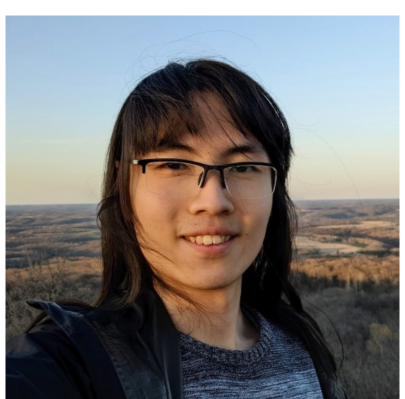
MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MIT STATISTICS + DATA SCIENCE CENTER

LABORATORY FOR INFORMATION & DECISION SYSTEMS
LIDS

Double trouble: Predicting new variant counts across two heterogeneous populations


Yunyi Shen
MIT


Lorenzo Masoero
Amazon Research


Joshua Schraiber
USC


Tamara Broderick
MIT

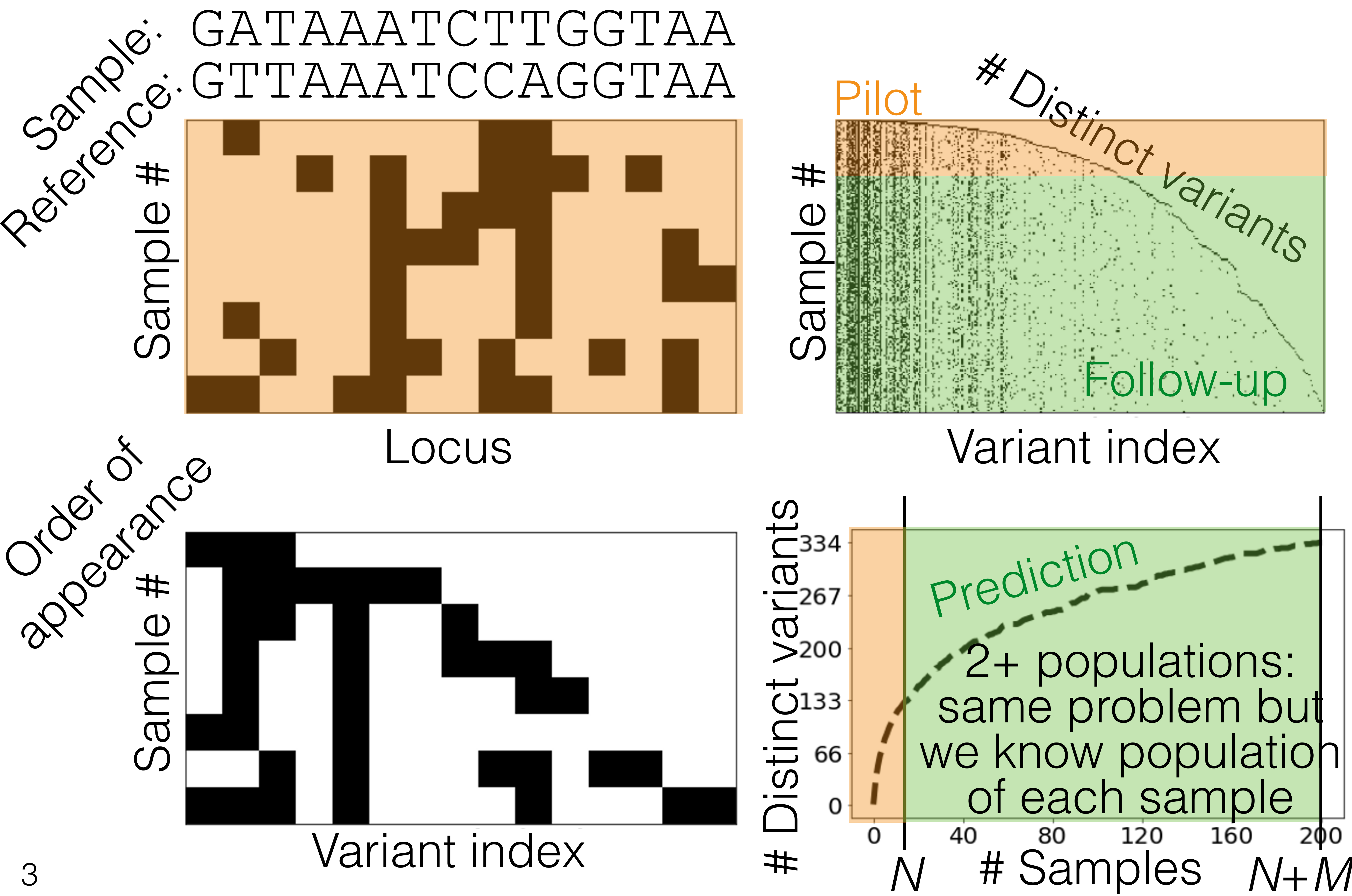
Planning a new genetics study

- Often want to collect genomics sequencing data across different populations
  - E.g. cases & controls to understand a disease
  - E.g. different cancer types
- Despite sequencing advances, scientists still often constrained by resources
- Would like to know how much we'll learn from a follow-up study given data from a (typically small) pilot study
  - Predict number of new genetic *variants* (points of difference relative to a reference genome)
- Lots of methods to predict in one population. But can't just group or separate two heterogeneous populations.
  [Camerlenghi+ 2024, Masoero+ 2022, Chakraborty+ 2019, Zou+ 2016, Gravel+ 2014, Ionita-Laza+ 2009]
- **We provide: the first method to predict the number of new variants across and between two populations**

Roadmap

- Setup: predicting the number of new variants
- A Bayesian framework for one population
- Natural extensions to two populations fail
- Our new model for two populations
  - Desirable theoretical properties
  - Good performance on real genetics data

