More for less: Predicting and maximizing genetic variant discovery via Bayesian nonparametrics

Lorenzo Masoero* Federico Camerlenghi† Stefano Favaro‡
Tamara Broderick§

Abstract

While the cost of sequencing genomes has decreased dramatically in recent years, this expense often remains non-trivial. Under a fixed budget, then, scientists face a natural trade-off between quantity and quality: spending resources to sequence a greater number of genomes (quantity) or spending resources to sequence genomes with increased accuracy (quality). Our goal is to find the optimal allocation of resources between quantity and quality. Optimizing resource allocation promises to reveal as many new variations in the genome as possible. In this paper, we introduce a Bayesian nonparametric methodology to predict the number of new variants in a follow-up study based on a pilot study. We validate our method on cancer and human genomics data. When experimental conditions are kept constant between the pilot and follow-up, we find that our prediction is competitive with the best existing methods. Unlike current methods, though, our new method allows practitioners to change experimental conditions between the pilot and the follow-up. We demonstrate how this distinction allows our method to be used for more realistic predictions and for optimal allocation of a fixed budget between quality and quantity.

1 Introduction

New genomics data promise to reveal more of the diversity, or variation, among organisms, and thereby new scientific insights. However, the process of collecting genetic data requires resources, and optimal allocation of these resources is typically a challenging task. Under a fixed budget constraint, there is often a natural trade-off between quality and quantity in genetic experiments. Sequencing genomes at a higher quality reveals more details about individual organisms' genomes but incurs a higher cost. Similarly, sequencing a greater number of genomes reveals more about variation across the population but also costs more to accomplish. It is then critical to understand how to

*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, lom@mit.edu
†Department of Economics, Management and Statistics, University of Milano-Bicocca
‡Department of Economic and Social Sciences, Mathematics and Statistics, University of Torino
§Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

optimally allocate a fixed budget between quality and quantity in genetic experiments, in the service of learning as much as possible from the experiment.

To maximize the amount learned from a genetic experiment, we first need to quantify a notion of "amount learned". Scientists use a reference genome for a species of interest in an experiment; a (genetic) *variant* is any difference in an observed genome relative to the reference genome. Variants facilitate understanding of evolution [Consortium, 2015, Mathieson and Reich, 2017], diversity of organisms [Consortium, 2015, Sirugo et al., 2019], oncology [Chakraborty et al., 2019], and disease [Cirulli and Goldstein, 2010, Zuk et al., 2014, Bomba et al., 2017]. Thus, the number of observed variants is a concrete metric of "amount learned" from a genetic experiment. For optimal budget allocation, then, we first predict the number of new variants in the follow-up study under different allocations of budget with respect to quality and quantity; next we choose the experimental setting that maximizes the number of new variants.

Optimal budget allocation supports scientists who face resource constraints. In research on non-human and non-model organisms, small sequencing studies are often conducted under limited budgets [da Fonseca et al., 2016]. The development of reliable and inexpensive sequencing pipelines is thus an active research area [Peterson et al., 2012, Souza et al., 2017, Aguirre et al., 2019]. Accurate prediction of the number of new variants can also be important for understanding the site of origin of cancers as well as the clonal origin of metastasis [Chakraborty et al., 2019]. And in precision medicine, accurate estimation of the number of new rare variants can aid effective study design and evaluation of the potential and limitations of genomic datasets [Momozawa and Mizukami, 2020, Zou et al., 2016]. We detail further potential applications in microbiome research, single-cell sequencing, and wildlife monitoring in Section 7.

There exists a rich statistical literature on prediction in a follow-up study, relative to a pilot study, when conditions do not change between the pilot and follow-up. We may think of each organism as belonging to multiple groups, where each group is defined by a variant, and the goal is to discover the number of new groups in a follow-up study. A simpler special case of this formulation occurs when each organism belongs to a single group, which is referred to as a species [Good and Toulmin, 1956, Efron and Thisted, 1976, Lijoi et al., 2007, Orlitsky et al., 2016]. In general, as in the context of genetic variation, organisms belong to multiple groups that we refer to as *features*. Researchers have developed a wide range of approaches for predicting the number of new features, often interpreted as amount of new genetic variation, in a follow-up study. These approaches include Bayesian methods [Ionita-Laza et al., 2009], jackknife-based estimators [Gravel, 2014], linear programming methods [Gravel, 2014, Zou et al., 2016], and variations on the classical Good-Toulmin estimator [Orlitsky et al., 2016, Chakraborty et al., 2019]. To the best of our knowledge, though, no existing work provides predictions when the experimental conditions may change between the pilot and follow-up study. And thus no existing work can be used directly for optimal allocation of a fixed budget in experimental design.

Moreover, while there is existing work in other forms of optimal experimental design, it does not fit our goals here. In pioneering work, Ionita-Laza and Laird [2010] propose how to allocate a fixed budget in a pilot study, before any data is observed. While their method treats every dataset the same, our method allows the different variation patterns in different datasets to inform the best follow-up design. Separately,

researchers have considered how to best choose samples among a number of subpopulations [e.g. Dumitrascu et al., 2018, Camerlenghi et al., 2020]. In this case, the trade-off is between uncertainty and reward, as in classic multi-armed bandit settings, rather than between quality and quantity.

In the present work, we propose a Bayesian nonparametric methodology to predict the number of new variants to be discovered in a follow-up study given observed data from a pilot study. Critically, our approach works when the experimental conditions change between the pilot and follow-up. We then demonstrate how to apply the proposed methodology for optimal budget allocation in the design of a follow-up study given data available from a pilot study. Here, for prediction, we build on a classic Bayesian nonparametric framework for feature allocations known as the beta-Bernoulli process [Hjort, 1990, Kim, 1999, Thibaux and Jordan, 2007, Teh and Gorur, 2009, Broderick et al., 2012]. The posterior distributions of all our predicted quantities, such as the number of new variants to be discovered, are available in closed-form expressions. Our corresponding Bayesian estimators are simple, computationally efficient, and scalable to massive datasets. In addition, our Bayesian nonparametric framework captures realistic power-law behaviors in genetic data. We will see that, when the pilot and follow-up studies are constrained to have the same experimental setup as in previous work, our predictions are competitive with the state-of-the-art and superior to a number of recent proposals. Most importantly, though, we demonstrate that our predictions maintain their accuracy when experimental conditions change between the pilot and follow-up. Finally, we give an empirical demonstration of how our predictions can be used for designing the follow-up study with an optimal allocation of a fixed budget between quality and quantity. We validate the proposed methodology on synthetic and real data, with a focus on human genomics. Specifically, we consider the TCGA and MSK-impact datasets [Cheng et al., 2015], as well as the recent gnomAD dataset of Karczewski et al. [2020].

2 Data and modeling assumptions

Modern high-throughput sequencing technologies allow accurate determination of an organism's genome [Reuter et al., 2015]. A reference genome serves as a fixed representative, and variants relative to the reference genome can take many forms, including deletions, inversions, translocations, and insertions; see Taylor and Taylor [2004] and references therein. In the present work, we do not distinguish between different forms of variants, though in Section 7 we briefly discuss how our framework could be extended to make this distinction. To establish notation and start building up to our Bayesian nonparametric model, we first assume that the process of observing variants is flawless; we develop a more realistic model for observations in Section 3.2.

Suppose there are J variants observed among the N pilot genomes, $0 \leq J < +\infty$, with ψ_j the label of the j-th variant in order of appearance. Let $x_{n,j}$ equal 1 if the variant with label ψ_j is observed for the n-th organism; otherwise, let $x_{n,j}$ equal 0. We collect data for the n-th organism in $X_n := \sum_{j=1}^{J} x_{n,j}\delta_{\psi_j}$, which pairs each variant observation with the corresponding variant label by putting a mass of size $x_{n,j}$ at location ψ_j. We use the notation $X_{N_1:N_2}$, where $N_1 \leq N_2$, to denote

$(X_{N_1}, X_{N_1+1}, X_{N_1+2}, \ldots, X_{N_2})$. Given the observable $X_{1:N}$, we consider a Bayesian approach to predict the number of variants in the follow-up study. Specifically, letting Θ be an appropriate latent parameter, we specify a generative model via a likelihood function $\mathrm{pr}(X_{1:N} \mid \Theta)$ and a prior distribution $\mathrm{pr}(\Theta)$. Technically there is a fixed, and finite, upper bound on the number of possible variants established by the size of any individual genome. But this bound is usually much larger, often by orders of magnitude, than the number of observed variants. Moreover, in practice, we expect that no study of any practical finite size N will reveal all possible variants, simply because some variants are so exceedingly rare. Bayesian nonparametric methods allow us to avoid hard-coding an unwieldy, large finite bound that may cause computational and modeling headaches. In particular, they allow the observed number of variants to be finite for any finite dataset and grow without bound, in such a way that computation typically scales closely with the actual number of variants observed. Formally, we imagine a countable infinity of latent variants, labelled as $\{\psi_j\}_{j\geq 1}$, and we write $X_n := \sum_{j\geq 1} x_{n,j}\delta_{\psi_j}$; since $x_{n,j} = 0$ for all unobserved variants, this equation reduces to the previous definition of X_n above.

Following existing methods for estimating new-variant cardinality [Ionita-Laza et al., 2009, Gravel, 2014, Zou et al., 2016, Orlitsky et al., 2016, Chakraborty et al., 2019], we assume that every variant appears independently of every other variant; that is, $x_{n,j}$ is independent of $x_{n,k}$ across all n for $j \neq k$. In reality, nearby positions on a genome can be highly correlated; this phenomenon is called linkage disequilibrium. However, our assumption has two principal advantages: (i) it makes our computations much easier; (ii) it is supported by our state-of-the-art empirical results in Section 6. We also make the milder assumption that organisms are (infinitely) exchangeable; roughly, we assume that the order in which we observe the sample organisms is immaterial for any sample size N. Since, for the moment, we assume variant observation is flawless, this assumption presently translates into an exchangeability assumption on the observed data. More precisely, let $[N] := \{1, \ldots, N\}$, and let σ_N represent a permutation of $[N]$. Then, for the variant with label ψ_j, for any N and any σ_N, we assume $\mathrm{pr}(x_{1,j}, \ldots, x_{N,j}) = \mathrm{pr}(x_{\sigma_N(1),j}, \ldots, x_{\sigma_N(N),j})$. Indeed, if we expected systematic variation among organisms in our population between earlier and later samples, we would find it difficult to predict future data from past data without knowing more about the nature of the variation.

Exchangeability of $\{x_{n,j}\}_{n\geq 1}$ implies the existence of a random variable θ_j, i.e. the variant's proportion, such that the $x_{n,j}$ are Bernoulli draws with parameter θ_j, independently and identically distributed across n [de Finetti, 1931]. We pair each θ_j with its variant's label ψ_j in a random measure $\Theta := \sum_{j\geq 1} \theta_j \delta_{\psi_j}$, and we assume the X_n's are conditionally independent and identically distributed given Θ. In addition, we make the following modeling assumptions: (i) the conditional distribution of X_n given Θ is the distribution of a Bernoulli process (BeP) with parameter Θ, and we write $X_n \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\Theta)$; (ii) the prior distribution on Θ is the law of the three-parameter beta process (3BP) [Teh and Gorur, 2009, Broderick et al., 2012]. In agreement with the assumption of independence for $\{x_{n,j}\}_{n\geq 1}$ across j, we can interpret the three-parameter beta process as a collection of independent priors on the θ_j such that it satisfies our goals: (G1) a finite number of observed variants in any finite sam-

ple; (G2) a number of observed variants that is unbounded as the number of samples grows. Furthermore, the three-parameter beta process is able to capture power-law behaviours [Teh and Gorur, 2009, Broderick et al., 2012], which are common in physical processes. The three-parameter beta process is characterized by: (i) a mass parameter α that scales the total number of variants observed; (ii) a discount parameter σ that controls the power-law growth in observed variant cardinality; (iii) a concentration parameter c that modulates the frequency of more widespread variants.

We say that the random measure Θ is distributed as a three-parameter beta process, $\Theta \sim 3\mathrm{BP}(\alpha, \sigma, c)$, if $\Theta = \sum_{j\geq 1} \theta_j \delta_{\psi_j}$, with $\{\theta_j\}$ drawn from a Poisson process with rate measure

$$\nu(\mathrm{d}\theta) = \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta^{-1-\sigma}(1-\theta)^{c+\sigma-1}\mathbf{1}(\theta \in [0,1])\mathrm{d}\theta,$$

where $\mathbf{1}(A)$ stands for the indicator function of the event A. The ψ_j's serve merely to distinguish the variants, so it is enough to ensure that they are all almost surely distinct. Thus we take $\psi_j \overset{iid}{\sim} \mathrm{Uniform}[0,1]$. The Poisson point process representation is convenient in our proofs. To meet goals G1 and G2, the three-parameter beta process hyperparameters must satisfy: $\alpha > 0$, $c > -\sigma$, and $\sigma \in [0,1)$ [Teh and Gorur, 2009, James, 2017, Broderick et al., 2018].

3 Predicting the number of new variants

3.1 Initial proposals for prediction

In Section 2 we introduced and motivated a Bayesian nonparametric model consisting of: (i) a Bernoulli process likelihood function, $X_n \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\Theta)$, for observed variants conditioned on variants' proportions; (ii) a three-parameter beta process prior, $\Theta \sim 3\mathrm{BP}(\alpha, \sigma, c)$ over variants' proportions. Now we use this model to predict the number, $U_N^{(M)}$, of new variants in a follow-up study of size $M \geq 1$ after an initial pilot study of size $N \geq 1$:

$$U_N^{(M)} := \sum_{j\geq 1} \mathbf{1}\left(\sum_{n=1}^{N} x_{n,j} = 0\right)\mathbf{1}\left(\sum_{m=1}^{M} x_{N+m,j} > 0\right).$$

We derive the posterior distribution of $U_N^{(M)}$ given $X_{1:N}$. So the expected value of the posterior distribution is a Bayesian nonparametric estimator of $U_N^{(M)}$ with respect to a squared loss function. With a slight abuse of notation, for any two random variables X and Y defined on the same probability space we let $X \mid Y$ denote the random variable whose distribution coincides with the conditional distribution of X given Y. We write $\mathcal{N}(\mu, \rho^2)$ for a Gaussian random variable with mean μ and variance ρ^2, and we let $(a)_{b\uparrow} := \prod_{i=1}^{b}(a+i-1)$ denote the rising factorial.

Proposition 1. *Let* $\Theta \sim 3\mathrm{BP}(\alpha, \sigma, c)$ *and* $X_n \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\Theta)$ *for* $n = 1, \ldots, N$ *and* $N \geq 1$. *Then,*

$$U_N^{(M)} \mid X_{1:N} \sim \text{Poisson}\left\{\alpha \sum_{m=1}^{M} \frac{(c+\sigma)_{(N+m-1)\uparrow}}{(c+1)_{(N+m-1)\uparrow}}\right\}. \tag{1}$$

From Theorem 1, the Bayesian nonparametric estimator of $U_N^{(M)}$ under squared loss is

$$P_N^{(M)} := E\left(U_N^{(M)} \mid X_{1:N}\right) = \alpha \sum_{m=1}^{M} \frac{(c+\sigma)_{(N+m-1)\uparrow}}{(c+1)_{(N+m-1)\uparrow}}.$$

$P_N^{(M)}$ predicts the number of new variants in a follow-up study. In the next result we show that the distribution of $U_N^{(M)} \mid X_{1:N}$ exhibits almost-sure power-law growth in the sample size N with power determined by the three-parameter beta process hyperparameters. We also characterize asymptotic noise around the posterior predictive mean. See Appendix A for proofs of Theorem 1 and Theorem 2.

Proposition 2. *Under the setting of Theorem 1,*

$$\frac{U_N^{(M)}}{M^\sigma} \,\Big|\, X_{1:N} \xrightarrow{a.s.} \xi \quad \text{as } M \to +\infty, \tag{2}$$

where $\xi := \frac{\alpha}{\sigma}\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}$. *The Equation* (2) *limit holds almost surely, conditionally given* $X_{1:N}$. *Also,*

$$\sqrt{M^\sigma}\left(\frac{U_N^{(M)}}{M^\sigma} - \xi\right) \,\Big|\, X_{1:N} \xrightarrow{d} \mathcal{N}(0, \xi) \quad \text{as } M \to +\infty, \tag{3}$$

where the limit in Equation (3) *holds true in distribution.*

Besides $U_N^{(M)}$, researchers may be interested in relatively rare new variants since rare variants are known to play a role in disease predisposition [Cirulli and Goldstein, 2010, Saint Pierre and Génin, 2014, Bomba et al., 2017]. In particular, let $U_N^{(M,r)}$ denote the number of new variants that occur exactly r times in the follow-up study, and let $U_N^{(M,\leq R)}$ denote the number of new variants that occur at most R times in the follow-up study. A suitably chosen small value of r or R encodes a notion of rareness for variants. See Theorem 4 and Theorem 5 for a characterization of the posterior distributions of $U_N^{(M,r)}$ and $U_N^{(M,\leq R)}$ given $X_{1:N}$.

Our propositions reveal key attributes of our Bayesian nonparametric estimators. First and foremost, the posterior distribution of $U_N^{(M)}$ depends on $X_{1:N}$ only via the initial sample size N. See Equation (A.11) for similar behavior in the posterior distributions of $U_N^{(M,r)}$ and $U_N^{(M,\leq R)}$. Moreover, from Theorem 2 we see that the large-M asymptotic behavior of the posterior distribution of $U_N^{(M)}$ is completely determined by hyperparameters of the three-parameter beta process; see Theorem 5 for similar behavior in the posterior distributions of $U_N^{(M,r)}$ and $U_N^{(M,\leq R)}$. Therefore, learning

hyperparameters of the three-parameter beta process from the observed data is critical. In Section 4, we propose an empirical Bayes procedure to this end.

Our Bayesian nonparametric approach above, like existing approaches for estimating the number of new variants in a follow-up study [Ionita-Laza et al., 2009, Gravel, 2014, Zou et al., 2016, Orlitsky et al., 2016, Chakraborty et al., 2019], relies on the assumption that variants are always observed under the same conditions. Moreover, none of these methods account for how improved variant observation quality may incur a larger cost. But conditions may change between pilot and follow-up experiments, and these changes may be informed by an experimental budget. We address these issues below. In Section 3.2, we show how our general Bayesian nonparametric framework can be adapted to the case where variants are not observed perfectly; in fact, we show how we can adapt to different experimental conditions between the pilot and follow-up. Then, in Section 5, we build on the work of Ionita-Laza and Laird [2010] to show that we can optimize for the best conditions, to yield the most variants, in the follow-up. That is, we next consider the challenging problem of optimal allocation of a fixed budget between quality and quantity in genomic experiments: spending resources for sequencing a greater number of genomes (quantity) or spending resources for sequencing with increased accuracy (quality).

3.2 Accounting for sequencing errors

We extend the Bayesian nonparametric estimator introduced in Section 3.1 to account for non-trivial sequencing error. In Section 3.1 we have assumed that if any organism exhibits a variant, that variant is detected, i.e., $x_{n,j} = 1$ for organism n. However, in practice, sequencing a genome is a complex and noisy process. Millions of reads of fragments of the same genomic sequence need to be aligned and compared to the reference genome. Every position j of the genome of individual n is read a random number $D_{n,j}$ of times. $D_{n,j}$ is the (random) sequencing depth of the process. Out of these $D_{n,j}$ times, $D_{n,j,\text{err}}$ reads give rise to an error, due to technological imperfections, and are discarded. Here, $0 \leq D_{n,j,\text{err}} \leq D_{n,j}$. The remaining $D_{n,j,\text{noerr}} = D_{n,j} - D_{n,j,\text{err}}$ reads are correctly processed, aligned to the reference genome, and recorded [Ionita-Laza and Laird, 2010]. Every error-free read can either agree with the reference genome, or disagree. We let $C_{n,j} \in \{0, 1, \ldots, D_{n,j,\text{noerr}}\}$ denote the number of times that reads are correctly processed and we observe disagreement with the reference genome. Finally, a variant is said to be called whenever some discrepancy criterion, i.e. the variant calling rule, is satisfied.

Following Ionita-Laza and Laird [2010], we focus on simple threshold variant calling rules. That is, a variant is called whenever a sufficient number of reads disagree with the reference genome. Given the threshold value $T > 0$, variation is declared if the count $C_{n,j}$ exceeds T, i.e. $x_{n,j} = \mathbf{1}(C_{n,j} \geq T)$. This threshold variant calling rule is a simplification of actual variant callers used in modern genomic pipelines; see e.g. [Xu, 2018] for a review of variant calling algorithms. While simplistic, the threshold rule has the benefit of being easy to interpret; by contrast, state-of-the-art alternatives are much more complex, to the point of being somewhat inscrutable by their users. In fact, understanding how to tailor the variant calling rule to the data-gathering process is itself an active area of research [Hwang et al., 2015, Cornish and Guda, 2015, Kumaran

et al., 2019].

In setting up our model to account for sequencing error, we make the following additional assumptions. (i) Following standard practice in the genetics literature [e.g., Lander and Waterman, 1988, Ionita-Laza and Laird, 2010, Sampson et al., 2011], we assume that the sequencing depth $D_{n,j}$ is a Poisson random variable with parameter λ, which we refer to as the sequencing quality. (ii) The reads are independent and identically distributed across individuals and positions. (iii) p_{err} is a fixed probability of reading error that depends on the sequencing technology. (iv) Conditionally on $D_{n,j}$ total reads, the number of error-free reads $D_{n,j,\text{noerr}}$ is a binomial random variable, with $D_{n,j}$ as the number of trials and $1 - p_{\text{err}}$ as the probability of success in a trial. Under these assumptions (i) – (iv), as showed in Theorem 6 in Appendix A.4, the probability of obtaining at least T successful reads at any position j for any individual n is

$$\phi(\lambda, T, p_{\text{err}}) := \sum_{t \geq T} \frac{e^{-\lambda}\lambda^t}{t!} \sum_{i=T}^{t} \binom{t}{i} (1 - p_{\text{err}})^i p_{\text{err}}^{t-i} = \sum_{t \geq T} \frac{e^{-\lambda(1-p_{\text{err}})}\{\lambda(1 - p_{\text{err}})\}^t}{t!}. \tag{4}$$

We still assume $\Theta \sim 3\text{BP}(\alpha, \sigma, c)$ for the prior distribution over variant proportions. As in Section 3.1, we draw whether organism n has variant with proportion θ_j according to $\text{Bernoulli}(\theta_j)$. If the organism does have the variant, we now draw whether we observe the variant according to $\text{Bernoulli}(\phi)$, with $\phi = \phi(\lambda, T, p_{\text{err}})$. Hence, the probability of declaring the presence of variant j is now given by $\text{pr}(C_{n,j} \geq T \mid \Theta) = \theta_j \phi$.

