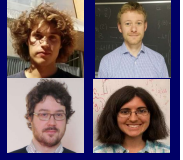
Getting the most bang for your buck:
Predicting and maximizing the number of new genomic variants in future experiments



Lorenzo Masoero[1], Federico Camerlenghi[2], Stefano Favaro[3], Tamara Broderick[1] ▲ [1]MIT, CSAIL, [2]University of Milano, [3]University of Torino

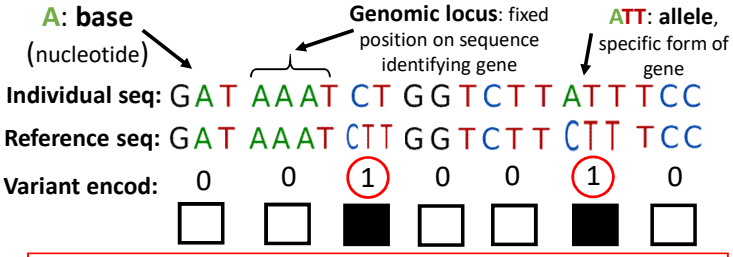
☺ Modern ML tackles increasingly complex real-world problems
☹ Often need extremely large amount of data
☹ Data-gathering can be expensive, requires careful planning

Need quantitative framework to know
1. How much can we ***expect to learn*** from new data
2. How to collect data to ***learn the most*** possible

We focus on **genomics** → *study* ***rare functional variants*** [rare diseases, precision medicine]

We provide:
1. Fast, accurate prediction of # new genomic variants to be seen in future samples
2. Optimal experimental design: how to design follow-up studies in genomics to learn the most

Data sketch & problem description

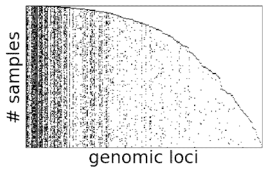


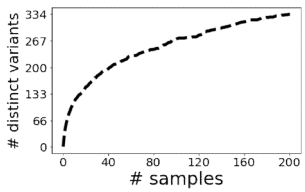
A **variant** is ``called'' whenever **allele differs from reference genome**

- **Goal**: design future experiments to find **rare functional variants**
- **Problem:** hard, and costly: present in few individuals, need a lot of data; data-gathering is expensive ☹
- **Available methods:** predict # new variants as we get more data, **but…**
  - ☹ Can be inaccurate (***Ionita-Laza et al. [2009]***, ***Zou et al. [2016]***)
  - ☹ No guidance on optimal sampling under a fixed budget (***Gravel [2014]***)
  - ☹ Can't accommodate changes in experimental conditions
- **Our method: BNP approach for optimal experimental design**
  - ☺ **Fast, accurate predictions for # new variants**
  - ☺ **Useful guidance on design of future experiments**
  - ☺ **Able to accommodate changes in experimental setup: useful for reliable optimal experimental design**

Bayesian nonparametric model

- BNP models grow in complexity as more data is collected
- We use the Indian buffet process [**IBP**]: prior for **binary matrices**





We fit this curve: **learn** hyperparameters of the IBP

- $\alpha > 0$: *mass parameter*: scales total # variants observed
- $\sigma \in [0,1)$: *discount parameter*: controls power law behavior
- $c > \sigma$: *concentration parameter*: modulates frequencies of widespread counts

Derive $U_N^{(M)}$: # new variants in M additional samples given N initial $X_{1:N}$

$$\underbrace{U_N^{(M)}}_{\text{\# new variants in M additional samples}} \mid \underbrace{X_{1:N}}_{\text{given N initial samples}} \sim Pois\Big(\underbrace{\alpha \sum_{m=1\dots M} \frac{(c+\sigma)_{N+m-1}}{(c+1)_{N+m-1}}}_{:= P_N^{(M)}}\Big)$$

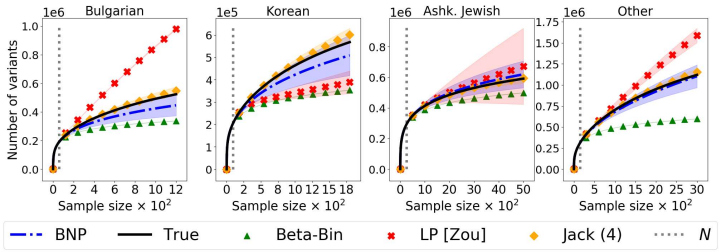
Use posterior mean $P_N^{(M)}$ as pointwise predictor via **empirical Bayes**:

$$\alpha^*, \sigma^*, c^* = argmin_{\alpha,\sigma,c}\Big[\sum_{m=1}^{N-n}\{P_n^{(m)}(\alpha,c,\sigma) - \big(U_n^{(m)} \mid X_{1:n}\big)\}\Big]$$

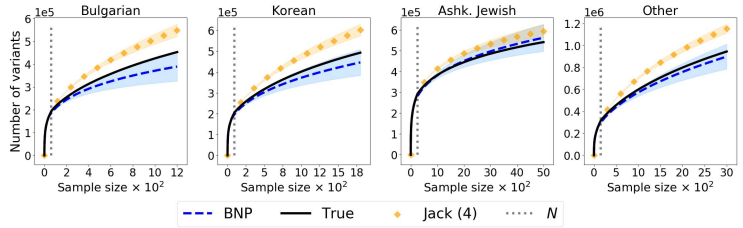
In practice: $n := \lfloor \frac{2N}{3} \rfloor$

Experiments: prediction and optimal design

- Sanity check: test predictor $P_N^{(M)}$ on synthetic and real data
- Performs comparably to best alternative when **experimental conditions don't change** between pilot and follow-up studies



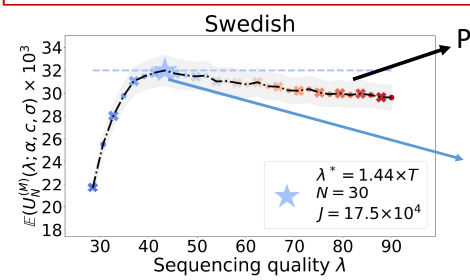
- # new variants depends on samples' quality
- when experimental conditions change:
  - **competing methods fail to capture change in samples' quality between pilot and follow up**
  - **BNP approach can successfully adapt and produce useful predictions**



Given a **pilot** study, design **follow-up** to maximize amount of new information (# new variants) under a fixed budget:
Under a **fixed budget,** solve **quality/quantity** tradeoff:
- **Quality**: sequencing depth λ
- **Quantity**: # new samples M

Solve **constrained optimization** using BNP predictor

$$M^*, \lambda^* = argmax_{M,\lambda} E[U_N^{(M)}(\lambda;\alpha,c,\sigma) \mid X_{1:N}]$$
$$\textbf{subject to: } Cost(M,\lambda) \le Budget$$



Points on the curve: feasible M, λ

M^*, λ^* maximizing $\boldsymbol{E[}U_N^{(M)}(\lambda;\alpha,c,\sigma)\boldsymbol{]}$ under a fixed budget