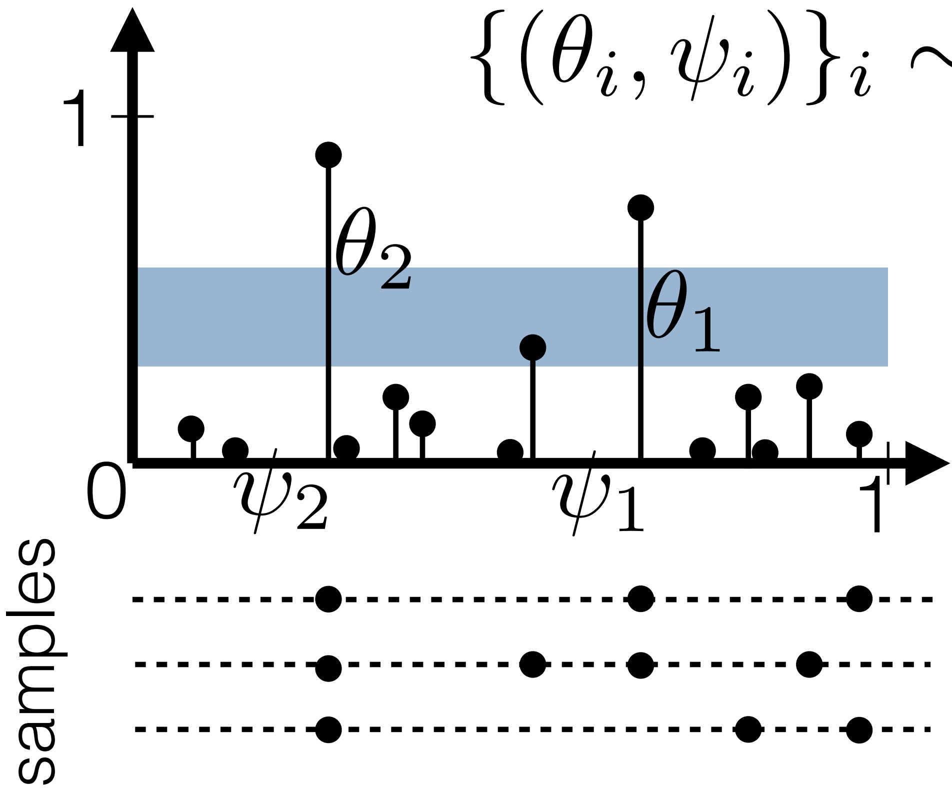
Predicting the number of new variants

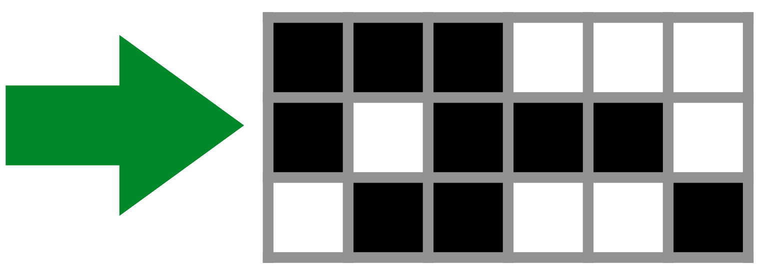


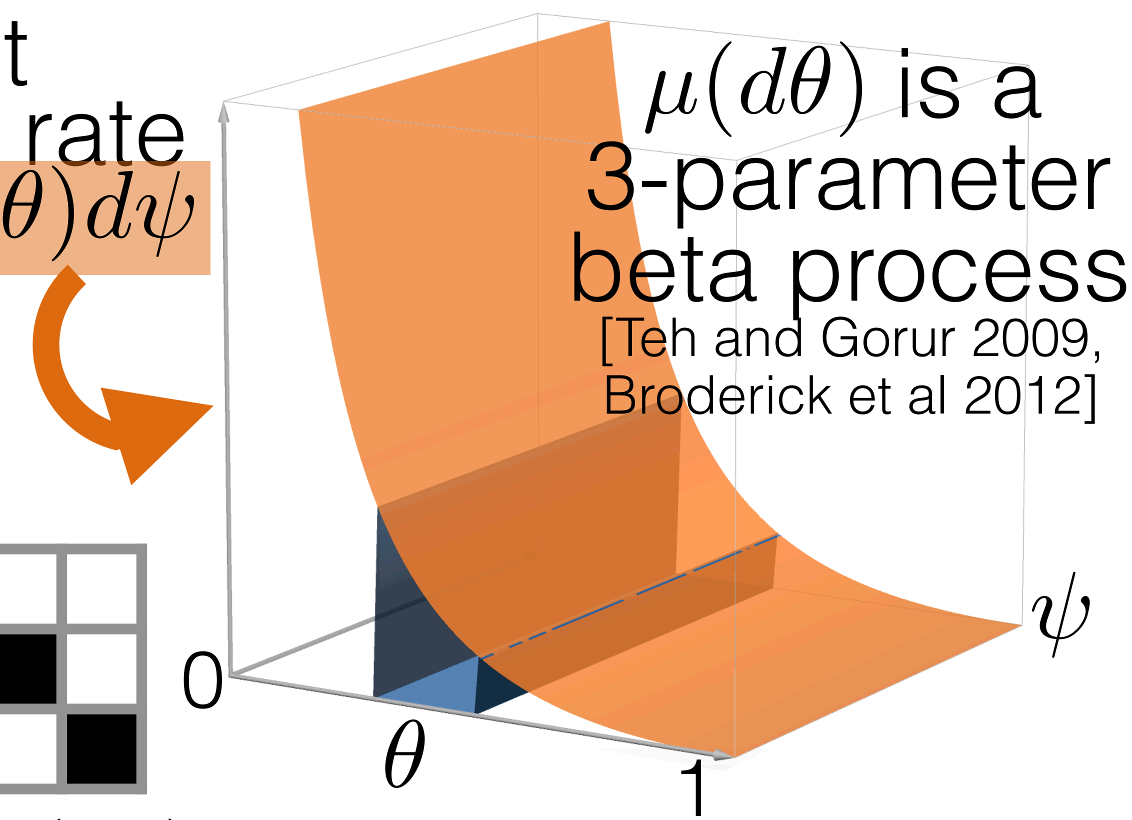
A Bayesian framework

- Masoero et al 2022: state-of-the-art prediction for the number of new variants in one population. Model:

$\{(\theta_i, \psi_i)\}_i \sim$ Poisson point process with rate measure $\mu(d\theta)d\psi$







- How to choose the rate measure $\mu(d\theta)$? Desiderata:
  - A finite number of variants per sample: $\int_0^1 \theta \mu(d\theta) < \infty$
  - There are always more variants to discover: $\int_0^1 \mu(d\theta) = \infty$
  - Power law growth (#variants/#samples*power* → 1 a.s.)
  - Conjugate rate measure for practical computation [Broderick et al 2018]
- Bonus benefits: can vary sequencing depth, tradeoff quality (depth) vs. quantity (samples) under a fixed budget

What about two+ populations?