Observe that ϕ is modulated by the parameter λ, which controls the sequencing depth and can be set by the practitioner. Ionita-Laza and Laird [2010] considered a setting with a single study, where that study is yet to be run. In this section, unlike the work of Ionita-Laza and Laird [2010], we assume that we have access to data from a pilot study when designing a follow-up study. We use subscripts to denote potentially different values of ϕ across experiments. For instance, the practitioner may choose a sequencing depth in the follow-up study that is different from the sequencing depth in the pilot study. Hence we write $\phi_{\text{pilot}} = \phi(\lambda_{\text{pilot}}, T, p_{\text{err}})$ for the pilot experiment and $\phi_{\text{follow}} = \phi(\lambda_{\text{follow}}, T, p_{\text{err}})$ for the follow-up. Our methods can be immediately extended to the case where there are multiple initial experiments with different ϕ values.

Proposition 3. *Let $\Theta \sim 3\text{BP}(\alpha, \sigma, c)$, that is $\Theta := \sum_{j \geq 1} \theta_j \delta_{\psi_j}$. Furthermore, let $X_n \mid \Theta \overset{iid}{\sim} \text{BeP}(\Theta_{pilot})$, where $\Theta_{pilot} := \sum_{j \geq 1} \phi_{\text{pilot}} \theta_j \delta_{\psi_j}$, for $n = 1, \ldots, N$ and $N \geq 1$, and let $X_{N+m} \mid \Theta \overset{iid}{\sim} \text{BeP}(\Theta_{follow})$, where $\Theta_{follow} := \sum_{j \geq 1} \phi_{\text{follow}} \theta_j \delta_{\psi_j}$, for $m = 1, \ldots, M$ and $M \geq 1$. Then,*

$$U_N^{(M)} \mid X_{1:N} \sim \text{Poisson}(\gamma), \tag{5}$$

with $\gamma := \alpha \phi_{\text{follow}} \sum_{m=1}^{M} E\{(1 - \phi_{\text{follow}} B)^{m-1} (1 - \phi_{\text{pilot}} B)^N\}$ and $B \sim \text{Beta}(1 - \sigma, c + \sigma)$.

The expected value of the posterior distribution in Theorem 3 provides a Bayesian nonparametric estimator, with respect to a squared loss function, of $U_N^{(M)}$. Namely,

this estimator is

$$\alpha\phi_{\text{follow}} \sum_{m=1}^{M} E\{(1-\phi_{\text{follow}}B)^{m-1}(1-\phi_{\text{pilot}}B)^{N}\},$$

where $B \sim \text{Beta}(1-\sigma, c+\sigma)$. This new estimator extends Section 3.1 to the case where sequencing error is taken into account. We defer the proof of Theorem 3 to Appendix A.

4 Empirics for the prediction

Our more realistic model of variant observation sets up a prediction framework for the number of new variants in a follow-up experiment. But without further development, we still face the difficulty that our predictor from Equation (1) does not use any information about the pilot experimental data except its cardinality. Recall that the hyperparameters α, σ, c control the behavior of the estimator (Theorem 2). So we will induce a dependency on the observed pilot data by fitting these hyperparameter values to the pilot data. One common approach in empirical Bayes is to maximize the probability of the data given the hyperparameters: $\arg\max_{\alpha,\sigma,c} \text{pr}(X_{1:N}|\alpha,\sigma,c)$ with $\text{pr}(X_{1:N}|\alpha,\sigma,c) = \int_{\Theta} \text{pr}(X_{1:N}|\Theta)\text{pr}(d\Theta|\alpha,\sigma,c)$. In the case without sequencing errors, this probability can be expressed in closed form as the exchangeable feature probability function (EFPF) [Broderick et al., 2013]. However, with sequencing errors, the integral can be very high-dimensional and expensive to compute with Markov chain Monte Carlo. Moreover, even without sequencing errors, the exchangeable feature probability function for the beta process is a complex function of sums, products, quotients, and exponentiation of gamma functions [Broderick et al., 2013, Eq. 8], which we find can lead to numerical instability in the optimization.

An easier choice is to treat the prediction from our model as a regression function with its own parameters α, σ, c. We can fit these parameters to the pilot project data by imagining subsets of the true pilot data as mini-pilot projects themselves and directly minimizing error in prediction on the remaining pilot data. In particular, consider index $n \in [N]$ as the size of the imagined mini-pilot. Then, by our earlier definition, $P_n^{(m)}$ is the prediction for the number of new variants in the next m data points given the first n data points. Here we write $P_n^{(m)}(\alpha,\sigma,c)$ to emphasize the hyperparameter dependence. Let $U_n^{(m)} \mid X_{1:N}$ be the true number of new variants in the next m data points (for m such that $n+m \leq N$) given the first n data points. Then we solve

$$\hat{\alpha},\hat{\sigma},\hat{c} := \underset{\substack{\alpha,\sigma,c:\\ \alpha>0,\ \sigma\in[0,1),c>-\sigma}}{\arg\min} \sum_{m=1}^{N-n} \left\{ P_n^{(m)}(\alpha,\sigma,c) - \left(U_n^{(m)} \mid X_{1:N}\right)\right\}^2. \qquad (6)$$

We set $n = \lfloor 2/3 \times N \rfloor$, a choice that works well across all applications we consider here. To find $\hat{\alpha},\hat{\sigma},\hat{c}$ we use the differential evolution algorithm [Storn and Price, 1997]. We also considered using multiple folds of the pilot study, in the style of cross validation, instead of a single train-test split. In our experiments, we did not observe

a noticeable difference between our proposal in Equation (6) and this more-involved procedure. We choose to minimize the 2-norm, but Equation (6) can be straightforwardly adapted for other standard choices of error (e.g., 1-norm). Finally, we use $\hat{P}_N^{(M)} := P_N^{(M)}(\hat{\alpha}, \hat{\sigma}, \hat{c})$ as our estimator for the number of new variants in the follow-up study of size M after observing data from the pilot study of size N.

5 Sequencing errors and optimal experimental design

Our goal is to maximize the number of variants we expect to observe under a fixed budget. To see how the budget comes into play, note two cost sources in the follow-up study. (i) It costs more to increase the number of samples M since sequencing each additional sample adds an additional cost. (ii) Likewise, it costs more to increase the quality of each sample, where increasing quality is accomplished by increasing the sequencing quality in the followup, λ_{follow}. We might encode the total cost as a function of these settings: $f(M, \lambda_{\text{follow}})$. Here, f is increasing in both of its arguments. Conversely, we expect to discover more variants as either of M or λ_{follow} increases and fewer variants as either quantity decreases. Therefore, we face a trade-off in where to best allocate experimental budget between M and λ_{follow}.

Our framework allows us to precisely quantify and optimize this trade-off. In particular, we now emphasize the dependence of $\hat{P}_N^{(M)}$ on λ_{follow}, via ϕ_{follow}, by writing $\hat{P}_N^{(M,\lambda_{\text{follow}})}$ for $\hat{P}_N^{(M)}$ computed with λ_{follow}. Since we can compute $\hat{P}_N^{(M,\lambda_{\text{follow}})}$ across values of M and λ_{follow} using Equation (5), we can optimize to find the maximum possible predicted variants under some budget C. We are interested in the experimental settings under which this maximum is achieved:

$$\underset{M,\lambda_{\text{follow}}}{\arg\max}\, \hat{P}_N^{(M,\lambda_{\text{follow}})} \quad \text{subject to} \quad f(M, \lambda_{\text{follow}}) \leq C. \tag{7}$$

To the best of our knowledge, no previous methods [Ionita-Laza et al., 2009, Ionita-Laza and Laird, 2010, Gravel, 2014, Zou et al., 2016, Orlitsky et al., 2016, Chakraborty et al., 2019] have been designed or modified to predict variants under different experimental conditions in a follow-up study given results from a pilot. We believe the Bayesian nonparametric framework we adopt here allows particularly straightforward handling of different sequencing depths, and more generally different experimental setups. Notably, Ionita-Laza and Laird [2010] consider experimental design, but only for a single future study, without observing any pilot data. Given its Bayesian grounding, their associated estimator might be adapted to our pilot and follow-up framework using similar techniques to those we introduce above. But we will see in Section 6 that the quality of their estimator is much worse than that of our method; any corresponding experimental design would therefore suffer. We suspect our gains are due to the flexibility of the Bayesian nonparametric framework and ability to capture power laws in the data.

Note that practitioners might instead be interested in maximizing the number of new rare variants in the follow-up study, i.e. variants that appear at most R times in the follow-up sample. In this case, we can still apply empirical Bayes estimates of

hyperparameters $\hat{\alpha}, \hat{\sigma}, \hat{c}$ obtained via Equation (6). In particular, let $\hat{P}_N^{(M,\leq R,\lambda_{\text{follow}})}$ be the Bayesian nonparametric estimator of the number of new rare variants, with hyperparameter values set to $\hat{\alpha}, \hat{\sigma}, \hat{c}$ and follow-up sequencing quality set to λ_{follow}. Then, to maximize the number of new rare variants, we solve the optimization problem in Equation (7) with $\hat{P}_N^{(M,\leq R,\lambda_{\text{follow}})}$ in place of $\hat{P}_N^{(M,\lambda_{\text{follow}})}$. We highlight that we still suggest learning the hyperparameters $\hat{\alpha}, \hat{\sigma}, \hat{c}$ via the original optimization problem, with the predictor $\hat{P}_N^{(M,\lambda_{\text{follow}})}$ of all new variants. We make this recommendation since rare variants may be sparser in the pilot study and thereby provide less information about these hyperparameters.

6 Experiments

6.1 Experimental setup

We evaluate our methods on both synthetic and real data. Code is available at `https://bitbucket.org/masoero/moreforless_bayesiandiscovery/src/master`. For real data, we use human cancer genomics datasets. In cancer genomics, rare variants may be useful in developing effective clinical procedures and understanding cancer biology, and researchers have recognized the importance of appropriate sequencing depth in the data-gathering process [Griffith et al., 2015, Rashkin et al., 2017]. Following the setup of Chakraborty et al. [2019], we consider the Cancer Genome Atlas (TCGA), a large and publicly available cancer genomics dataset. It contains somatic mutations from $N = 10{,}295$ patients and spans 33 different cancer types. See Appendix F for more details on the data. In what follows, we show that our method produces accurate predictions when the sequencing depth is kept constant (Section 6.2); we show it is the only method that can produce accurate predictions under changing conditions (Section 6.3) and the only method that can inform optimal design of experiments (Section 6.4). In Appendix F we report additional cancer genomics results, including with the MSK-impact database, a targeted sequencing study also used by Chakraborty et al. [2019].

In Appendix G, we report results for the Genome Aggregation Database [Karczewski et al., 2020], a recent extension of the Exome Aggregation Consortium data set [Lek et al., 2016] and the largest publicly available human genomic dataset. We include additional experiments on synthetic data in Appendix H, to illustrate when and why different methods may fail.

6.2 Prediction with no sequencing errors

Researchers have developed several approaches for predicting the number of new variants in a follow-up study under the assumption of perfect recovery of variants: e.g., parametric Bayesian methods [Ionita-Laza et al., 2009], linear programming methods [Gravel, 2014, Zou et al., 2016], a harmonic jackknife [Gravel, 2014], and a smoothed version of the classic Good-Toulmin estimator [Chakraborty et al., 2019]. To assess the prediction error under constant sequencing conditions, we focus on the TCGA dataset.

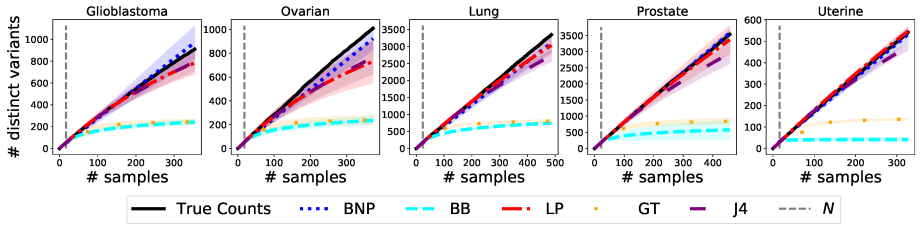

Figure 1: TCGA dataset: predicting the number of new variants. The true number of distinct variants (black) is compared to: our method (blue, BNP); Bayesian parametric (cyan, BB); linear program (red, LP); Good-Toulmin (orange, GT); 4th order jackknife (purple, J4). Shaded regions represent one standard deviation.

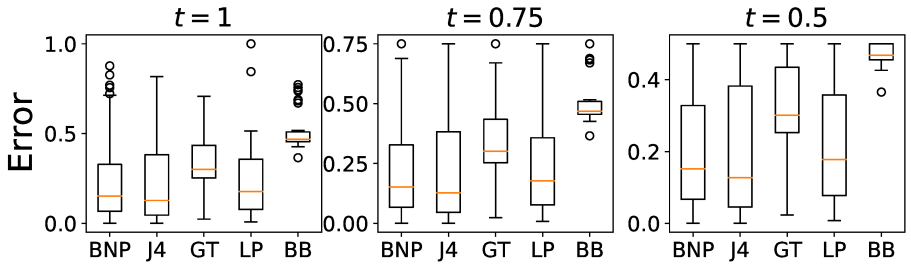

Figure 2: Trimmed percent prediction error on the TCGA data across all 33 cancer types and 20 folds, for different trimming thresholds t (Equation (F.1)). We compare our method (BNP) to Bayesian parametric (BB), linear program (LP), Good-Toulmin (GT), 4-th order jackknife (J4).

We partition the samples in the dataset into 33 datasets according to cancer-type annotation of each patient. For each cancer type, we predict the number of new variants that will be observed in a follow-up sample given a pilot sample. To do so, we use an approach akin to cross validation; namely we treat each cancer type as a dataset. We divide the dataset into 20 folds of equal size; we consider each fold in turn as a pilot study and treat the remaining folds as the follow-up. A smaller number of folds corresponds to a larger pilot study. All methods improve when the pilot study is increased substantially in size, i.e. when there is more information in the pilot. We find that the choice of 20 folds creates a challenging scenario with a small amount of pilot information. Nonetheless, our method, the harmonic jackknife, and the linear program all still perform well in these conditions.

We follow Zou et al. [2016] to visualize our results for five cancer types in Figure 1; namely, we plot the number of total predicted variants, averaged across folds, as a function of total data points (pilot plus follow-up). A vertical dashed line marks the pilot size; non-trivial predictions are to the right of this line. Shaded regions indicate one empirical standard deviation, measured across the folds. Figure 1 demonstrates that our predictor matches the true number of variants much more closely than the parametric

Bayesian method and smoothed Good-Toulmin estimator. In this case without sequencing error, our method has roughly the same performance as the harmonic jackknife and linear programming.

To more directly compare performance of the methods across all 33 cancer types, we calculate the error of each method across all types and all folds within each type; see Figure 2. More precisely, for each of the 33 cancer types and for each of the 20 folds we compute the trimmed absolute percentage prediction error incurred by the five methods at the largest possible extrapolation value. See Equation (F.1) in Appendix F. In Figure 2, we summarize these $33 * 20 = 660$ error values for each method in a boxplot. Lower errors are better. We find that our Bayesian nonparametric methods performs similarly to the linear programming method and to the harmonic jackknife. Our method outperforms the smoothed Good-Toulmin estimator and the parametric Bayesian approach. In Appendix F, we also follow Chakraborty et al. [2019] and run an experiment with an entirely separate pilot and follow-up study. In terms of comparison among estimators, these additional experiments lead to similar conclusions.

We performed additional experiments to better understand how our method compares to existing methods. In Appendix H.1 we run both the Bayesian parametric approach and our method on data simulated (a) under the parametric Bayesian model used by Ionita-Laza et al. [2009] and (b) under our own 3-parameter beta process model. We find that the approach of Ionita-Laza et al. [2009] works well with data simulated from their model but poorly with the three-parameter beta process data. Our results suggest that the parametric Bayesian method [Ionita-Laza et al., 2009] struggles with data exhibiting power laws, which we expect in real life.

While the method of Zou et al. [2016] performs well in our experiments above, we found serious numerical issues in other cases. In particular, Zou et al. [2016] exploits a linear programming approach to estimate rare variant proportions; the authors approximate proportions of common variants with the corresponding empirical frequencies. The authors define a variant as "rare" if it has frequency less than $\kappa/100$, for a user-defined threshold $\kappa \in (0, 100)$, interpreted as a percent. In practice, we found that the output of the algorithm is very sensitive to the choice of κ; see Appendix H.3. The authors suggest $\kappa = 1$ as a default setting, but we observed numerical instability and poor predictive performance for this value. This observation holds especially when the pilot size N is small, which we believe to be a particular case of interest in designing experiments for further data collection (i.e., for the follow-up study). For instance, we expect the small-N case to arise in the study of non-model organisms [Russell et al., 2017]. In Figure 1, we chose $\kappa = 20$, which led to convergence of the optimization algorithm in all cases. We explore other values of κ in Appendix H.3. Beyond these issues, we sometimes found that the method of Zou et al. [2016] failed to converge. While the convergence issue did not arise for our experiments in this section, it did arise for another analysis of the TCGA data; see Appendix F.2.

The Good-Toulmin method used in Chakraborty et al. [2019] performs poorly in our experiments above (Figure 1 and Figure 2), as well as in our further real-data experiments in Appendix G. However, we find that this method seems competitive with the best alternative on other cancer genomics data; see Appendix F.2. Further understanding of the variable performance of this estimator would be an important first step before any potential future use. By contrast, we find that jackknife Gravel

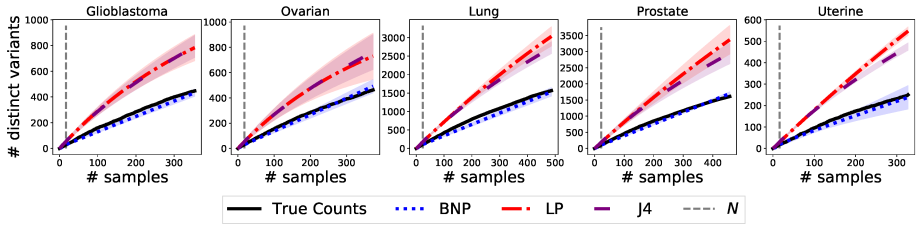

Figure 3: TCGA dataset: predicting the number of new variants. The true number of distinct variants (black) is compared with: our method (blue, BNP); linear program (red, LP); 4th order jackknife (purple, J4). Shaded bands represent one standard deviation.

[2014], with an appropriate hyperparameter calibration, performs well across all of our experiments when conditions are kept constant between the pilot and follow-up.

6.3 Prediction under different experimental conditions

We now turn to the case where there may be sequencing errors in the pilot study, in the follow-up study, or both. Moreover, the sequencing quality may differ between the pilot study and the follow-up study. To the best of our knowledge, no existing methods work in this case. We believe that the parametric Bayesian method of Ionita-Laza et al. [2009], the smoothed Good-Toulmin estimator of Chakraborty et al. [2019], and the linear programming method of Zou et al. [2016] could all be adapted to take sequencing errors into account. However, we have seen that the parametric Bayesian and Good-Toulmin methods already struggle when there are no sequencing errors. And the linear programming method suffers from numerical instability when the training sample size is small (the case of most interest). While the harmonic jackknife of [Gravel, 2014] performs well when there are no sequencing errors, we do not think it will be straightforward to adapt it to the case where sequencing quality may change between the pilot and follow-up.

In Figure 3 we see that there is indeed a noticeable difference in the number of observed variants when the experimental conditions change between the pilot and follow-up. In particular, we consider a pilot sequencing quality $\lambda_{\text{pilot}} = 100$ and a follow-up sequencing quality $\lambda_{\text{follow}} = 50$. We use a fixed threshold $T = 45$, a realistic coverage value in human genomic experiments [Karczewski et al., 2020], and the same five cancer types as in Figure 1. To represent this change between studies, we use the TCGA data as in Section 6.2 but apply additional thinning to simulate imperfect observation due to sequencing depth; see Appendix F.1 for additional details. Since the harmonic jackknife is not able to use information about the changing sequencing depth, we expect our Bayesian nonparametric method to deliver superior predictive performance when sequencing quality changes. This behavior is exactly what we see in Figure 3.

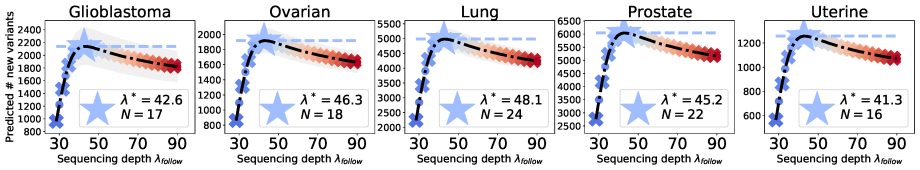

Figure 4: TCGA dataset: designing an experiment to maximize the number of new variants in a follow-up study.

6.4 Designing experiments to maximize the number of observed variants

We show that our method can be used for experimental design in practice. Our procedure consists of three steps. (i) Given the pilot data and sequencing quality λ_{pilot}, we minimize Equation (6) to estimate the parameters c, σ, α. (ii) We consider a range of values of the follow-up sequencing quality λ_{follow}; for each λ_{follow}, we choose the maximum follow-up size M that stays within our budget C, and we use the learned values of c, σ, α to predict the number of new variants in each case. (iii) We choose the settings of λ_{follow} and M that maximize the number of new variants. We illustrate this procedure in Figure 4. In our experiments, we set the cost function $f(M, \lambda_{\text{follow}}) = M \log(\lambda_{\text{follow}})$ as in Ionita-Laza and Laird [2010]. For every cancer type, we retain 5% of the observations as a pilot study. We set a budget C such that we can sample at full depth, i.e. coverage of 100x, only half of the total remaining 95% of the observations. We set variant calling rule threshold $T = 45$, error $p_{\text{err}} = 0.01$, and $\lambda_{\text{pilot}} = 100$. We run the procedure over all folds. We plot the predicted number of observed variants in the follow-up by maximizing M under the budget C and quality λ_{follow}, $\hat{P}_N^{(M,\lambda_{\text{follow}})}$; the shaded region in Figure 4 illustrates one standard deviation. We see a trade-off in quality and quantity. Namely, maximizing quantity M leads to very small values of λ_{follow} to maintain the budget C. With sufficiently low quality, though, fewer variants are discovered. Conversely, when λ_{follow} is set very high, we require a very small M to maintain the budget C, and not many variants are discovered. Intermediate values of λ_{follow} and M serve to maximize the number of variants discovered under a fixed budget.

7 Discussion

We have presented a Bayesian nonparametric method for predicting the number of variants in a follow-up study using information from a pilot study. Our method works even when the follow-up study has different experimental conditions from the pilot study, and can be used for optimal design of the follow-up study.

