus
- University of Washington Genetic Analysis Center

We are recruiting research scientists –
– email Cathy Laurie: cclaurie@uw.edu... **fast!**