- Idea: treat the two populations as disjoint, with no shared variants. Apply one-population methods separately.
  - Problem: In real-life, there are shared variants. In fact, we'd like to predict how many in future samples.
- Idea: group everything into a single population.
  - Problem: Populations exhibit different growth rates.
- Idea: take an approach analogous to previous slide
  - A variant's frequency in two populations: $\boldsymbol{\theta}_i = (\theta_{i,1}, \theta_{i,2})$
  - Draw the tuples of variant frequencies from a Poisson point process with rate measure $\nu(d\boldsymbol{\theta})$
  - A sample in population *p* exhibits variant *i* with probability equal to $\theta_{i,p}$
- But how to choose $\nu(d\boldsymbol{\theta})$?
  - A natural idea: $\nu(d\boldsymbol{\theta}) = \mu_1(d\theta_1)\mu_2(d\theta_2)$

The factorized extension fails

- Desiderata:
  A. Finite number of variants per sample.
  B. Always more variants to discover in either population.
- **Theorem**: Assume we use the two-population framework on the previous slide. We can't satisfy Desiderata A and B and factorize $\nu(d\boldsymbol{\theta}) = \mu_1(d\theta_1)\mu_2(d\theta_2)$
- **Rough proof intuition**:
  - By the factorization & Desideratum B, at least one direction (let's say population 1) has infinite mass.
    $$\infty = \int \nu(d\boldsymbol{\theta}) = \int \mu_1(d\theta_1) \int \mu_2(d\theta_2)$$
  - To find the expected number of variants in population 2:
    - Given the factorization, we directly take the integral of population 1, which has infinite mass.
      $$\int \theta_2 \nu(d\boldsymbol{\theta}) = \int \mu_1(d\theta_1) \int \theta_2 \mu_2(d\theta_2)$$
    - So the expectation is infinite, a contradiction with A.

Benefits of our new model

- We propose a new rate measure that doesn't factorize (exact rate measure form on next slide)
- We show that our new proposed rate measure:
  - (**Proposition**) Satisfies Desiderata A & B
    - A: Finite number of variants per sample
    - B: Always more variants to discover
  - (**Theorem**) Exhibits desirable power-law behavior
    - Consider projection to one population or proportional sampling of populations.
    - Our theory on arXiv is rough; better results on the way!
  - (**Proposition**) Is conjugate.
    - Not as nice computationally as the one-population beta process though.
    - Admits a feasible hyperparameter-selection algorithm.

Our new rate measure

- Review: One version of a 3-parameter beta process:

$$\mu(d\theta) \propto \alpha\theta^{-1-\sigma}(1-\theta)^{c-1}\, d\theta$$

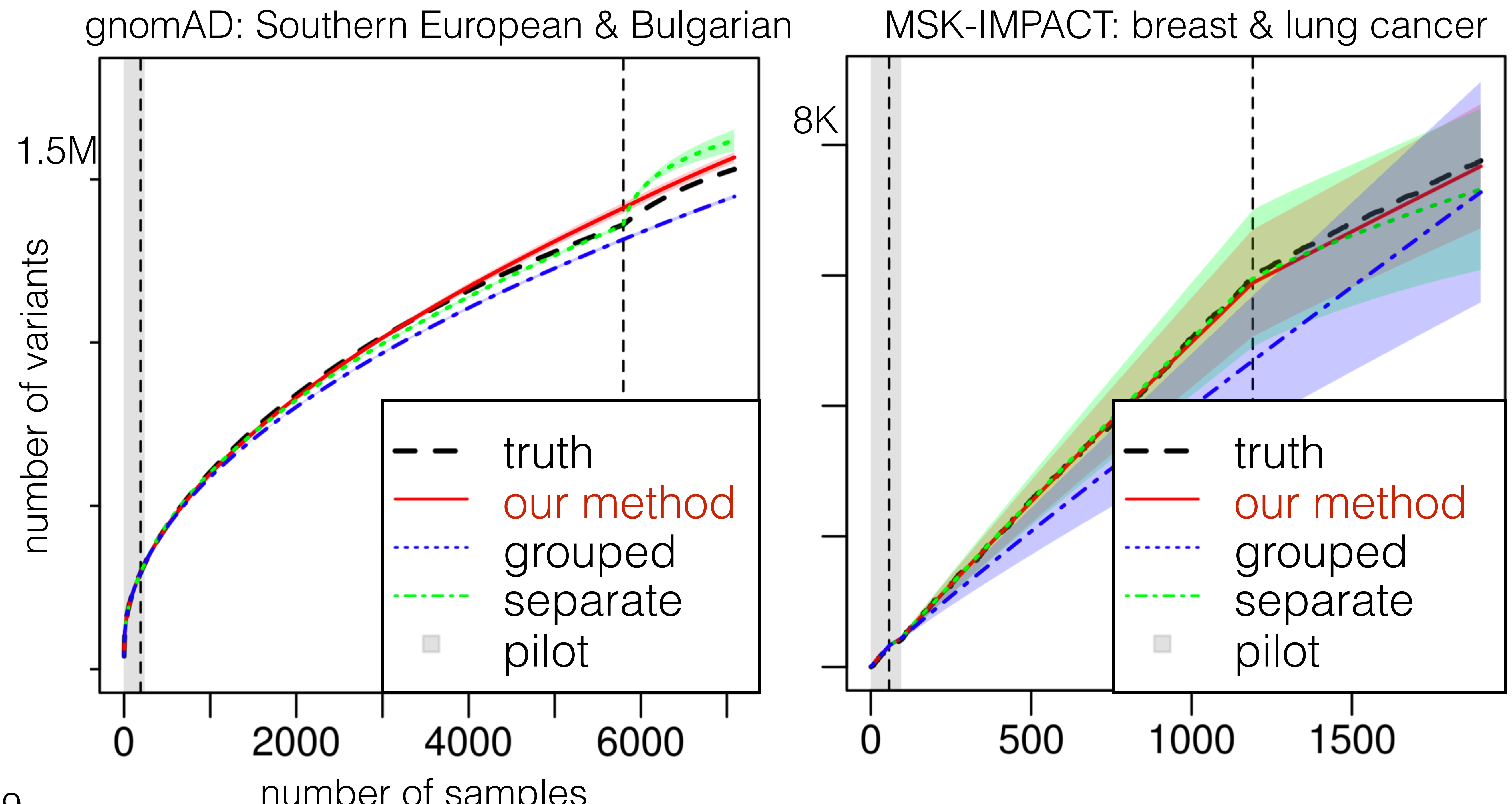
  - Improper beta distribution (Desiderata A,B & conjugacy)
  - Rate parameter $\sigma \in (0,1)$ controls power-law rate
  - Mass parameter α scales expected total # variants
  - Concentration c controls common-variant frequencies
- Our rate measure for two populations (better options?)

$$\nu(d\boldsymbol{\theta}) \propto \alpha \frac{(\theta_1 + \theta_2^{\sigma_2/\sigma_1})^{-\sigma_1}}{(\theta_1+\theta_2)^{\gamma_1+\gamma_2}} \cdot \theta_1^{\gamma_1-1}(1-\theta_1)^{c_1-1} \cdot \theta_2^{\gamma_2-1}(1-\theta_2)^{c_2-1}\, d\boldsymbol{\theta}$$

  - Two proper beta distributions times a non-factorizable term that makes the density improper (A,B,conjugacy)
  - Unique parameter in each population: rate σ_p, concentration c_p, (new) correlation γ_p
  - Single mass parameter α
  - If $\sigma_1 = \sigma_2, \theta_1 = \rho\theta_2$ ➡ $\nu(d\boldsymbol{\theta}) \propto \alpha\theta_1^{-1-\sigma_1}(1-\theta_1)^{c_1+c_2-1}\, d\boldsymbol{\theta}$

Predicting number of new variants

- Our method improves on (1) treating the two populations as disjoint, with no shared variants, or (2) grouping everything into a single population



Conclusions

- We predict the number of new genetic variants for a follow-up study given a pilot study (both the total number and the shared number). **We provide the first predictor that can handle heterogeneity in multiple populations.**
    - Y Shen, L Masoero, J Schraiber, T Broderick. Double trouble: Predicting new variant counts across two heterogeneous populations. ArXiv.

See also:

- Masoero, Camerlenghi, Favaro, Broderick. More for less: predicting & maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika*, 2022.
- Broderick, Wilson, Jordan. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 2018.
- Broderick, Jordan, Pitman. Beta processes, stick-breaking, and power laws. *Bayesian Analysis*, 2012.
- Campbell, Cai, Broderick. Exchangeable trait allocations. *Electronic Journal of Statistics*, 2018.
- Broderick, Pitman, and Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 2013.