Though our experiments here focus on rare variants from bulk studies in human genetics, we briefly describe further potential applications to emphasize the generality of our framework. First, in microbiome research, there is an increasing interest in (i) devising low-cost pipelines for efficient sequencing [Rajan et al., 2019, Sanders et al.,

2019], as well as (ii) defining best-practice protocols for data collection processes [Hillmann et al., 2018, Bharti and Grimm, 2021]. Indeed, scientists have already expressed an interest in optimal allocation of a budget given information from a pilot experiment [Zaheer et al., 2018, Pereira-Marques et al., 2019]. Second, in single-cell sequencing, scientists are interested in reliably estimating important gene properties. In this case, there exists a vast and growing literature that highlights the importance of establishing the optimal trade-off between the quality (sequencing depth) of the experiment, and the number of cells to be sequenced. See, for example, Bacher and Kendziorski [2016], Li and Li [2018], Zhang et al. [2020]. Third, it is becoming common practice to use modern, non-invasive approaches for surveying wildlife populations, such as camera-traps [Tarugara et al., 2019, Welbourne et al., 2020]. Accurate estimation of the living population and timely adoption of preventive measures are crucial for the survival of endangered species [Johansson et al., 2020]. But conservation groups often face a limited budget. These groups might benefit from trading off equipment density and quality.

While the present paper has focused on data that can be represented as collections of binary features (e.g. variants and non-variants), our method may be extended to the case in which the observations are vectors of counts, as well as the case in which there exist multiple categories for each feature (e.g. different types of variants). In particular, by means of the Bayesian nonparametric conjugacy framework of James [2017], Broderick et al. [2018], we may extend our method to use a categorical (or multinomial) likelihood process with a conjugate Bayesian nonparametric prior for the now-multiple frequencies per variant location. Our Bayesian nonparametric method may also be easily extended to accommodate multiple different pilot studies. For the latter extension, we would still generate variant proportions according to the three-parameter beta process; we would then generate variants in each pilot study according to different damped Bernoulli processes. The ultimate effect would be to introduce more distinct, but workable, Bernoulli terms in Equation (5). Moreover, in this work we have focused on threshold variant calling rules, which are a simplification of state-of-the art variant callers [Xu, 2018]. Extending our framework to encompass more realistic variant calling rules is an interesting future research direction. An important practical challenge in this case will be even specifying a formula or series of formulas to describe how popular variant callers work.

Acknowledgments

The authors are grateful to the Editor, the Associate Editor, and two anonymous Referees for their comments, corrections, and suggestions, which have greatly improved the paper. The results shown in the present paper are in whole or part based upon data generated by the TCGA Research Network. The authors thank Boyu Ren, Joshua Schraiber, Michael Hoffman, and Brian Trippe for useful discussions and comments. The authors are also grateful to Boyu Ren for help working with the gnomAD dataset. Federico Camerlenghi and Stefano Favaro received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257, and the Italian Ministry of Education,

University and Research, "Dipartimenti di Eccellenza" grant 2018–2022. Lorenzo Masoero and Tamara Broderick were supported in part by the DARPA I2O LwLL program, the CSAIL-MSR Trustworthy AI Initiative, an NSF CAREER Award, a Sloan Research Fellowship, and ONR.

References

N. C. Aguirre, C. V. Filippi, G. Zaina, J. G. Rivas, C. V. Acuña, P. V. Villalba, M. N. García, S. González, M. Rivarola, M. C. Martínez, et al. Optimizing ddRAD-seq in non-model species: A case study in *Eucalyptus dunnii* Maiden. *Agronomy*, 9(9): 484, 2019.

R. Bacher and C. Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1):1–14, 2016.

P. Berti, I. Crimaldi, L. Pratelli, and P. Rigo. Central limit theorems for an Indian buffet model with random weights. *The Annals of Applied Probability*, 25(2):523–547, 2015.

R. Bharti and D. G. Grimm. Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, 22(1):178–193, 2021.

L. Bomba, K. Walter, and N. Soranzo. The impact of rare and low-frequency genetic variants in common disease. *Genome biology*, 18(1):77, 2017.

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476, 2012.

T. Broderick, J. Pitman, and M. I. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

T. Broderick, A. C. Wilson, and M. I. Jordan. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 24(4B):3181–3221, 2018.

K. P. Burnham and W. S. Overton. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65(3):625–633, 1978.

F. Camerlenghi, B. Dumitrascu, F. Ferrari, B. E. Engelhardt, S. Favaro, et al. Nonparametric Bayesian multiarmed bandits for single-cell experiment design. *Annals of Applied Statistics*, 14(4):2003–2019, 2020.

S. Chakraborty, A. Arora, C. B. Begg, and R. Shen. Using somatic variant richness to mine signals from rare variants in the cancer genome. *Nature Communications*, 10: 5506, 2019.

D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets. *The Journal of Molecular Diagnostics*, 17(3):251–264, 2015.

E. T. Cirulli and D. B. Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415, 2010.

A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

. G. P. Consortium. A global reference for human genetic variation. *Nature*, 526(7571): 68, 2015.

A. Cornish and C. Guda. A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International*, 2015.

R. R. da Fonseca, A. Albrechtsen, G. E. Themudo, J. Ramos-Madrigal, J. A. Sibbesen, L. Maretty, M. L. Zepeda-Mendoza, P. F. Campos, R. Heller, and R. J. Pereira. Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, 30:3–13, 2016.

B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Accademia Nazionale dei Lincei, Ser. 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturali 4*, pages 251–299, 1931.

B. Dumitrascu, K. Feng, and B. E. Engelhardt. Gt-ts: Experimental design for maximizing cell type discovery in single-cell data. *bioRxiv:10.1101/386540*, 2018. doi: 10.1101/386540. URL `https://www.biorxiv.org/content/early/2018/08/07/386540`.

B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

I. Good and G. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

S. Gravel. Predicting discovery rates of genomic features. *Genetics*, 197(2):601–610, 2014.

S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante, and D. L. Altshuler. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

M. Griffith, C. A. Miller, O. L. Griffith, K. Krysiak, Z. L. Skidmore, A. Ramu, J. R. Walker, H. X. Dang, L. Trani, D. E. Larson, et al. Optimizing cancer genome sequencing and analysis. *Cell Systems*, 1(3):210–223, 2015.

B. Hillmann, G. A. Al-Ghalith, R. R. Shields-Cutler, Q. Zhu, D. M. Gohl, K. B. Beckman, R. Knight, and D. Knights. Evaluating the information content of shallow shotgun metagenomics. *MSystems*, 3(6), 2018.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294, 1990.

S. Hwang, E. Kim, I. Lee, and E. M. Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5 (1):1–8, 2015.

I. Ionita-Laza and N. M. Laird. On the optimal design of genetic variant discovery studies. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

I. Ionita-Laza, C. Lange, and N. M. Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13): 5008–5013, 2009.

L. F. James. Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. *The Annals of Statistics*, 45(5):2016–2045, 2017.

Ö. Johansson, G. Samelius, E. Wikberg, G. Chapron, C. Mishra, and M. Low. Identification errors in camera-trap studies result in systematic population overestimation. *Scientific Reports*, 10(1):1–10, 2020.

K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

Y. Kim. Nonparametric Bayesian estimators for counting processes. *Annals of Statistics*, pages 562–588, 1999.

M. Kumaran, U. Subramanian, and B. Devarajan. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics*, 20(1):1–11, 2019.

E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, 1988.

M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, and B. B. Cummings. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

W. V. Li and J. J. Li. An accurate and robust imputation method scimpute for single-cell RNA-seq data. *Nature Communications*, 9(1):1–9, 2018.

A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

I. Mathieson and D. Reich. Differences in the rare variant spectrum among human populations. *PLoS Genetics*, 13(2):e1006581, 2017.

Y. Momozawa and K. Mizukami. Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics*, pages 1–13, 2020.

A. Orlitsky, A. T. Suresh, and Y. Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

J. Pereira-Marques, A. Hout, R. M. Ferreira, M. Weber, I. Pinto-Ribeiro, L.-J. van Doorn, C. W. Knetsch, and C. Figueiredo. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in Microbiology*, 10:1277, 2019.

B. K. Peterson, J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, 7(5), 2012.

S. K. Rajan, M. Lindqvist, R. J. Brummer, I. Schoultz, and D. Repsilber. Phylogenetic microbiota profiling in fecal samples depends on combination of sequencing depth and choice of NGS analysis method. *PLoS One*, 14(9):e0222171, 2019.

S. Rashkin, G. Jun, S. Chen, G. R. Abecasis, Genetics, E. of Colorectal Cancer Consortium, et al. Optimal sequencing strategies for identifying disease-associated singletons. *PLoS Genetics*, 13(6), 2017.

J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597, 2015.

J. J. Russell, J. A. Theriot, P. Sood, W. F. Marshall, L. F. Landweber, L. Fritz-Laylin, J. K. Polka, S. Oliferenko, T. Gerbich, A. Gladfelter, et al. Non-model model organisms. *BMC Biology*, 15(1):55, 2017.

A. Saint Pierre and E. Génin. How important are rare variants in common disease? *Briefings in Functional Genomics*, 13(5):353–361, 2014.

J. Sampson, K. Jacobs, M. Yeager, S. Chanock, and N. Chatterjee. Efficient study design for next generation sequencing. *Genetic Epidemiology*, 35(4):269–277, 2011.

J. G. Sanders, S. Nurk, R. A. Salido, J. Minich, Z. Z. Xu, Q. Zhu, C. Martino, M. Fedarko, T. D. Arthur, F. Chen, et al. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology*, 20(1): 1–14, 2019.

A. N. Shiryaev. *Probability (2Nd Ed.)*. Springer-Verlag, Berlin, Heidelberg, 1995. ISBN 0-387-94549-0.

G. Sirugo, S. Williams, and S. Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

C. A. Souza, N. Murphy, C. Villacorta-Rath, L. N. Woodings, I. Ilyushkina, C. E. Hernandez, B. S. Green, J. J. Bell, and J. M. Strugnell. Efficiency of ddRAD target enriched sequencing across spiny rock lobster species (Palinuridae: *Jasus)*. *Scientific Reports*, 7(1):1–14, 2017.

R. Storn and K. Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

A. Tarugara, B. W. Clegg, E. Gandiwa, and V. K. Muposhi. Cost-benefit analysis of increasing sampling effort in a baited-camera trap survey of an African leopard (*Panthera pardus*) population. *Global Ecology and Conservation*, 18:e00627, 2019.

C. F. Taylor and G. R. Taylor. Current and emerging techniques for diagnostic mutation detection. In *Molecular Diagnosis of Genetic Diseases*, pages 9–44. Springer, 2004.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, pages 1838–1846, 2009.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571, 2007.

F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142, 1951.

D. J. Welbourne, A. W. Claridge, D. J. Paull, and F. Ford. Camera-traps are a cost-effective method for surveying terrestrial squamates: A comparison with artificial refuges and pitfall traps. *PLoS One*, 15(1):e0226913, 2020.

C. Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24, 2018.

R. Zaheer, N. Noyes, R. O. Polo, S. R. Cook, E. Marinier, G. Van Domselaar, K. E. Belk, P. S. Morley, and T. A. McAllister. Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, 8(1):1–11, 2018.

M. J. Zhang, V. Ntranos, and D. Tse. Determining sequencing depth in a single-cell RNA-seq experiment. *Nature Communications*, 11(1):1–11, 2020.

J. Zou, G. Valiant, P. Valiant, K. Karczewski, S. O. Chan, K. Samocha, M. Lek, S. Sunyaev, M. Daly, and D. G. MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7:13293, 2016.

O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.

Appendix

This document contains the supplementary material for "*More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics*". In Appendix A we present the proofs of the results presented in Section 2. We next provide detail about the competing methods we considered. In Appendix B the Bayesian parametric estimator of Ionita-Laza et al. [2009], in Appendix C the linear program proposed by Zou et al. [2016], in Appendix D the Jackknife estimator used in Gravel [2014], and in Appendix E the Good-Toulmin estimator used in Chakraborty et al. [2019]. We conclude providing additional experimental results. In Appendix F we present additional detail about the data used in Chakraborty et al. [2019], and considered in the analysis in the main text. In Appendix G we report results for the gnomAD project [Karczewski et al., 2020], an extension of the datasets previously considered in Gravel [2014], Zou et al. [2016]. We conclude with extensive experiments on simulated data in Appendix H.

A Additional results and proofs

Proof of Theorem 1

Proof. By construction, the variant frequencies $\{\theta_j\}$ are formed from a Poisson point process with rate measure

$$\nu(\mathrm{d}\theta) = \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta^{-1-\sigma}(1-\theta)^{c+\sigma-1} \mathbf{1}_{[0,1]}(\theta)\mathrm{d}\theta. \tag{A.1}$$

Recall that a variant with frequency θ_j appears in organism n with Bernoulli probability θ_j, independently across n. Therefore, the collection of variant frequencies whose corresponding variants have not yet appeared after N organisms comes from a thinned Poisson point process relative to the original Poisson point process generating the $\{\theta_j\}$; the thinned process has rate measure $\nu(\mathrm{d}\theta) \cdot \mathrm{Bernoulli}(0|\theta)^N$ and is independent of the collection of frequencies that did appear in the first N organisms. Similarly, the collection of variant frequencies corresponding to variants that did not appear in the first N organisms but then did appear in the first follow-up organism comes from a thinned Poisson point process with rate measure $\nu(\mathrm{d}\theta) \cdot \mathrm{Bernoulli}(0|\theta)^N \cdot \mathrm{Bernoulli}(1|\theta)$ and is independent of the collection of frequencies that did not appear in the first $N+1$ organisms. Recursively, for $m \geq 1$, the collection of variant frequencies corresponding to variants that did not appear in the first $N+m-1$ organisms but then did appear in the mth follow-up organism comes from a thinned Poisson point

process with rate measure

$$
\begin{aligned}
&\nu(\mathrm{d}\theta)\mathrm{Bernoulli}(0|\theta)^{N+m-1}\mathrm{Bernoulli}(1|\theta) \\
&= \alpha\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\theta^{-1-\sigma+1}(1-\theta)^{c+\sigma-1+N+m-1}\mathbf{1}_{[0,1]}(\theta)\mathrm{d}\theta \\
&= \alpha\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\cdot\frac{\Gamma(1-\sigma)\Gamma(c+\sigma-1+N+m)}{\Gamma(c+N+m)} \\
&\quad\cdot\mathrm{Beta}(\theta\mid 1-\sigma, c+\sigma-1+N+m)\mathrm{d}\theta \\
&= \alpha\frac{(c+\sigma)_{(N+m-1)\uparrow}}{(1+c)_{(N+m-1)\uparrow}}\mathrm{Beta}(\theta\mid 1-\sigma, c+\sigma-1+N+m)\mathrm{d}\theta.
\end{aligned}
$$

Finally, we observe that the number of points in a Poisson point process is Poisson distributed with mean equal to the integral of its rate measure. Each of these Poisson point processes is independent, and the sum of independent Poissons is Poisson with mean equal to the sum of the means. So, since $U_N^{(M)}$ is the sum of points in these M Poisson point processes with $m \in [M]$, we have $U_N^{(M)}$ is Poisson with mean

$$
\begin{aligned}
&\sum_{m=1}^{M}\int_0^1 \alpha\frac{(c+\sigma)_{(N+m-1)\uparrow}}{(1+c)_{(N+m-1)\uparrow}}\mathrm{Beta}(\theta|1-\sigma, c+\sigma-1+N+m)\mathrm{d}\theta \\
&\qquad = \sum_{m=1}^{M}\alpha\frac{(c+\sigma)_{(N+m-1)\uparrow}}{(1+c)_{(N+m-1)\uparrow}},
\end{aligned}
$$

as was to be shown.

□

Proof of Theorem 2

In the following we make use of the O notation, indeed we will write $f(x) = O(g(x))$ to mean that the ratio $|f(x)/g(x)|$ is a bounded function of the variable x; we also write $f(x) = o(g(x))$ as $x \to x_0$ (little o notation) to mean that $\lim_{x\to x_0} f(x)/g(x) = 0$. A preliminary result is needed.

Lemma 1. *For any $c > 0$, $N \geq 1$ and $\sigma \in (0,1)$ we have that*

$$
\frac{1}{M^\sigma}\sum_{m=1}^{M}\frac{\Gamma(c+N+m-1+\sigma)}{\Gamma(c+N+m)} = \frac{1}{\sigma} + O(M^{-\sigma}) \tag{A.2}
$$

is satisfied as M grows to $+\infty$.

Proof. As in the proof of Berti et al. [2015, Lemma 2], we know that for any $x > 0$,

$$
\frac{\Gamma(x+\sigma)}{\Gamma(x+1)} = x^{\sigma-1}(1+g(x)),
$$

where $g : (0,+\infty) \to R$ is such that $\sup_{x\geq 0}|g(x)x| < +\infty$. Putting $x = c+N+m-1$, where $m \geq 1$ while c and N are fixed constants, this very last condition on

g is equivalent to $\sup_{y\geq c+N-1}|g(y)y|<+\infty$. Hence there exists $K>0$ such that $|g(y)|\leq K/y$ for any $y\geq c+N-1$. With this in mind we focus on the left hand side of Equation (A.2) in the statement of Lemma 1:

$$\frac{1}{M^\sigma}\sum_{m=1}^{M}\frac{\Gamma(c+N+m-1+\sigma)}{\Gamma(c+N+m)} \tag{A.3}$$

$$=\frac{1}{M^\sigma}\sum_{m=1}^{M}(c+N+m-1)^{\sigma-1}(1+g(c+N+m-1))$$

$$=\frac{1}{M^\sigma}\sum_{m=1}^{M}(c+N+m-1)^{\sigma-1} \tag{A.4}$$

$$+\frac{1}{M^\sigma}\sum_{m=1}^{M}(c+N+m-1)^{\sigma-1}g(c+N+m-1). \tag{A.5}$$

As for the sum of Equation (A.4) note that the following inequalities hold true

$$\begin{aligned}\frac{(c+N+M)^\sigma-(c+N)^\sigma}{\sigma M^\sigma}&=\int_1^{M+1}\frac{(c+N+m-1)^{\sigma-1}}{M^\sigma}\mathrm{d}m\\&\leq\sum_{m=1}^{M}\frac{(c+N+m-1)^{\sigma-1}}{M^\sigma}\\&\leq\int_0^{M}\frac{(c+N+m-1)^{\sigma-1}}{M^\sigma}\mathrm{d}m\\&=\frac{(c+N+M-1)^\sigma-(c+N-1)^\sigma}{\sigma M^\sigma},\end{aligned} \tag{A.6}$$

where we have used the fact that $(c+N+m-1)^{\sigma-1}$ is decreasing in m, and used the corresponding integrals to bound the sum. We can use an asymptotic expansion of the upper and the lower bound in Equation (A.6) to get

$$\begin{aligned}&\frac{1}{\sigma}\left(\frac{\sigma(c+N)}{M}+o\Big(\frac{1}{M}\Big)-\frac{(c+N)^\sigma}{M^\sigma}\right)\leq\sum_{m=1}^{M}\frac{(c+N+m-1)^{\sigma-1}}{M^\sigma}-\frac{1}{\sigma}\\&\quad\leq\frac{1}{\sigma}\left(\frac{\sigma(c+N-1)}{M}+o\Big(\frac{1}{M}\Big)-\frac{(c+N-1)^\sigma}{M^\sigma}\right),\end{aligned}$$

which entails that

$$\sum_{m=1}^{M}\frac{(c+N+m-1)^{\sigma-1}}{M^\sigma}=\frac{1}{\sigma}+O\Big(\frac{1}{M^\sigma}\Big). \tag{A.7}$$

As for Equation (A.5), we exploit the properties of g to get

$$\begin{aligned}
&\Big|\frac{1}{M^\sigma}\sum_{m=1}^{M}(c+N+m-1)^{\sigma-1}g(c+N+m-1)\Big| \\
&\quad\leq \frac{K}{M^\sigma}\sum_{m=1}^{M}(c+N+m-1)^{\sigma-2} \\
&\quad\leq \frac{K}{M^\sigma}\int_{c+N-1}^{c+N+M-1}\frac{1}{x^{2-\sigma}}\mathrm{d}x \\
&\quad= \frac{K}{M^\sigma(1-\sigma)}\left\{\frac{1}{(c+N-1)^{1-\sigma}}-\frac{1}{(c+N+M-1)^{1-\sigma}}\right\}
\end{aligned}$$

The last inequality implies that

$$\Big|\frac{1}{M^\sigma}\sum_{m=1}^{M}(c+N+m-1)^{\sigma-1}g(c+N+m-1)\Big| = O\Big(\frac{1}{M^\sigma}\Big). \tag{A.8}$$

Putting Equation (A.6) and Equation (A.8) in Equation (A.4) and Equation (A.5) the thesis follows. □

If X is a real valued random element, we denote by $\Phi_X(t) = E(e^{itX})$ its characteristic function, where i is the imaginary unit. We also assume that all the random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathrm{pr})$, and we denote by pr_N the probability pr given $X_{1:N}$; E_N and var_N will stand for the expected valued and the variance given $X_{1:N}$, respectively.

of Theorem 2. We start by showing the strong law of large numbers of Equation (2) in the main text. From Lemma 1 we deduce that

$$\frac{E_N(U_N^{(M)})}{M^\sigma} = \frac{\alpha}{M^\sigma}\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}\sum_{m=1}^{M}\frac{\Gamma(c+\sigma+N+m-1)}{\Gamma(c+N+m)} \to \frac{\alpha\Gamma(c+1)}{\sigma\Gamma(c+\sigma)} \tag{A.9}$$

as $M \to +\infty$. We observe that $U_N^{(M)} = H_N^{(1)} + \cdots + H_N^{(M)}$, where $H_N^{(m)}$ are independent Poisson random variables with mean

$$\frac{\alpha(c+\sigma)_{(N+m-1)\uparrow}}{(c+1)_{(N+m-1)\uparrow}},$$

for $m = 1, \ldots, M$ and M is arbitrary large. $H_N^{(m)}$ is the number of new variants that have been observed in the $(N+m)$-th individual, conditionally on the first N individuals. As a consequence we may write

$$\frac{U_N^{(M)} - E_N\left(U_N^{(M)}\right)}{M^\sigma} = \frac{H_N^{(1)} - E_N(H_N^{(1)}) + \cdots + H_N^{(M)} - E_N(H_N^{(M)})}{M^\sigma}.$$

The Kronecker's lemma [Shiryaev, 1995, Lemma IV.3.2] implies that

$$\lim_{M\to+\infty} \frac{U_N^{(M)} - E_N\left(U_N^{(M)}\right)}{M^\sigma} = 0 \quad \mathrm{pr}_N - \text{almost surely},$$

provided that the following condition is satisfied

$$\sum_{m=1}^{+\infty} \frac{\mathrm{var}_N(H_N^{(m)})}{m^{2\sigma}} < +\infty. \tag{A.10}$$

This may be easily verified as follows:

$$\begin{aligned}\sum_{m=1}^{+\infty} \frac{\mathrm{var}_N(H_N^{(m)})}{m^{2\sigma}} &= \sum_{m=1}^{+\infty} \frac{\alpha}{m^{2\sigma}} \frac{(c+\sigma)_{N+m-1\uparrow}}{(c+1)_{N+m-1\uparrow}} \\ &= \alpha \frac{\Gamma(c+1)}{\Gamma(c+\sigma)} \sum_{m=1}^{+\infty} \left\{ \frac{1}{m^{2\sigma}} \frac{\Gamma(c+\sigma+N+m-1)}{\Gamma(c+N+m)} \right\} < +\infty.\end{aligned}$$

The series turns out to be convergent because the following asymptotic relation holds true:

$$\frac{\Gamma(c+\sigma+N+m-1)}{\Gamma(c+N+m)} \sim \frac{\alpha}{m^{1+\sigma}} \frac{\Gamma(c+1)}{\Gamma(c+\sigma)}.$$

Hence Equation (A.10) is satisfied, so we conclude that

$$\lim_{M\to+\infty} \frac{U_N^{(M)} - E_N\left(U_N^{(M)}\right)}{M^\sigma} = 0 \quad \text{almost surely},$$

which is equivalent to the thesis thanks to Equation (A.9).

We now prove the central limit theorem stated in Equation (3) in the main text. We prove the result using the convergence of characteristic functions. We use the fact that the posterior distribution of $U_N^{(M)}$ is Poisson to evaluate the characteristic function *a posteriori*: for convenience, let $\tilde{U}_N^{(M)} := \sqrt{M^\sigma}\left(\frac{U_N^{(M)}}{M^\sigma} - \xi\right)$, where we recall that ξ is defined as

$$\xi := \frac{\alpha}{\sigma} \frac{\Gamma(c+1)}{\Gamma(c+\sigma)}.$$

Then,

$$\begin{aligned}&\Phi_{\tilde{U}_N^{(M)}|X_{1:N}}(t) = E_N\left[\exp\left\{it\tilde{U}_N^{(M)}\right\}\right] \\ &= \exp\left\{-it\xi\sqrt{M^\sigma} + \alpha(e^{it/\sqrt{M^\sigma}} - 1) \sum_{m=1}^{M} \frac{(c+\sigma)_{N+m-1\uparrow}}{(c+1)_{N+m-1\uparrow}}\right\} \\ &= \exp\left\{-it\xi\sqrt{M^\sigma} + (e^{it/\sqrt{M^\sigma}} - 1) \frac{\alpha\Gamma(c+1)}{\Gamma(c+\sigma)} \sum_{m=1}^{M} \frac{\Gamma(c+N+m-1+\sigma)}{\Gamma(c+N+m)}\right\}.\end{aligned}$$

We now use Lemma 1 and the asymptotic expansion of the exponential function to get

$$
\begin{aligned}
&\Phi_{\tilde{U}_N^{(M)}\big|X_{1:N}}(t) = \\
&= \exp\left\{-it\xi\sqrt{M^\sigma} + \frac{\alpha\Gamma(c+1)}{\Gamma(c+\sigma)}\left(\frac{it}{\sqrt{M^\sigma}} - \frac{t^2}{2M^\sigma} + O(M^{-\frac{3}{2}\sigma})\right)\left(\frac{M^\sigma}{\sigma} + O(1)\right)\right\} \\
&= \exp\left\{-it\xi\sqrt{M^\sigma} + \frac{\alpha}{\sigma}\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}\left(it\sqrt{M^\sigma} - \frac{t^2}{2} + O\left(\sqrt{M^\sigma}\right)\right)\right\} \\
&= \exp\left\{-it\xi\sqrt{M^\sigma} + \xi(it\sqrt{M^\sigma} - \frac{t^2}{2} + O\left(\sqrt{M^\sigma}\right))\right\} \\
&= \exp\left\{-\frac{\xi t^2}{2} + O\left(\sqrt{M^\sigma}\right)\right\},
\end{aligned}
$$

where in the penultimate line we substituted

$$\xi = \frac{\alpha}{\sigma}\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}.$$

Therefore, as M grows to infinity, we get

$$\Phi_{\sqrt{M^\sigma}\left(\frac{U_N^{(M)}}{M^\sigma} - \xi\right)\big|X_{1:N}}(t) \longrightarrow \exp\left\{-\frac{\xi t^2}{2}\right\},$$

and the thesis follows. □

Proof of Theorem 3

Proof. To see the almost sure finiteness of the Poisson parameter and hence of the random variables $U_N^{(M)}$ and of $U_N^{(M,r)}$, and $U_N^{(M,\leq R)}$, note that the parameter constraints for the three-parameter beta process are specifically constructed so that $\theta\nu(\mathrm{d}\theta)$ is a proper beta distribution; see the end of section 2 and James [2017], Broderick et al. [2018]. The θ factor will arise from $\mathrm{Bernoulli}(1 \mid \phi_{\mathrm{follow}}\theta)$.

The exact form of the Poisson parameter in Equation (5) arises by following the same thinning argument as in the proof of Theorem 1. To see the beta representation,

$$
\begin{aligned}
&\mathrm{Bernoulli}(1 \mid \phi_{\mathrm{follow}}\theta)\mathrm{Bernoulli}(0 \mid \phi_{\mathrm{follow}})^{m-1}\mathrm{Bernoulli}(0 \mid \phi_{\mathrm{pilot}})^N\nu(\mathrm{d}\theta) \\
&= \alpha\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\theta^{-1-\sigma}((1-\theta)^{c+\sigma-1}\times \\
&\times (\phi_{\mathrm{follow}}\theta)(1-\phi_{\mathrm{follow}}\theta)^{m-1}(1-\phi_{\mathrm{pilot}}\theta)^N\mathbf{1}_{[0,1]}(\theta)\mathrm{d}\theta \\
&= \alpha\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\phi_{\mathrm{follow}}(1-\phi_{\mathrm{follow}}\theta)^{m-1}(1-\phi_{\mathrm{pilot}}\theta)^N \cdot \frac{\Gamma(1-\sigma)\Gamma(c+\sigma)}{\Gamma(c+1)} \\
&\quad \cdot \mathrm{Beta}(\theta \mid 1-\sigma, c+\sigma)\mathrm{d}\theta \\
&= \alpha\phi_{\mathrm{follow}}\mathrm{Beta}(\theta \mid 1-\sigma, c+\sigma)\mathrm{d}\theta.
\end{aligned}
$$

The exact form of the Poisson parameter γ_r in Equation (A.24) arises by following the same thinning argument as in the proof of Theorem 4. To see the beta representation,

$$\begin{aligned}
&\mathrm{Bernoulli}(1 \mid \phi_{\mathrm{follow}}\theta)^r \mathrm{Bernoulli}(0 \mid \phi_{\mathrm{follow}}\theta)^{M-r} \mathrm{Bernoulli}(0 \mid \phi_{\mathrm{pilot}}\theta)^N \nu(\mathrm{d}\theta) \\
&= \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta^{-1-\sigma}((1-\theta)^{c+\sigma-1} \times \\
&\times (\phi_{\mathrm{follow}}\theta)^r (1-\phi_{\mathrm{follow}}\theta)^{M-r} (1-\phi_{\mathrm{pilot}}\theta)^N \mathbf{1}_{[0,1]}(\theta) \mathrm{d}\theta \\
&= \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \phi_{\mathrm{follow}}^r (1-\phi_{\mathrm{follow}}\theta)^{M-r} (1-\phi_{\mathrm{pilot}}\theta)^N \cdot \frac{\Gamma(r-\sigma)\Gamma(c+\sigma)}{\Gamma(c+r)} \\
&\quad \cdot \mathrm{Beta}(\theta \mid r-\sigma, c+\sigma) \mathrm{d}\theta \\
&= \alpha \phi_{\mathrm{follow}}^r \frac{(1+c)_{(r-1)\uparrow}}{(1-\sigma)_{(r-1)\uparrow}} \cdot \mathrm{Beta}(\theta \mid r-\sigma, c+\sigma) \mathrm{d}\theta.
\end{aligned}$$

□

A.1 Number of new rare variants

Proposition 4 (Number of new rare variants)**.** *Assume the model in Theorem 1. Let $U_N^{(M,r)}$ represent the number of new variants that occur r times in a follow-up sample of size M after a preliminary study of size N. I.e., we count the variants that do not occur in the preliminary N samples but then occur r times in the follow-up M samples. Let $U_N^{(M,\leq R)}$ similarly represent the number of new variants that occur at most R times. Here $r, R \in [M]$. Define*

$$U_N^{(M,r)} := \sum_{j=1}^{\infty} \mathbf{1}\left(\sum_{n=1}^{N} x_{n,j} = 0\right) \mathbf{1}\left(\sum_{m=1}^{M} x_{N+m,j} = r\right).$$

Then

$$U_N^{(M,r)} \mid X_{1:N} \sim \mathrm{Poisson}(\lambda_r), \tag{A.11}$$

for $\lambda_r := \alpha \binom{M}{r} \frac{(1-\sigma)_{(r-1)\uparrow}(c+\sigma)_{(N+M-r)\uparrow}}{(c+1)_{(N+M-1)\uparrow}}$. Moreover, for

$$U_N^{(M,\leq R)} := \sum_{r=1}^{R} U_N^{(M,r)},$$

it holds

$$U_N^{(M,\leq R)} \mid X_{1:N} \sim \mathrm{Poisson}\left(\sum_{r=1}^{R} \lambda_r\right).$$

Just like for $U_N^{(M)}$, the Bayesian nonparametric predictors of $U_N^{(M,r)}$ and $U_N^{(M,\leq R)}$ correspond to the expected values of $U_N^{(M,r)} \mid X_{1:N}$ and $U_N^{(M,\leq R)} \mid X_{1:N}$, respectively, i.e. the parameters of the posterior predictive Poisson distributions displayed in (A.11). Similarly to Proposition (2), the large M asymptotic behaviour of $U_N^{(M,r)} \mid X_{1:N}$ and $U_N^{(M,\leq R)} \mid X_{1:N}$ display very specific power law behavior almost surely.

Proof. Analogous to the proof of Theorem 1, we consider the Poisson point process $\{\theta_j\}$ and thin it to those frequencies corresponding to variants chosen no times in the preliminary N samples and chosen exactly r times out of the follow-up M samples. The probability of being chosen to be thinned, then, is $\mathrm{Bernoulli}(0 \mid \theta)^N \cdot \binom{M}{r} \cdot \mathrm{Bernoulli}(0 \mid \theta)^{M-r} \cdot \mathrm{Bernoulli}(1 \mid \theta)^r$. The thinned process therefore has rate measure

$$\begin{aligned}
&\nu(\mathrm{d}\theta) \cdot \mathrm{Bernoulli}(0 \mid \theta)^N \cdot \binom{M}{r} \cdot \mathrm{Bernoulli}(0 \mid \theta)^{M-r} \cdot \mathrm{Bernoulli}(1 \mid \theta)^r \\
&= \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta^{-1-\sigma+r}(1-\theta)^{c+\sigma-1+N+M-r}\mathrm{d}\theta \\
&= \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \frac{\Gamma(r-\sigma)\Gamma(c+\sigma+N+M-r)}{\Gamma(c+N+M)} \times \\
&\qquad \times \mathrm{Beta}(r-\sigma, c+\sigma+N+M-r)\mathrm{d}\theta \\
&= \alpha \frac{(1-\sigma)_{(r-1)\uparrow}(c+\sigma)_{(N+M-r)\uparrow}}{(1+c)_{(N+M-1)\uparrow}} \mathrm{Beta}(r-\sigma, c+\sigma+N+M-r)\mathrm{d}\theta.
\end{aligned}$$

Since $U_N^{(M,r)}$ counts the thinned atoms, it has Poisson distribution with mean equal to the integral of the rate measure, i.e. mean equal to

$$\alpha \frac{(1-\sigma)_{(r-1)\uparrow}(c+\sigma)_{(N+M-r)\uparrow}}{(1+c)_{(N+M-1)\uparrow}}, \tag{A.12}$$

as was to be shown.

The distribution of $U_N^{(M,\leq R)}$ follows immediately from the observation that the $U_N^{(M,r)}$ are independent Poisson random variables, where the independence is inherited from the independent thinned Poisson point processes.

□

A.2 Asymptotics for number of new rare variants

Proposition 5 (Asymptotics for number of new rare variants)**.** *Under the setting of Theorem 4,*

$$\frac{U_N^{(M,r)}}{M^\sigma} \Bigg| X_{1:N} \xrightarrow{a.s.} \xi_r \ \text{ as } M \to \infty, \tag{A.13}$$

and

$$\frac{U_N^{(M,\leq R)}}{M^\sigma} \Bigg| X_{1:N} \xrightarrow{a.s.} \sum_{r=1}^{R} \xi_r \ \text{ as } M \to \infty, \tag{A.14}$$

where $\xi_r := \frac{\alpha}{r!}(1-\sigma)_{(r-1)\uparrow}\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}$*, and the limits hold true almost surely, conditionally given* $X_{1:N}$*. In addition,*

$$\sqrt{M^\sigma}\left(\frac{U_N^{(M,r)}}{M^\sigma} - \xi_r\right) \Bigg| X_{1:N} \xrightarrow{d} \mathcal{N}(0, \xi_r) \quad \text{as } M \to \infty \tag{A.15}$$

and

$$\sqrt{M^\sigma}\left(\frac{U_N^{(M,\leq R)}}{M^\sigma}-\sum_{r=1}^{R}\xi_r\right)\Big|X_{1:N}\xrightarrow{d}\mathcal{N}\left(0,\sum_{r=1}^{R}\xi_r\right)\quad \textit{as } M\to\infty. \tag{A.16}$$

The limits displayed in Equation (A.15) *and Equation* (A.16) *hold in distribution conditionally given* $X_{1:N}$.

Proof. We start by proving Equation (A.14) in the main text, but in order to do this we have to define some other statistics:

$$U_N^{(M,\leq R)}:=\sum_{r=1}^{R}U_N^{(M,r)}\quad\text{and}\quad U_N^{(M,\geq R)}:=\sum_{r=R}^{M}U_N^{(M,r)} \tag{A.17}$$

which have to be respectively interpreted as the number of new genomic variants observed at most R times and the number of new genomic variants observed at least R times. Our strategy is the following: we prove that $U_N^{(M,\geq R)}/M^\sigma$ converges almost surely to a constant and then we use the relation

$$U_N^{(M,R)}=U_N^{(M,\geq R)}-U_N^{(M,\geq R+1)} \tag{A.18}$$

to prove the convergence of $U_N^{(M,R)}$.
We evaluate the fist moment of $U_N^{(M,\geq R)}/M^\sigma$ a posteriori: for notation purpose, let $\tilde{U}_N^{(M)}$

$$\begin{aligned}\frac{1}{M^\sigma}E_N\left(U_N^{(M,\geq R)}\right)&=\frac{E_N[U_N^{(M)}]}{M^\sigma}-\sum_{r=1}^{R-1}\frac{E_N\left(U_N^{(M,r)}\right)}{M^\sigma}\\&\longrightarrow\frac{\alpha\Gamma(c+1)}{\sigma\Gamma(c+\sigma)}-\alpha\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}\sum_{r=1}^{R-1}\frac{(1-\sigma)_{R-1\uparrow}}{R!},\end{aligned} \tag{A.19}$$

as $M\to\infty$, where we have used Equation (A.23) and Equation (A.9). It then follows that

$$E_N[U_N^{(M,r)}]\asymp c_1M^\sigma$$

for some positive constant $c_1>0$. Besides for the variance of $U_N^{(M,\geq R)}$ we get

$$\begin{aligned}&\text{var}_N\left(U_N^{(M,\geq R)}\right)=\text{var}_N\left(U_N^{(M)}-U_N^{(M,\leq R-1)}\right)\\&=E_N\left\{U_N^{(M)}-\sum_{r=1}^{R-1}\left(U_N^{(M,r)}\right)-E_N\left(U_N^{(M)}\right)+\sum_{r=1}^{R-1}E_N(U_N^{(M,r)})\right\}^2\\&\leq E_N\left\{\left|U_N^{(M)}-E_N\left(U_N^{(M)}\right)\right|+\sum_{r=1}^{R-1}\left|U_N^{(M,r)}-E_N\left(U_N^{(M,r)}\right)\right|\right\}^2\\&\leq R\cdot E_N\left\{\left|U_N^{(M)}-E_N\left(U_N^{(M)}\right)\right|^2+\sum_{r=1}^{R-1}\left|U_N^{(M)}-E_N\left(U_N^{(M,r)}\right)\right|^2\right\}\end{aligned}$$

where the last inequality follows by a simple application of the discrete version of the Hölder's inequality. Then, using also the fact that we get that $U_N^{(M)}$ and $U_N^{(M)}$ are Poisson random variable a posteriori, we obtain:

$$\begin{aligned}\text{var}_N(U_N^{(M,\geq R)}) &\leq r\left(\text{var}_N(U_N^{(M)}) + \sum_{r=1}^{R-1}\text{var}_N(U_N^{(M,r)})\right)\\ &= R\left\{E_N[U_N^{(M)}] + \sum_{r=1}^{R-1}E_N\left(U_N^{(M,r)}\right)\right\} \asymp c_2 M^\sigma\end{aligned}$$

where $c_2 > 0$ is a positive constant. From all the previous considerations and by an application of the Markov inequality we obtain that for any $\varepsilon > 0$

$$\begin{aligned}\text{pr}_N\left\{\Big|\frac{U_N^{(M,\geq R)}}{E_n\left(U_N^{(M,\geq R)}\right)} - 1\Big| \geq \varepsilon\right\} &\leq \frac{\text{var}_N\left(U_N^{(M,\geq R)}\right)}{\varepsilon^2\left\{E_N\left(U_N^{(M,\geq R)}\right)\right\}^2}\\ &\lesssim \frac{c_2 M^\sigma}{\varepsilon^2(c_1 M^\sigma)^2} \asymp \frac{1}{M^\sigma}\end{aligned} \tag{A.20}$$

hence we can conclude that the ratio

$$\frac{U_N^{(M,\geq R)}}{E_N\left(U_N^{(M,\geq R)}\right)}$$

converges in probability to 1. Besides if we choose the subsequence $M_k := k^{2/\sigma}$, as $k = 1, 2, \ldots$, an application of the first Borel-Cantelli lemma leads us to state that the ratio converges to 1 almost surely. Since $U_N^{(M,\geq R)}$ is an increasing process as M increases, for any M in the interval $\{\lfloor m_k \rfloor, \ldots, \lfloor m_{k+1} \rfloor\}$ we have that

$$U_N^{(\lfloor m_k \rfloor,\geq R)} \leq U_N^{(M,\geq R)} \leq U_N^{(\lfloor m_{k+1} \rfloor,\geq R)}$$

where $\lfloor x \rfloor$ denotes the integer part of x. Hence we also have that

$$\frac{U_N^{(\lfloor m_k \rfloor,\geq R)}}{E_N\left(U_N^{(\lfloor m_{k+1} \rfloor,\geq R)}\right)} \leq \frac{U_N^{(M,\geq R)}}{E_N\left(U_N^{(M,\geq R)}\right)} \leq \frac{U_N^{(\lfloor m_{k+1} \rfloor,\geq R)}}{E_N\left(U_N^{(\lfloor m_k \rfloor,\geq R)}\right)}.$$

Leveraging the fact that the lower and upper bound of the central term converge to 1 as $k \to \infty$,

$$\frac{U_N^{(M,\geq R)}}{E_N[U_N^{(M,\geq R)}]} \to 1,$$

in an almost sure sense as $M \to +\infty$. In other words, using Equation (A.19), we have just proved that

$$\frac{U_N^{(M,\geq R)}}{M^\sigma} \longrightarrow \frac{\alpha\Gamma(c+1)}{\sigma\Gamma(c+\sigma)} - \alpha\frac{\Gamma(c+1)}{\Gamma(c+\sigma)}\sum_{r=1}^{R-1}\left(\frac{(1-\sigma)_{r-1\uparrow}}{r!}\right), \tag{A.21}$$

pr_N-almost-surely as $M \to \infty$.

The thesis now follows by Equation (A.21) and the Equation (A.18), indeed:

$$\frac{U_N^{(M,R)}}{M^\sigma} = \frac{U_N^{(M,\geq R)}}{M^\sigma} - \frac{U_N^{(M,\leq R+1)}}{M^\sigma} \to \alpha \frac{\Gamma(c+1)}{\Gamma(c+\sigma)} \frac{(1-\sigma)_{r-1\uparrow}}{r!},$$

P_N-almost surely as M grows.

To prove Equation (A.15) in the main text, one has to prove that the characteristic functions converge, more precisely

$$\Phi_{\sqrt{M^\sigma}\left(\frac{U_N^{(M,R)}}{M^\sigma} - \xi_R\right) \Big| X_{1:N}}(t) \longrightarrow \exp\left(-\frac{t^2 \xi_R}{2}\right) \quad \text{for any } t \in R,$$

as M goes to infinity.

First of all observe that the (posterior) expectation of $U_N^{(M,R)}$ is such that

$$\begin{aligned} E_N\left(U_N^{(M,R)}\right) &= \alpha \binom{M}{R} \frac{(1-\sigma)_{R-1\uparrow}(c+\sigma)_{N+M-R}}{(c+1)_{N+M-1\uparrow}} \\ &= \left(\frac{\Gamma(M+1)}{\Gamma(M-R+1)} \frac{\Gamma(c+\sigma+N+M-R)}{\Gamma(c+N+M)}\right) \times \qquad \text{(A.22)} \\ &\times \alpha \frac{(1-\sigma)_{R-1\uparrow}}{\Gamma(R+1)} \frac{\Gamma(c+1)}{\Gamma(c+\sigma)} \\ &= M^\sigma \left(1 + o(M^{-1})\right) \alpha \frac{(1-\sigma)_{R-1\uparrow}}{\Gamma(R+1)} \frac{\Gamma(c+1)}{\Gamma(c+\sigma)}, \qquad \text{(A.23)} \end{aligned}$$

where we have used the asymptotic expansion of ratios of gamma functions given by Tricomi and Erdélyi [1951]. Let $\tilde{U}_N^{(M,R)} := \sqrt{M^\sigma}\left(\frac{U_N^{(M,R)}}{M^\sigma} - \xi_R\right)$. Using the expansion given in Equation (A.23) it is easy to see that

$$\begin{aligned} &\Phi_{\tilde{U}_N^{(M,R)} \Big| X_{1:N}}(t) = \\ &= \exp\left\{-it\sqrt{M^\sigma}\xi_R + M^\sigma \xi_R \left(1 + o(M^{-1})\right)\left(\frac{it}{\sqrt{M^\sigma}} + \frac{t^2}{2M^\sigma} + o(M^{-1+\sigma})\right)\right\} \\ &= \exp\left\{-\frac{t^2 \xi_R}{2} + o(1)\right\}, \end{aligned}$$

therefore the thesis follows. □

A.3 Number of new rare variants in presence of noise

Similarly, for $U_N^{(M,r)}$ and $U_N^{(M,\leq R)}$ defined in Theorem 4 with $r, R \in [M]$, we have that these quantities are almost surely finite with respective distributions

$$U_N^{(M,r)} \mid X_{1:N} \sim \text{Poisson}\left(\gamma_r\right), \quad U_N^{(M,\leq R)} \mid X_{1:N} \sim \text{Poisson}\left(\sum_{r=1}^{R} \gamma_r\right), \quad \text{(A.24)}$$

where

$$\gamma_r := \binom{M}{r} \int_{\theta=0}^{1} \text{Bernoulli}(1 \mid \phi_{\text{follow}}\theta)^r \text{Bernoulli}(0 \mid \phi_{\text{follow}}\theta)^{M-r} \\ \text{Bernoulli}(0 \mid \phi_{\text{pilot}}\theta)^N \nu(\mathrm{d}\theta) \\ = \alpha \binom{M}{r} \phi_{\text{follow}}^r \frac{(1+c)_{(r-1)\uparrow}}{(1-\sigma)_{(r-1)\uparrow}} E_B\{(1-\phi_{\text{follow}}B)^{M-r}(1-\phi_{\text{pilot}}B)^N\},$$

for $B \sim \text{Beta}(\theta \mid r-\sigma, c+\sigma)$.

A.4 Proof of equality in Equation (4)

To show that $\phi(\lambda, T, p_{err})$ is the right tail of a Poisson distribution, we recur to the Binomial thinning of Poisson random variables.

Proposition 6 (Binomial thinning of Poisson random variables)**.** *Let $N \sim \text{Poisson}(\lambda)$. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(q)$ independently and identically distributed. Then, $S_N := X_1 + \ldots X_N \sim \text{Poisson}(\lambda q)$, and*

$$\Pr(X_N \geq T) = \sum_{t \geq T} \frac{e^{-\lambda q}\lambda^q}{t!}. \tag{A.25}$$

Proof. Let $S_n \sim \text{Binomial}(n, q)$ be a binomial random variable with success probability q and n draws. The moment generating function of the binomial distribution is

$$E\left[t^{S_n}\right] = (1-q+qt)^n,$$

while the moment generating function of the Poisson distribution with parameter $\lambda > 0$ is

$$E\left[t^N\right] = \sum_{k \geq 0} \frac{(\lambda t)^k}{k!} e^{-\lambda} = \exp\{\lambda t - \lambda\}.$$

Hence,

$$E\left[t^{S_N}\right] = \sum_{n \geq 0} \frac{E\left[t^{S_n}\right] \lambda^n e^{-\lambda}}{n!} = \sum_{n \geq 0} \frac{(1-q+qt)^n \lambda^n e^{-\lambda}}{n!} \\ = \exp\left\{\lambda(1-q+qt) - \lambda\right\} = \exp\left\{\lambda q t - \lambda q\right\}, \tag{A.26}$$

which implies $S_N \sim \text{Poisson}(\lambda q)$. □

In light of this proposition, the equality in Equation (4) follows.

B Bayesian prediction with the Beta-Bernoulli product model

We here review the approach proposed by Ionita-Laza et al. [2009]. The authors consider the same problem of genomic variation described in Section 2.

Ionita-Laza et al. [2009] assume that there exists a finite, albeit unknown, number of loci at which genomic variation can be observed. We denote such quantity with the letter K. Given a pilot study $X = X_{1:N}$ with J distinct variants, we can obtain the site-frequency-spectrum (or fingerprint) of the sample,

$$\boldsymbol{f}_N = [f_{N,1} \ldots, f_{N,J}] \quad \text{with} \quad f_{N,j} = \sum_{\ell=1}^{J} \mathbf{1}\left(\sum_{n=1}^{N} x_{n,\ell} = j\right), \tag{B.1}$$

so that $f_{N,1}$ counts the number of variants observed only once among the N samples, $f_{N,2}$ the number of variants observed in exactly two samples etc. The input data $X_{1:N}$ is here viewed as a binary matrix, $X_{1:N} \in \{0,1\}^{N \times J}$, in which all positions at which variation is not observed are discarded, and the order of the columns is immaterial. This binary matrix is modeled via a parametric beta-Bernoulli model: the authors assume that there exists a fixed, unknown number $K < \infty$ of loci at which variation can be observed. For each $j \in [K]$, they assume that there exists an associated variant, labelled by index j, displayed by any observation (row) with probability $\theta_j \in [0,1]$. The frequencies θ_j, $j = 1, \ldots, K$ are distributed according to a beta distribution with parameters a, b, i.e.

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \ldots & \theta_K \end{bmatrix}, \quad \text{with} \quad \theta_j \sim \text{Beta}(a,b)\ \forall j,$$

independently and identically distributed. Conditionally on $\boldsymbol{\theta}$,

$$X_n = \begin{bmatrix} x_{n,1} & \ldots & x_{n,K} \end{bmatrix}, \quad \text{with} \quad x_{n,j} \sim \text{Bernoulli}(\theta_j).$$

Therefore, the columns of the matrix $X_{1:N}$ are independently and identically distributed, while the rows are made of independent, but not identically distributed entries. Under this model, the number of counts of each variant is binomially distributed, conditionally on the latent frequency of such variant, i.e.

$$z_{N,j} \mid \theta_j := \sum_{i=1}^{n} x_{n,j} \mid \theta_j \sim \text{Binomial}(N, \theta_j).$$

Recalling that $f_{N,j} = \sum_{\ell=1}^{J} \mathbf{1}(z_{N,\ell} = j)$ is the number of variants which appear exactly j times among the first N samples, and letting $g(x;a,b)$ be the density function of a beta random variable with parameters a, b evaluated at x,

$$g(x;a,b) = \frac{x^{a-1}(1-x)^{b-1}}{\boldsymbol{B}(a,b)} \mathbf{1}_{[0,1]}(x),$$

with $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, then the probability that exactly j of the N individuals show variation at a given site is given by

$$\begin{aligned} p_{N,j} &= \int_0^1 \binom{N}{j} \theta^k (1-\theta)^{N-j} g(\theta;a,b)\mathrm{d}\theta \\ &= \binom{N}{j} \int_0^1 \frac{\theta^{N+a-1}(1-\theta)^{N-j+b-1}}{\boldsymbol{B}(a,b)}\mathrm{d}\theta = \binom{N}{j} \frac{(a)_{j\uparrow}(b)_{N-j\uparrow}}{(a+b)_{N\uparrow}}. \end{aligned} \tag{B.2}$$

Because we can't observe more than N variants in N trials, and since we don't know anything about variants which are yet to be observed, probabilities of Equation (B.2) are then normalized as follows:

$$\lambda_{N,j} = \frac{p_{N,j}}{\sum_{\ell=1}^{N} p_{N,\ell}} = \frac{\binom{N}{j}(a)_{j\uparrow}(b)_{N-j\uparrow}}{\sum_{\ell=1}^{N}\binom{N}{\ell}(a)_{\ell\uparrow}(b)_{N-\ell\uparrow}},$$

for all $\ell = 1, \ldots, N$. It follows that the log likelihood for the observed data $X_{1:N}$ is given by

$$\ell_{a,b}^{\text{BBPM}}(X_{1:N}) = \log\left(\prod_{j=1}^{N} \lambda_{N,j}^{f_{N,j}}\right) = \sum_{j=1}^{N} f_{N,j}\log(\lambda_{N,j}).$$

Notice that the expected number of variants appearing exactly once in a sample of N observations can be computed in closed form,

$$\begin{aligned}
\eta_{N,1} := E[f_{N,1}] &= E\left\{\sum_{j=1}^{K}\binom{N}{1}\mathbf{1}\left(\sum_{n=1}^{N} x_{n,j} = 1\right)\right\} \\
&= KN\int_{[0,1]} \frac{\theta^{a-1}(1-\theta)^{b-1}}{\boldsymbol{B}(a,b)}\theta(1-\theta)^{N-1}\mathrm{d}\theta \\
&= KN\frac{\boldsymbol{B}(a+1, N+b-1)}{\boldsymbol{B}(a,b)} = \frac{aKN}{N+b-1}\frac{\boldsymbol{B}(a,N+b)}{\boldsymbol{B}(a,b)},
\end{aligned}$$

where we used independence of the variants, linearity of the expectation operator and the properties of the beta function. Letting $M = tN$ be the number of additional samples to be observed, we can compute the expected number of hitherto unseen variants, to be observed in additional M samples after N samples have been collected as

$$\begin{aligned}
\Delta_N(M) &= E\left\{\sum_{j=1}^{K}\mathbf{1}\left(\sum_{m=1}^{M} x_{m,j} > 0\right)\mathbf{1}\left(\sum_{n=1}^{N} x_{n,j} = 0\right)\right\} \\
&= \frac{K}{\boldsymbol{B}(a,b)}\int_{[0,1]} (1-(1-\theta)^{(t+1)N}) - (1-(1-\theta)^{N})\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta \\
&= \frac{K}{\boldsymbol{B}(a,b)}\int_{[0,1]} \left\{(1-\theta)^{N} - (1-\theta)^{(t+1)N}\right\}\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta \\
&= K\frac{\boldsymbol{B}(a,N+b)}{\boldsymbol{B}(a,b)} - K\frac{\boldsymbol{B}(a,N(t+1)+b)}{\boldsymbol{B}(a,b)}
\end{aligned}$$

Now, noting that

$$K\frac{\boldsymbol{B}(a,N+b)}{\boldsymbol{B}(a,b)} = \frac{\eta_{N,1}}{a}\frac{N+b-1}{N}$$

and

$$K\frac{\boldsymbol{B}(a, N(t+1)+b)}{\boldsymbol{B}(a,b)} = \frac{\eta_{N,1}}{a}\frac{N+b-1}{N}\frac{\boldsymbol{B}(a, N(t+1)+b)}{\boldsymbol{B}(a, N+b)},$$

it follows that

$$\Delta_N(M) = \frac{\eta_{N,1}}{a}\frac{N+b-1}{N}\left\{1 - \frac{\boldsymbol{B}(a, N(t+1)+b)}{\boldsymbol{B}(a, N+b)}\right\}. \tag{B.3}$$

Importantly, $\Delta_N(M)$ depends on K only via $\eta_{N,1}$. To use the estimator $\Delta_N(M)$, Ionita-Laza et al. [2009] substitute $\eta_{N,1}$ with its empirical counterpart $f_{N,1}$, the number of variants which have been observed exactly once in the sample $X_{1:N}$. The parameters a, b are found via maximization of the log-likelihood of the model,

$$\{a^*, b^*\} = \underset{a>0, b>0}{\arg\max} \left\{\ell_{a,b}^{\text{BBPM}}(X_{1:N})\right\}$$

Remark 7. *The estimator obtained in Equation* (B.3) *crucially relies on the empirical frequency of variants observed once among the first N draws, $f_{N,1}$. For example, if a dataset had $f_{N,1} = 0$, $\Delta_N(M) = 0$ for every $M > 0$.*

C Linear program to estimate the frequencies of frequencies

Zou et al. [2016] assume, in the same way as Ionita-Laza et al. [2009], that there exists a finite albeit unknown number of sites at which variants can be observed. They formalize the problem of hitherto unseen variants prediction as that of recovering the distribution of frequencies of all the genetic variants in the population, including those variants which have not yet been observed.

They assume that each possible variant in a sample is independent of the other variants, and that the j-th variant appears with a given probability θ_j conditionally independently and identically distributed across all the individuals observed - i.e. the θ_j are parameters of independent Bernoulli random variables $x_{n,j}$ for all $n \geq 1$ and j. Therefore the pilot study $X_{1:N}$ is modeled by a collection of independent Bernoulli random variables, which are also identically distributed along each column, and the sum $z_{N,j} \mid \theta_j := \sum_{n=1}^{N} x_{n,j} \mid \theta_j \sim \text{Binomial}(N, \theta_j)$. From the frequencies $z_{N,1}, \ldots, z_{N,J}$ of the J variants observed among the first N samples, it is possible to compute the fingerprint of the sample, $\boldsymbol{f}_N$. Given the fingerprint, the goal is to recover the population's histogram, which is a map quantifying, for every $\theta \in [0, 1]$, the number of variants such that $\theta_j = \theta$. Formally, learn a map h from the distribution of frequencies to integers

$$h : (0, 1] \to \mathbb{N} \cup \{0\} \tag{C.1}$$

Because for N large enough the empirical frequencies associated to common variants should be well approximated by their empirical counterpart, Zou et al. [2016]

only consider the problem of estimating the histogram from the truncated fingerprint $\boldsymbol{f}_N^{(\kappa)} = \{f_{N,j} \ : \ j/N \leq 100 \times \kappa\}$. In their analysis, the authors only consider $\kappa = 1$, i.e. they consider "common" variants all those variants that appear in more than 1% of the sample elements. Moreover, rather than learning a continuous function as described by Equation (C.1), they solve a discretized version of the problem. They fix a discretization factor $\delta \geq 1$, and then set up a linear program in which the goal is to correctly estimate the population histogram associated to the frequencies in the set $\mathcal{S} = \left\{\frac{1}{1000N}, \delta\frac{1}{1000N}, \ldots, \delta^i\frac{1}{1000N}, \ldots, \kappa\right\}$. The value δ, given κ, determines how many frequencies are going to be estimated in $(0, \kappa]$: the lower δ, the finer the discretization. The authors suggest using $\delta = 1.05$. In our experiments, we set $\delta = 1.01$, for which we find the method to produce better results, at the cost of a small additional computational effort. Finally, the problem of recovering the histogram is solved through the following optimization:

$$\min_{h(\theta), \theta \in \mathcal{S}} \sum_{j: j \leq N\kappa} \frac{1}{1 + f_{N,j}} \left| f_{N,j} - \sum_{\theta \in \mathcal{S}} h(\theta) \mathrm{Binomial}(N, \theta, j) \right|$$

subject to

$$h(\theta) \geq 0, \sum_{\theta \in \mathcal{S}} h(\theta) \leq K, \sum_{\theta \in \mathcal{S}} \theta \cdot h(\theta) + \sum_{j: j > N\kappa}^{J} \frac{j}{N} f_{N,j} = \frac{J}{N},$$

where K is an upper bound on the total number of variants, and $\mathrm{Binomial}(N, \theta, j)$ is the probability that a Binomial draw with bias θ and N rounds is equal to j.

Given the histogram $\hat{h}$ which solves the linear program above, one can obtain an estimate of the number of unique variants at any sample size M using

$$V(\hat{h}, M) = \sum_{\theta: \hat{h}(\theta) > 0} \hat{h}(\theta)(1 - (1 - \theta)^M).$$

Following Zou et al. [2016], we refer to this estimator as the "unseenEST" estimator.

D Jackknife estimators

Jackknife estimators for predicting the number of hitherto unseen species were first introduced by in the capture-recapture literature by Burnham and Overton [1978].

Given $X_{1:N} \overset{iid}{\sim} F(\psi)$ for some distribution F and some parameter ψ, let $\hat{\psi}_N = \hat{\psi}_N(X_{1:N})$ be an estimator of ψ with the property that

$$E[\hat{\psi}_N] = \psi + \frac{a_1}{N} + \frac{a_2}{N^2} + \ldots, \tag{D.1}$$

for fixed constants $a_1, a_2, \ldots$. Without loss of generality assume $\hat{\psi}_N$ to be symmetric in its inputs $X_{1:N}$, and denote with $\mathcal{I} \subset [N]$ a subset of given size p, let $\hat{\psi}_{N-p,\mathcal{I}}$ be the

estimate obtained by dropping the observations whose indices are in $\mathcal{I}$. Similarly, let

$$\hat{\psi}_N^{(p)} = \binom{N}{p}^{-1} \sum_{\mathcal{I}:|\mathcal{I}|=p} \hat{\psi}_{N-p,\mathcal{I}} \tag{D.2}$$

The idea of the Jackknife estimator is that, if the assumption of Equation (D.1) holds, we can improve over $\hat{\psi}_N$ by using a correction originating from Equation (D.2). The p-th order Jackknife estimator is defined as

$$\hat{\psi}_N^{J_p} = \frac{1}{p} \sum_{\ell=0}^{p} \left\{ (-1)^\ell \binom{p}{\ell} (N-\ell)^p \hat{\psi}_N^{(\ell)} \right\}. \tag{D.3}$$

Under the assumption of Equation (D.1), the estimator of Equation (D.3) has bias approaching zero polynomially fast in the correction order, $\text{Bias}(\hat{\psi}_N^{J_p}) \sim N^{-p-1}$.

D.1 An estimator for the population size

Burnham and Overton [1978] introduced a nonparametric procedure to estimate the total number of animals present in a closed population when capture-recapture data is available. Assume that there is a fixed, but unknown number K of total species. Over the course of N repeated observational experiments, $J \leq K$ distinct species are observed.

Let $X_{1:N}$ be the collection of available data, in which $X_n = [x_{n,1}, \ldots, x_{n,J}]$, with $x_{n,j} = 1$ if species j has been observed on the n-th experiment, and 0 otherwise. Moreover, assume that each species $j \in [K]$ has a fixed, but unknown probability $\theta_j \in (0, 1]$ of being observed.

Notice that while Burnham and Overton [1978] developed the estimator having in mind a fixed and finite population of animals, we can also think of each sample X_n as a genomic sequence characterized by the presence or absence of genetic variants at different sites.

The nonparametric MLE for the total support size K is given by $\hat{K}^{\text{MLE}}(X_{1:N}) = \hat{K}_N^{\text{MLE}} = J$. Clearly $J \leq K$, therefore J is a biased estimate for K. If one assumes, in a similar spirit to Equation (D.1), that

$$E[\hat{K}_N^{\text{MLE}}] = K + \frac{a_1}{N} + \frac{a_2}{N^2} + \ldots, \tag{D.4}$$

then one could use the jackknife estimator of Equation (D.3) to estimate K. This requires computing $\hat{\psi}_N^{(\ell)}$ for $\ell = 1, \ldots, p$, which are linear functions of the observed fingerprint $\boldsymbol{f}_N$.

The case $p = 1$: We outline the approach for $p = 1$. Let $q_{N,n}$ be the number of animals which have been observed only once out of the N trials, exactly on the n-th,

$$q_{N,n} = \sum_{j \geq 1} \mathbf{1}(x_{n,j} = 1) \mathbf{1}\left(\sum_{n' \neq n} x_{n',j} = 0 \right)$$

Then, because $q_{N,1} + \ldots + q_{N,N} = f_{N,1}$ by construction,

$$\hat{K}_N^{(1,\setminus n)} = J - q_{N,n} \quad \text{and} \quad \hat{K}_N^{(1)} = \frac{1}{N}\sum_{n=1}^{N} \hat{K}_N^{(1,\setminus n)} = J - \frac{f_{N,1}}{N}. \tag{D.5}$$

Therefore, the order 1 jackknife estimator for the total population size is obtained by plugging in $\hat{\psi}_N^{(0)} = J$ and $\hat{\psi}_N^{(1)} = J - \frac{f_{N,1}}{N}$ in Equation (D.3):

$$\hat{K}_N^{J_1} = J + \frac{N-1}{N} f_{N,1} \tag{D.6}$$

The case for general p: For any $p \leq N$, it always holds that

$$\hat{K}_N^{(p)} = J - \binom{N}{p}^{-1} \sum_{\ell=1}^{p} \binom{N-\ell}{p-\ell} f_{N,\ell} \tag{D.7}$$

This formula allows to obtain the general Jackknife estimator of order p, which is a linear function of the observed number of species J and correction terms which depend on the fingerprint $\boldsymbol{f}_N$,

$$\hat{K}_N^{J_p} = \sum_{\ell=1}^{p} a_{N,\ell}^{(p)} f_{N,\ell}.$$

D.2 Estimators for the number of hitherto unseen genomic variants

Taking inspiration from the approach of Burnham and Overton [1978], Gravel et al. [2011] and Gravel [2014] developed Jackknife estimators for the number of hitherto genomic variants which are going to be observed in M additional samples given N initial ones. Let $V(N)$ denote the total number of variants observed in N samples, and let $\Delta(N+M, N) := H_{N+M-1} - H_{N-1} = \sum_{\ell=N}^{M+N-1} 1/\ell$, where

$$H_N \;=\; 1 + 1/2 + \ldots + 1/N$$

is the N-th harmonic number. To derive their estimators, the authors use the assumption that for a given order $p \geq 1$ the total number of variants present in $N+M$ samples can be estimated as follows:

$$\hat{V}_N^{(M)} = V(N) + \sum_{\ell=1}^{p} a_{N,\ell}^{(p)} \Delta(N+M, N)^{\ell}, \tag{D.8}$$

where $\boldsymbol{a}_N^{(p)} = [a_{N,1}^{(p)}, \ldots, a_{N,p}^{(p)}]$ are constants which depend on the initial sample size N, on the order p and on the fingerprint of the sample $\boldsymbol{f}_N$. This assumption is exact in the case of a constant size and neutrally evolving population (Gravel et al. [2011]). For a given order p the unknown coefficients are obtained by solving the following system of equations:

$$\hat{V}_N^{(M)} = \hat{V}_{N-1}^{(M)} = \ldots = \hat{V}_{N-p}^{(M)}. \tag{D.9}$$

Equating $\hat{V}_N^{(M)}$ to $\hat{V}_{N-j}^{(M)}$ using Equation (D.8) for $j = 1, \ldots, p$, we obtain a system of $p-1$ equations of the form

$$V(N) - V(N-j) = \sum_{\ell=1}^{p} a_{N,\ell}^{(p)} (\Delta(N+M, N-j)^{\ell} - \Delta(N+M, N)^{\ell}). \quad \text{(D.10)}$$

Using the additional equality

$$V(N) - V(N-\ell) = \sum_{j=1}^{\ell} \frac{\binom{\ell}{j}}{\binom{N}{j}} f_{N,j}. \quad \text{(D.11)}$$

we can solve for $a_{N,\ell}^{(p)}$ and express these in terms of N, $\Delta(N+M, N)$ and the fingerprint $\boldsymbol{f}_N$, and the final estimator is a linear function of the fingerprint $\boldsymbol{f}_N$.

D.3 Choice of the jackknife order

As pointed out in Burnham and Overton [1978], the optimal order p of the jackknife estimator heavily depends on the data under consideration. It is therefore desirable to obtain a procedure which uses the data to guide the choice of such order. Burnham and Overton [1978] phrase this decision problem as a sequential hypothesis test, in which one keeps increasing the order of the jackknife until the data suggests that the drop in bias obtained by increase the jackknife order is exceeded by the gain in variance. Precisely, for $p = 1, 2, \ldots$ one sequentially performs the following test:

$$H_{0,p}: \; E(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p}) = 0 \quad \text{versus} \quad H_{a,p}: \; E(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p}) \neq 0. \quad \text{(D.12)}$$

If $H_{0,p}$ is rejected, this has to be interpreted as evidence that the bias reduction provided by the $p+1$-th order (with respect to the p-th) is larger than the associated increase in variance, and $p+1$-th order should be preferred to the p-th order [Burnham and Overton, 1978]. The first order p for which the test fails to reject the null hypothesis is chosen as the jackknife order.

The test relies on the following observation: the difference between two jackknife estimators of different orders $p+1$ and p is given by

$$\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p} = \sum_{\ell=1}^{p+1} \tilde{a}_{N,\ell}^{(p+1,p)} f_{N,p}, \quad \text{(D.13)}$$

again a linear combination of the fingerprint. Because the conditional distribution of the fingerprint is independent of K given J, the minimum variance estimator of the conditional variance is given by

$$\text{est var}(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p} \mid J) = \frac{J}{J-1} \left\{ \sum_{\ell=1}^{p} (\tilde{a}_{N,\ell}^{(p+1,p)})^2 f_{N,\ell} \frac{(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p})^2}{J} \right\}. \quad \text{(D.14)}$$

Under $H_{0,p}$, the test statistic

$$T_p = \frac{\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p}}{\sqrt{\text{est var}(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p} \mid J)}} \tag{D.15}$$

is approximately normally distributed.

For a given extrapolation size M, we can apply the same procedure to the estimators derived in Gravel et al. [2011] and Gravel [2014], which are again linear combinations of the fingerprint.

E Good-Toulmin estimators

In recent work, Chakraborty et al. [2019] used the classic smoothed Good-Toulmin estimator [Good and Toulmin, 1956, Efron and Thisted, 1976, Orlitsky et al., 2016] in the context of rare variants prediction. Under the same sampling model assumed by Zou et al. [2016], this method allows to predict the number of additional variants that will be observed in M additional samples by using the formula

$$\Delta_N(M) \mid X_{1:N} = \begin{cases} \sum_{r=1}^{\infty} (-1)^{r+1} \left(\frac{M}{N}\right)^r f_r & \text{if } M/N \leq 1 \\ \sum_{r=1}^{\infty} (-1)^{r+1} \left(\frac{M}{N}\right)^r f_r P(M,N,r) & \text{if } M/N > 1 \end{cases}, \tag{E.1}$$

where

$$P(M,N,r) = \Pr(\text{Binomial}(\kappa(M,N)), \theta(M,N) \geq r) \tag{E.2}$$

where the smoothing parameters κ and θ can take two different forms: either

$$\kappa(M,N) = \lfloor 0.5 \log_2((M^2/N)/(M/N-1)) \rfloor \quad \text{and} \quad \theta(M,N) = 1/(M/N+1) \tag{E.3}$$

or

$$\kappa(M,N) = \lfloor 0.5 \log_3((M^2/N)/(M/N-1)) \rfloor \quad \text{and} \quad \theta(M,N) = 2/(M/N+2). \tag{E.4}$$

F Additional details and experiments on the TCGA and MSK-impact dataset

F.1 Details about the experimental setup

The TCGA and the MSK-impact datasets are two publicly available cancer genomics datasets, containing somatic variants from $N = 10{,}275$ and $N = 9{,}091$ samples respectively. In both datasets, for each patient-id, we have access to a list of recorded variants, together with (i) the gene at which the variant was observed, (ii) and the type of cancer the patient was diagnosed with. The TCGA dataset contains variants from

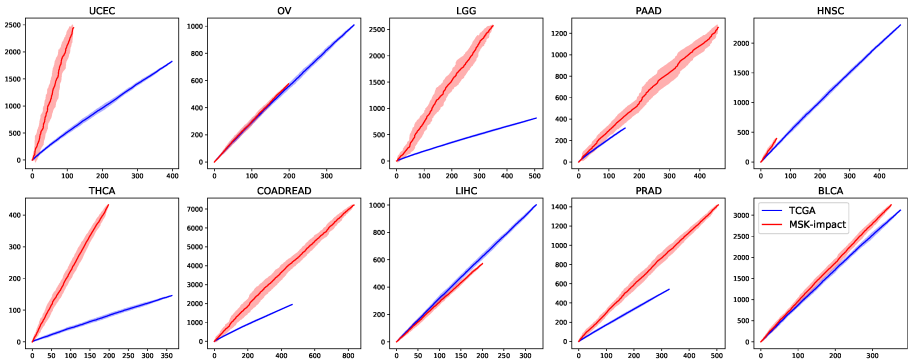

Figure 5: For ten different cancer subtypes, we plot the rate of growth of the number of distinct variants as a function of the sample size for TCGA and MSK-impact. We shuffle the observations according to 20 different permutations, the solid line represents the average number observed across all permutations, and the shaded regions represent one standard deviation above/below. For both datasets, we remove hypermutated outliers - we drop those samples that show more than 3 times the median number of variants observed in the tumor subtype dataset.

the whole exome (variants are recorded across a total of $G = 19{,}441$ genes) for 33 different cancer types. The MSK-impact is a hybridization capture-based NGS clinical assay that is capable of detecting mutations in all exons and selected introns and promoter mutations in $G_1 = 412$ cancer-associated genes [Chakraborty et al., 2019] across a finer classification of 329 different cancer types.

We thank an anonymous reviewer for pointing out that, as in Chakraborty et al. [2019], this data might suffer from normal cell contamination and tumor heterogeneity. We believe future work might address this issue by introducing an additional parameter variable describing what fraction of each sample belongs to the tumor, but this extension is beyond the scope of the present work.

From the data, it is natural to obtain a binary encoding of variation as described in Section 2 either for specific cancer types (i.e. across all genes, restrict our attention to patients who got diagnosed with the same cancer type) or to specific genes. That is to say, use the machinery of Section 3 to either predict (A) how many new variants we are going to observe from new samples that have been classified with a specific cancer type or (B) how many new variants we are going to observe from new samples in a specific gene (for any cancer type).

When compared across tumors, viceversa, the TCGA and MSK-impact, even when restricting the attention to the same targeted genes, can show substantially different behavior (see Figure 5). Therefore, we don't try to predict the rate of growth of new variants observed for cancer types using, e.g. the TCGA dataset as a pilot study and the MSK-impact as a follow-up study.

Across genes, the TCGA and MSK-impact are relatively similar in terms of number of variants observed per gene (see Figure 6 and Figure 7).

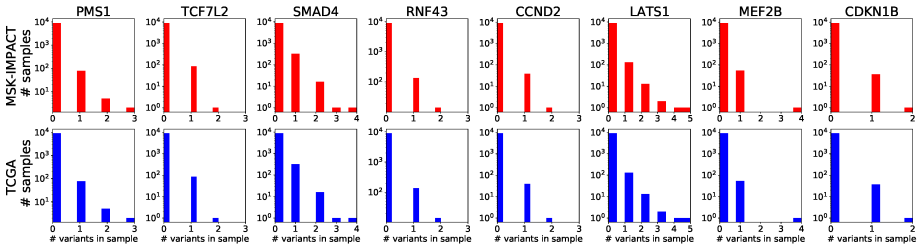

Figure 6: Comparing the number of variants observed for a given gene across the samples (top row: IMPACT, bottom row: TCGA; different columns are different genes).

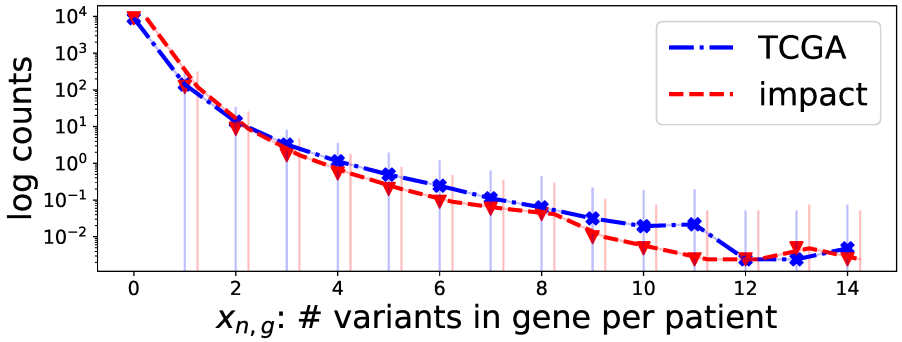

Figure 7: For every patient n and every gene g, we let $x_{n,g}$ be the number of variants recorded for patient n in gene g. This number ranges from 0 to 14 in both TCGA and MSK-impact. We plot the histogram of the $x_{n,g}$ for both the MSK-impact (red) and the TCGA (blue), y-axis in log-scale.

For the experiments under changing experimental conditions, i.e. in the setting in which samples are noisy, we perform further thinning to generate the data. That is to say, given K variants across samples, each with empirical frequency $\hat{\theta}_k$, $k = 1, \dots, K$ and for a given choice of T, sampling error p_{err} and sequencing quality λ, we obtain the associated probability ϕ that at least T successful reads are obtained at any position k for any individual n, i.e.

$$\phi(\lambda, T, p_{\text{err}}) := \sum_{t \geq T} \frac{1}{t!} e^{-\lambda(1-p_{\text{err}})} \{\lambda(1 - p_{\text{err}})\}^t.$$

Then, an individual observation $X_n = [x_{n,1}, \dots, x_{n,K}]$ is obtained by independently sampling Bernoulli random variables,

$$x_{n,k} \mid \hat{\theta}_k, \phi(\lambda, T, p_{\text{err}}) \sim \text{Bernoulli}(\hat{\theta}_j \phi(\lambda, T, p_{\text{err}})).$$

F.2 Prediction across genes with the same experimental conditions

We replicate the setup of Chakraborty et al. [2019] and use the TCGA dataset as a pilot study and the MSK-impact as a follow-up study. We restrict our attention to the 412 targeted genes in the MSK-impact. For each targeted gene, in a similar way as to what done for the experiments in Section 6, we create ten different folds of the data, by sampling (without replacement) for each fold a random subset of 80% of the data. We train on these folds of the TCGA dataset our BNP predictor, as well as the Good-Toulmin method used in Chakraborty et al. [2019] and the Jackknife estimators proposed by Gravel [2014] to predict both (1) the expected number of new variants in a single new sample in MSK-impact and (2) the total number of new variants we expect to see in a total cohort of the same size as the MSK-impact ($M = 9{,}091$ samples). The linear programming of Zou et al. [2016] failed to provide reliable prediction, especially for those genes in which very few observations showed variation (i.e. those genes with few "active" patients). We therefore we excluded it in our analysis. As shown in Figure 8, the three methods considered (Good-Toulmin, fourth order Jackknife and our BNP predictor) performed similarly.

To quantify the predictive performance of the different methods we used the following setup: for gene g, let $Z^{\text{TCGA},g} \in \{0,1\}^{N \times K}$ be the binary matrix of variation in gene g in the TCGA and similarly $Z^{\text{MSK},g} \in \{0,1\}^{M \times K}$ the corresponding matrix for the MSK-impact. Let $p_{1,g}$ be, across all patients in the MSK-impact, the average observed number of new variants displayed by one individual in gene g which are not displayed by any patient in the TCGA dataset in gene g:

$$p_{1,g} := \frac{1}{M} \sum_m \left[\sum_k 1(Z^{\text{MSK},g}_{m,k} = 1) 1 \left\{ \sum_n \left(Z^{\text{TCGA},g}_{n,k} = 0 \right) \right\} \right].$$

Similarly, let $p_{M,g}$ be the total number of new variants displayed in the MSK-impact that were not present in the TCGA dataset for gene g:

$$p_{M,g} = \sum_k \left\{ 1 \left(\sum_m Z_{m,k} > 0 \right) 1 \left(\sum_n Z_{n,k} = 0 \right) \right\}.$$

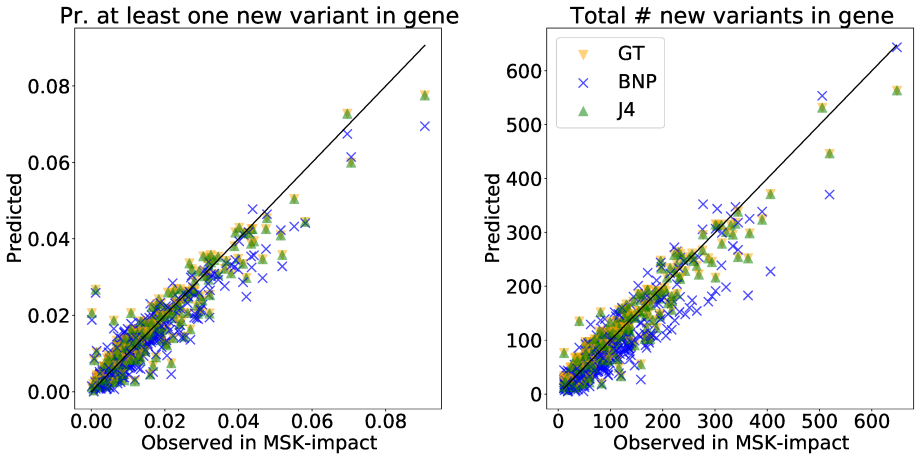

Figure 8: For every targeted gene present in the MSK-impact dataset, we use samples from the TCGA dataset to predict the expected number of new variants in that gene (1, left) from a new sample and (2, right) in a new cohort of the same size as the MSK-impact.

For a given method $\mathcal{A} \in \{\text{BNP}, \text{GT}, \text{J4}\}$ (our Bayesian nonparametric method, the Good-Toulmin method used in Chakraborty et al. [2019] and the fourth order Jackknife proposed in Gravel [2014]), let $\hat{p}^{\mathcal{A}}_{1,g}$ and $\hat{p}^{\mathcal{A}}_{M,g}$ be the prediction of $p_{1,g}$ and $p_{M,g}$ respectively. Given a threshold t and method $\mathcal{A}$, for $m \in \{1, M\}$

$$\ell(t; g, m) = \min \left\{ \left| \frac{\hat{p}_{m,g} - p_{m,g}}{p_{m,g}} \right|, t \right\}, \tag{F.1}$$

and, summing over all genes g,

$$\ell(t; m) = \sum_{g} \ell(t; g, m). \tag{F.2}$$

For $\alpha = 1$ and $t = \infty$, Equation (F.2) quantifies the absolute average percentage error. Setting $t < \infty$, instead, implies a Huber loss, in which we truncate the (percentage) loss in case it is larger than the threshold t. In practice, in our experiments, for every gene g we have ten different folds of the data, which lead, for every predictor, to ten different prediction values. For every gene, and for every prediction method, we retain the median predicted value across the folds, and compute the error (Equation (F.2)) relative to that fold.

F.3 Predictions across genes with different experimental conditions

We then move to prediction under changing experimental conditions. Given the original MSK-impact cohort (sampled at an average of 480x depth), we generate pseudo-observations by subsampling data at 100x. To do so, we assume that each patient-locus pair for which variation is observed, $X_{n,k} = 1$, in the final dataset is generated

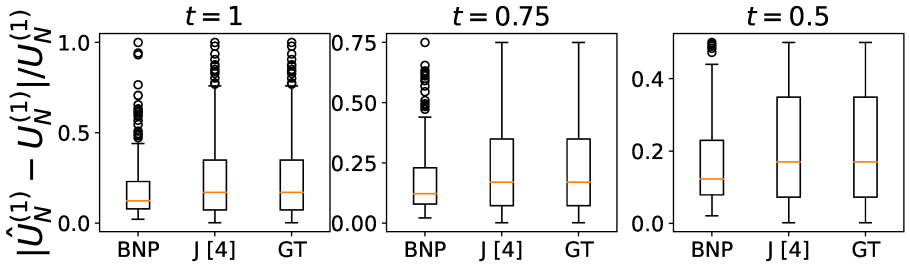

Figure 9: For the one-step-ahead prediction problem, we use the TCGA dataset (on each targeted gene) to predict the expected number of new variants in one additional sample in the MSK-impact dataset using our BNP predictor, the (smoothed) Good-Toulmin predictor, as well as the fourth order jackknife. We report, for each fold in the data and for each gene, the loss introduced in Equation (F.2) for $t \in \{1, 0.75, 0.5\}$ through a boxplot.

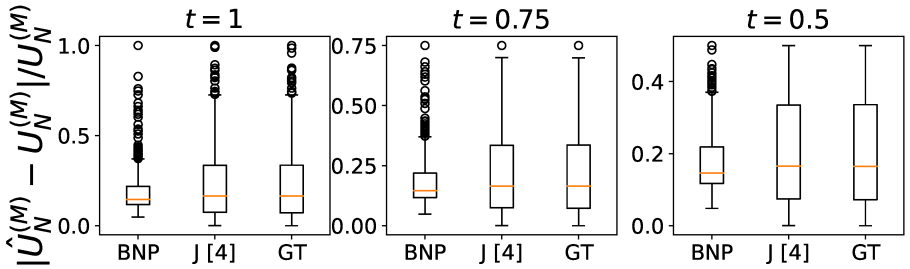

Figure 10: For the many-steps-ahead prediction problem, we use the TCGA dataset (on each targeted gene) to predict the expected number of new variants in a new dataset of the size of the MSK-impact dataset using our BNP predictor, the (smoothed) Good-Toulmin predictor, as well as the fourth order jackknife.We report, for each fold in the data and for each gene, the loss introduced in Equation (F.2) for $t \in \{1, 0.75, 0.5\}$ through a boxplot.

as follows: first a Poisson random variable $Y_{n,k} \sim \text{Poisson}(\lambda)$ is drawn, and then $X_{n,k} = 1(Y_{n,k} \geq T)$ is kept. We pick $\lambda = 100$ and $T = 90$ in our experiment. On the new dataset, again, our predictor outperforms alternative methods (see Figure 11). The explanation follows the one given in Section 6: our method can adapt to the changing conditions whereas competing methods cannot.

F.4 Optimal design of experiments

Last, we use prediction under changing conditions to inform the (optimal) design of a follow up study. We choose again the cost function $C(m, \lambda) = m \log \lambda$, to sample m new observations at depth λ. We fix a budget which allows us to sample at full depth ($\lambda = 480$) only $M' = M/2$ observations, half of the MSK-impact sample size.

We find that the same trade-off observed in Section 6.4 is also present here. Across genes, we can find a configuration of the sequencing depth ($\lambda = 62$) that leads to a median gain of 6 additional new variants discovered (per gene), that is an average increase of 10.56% with respect to the number of variants we would have discovered if we had used the full-sequencing depth available alternative under the same budget constraint and cost function.

G Additional experimental results: GnomAD data

G.1 Experimental setup

In order to run our experiments, we use data from the gnomAD (genome aggregation dataset) discovery project [Karczewski et al., 2020], the largest and most comprehensive publicly available human genome dataset. This dataset contains 125'748 exomes sequences (i.e. protein-coding regions of the genome), from 8 main populations (African American, Latino, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, South Asian, Other[1]). Sample size varies widely across sub populations, e.g. the "Other" subgroup counts only 3'070 observations, while "Non-Finnish European" contains 56'885 individuals. Moreover, some of these main populations are further split into additional sub populations, e.g. "Non-Finnish European" contains the "Bulgarian", "Estonian", "Northern European", "Southern European", "Swedish", "Other European" sub populations, while the "East Asian" sub population is further split into the "Korean", "Japanese" and "Other East Asian" sub populations (see Karczewski et al. [2020] for additional details). We ran our analysis on all populations and sub populations.

Because for privacy reasons not all individual sequences are accessible, in order to run our analysis we generate synthetic data which closely resembles the true data as follows. For every subpopulation with N individuals and every position $j = 1, \ldots, K$ in the exome, we have access to the total number of individuals N_j showing variation at position j. We compute the empirical frequency of variation at site j, $\hat{\theta}_j := N_j/N$ for all $j = 1, \ldots, K$. Our data is then generated by sampling independent Bernoulli

[1]The "Other" subgroup contains all "*individuals were classified as "other" if they did not unambiguously cluster with the major populations in a principal component analysis*"

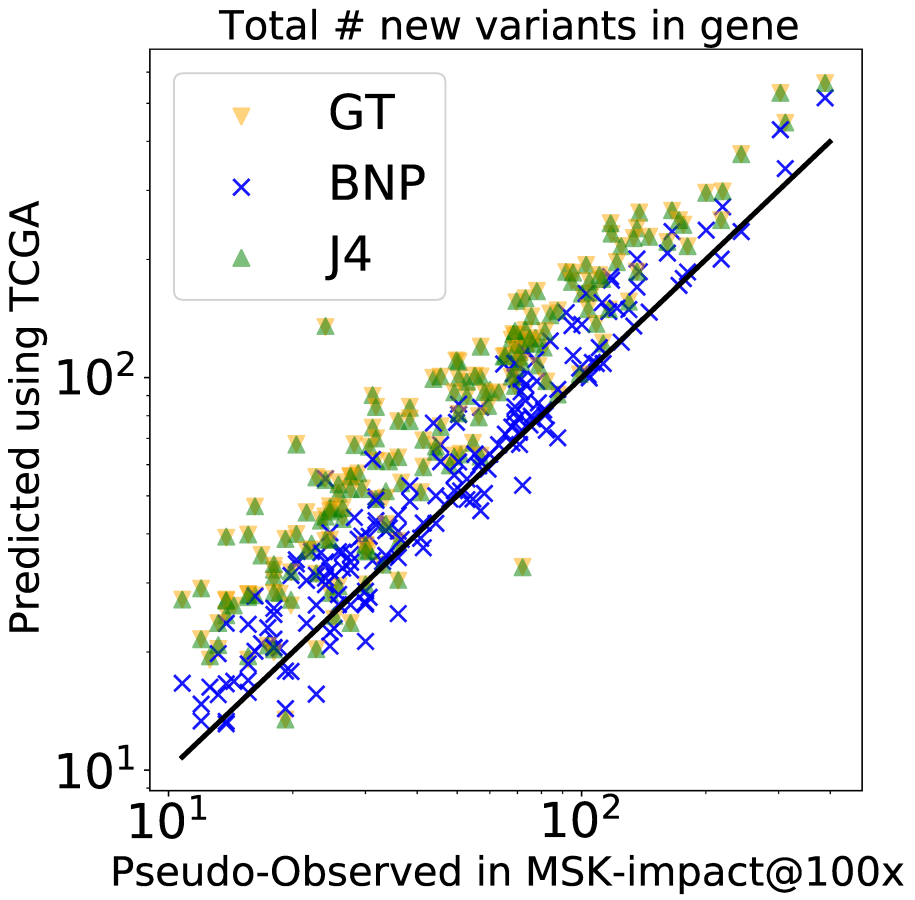

Figure 11: For every targeted gene present in the MSK-impact dataset, we use samples from the TCGA dataset to predict the expected number of new variants in that gene in a new cohort of the same size as the MSK-impact, but now assuming the MSK is sampled at a different sequencing depth than the TCGA.

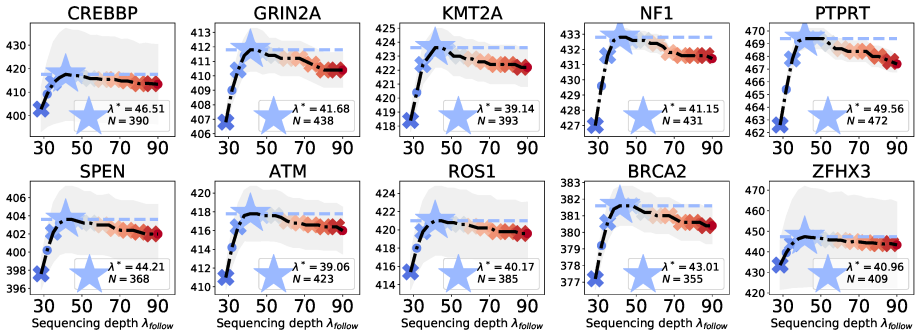

Figure 12: For every targeted gene present in the MSK-impact dataset, we use samples from the TCGA dataset to find the optimal configuration of the sequencing depth in the follow up study so maximize the number of new variants discovered. Here we display results for the 10 genes with the largest number of variants.

random vectors $X_1, \ldots, X_N$, with $X_n = [x_{n,1}, \ldots, x_{n,K}]$. The entries in the vector are independent Bernoulli random variables, $x_{n,j} \sim \text{Bernoulli}(\hat{\theta}_j)$.

For the prediction experiments under changing conditions, i.e. in the setting in which samples are noisy, we perform further thinning to generate the data. That is to say, given the empirical frequencies $\{\hat{\theta}_j\}$, $j = 1, \ldots, K$ and for a given choice of T, sampling error p_{err} and sequencing quality λ, we obtain the associated probability ϕ that at least T successful reads are obtained at any position j for any individual n, i.e.

$$\phi(\lambda, T, p_{\text{err}}) := \sum_{t \geq T} \frac{1}{t!} e^{-\lambda(1-p_{\text{err}})} \{\lambda(1 - p_{\text{err}})\}^t.$$

Then, an individual observation $X_n = [x_{n,1}, \ldots, x_{n,K}]$ is obtained by independently sampling Bernoulli random variables,

$$x_{n,j} \mid \hat{\theta}_j, \phi(\lambda, T, p_{\text{err}}) \sim \text{Bernoulli}(\hat{\theta}_j \phi(\lambda, T, p_{\text{err}})).$$

G.2 Prediction with no sequencing errors

Various existing methods predict the number of new variants in a follow-up study under the assumption that experimental conditions remain constant between the pilot and follow-up. These approaches use, respectively, parametric Bayesian methods [Ionita-Laza et al., 2009], linear programming [Gravel, 2014, Zou et al., 2016], a harmonic jackknife [Gravel, 2014], and a smoothed version of the classic Good-Toulmin estimator [Orlitsky et al., 2016, Chakraborty et al., 2019]. We encountered numerical issues with the linear programming method of Gravel [2014] and so, like Zou et al. [2016], we do not include it in our comparison. Even in their original paper, Gravel [2014] did not report superior performance of their linear programming approach over their other method, the harmonic jackknife [Gravel, 2014], which we do include in our comparison. To assess prediction error in each case, we use an approach akin to cross

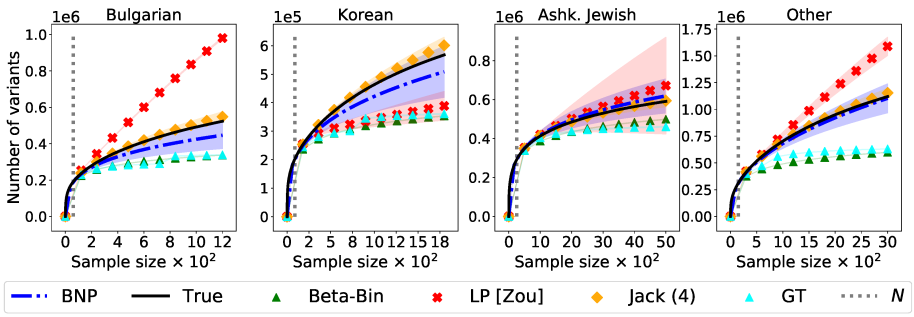

Figure 13: Predicting the number of new variants in a follow-up study under constant experimental conditions. The solid black line displays the true number of distinct variants (vertical axis) as the sample size increases (horizontal axis). The dotted vertical line indicates the pilot study sample size. Lines for each method are averaged across all folds; see Appendix G.2 (blue: our method, Bayesian nonparametric (BNP), Eq. (1); green: Ionita-Laza et al. [2009]; red: Zou et al. [2016]; orange: Gravel [2014] (4th order), cyan: Chakraborty et al. [2019]). Shaded regions show one standard deviation across data folds.

validation. Namely, we treat each subpopulation as a dataset. We divide the subpopulation into 33 folds of equal size. For a smaller number of folds, each fold represents a larger pilot study. All methods improve when the pilot study is increased substantially in size, i.e. when there is more information in the pilot. We find that the choice of 33 folds creates a challenging scenario with a small amount of pilot information. Nonetheless, both our method and the harmonic jackknife still perform well in these conditions. We consider each fold in turn as data from the pilot study and treat the remaining data (i.e., the data not in this fold) as the follow-up. We follow Zou et al. [2016] to make a visual summary of our results; namely, we plot the mean number of variants across all folds as a function of dataset size in the pilot, and we plot the mean number of total predicted variants (across both pilot and follow-up) as a function of dataset size in the same plot. A vertical dashed line marks the pilot size. Shaded regions indicate one empirical standard deviation, measured across the folds. We include the exact values in the plot for comparison. Figure 13 demonstrates that our method matches the exact value more closely than the parametric Bayesian approach [Ionita-Laza et al., 2009], the linear programming approach [Zou et al., 2016] and the nonparametric smoothed Good-Toulmin method [Orlitsky et al., 2016, Chakraborty et al., 2019]. And our method has roughly the same performance as the jackknife approach [Gravel, 2014] when the pilot and follow-up have the same experimental conditions. We next explain the relative performance of all methods in more detail.

In Appendix H we run both the Bayesian parametric approach and our method on data simulated under the parametric Bayesian model used by Ionita-Laza et al. [2009]. We also run both methods on data simulated under the 3-parameter beta process model we propose above; see Appendix H.1. The approach of Ionita-Laza et al. [2009] provides excellent predictions when the data is generated under their assumed model, but

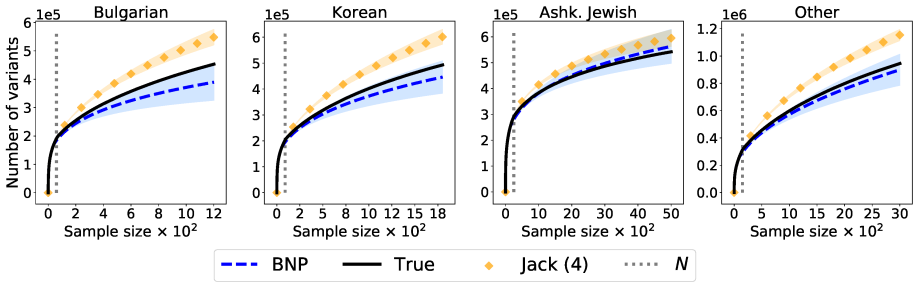

Figure 14: Predicting the number of new variants under different experimental conditions between the pilot and follow-up. Same four subpopulations (gnomAD). Pilot sequencing quality is $\lambda_{\text{pilot}} = 45$. Follow-up sequencing quality is $\lambda_{\text{follow}} = 32$. Horizontal axis is the number of samples. Vertical axis is the number of total observed variants across both pilot and follow-up. The threshold is $T = 30$.

deteriorates in performance under the simulated 3-parameter beta process data, as for real data. Therefore, we believe the parametric Bayesian method suffers on real data due to the assumed Ionita-Laza et al. [2009] model being ill adapted for real-life power laws. Similarly, we show in Appendix H.5 that the smoothed Good-Toulmin predictor [Orlitsky et al., 2016, Chakraborty et al., 2019] performs well for power laws with low exponent values, as expected. However, we also see that this estimator performs poorly for power laws with high exponent values. We verify in Appendix H.5 that the gnomAD data exhibits high exponent values. This behavior explains the underperformance of the smoothed Good-Toulmin predictor in our real-data experiments.

Zou et al. [2016] use a linear program to estimate *rare* variant frequencies; they approximate frequencies of common variants with the empirical frequency. "Rare" is defined to be any frequency less than $\kappa/100$, for some user-defined threshold $\kappa \in (0, 100)$, interpreted as a percent. In practice, we found that the output of the algorithm is very sensitive to the choice of κ (see Appendix H.3). The authors suggest $\kappa = 1$ as a default setting, but we observed numerical instability and poor predictive performance for this choice. This observation holds especially when the pilot size N is small, which we believe to be a particular case of interest in designing experiments for further data collection (i.e., for the follow-up study). For instance, we expect the small-N case to arise frequently in the study of non-model organisms [Russell et al., 2017]. In Figure 13, we chose $\kappa = 20$, which led to convergence of the optimization algorithm in all cases. We explore other values of κ in Appendix H.3.

G.3 Prediction under different experimental conditions

We now turn to the case where there may be sequencing errors in the pilot, in the follow-up, or both. And the sequencing quality may differ between the pilot and the follow-up. No existing method works in this case. Since, like our method, the method of Ionita-Laza et al. [2009] is Bayesian, we believe it could be straightforwardly adapted using similar ideas to the ones we present here. But we have already

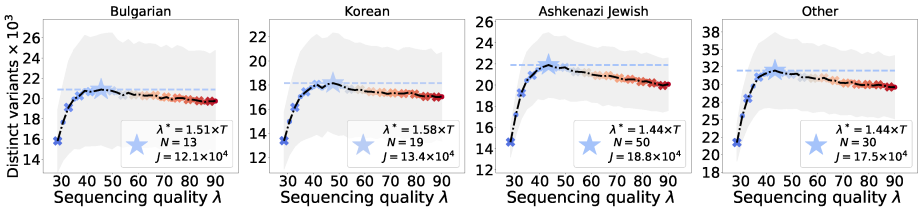

Figure 15: Designing an experiment to maximize the number of new variants in a follow-up study. Same four subpopulations (gnomAD). Horizontal axis is the follow-up sequencing quality λ_{follow}. Vertical axis is the predicted number of observed variants in the follow-up by maximizing M under the budget C and quality λ_{follow}.

seen that our Bayesian nonparametric method provides much more accurate predictions in the case of no sequencing errors, so we believe it is more fruitful to develop the Bayesian nonparametric approach. Similarly, we believe the linear programming approach [Zou et al., 2016] might be adapted to the special case where the follow-up is more error prone than the pilot. But even this development would still leave open the case where the follow-up might be made more accurate by increasing sequencing quality, and we have already observed that the Bayesian nonparametric approach provides better, more automatic predictions in the case of error-free observation. Finally, while the jackknife approach performs very well in the case with no sequencing errors, we do not think it will be as straightforward to adapt to the case where sequencing quality may change between the pilot and follow-up.

In Figure 14 we see that there is indeed a noticeable difference in the number of observed variants when the experimental conditions change between the pilot and follow-up. In particular, we consider a pilot sequencing quality $\lambda_{\text{pilot}} = 45$ and a follow-up sequencing quality $\lambda_{\text{follow}} = 32$. We use a fixed threshold $T = 30$, a standard coverage value in human genomic experiments [Karczewski et al., 2020]. To represent this change between studies, we use the gnomAD data as in Appendix G.2 but apply additional thinning to simulate imperfect observation due to sequencing depth; see Appendix G.1 for additional details. Since the jackknife is not able to use information about the changing sequencing depth, we expect our Bayesian nonparametric method should deliver superior predictive performance when sequencing quality changes. This behavior is exactly what we see in Figure 14.

G.4 Designing experiments to maximize the number of observed variants

Finally, we demonstrate that our Bayesian nonparametric predictor can be used for experimental design in practice. Our procedure consists of three steps. (1) Given the pilot data and sequencing quality λ_{pilot}, we minimize Equation (6) to estimate the parameters c, σ, α. (2) Next, we consider a range of values of the follow-up sequencing quality λ_{follow}. For each λ_{follow}, we choose the maximum follow-up size M that stays within our budget C. And we use the learned values of the parameters c, σ, α to predict

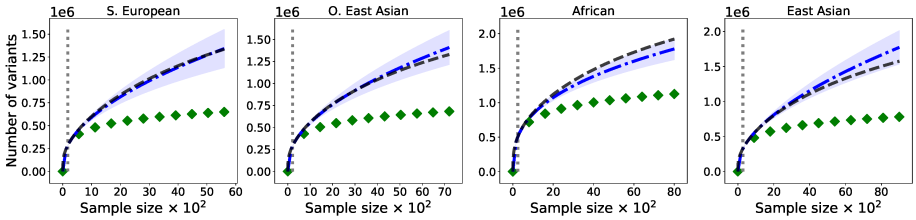

Figure 16: The blue dotted line is the posterior predictive mean of the number of distinct variants $U_N^{(m)}$ observed according to the Bayesian nonparametric predictor, averaged across 33 samples of size N. The green diamonds report the posterior predictive mean of the Bayesian parametric estimator of Ionita-Laza et al. [2009], averaged across the same subsets of the original data. The shaded blue and green regions report the prediction error by covering one standard empirical deviation for the two predictors.

the number of new variants in each case. (3) We choose the settings of λ_{follow} and M that maximize the number of new variants.

We illustrate this procedure in Figure 15.

In our experiments, we set the cost function $f(M, \lambda_{\text{follow}}) = M \log(\lambda_{\text{follow}})$ as in Ionita-Laza and Laird [2010], budget $D = 3000$, threshold $T = 30$, error $p_{\text{err}} = 0.01$, and $\lambda_{\text{pilot}} = 40$. We run the procedure over all folds, plot the empirical mean line, and plot the shaded region to illustrate one standard deviation. We see a trade-off in quality and quantity in Figure 15. Namely, maximizing quantity M leads to very small values of λ_{follow} to maintain the budget C. With sufficiently low quality, though, fewer variants are discovered. Conversely, when λ_{follow} is set very high, we require a very small M to maintain the budget C, and not many variants are discovered. Intermediate values of λ_{follow} and M serve to maximize the number of variants discovered under a fixed budget.

In Figure 16, Figure 17 and Figure 18 we report results of the prediction of the number of new variants on some sub populations of the gnomAD dataset. We consider the Bulgarian, South Korean, Other East Asian, African and East Asian subpopulations. The x-axis displays the total number of samples collected. On the y-axis, we plot the number of distinct genomic variants. The solid black line displays the true number of distinct variants, the vertical grey line is placed in correspondence of the training sample size N (left: $N \in \{42, 61, 228, 257, 291\}$).

H Additional experimental results: results on synthetic data

H.1 Synthetic data from the Indian buffet process

In this section, we provide experimental results for data drawn from the three parameters Indian buffet process. When the data is drawn from the true model, we expect the Bayesian nonparametric estimators of Section 3 to work particularly well. We test

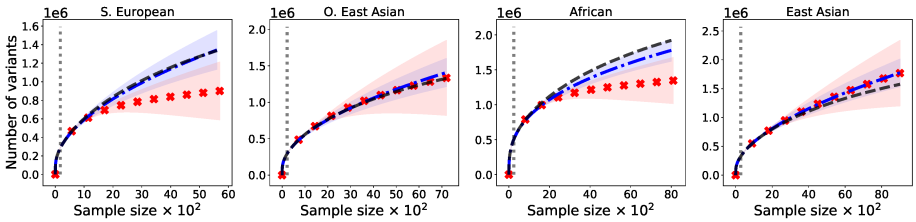

Figure 17: Results of the estimation of the number of new variants on some sub populations of the gnomAD dataset. The x-axis displays the total number of samples collected. On the y-axis, we plot the number of distinct genomic variants observed. The solid black line keeps track of the true number of distinct variants, the vertical grey line is placed in correspondence of the training sample size N. The blue dotted line is the posterior predictive mean of the number of distinct variants $U_N^{(m)}$ observed according to the Bayesian nonparametric predictor, averaged across 33 samples of size N. The dotted red line is the empirical mean of the UnseenEST estimator of Zou et al. [2016] across the same samples. The shaded blue and red regions report the prediction error by covering one standard empirical deviation for the two predictors. Here, we fix $\kappa = 1\%$, the value considered in Zou et al. [2016].

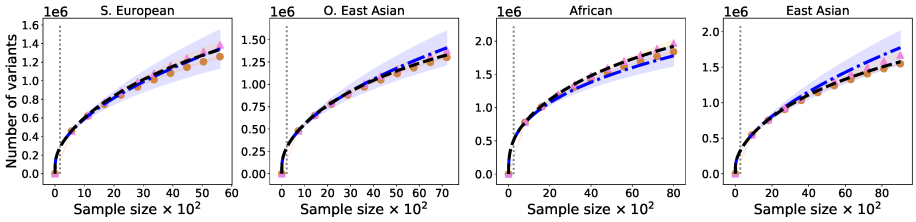

Figure 18: Again for the same sub populations considered in Figure 16 and Figure 17, we compare the Bayesian nonparametric estimator to the Jackknife estimator proposed in Gravel [2014], for the third and fourth orders. Lower order consistently underestimate the number of distinct variants

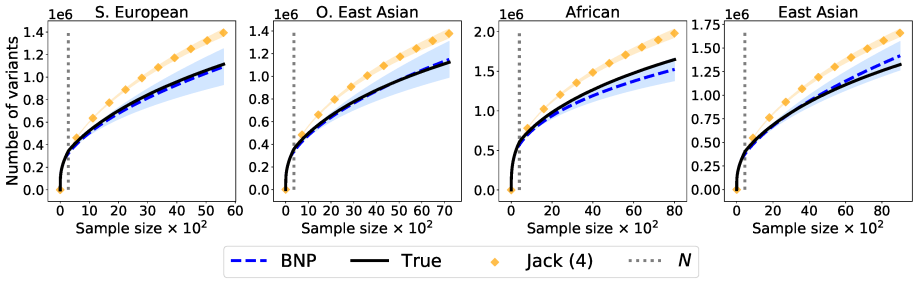

Figure 19: Prediction of the number of yet unseen variants for five subpopulations (gnomAD). For each subpopulation, we assume to have access to a small number of high quality genetic samples N in which variation is observed at J distinct loci, collected at initial sequencing depth $\lambda_{\text{pilot}} = 45$ and threshold $T = 30$. We imagine that the follow-up is performed at a different sequencing depth, $\lambda_{\text{follow}} = 32$.

against a large collection of parameters $\alpha > 0$, $\sigma \in [0, 1)$ and $c > -\sigma$. We report here results for different configurations. In all cases, the optimization procedure outlined in Section 6 recovers the rate of growth of the distinct variants. Interestingly, in some instances, the optimization recovers parameters that differ from the true parameters that generate the process, but still have good predictive performance (see Figure 20).

We also tested the performance of the predictor $P_N^{(M,r)}$ for the number of new variants that are going to appear a given number r of times as the initial sample of size N is enlarged with M additional observations. We found the performance of the predictor, in this case, to be very sensitive to the value of σ. In particular, while we expect the estimator to be exact as the extrapolation size M diverges, we observe that when σ is close to 0, the performance degrades for small values of r.

H.2 Synthetic data from the beta-Bernoulli model

Ionita-Laza et al. [2009] under the true model: We first consider the case in which the variants frequencies $\theta_1, \ldots, \theta_K$ are independently and identically distributed draws from a beta distribution, i.e. for some parameters $\alpha > 0$ and $\beta > 0$, independently and identically distributed across $j = 1, \ldots, K$ it holds

$$\theta_j \sim f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbf{1}_{(0,1)}(\theta). \tag{H.1}$$

Conditionally on the variants $\theta_1, \ldots, \theta_K$, each observation X_n is a binary vector of independent Bernoulli random variables, $x_{n,j} \mid \theta_j \sim \text{Bernoulli}(\theta_j)$. This is exactly the model considered by Ionita-Laza et al. [2009]. Therefore we are not surprised to verify in Figure 24 that the predictor derived in Appendix B outperforms the Bayesian nonparametric counterpart when the variants comes from the model of Equation (H.1).

Ionita-Laza et al. [2009] under misspecification: the case of power laws: Here we consider the case in which the variants frequencies $\theta_1, \ldots, \theta_K$ are independently

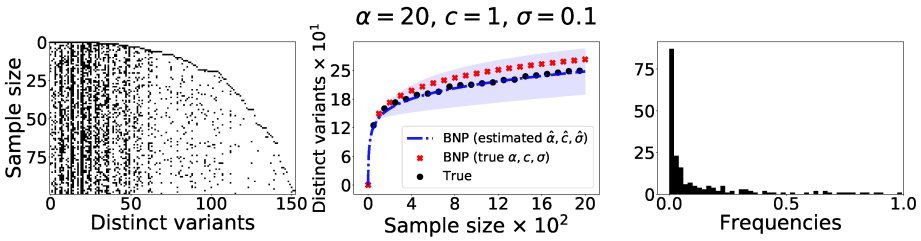

Figure 20: A draw from a three-parameter Indian buffet process. Here, $\alpha = 20$, $c = 1$, $\sigma = 0.1$. In the left panel, we see the binary matrix $\boldsymbol{X}$ containing the first $N = 100$ samples (x-axis) from the process, in its left-ordered-form (lof), i.e. variants (y-axis) are sorted by the order of appearance, so that as more points are added to the dataset, more columns contain nonzero entries. In the central panel, we plot the number of distinct variants (y-axis) as a function of the sample size (x-axis), extrapolating up to $M = 1900$ additional samples. Last, on the right panel, we plot the empirical distribution of frequencies among the first N samples.

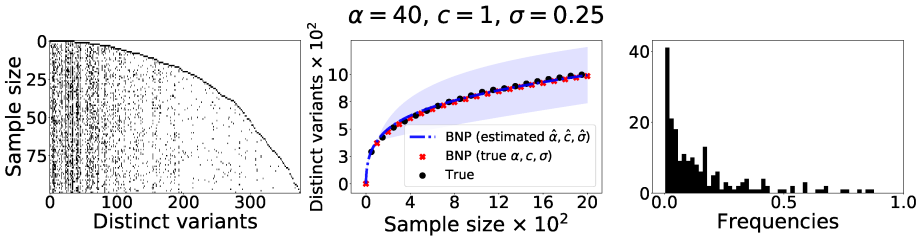

Figure 21: In this figure, we reproduce the visualizations explained in Figure 20 for a draw from a three-parameter Indian buffet process with parameters $\alpha = 40$, $c = 1$ and $\sigma = 0.25$

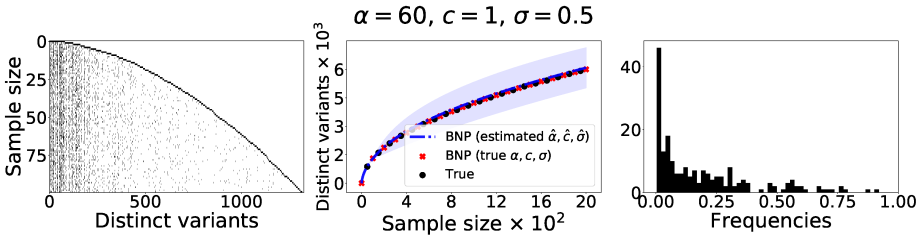

Figure 22: In this figure, we reproduce the visualizations explained in Figure 20 for a draw from a three-parameter Indian buffet process with parameters $\alpha = 60$, $c = 1$ and $\sigma = 0.5$

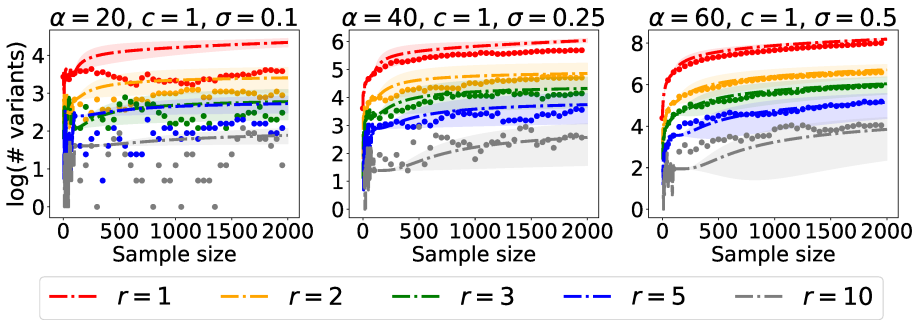

Figure 23: In this figure, for the three configurations of parameters α, c, σ considered in Figure 20, Figure 21 and Figure 22, we plot the performance of the estimators $P_N^{(M,r)}$ for $M = 1, \ldots, 1900$ and $r = 1, 2, 3, 5, 10$. Dotted line show the performance of the estimators, while points show the true values of the process.

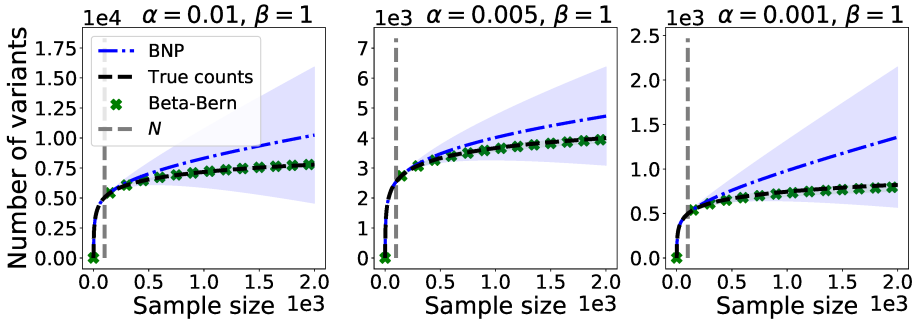

Figure 24: Performance of the beta-Bernoulli predictor (green crosses) proposed by Ionita-Laza et al. [2009] and of the nonparametric Bayesian predictor (dotted blue line) on three different datasets (each panel represents a different dataset). Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimension $K = 10^4$. The K coordinates are independently and identically distributed draws from a beta distribution. Conditionally on $\boldsymbol{\theta}$, we draw a random matrix $\boldsymbol{X}$ with $N = 2000$ rows and K columns. The (n, j)-th entry $x_{n,j}$ is Bernoulli distributed with mean θ_j, so that the columns of $\boldsymbol{X}$ are independently and identically distributed. We retain the first $N = 200$ rows as training set and obtain the two estimators. We project up to $N + M = 2000$ observations. We repeat the procedure over ten distinct folds of the data of the same size N to produce estimates of the prediction error. This estimate of the error is displayed by plotting one empirical standard deviation across the ten predicted values across the different folds, for each extrapolation value $\ell = 201, \ldots, 2000$. From left to right, we vary the first shape parameter of the beta distribution $\alpha \in \{10^{-1}, 10^{-2}, 10^{-2} \times 2^{-1}\}$, driving the mean of the distribution to zero, while keeping the second parameter $\beta = 1$ fixed.

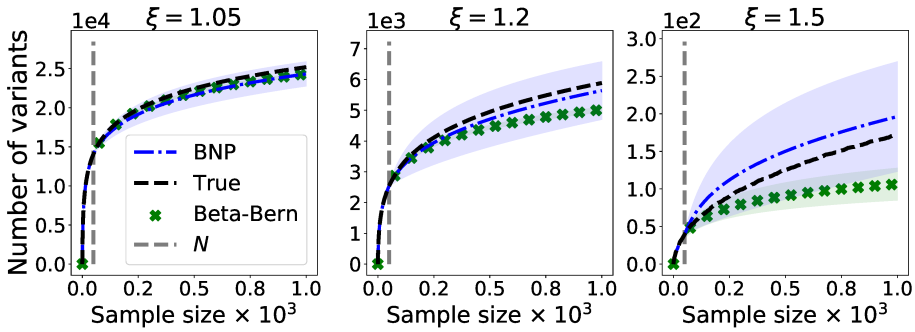

Figure 25: Performance of the beta-Bernoulli predictor (green solid line) proposed by Ionita-Laza et al. [2009] versus the nonparametric Bayesian predictor (dotted blue line) on three different datasets (each panel represents a different dataset). Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimensions $K = 10^4$. The K coordinates are independently and identically distributed draws from a power law distribution as described in Equation (H.2). Conditionally on $\boldsymbol{\theta}$, we draw a random matrix $\boldsymbol{X}$ with $N = 1000$ rows and K columns. The (n, j)-th entry $x_{n,j}$ is Bernoulli distributed with mean θ_j, so that the columns of $\boldsymbol{X}$ are independently and identically distributed. We retain the first $N = 50$ rows as training set and obtain the two predictors. We project up to $N + M = 1000$ observations. We repeat the procedure over ten folds of the same data of same size $N = 50$. We repeat the procedure over ten folds of the same data to produce estimates of the prediction error. This estimate of the error is displayed by plotting one empirical standard deviation across the ten predicted values across the different folds, for each extrapolation value $\ell = 51, \ldots, 1000$. From left to right, we vary the exponent of the power law distribution (left, $\xi = 1.05$, center, $\xi = 1.2$, right $\xi = 1.5$).

and identically distributed draws from a power law distribution, i.e. for some tail exponent $\xi \geq 0$

$$\theta_j \sim f(\theta) \propto \theta^{-\xi} \mathbf{1}_{(0,1)}(\theta). \tag{H.2}$$

The parameter ξ controls the left tail of the distribution: for $\xi = 0$, the distribution is uniform over the support $[0, 1]$. The larger the value of ξ, the more mass we put over rare frequencies. Power laws arise in a vast number of natural phenomena, including ecology, biology, physical and social sciences [Clauset et al., 2009]. Therefore, having an estimator that is effective when frequencies exhibit a power law behavior is desirable for virtually any applied scenario. In our experiments, the Bayesian parametric approach works well for moderate exponents, i.e. when the power law behavior is relatively mild. However, as soon as the exponent ξ becomes large, the parametric model fails to deliver consistent results (see Figure 25). Conversely, the Bayesian nonparametric estimator performs reasonably well.

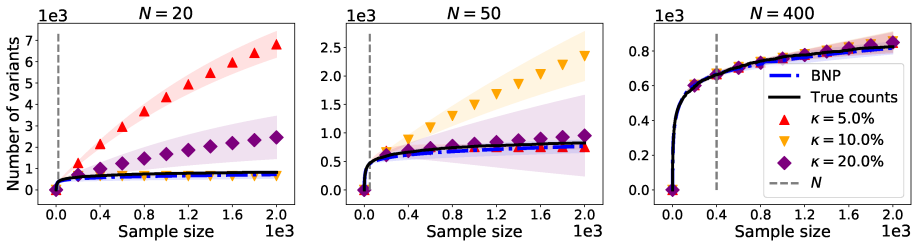

Figure 26: Comparison of the Bayesian nonparametric estimator (blue dotted line) to the frequentist nonparametric estimator proposed by Zou et al. [2016]. We generate synthetic datasets as follows: we first draw a random vector $\boldsymbol{\theta}$ of $K = 10^4$ independently and identically distributed beta random variables with parameters $\alpha = 0.001$ and $\beta = 1$. Conditionally on $\boldsymbol{\theta}$, we draw a random matrix $\boldsymbol{X}$ with $N = 2000$ rows and K columns. In each subplot, we retain a different fraction of rows of $\boldsymbol{X}$ to be used as training set (from left to right, $N \in \{20, 50, 400\}$). For each value of N, we compute the Bayesian nonparametric estimator, as well as the frequentist nonparametric estimator, varying the threshold parameter $\kappa \in \{5\%, 10\%, 20\%\}$ (red $(+)$, orange $(\star)$, purple $(\diamond)$) respectively. We highlight how the performance of the frequentist nonparametric estimator, especially when N is small, highly depends on the choice of κ, in an counterintuitve and somewhat unpredictable way. For example, when $N = 20$, choosing $\kappa = 10\%$ provides much better results than $\kappa = 5\%$ or $\kappa = 20\%$. However, for $N = 50$, both $\kappa = 5\%$ and $\kappa = 20\%$ perform much better than $\kappa = 10\%$. As N increases, the performance of the nonparametric frequentist estimator stabilizes, and becomes less sensitive to the choice of the parameter κ.

H.3 The choice of the hyperparameter κ for the frequentist nonparametric estimator proposed by Zou et al. [2016]

Choosing the parameter κ is particularly challenging when the sample size N is small relative to the total number of frequencies - as in the genomics application we consider. As a general principle, in order to avoid numerical instability, the input size has to be sufficiently large. For example, given a sample of $N = 100$ observations, if one sets $\kappa = 1$, the algorithm will take as an input only the number of variants which have been observed once. This will typically lead to numerical instability, which will not arise for larger values of κ (e.g. $\kappa \geq 10$). A general rule of thumb one could follow is to decrease κ as a function of the training sample size N: the larger N, the smaller κ. While this intuition seems to work on some instances, we found cases in which unpredictable behaviors can affect the quality of the predictions (see Figure 26 and Figure 27).

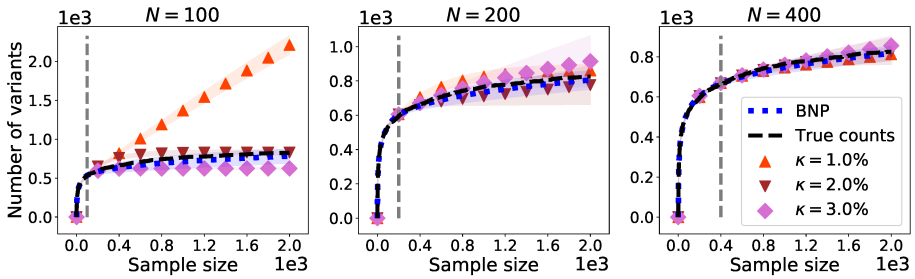

Figure 27: Comparison of the Bayesian nonparametric estimator (blue dotted line) to the frequentist nonparametric estimator of Zou et al. [2016]. We use the same data showed in Figure 26 but using much smaller values of $\kappa \in \{1\%, 2\%, 3\%\}$. Trying to run the linear program for these values of κ and $N < 100$ causes issues in the optimization routine, and therefore we only test it for N sufficiently large. We notice that for both $N = 100$ and $N = 200$, the suggested value of $\kappa = 1\%$ provides worse results than choosing a larger value of κ, whereas for $N = 400$, the performance of the estimator becomes less sensitive to the choice of κ.

H.4 Bias variance trade-off for the Jackknife estimator and optimal choice of the order p

As discussed in Burnham and Overton [1978], Gravel et al. [2011], Gravel [2014], and briefly in Appendix D.3, the prediction quality of jackknife estimators crucially depends on the *order* chosen. Lower orders can suffer from large bias, but have small variance, while higher orders incur in less bias at the cost of higher variance. On different datasets, the accuracy of different orders can vary dramatically. In this section we provide some experimental results (see Figure 28) to illustrate this trade-off.

In this section we provide some experimental results (see Figure 28) to illustrate this trade-off.

H.5 Analysis of the Good-Toulmin estimator and the case of power laws

Last, we performed synthetic experiments to understand the behavior of the smoothed Good-Toulmin estimator proposed by Orlitsky et al. [2016] and recently used by Chakraborty et al. [2019]. In our experiments, we considered differed regimes for the data generating process and found that the estimator performs very well when the distribution over variants' frequencies does not put too much mass on very small variants. This holds true even for moderate and small sample size N (see Figure 29) and well beyond the $M = N \log N$ extrapolation limit. However, when the vast majority of variants' are very rare, the estimator struggles to produce reliable results (see Figure 30). In all our experiments, we consider the two different smoothing choices suggested in Chakraborty et al. [2019] (GT 1 corresponds to Equation (E.3) and GT 2 corresponds to Equation (E.4) in Appendix E).

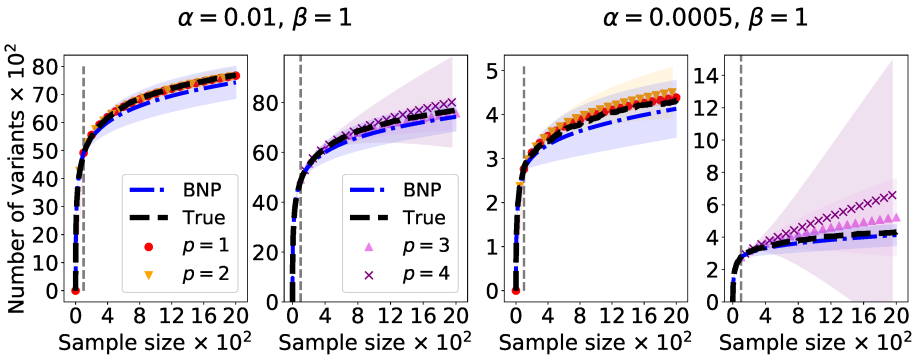

Figure 28: Comparison of the Bayesian nonparametric estimator (blue dotted line) to the jackknife estimator of Gravel [2014] for different choices of the order p. We generate two datasets as follows: for $\alpha \in \{0.01, 0.0005\}$ and $\beta = 1$, we generate two sets of $K = 10^4$ independently and identically distributed beta distributed draws $\boldsymbol{\theta}$ with parameters α, β. We then draw a random matrix $\boldsymbol{X}$ with $N = 2000$ rows, in which each entry $x_{n,j}$ is Bernoulli distributed with mean θ_j. We retain $N = 100$ rows for training. The two left panels show results for the dataset obtained when $\alpha = 0.01, \beta = 1$ across different choices of the jackknife order p. The two right panels show the same results for the dataset obtained when $\alpha = 0.0005$. Lower order jackknife estimators perform extremely well, and have little variance, while higher order jackknife estimators have worse performance, and higher variance. Such behavior worsens as α gets smaller, i.e. when the mean of the beta draws approach 0.

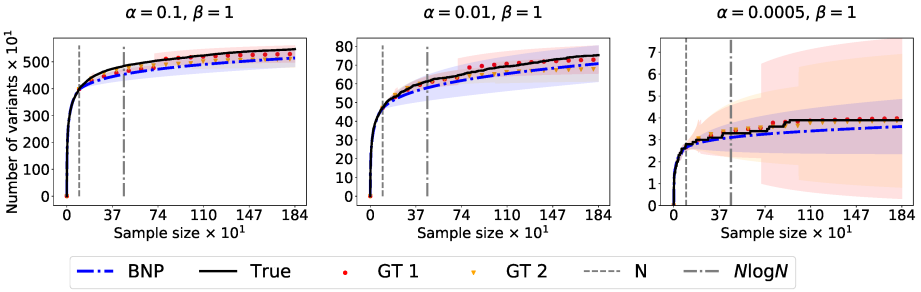

Figure 29: Prediction of the number of yet unseen variants for three synthetic datasets. Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimension $K = 10^4$. The K coordinates are independently and identically distributed draws from a beta distribution. Conditionally on $\boldsymbol{\theta}$, we draw a random matrix $\boldsymbol{X}$ with $N = 2000$ rows and K columns. The (n, j)-th entry $x_{n,j}$ is Bernoulli distributed with mean θ_j, so that the columns of $\boldsymbol{X}$ are independently and identically distributed. We retain the first $N = 400$ rows as training set and obtain the two estimators. We project up to $N + M = 2000$ observations. We repeat the procedure over ten distinct folds of the data of the same size N to produce estimates of the prediction error. This estimate of the error is displayed by plotting one empirical standard deviation across the ten predicted values across the different folds, for each extrapolation value $\ell = 401, \ldots, 2000$. From left to right, we vary the first shape parameter of the beta distribution $\alpha \in \{10^{-1}, 10^{-2}, 10^{-2} \times 2^{-1}\}$, driving the mean of the distribution to zero, while keeping the second parameter $\beta = 1$ fixed.

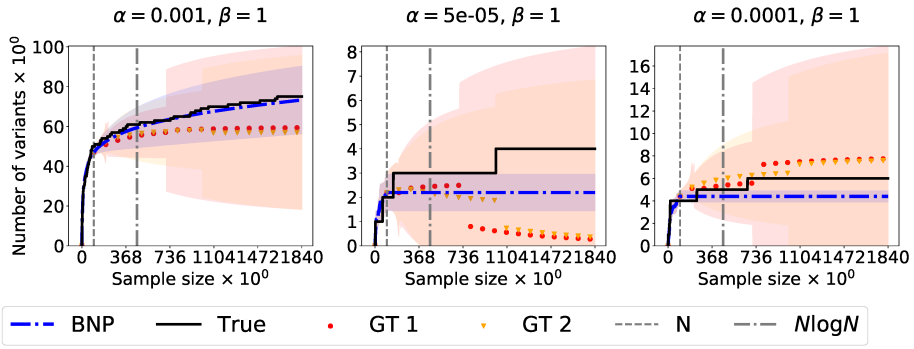

Figure 30: Prediction of the number of yet unseen variants for three synthetic datasets. Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimension $K = 10^4$. The K coordinates are independently and identically distributed draws from a beta distribution. Conditionally on $\boldsymbol{\theta}$, we draw a random matrix $\boldsymbol{X}$ with $N = 2000$ rows and K columns. The (n, j)-th entry $x_{n,j}$ is Bernoulli distributed with mean θ_j, so that the columns of $\boldsymbol{X}$ are independently and identically distributed. We retain the first $N = 400$ rows as training set and obtain the two estimators. We project up to $N+M = 2000$ observations. We repeat the procedure over ten distinct folds of the data of the same size N to produce estimates of the prediction error. This estimate of the error is displayed by plotting one empirical standard deviation across the ten predicted values across the different folds, for each extrapolation value $\ell = 401, \ldots, 2000$. From left to right, we vary the first shape parameter of the beta distribution $\alpha \in \{10^{-3}, 5 \times 10^{-5}, 10^{-4}\}$, driving the mean of the distribution to zero, while keeping the second parameter $\beta = 1$ fixed.

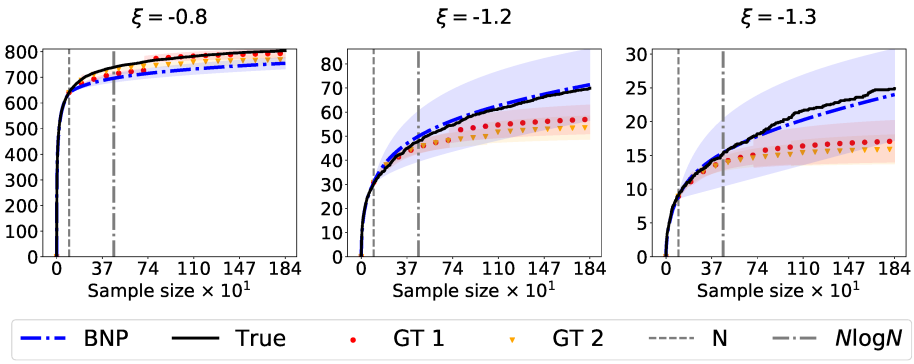

Figure 31: Performance of the smoothed Good-Toulmin predictor proposed by Orlitsky et al. [2016] versus the nonparametric Bayesian predictor (dotted blue line) on three different datasets (each panel represents a different dataset). Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimensions $K = 10^4$. The K coordinates are independently and identically distributed drawn from a power law distribution as described in Equation (H.2). Conditionally on $\boldsymbol{\theta}$, we draw a random matrix $\boldsymbol{X}$ with $N = 2000$ rows and K columns. The (n, j)-th entry $x_{n,j}$ is Bernoulli distributed with mean θ_j, so that the columns of $\boldsymbol{X}$ are independently and identically distributed. We retain the first $N = 400$ rows as training set and obtain the two predictors. We project up to $N + M = 2000$ observations. We repeat the procedure over ten folds of the same data of same size $N = 400$. We repeat the procedure over ten folds of the data to produce estimates of the prediction error. This estimate of the error is displayed by plotting one empirical standard deviation across the ten predicted values across the different folds, for each extrapolation value $\ell = 401, \ldots, 2000$. From left to right, we vary the exponent of the power law distribution (left, $\xi = 1.05$, center, $\xi = 1.2$, right $\xi = 1.3$).

We also considered the case in which the variants follow a power law distribution. In this case, we find that when the exponent of the power law is not too large (in absolute value), then the estimator performs well. However, when the absolute value of the exponent satisfies $|\xi| > 1$, the estimator systematically underestimates the number of new variants to be seen, in the same way observed for real data (see Figure 31). Importantly, we verify that the empirical distribution of variants' frequencies across the datasets considered is indeed explained by power laws with exponent $|\xi| > 1$.

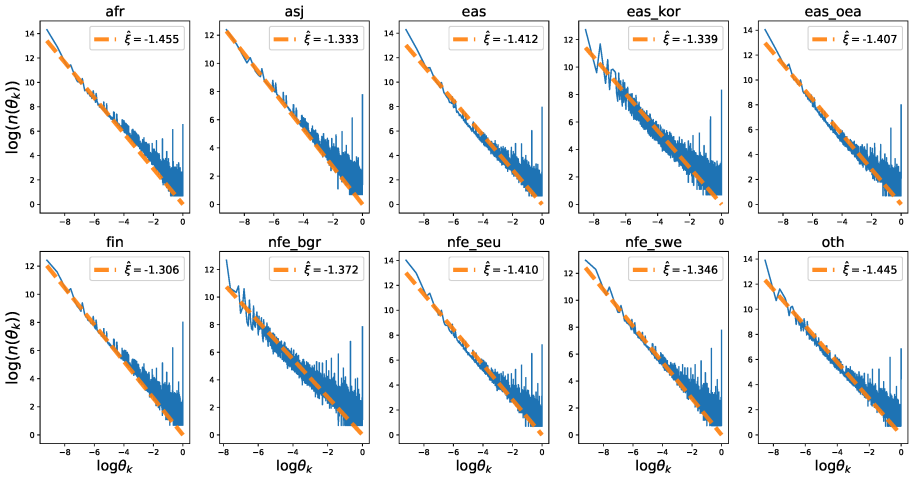

Figure 32: Fitting a power-law distribution to the empirical variants' frequencies of the gnomAD dataset. Following the method outlined in Goldstein et al. [2004], we fit a linear regression to the log-log plot of the binned empirical variants' frequencies to determine the exponent of the power law distribution. We only consider the 20 rarest frequencies, as suggested in Goldstein et al. [2004]. In all the datasets considered, we find that the estimate of ξ is larger than 1, the regime in which the smoothed Good-Toulmin provides systematic underestimation even on synthetic data.