Genomic variety prediction via Bayesian nonparametrics

Lorenzo Masoero LOM@MIT.EDU
Federico Camerlenghi FEDERICO.CAMERLENGHI@UNIMIB.IT
Stefano Favaro STEFANO.FAVARO@UNITO.IT
Tamara Broderick TBRODERICK@CSAIL.MIT.EDU

Abstract

Despite the advent of Big Data, data-gathering in many domains can still be an expensive process that necessitates careful planning. For instance, in genomics, researchers can spend money and time to sequence a greater number of individual genomes – or alternatively they can spend these resources to sequence individual genomes with increased accuracy. In either case, spending resources has the potential to reveal new variations in the genome and thereby new genetic insights. We consider the case where scientists have already conducted a pilot study to reveal some variants in a genome and are contemplating a follow-up study. We provide a novel prediction method, using Bayesian nonparametric [BNP] methods, for how many variants scientists can expect to find in the follow-up based on the information in the pilot. When sequencing accuracy is kept constant between the pilot and follow-up, we demonstrate on (real) data from the gnomAD project (Karczewski et al., 2019) that our prediction is more accurate than two recent proposals – and as accurate as a more classic proposal. Unlike other existing methods though, our method allows practitioners to change the sequencing accuracy between the pilot and the follow-up. We demonstrate how this distinction allows our method to be used both for more realistic predictions as well as for optimal experimental design of the follow-up study under a resource budget.

1. Data and Model

Modern high throughput sequencing technologies allow accurate determination of an organism's genome (Reuter et al., 2015). Typically for some population of interest, researchers define a *reference genome*, which serves as a fixed representative for this population. We say that a *variant* is observed wherever a sequenced genome differs from the reference genome. We consider the case where the variants in a sample of N individuals have already been observed in a pilot study, and we wish to predict the number of new, hitherto unseen variants that we will observe if we were to take a new, follow-up sample of size M.

Suppose there are J variants observed among the N genomes in the pilot study, with $0 \leq J < \infty$. Let $\{\psi_j\}_{j\geq 1}$ denote a collection of distinct labels. Each ψ_j distinguishes the jth variant among other possible variants. For ease of representation, we assume that variants are ordered as they appear in the sample; i.e., by *order of appearance*. Let $x_{nj} = 1$ if the variant with label ψ_j is observed for the nth organism; otherwise, let $x_{nj} = 0$. We collect all the variant information for the nth organism in the measure $X_n := \sum_{j=1}^{J} x_{nj}\delta_{\psi_j}$, which pairs each variant observation with the corresponding label by putting a mass of size x_{nj} at location ψ_j. We write $X_{N_1:N_2}$ to denote the set $\{X_{N_1}, X_{N_1+1}, X_{N_1+2}, \ldots, X_{N_2}\}$, $N_1 \leq N_2$.

We here take a Bayesian approach. To do so, we must choose an appropriate latent parameter Θ. We specify our generative model via a likelihood $\text{pr}(X_{1:N}|\Theta)$ and a prior $\text{pr}(\Theta)$. Bayes Theorem then yields the posterior $\text{pr}(\Theta|X_{1:N})$, which encodes what we know about the latent parameter after having observed N data points. The posterior predictive $\text{pr}(X_{N+1:N+M}|X_{1:N})$ follows. Finally, since the number of new variants in the follow-up study is a function of $X_{M:M+N}$ and $X_{1:N}$, we can compute its distribution conditional on $X_{1:N}$ from the posterior predictive. It thus remains to specify our model.

Technically there is a fixed, finite upper bound on the number of possible variants established by the (necessarily finite) size of any individual genome. But this bound is usually much larger than the number of observed variants. In practice, moreover, we typically 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 [BNP] methods allow us to avoid hard-coding an unwieldy, large finite bound that may cause computational and modeling headaches. In particular, BNP methods allow the observed number of variants to be finite for any finite data set and grow without bound – in such a way that computation typically scales closely with the actual number of variants observed. The mechanism by which BNP methods work is that we imagine a countable infinity of *latent*, unseen variants; thus, for any N, there are always more latent variants to draw on in future. To use these BNP methods, then, we provide a label to each of the latent infinity of variants: $\{\psi_j\}_{j=1}^{\infty}$. And we write $X_n := \sum_{j=1}^{\infty} x_{nj}\delta_{\psi_j}$; since $x_{nj} = 0$ for all of the unobserved variants, this equation can be considered equivalent to the previous definition of X_n above.

In reality, nearby positions on a genome can be highly correlated; this phenomenon is called *linkage disequilibrium*. We make the simplifying assumption that in fact every variant appears independently of every other variant; that is, $\{x_{nj}\}$ is independent of $\{x_{nk}\}$ across all n for $j \neq k$. This assumption makes inference computationally efficient and is common to other methods (Ionita-Laza et al., 2009; Gravel, 2014; Zou et al., 2016). We assume that $\{x_{nj}\}_{n=1}^{\infty}$ are (infinitely) exchangeable. In turn, this implies, by de Finetti (1931), the existence of a latent, random variant frequency θ_j such that the x_{nj} are Bernoulli draws with frequency θ_j, identically and independently distributed across n. We can collect the pairs of each θ_j together with its associated variant label ψ_j in a measure $\Theta := \sum_{j=1}^{\infty} \theta_j \delta_{\psi_j}$. We assume the X_n are independent and identical conditional on Θ. We say that X_1 given Θ is described by a *Bernoulli process* (BeP) with parameter Θ, and we write $X_n|\Theta \overset{iid}{\sim} \text{BeP}(\Theta)$. This final equation serves as the likelihood in our Bayesian model.

We model the variants' frequencies through the jumps of a three-parameter beta process. This prior guarantees that a finite number of variants is observed in any finite sample and that the number of observed variants is unbounded as the number of samples grows. We choose the *three-parameter beta process* (3BP) model for its ability to capture realistic power laws. In particular, its three parameters are (1) a *mass parameter* α that scales the total number of variants, (2) a *discount parameter* σ that controls the power law of the growth, and (3) a *concentration parameter* c that modulates the frequency of more widespread variants. See Teh and Gorur (2009) and the Appendix for more details.

2. Predicting the number of new variants

We proposed and justified a Bayesian model consisting of (1) a prior $\Theta \sim \mathrm{3BP}(\alpha, \sigma, c)$ over variant frequencies and (2) a likelihood $X_n \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\Theta)$ for observed variants conditioned on the variant frequencies. Because the Bayes decision rule that minimizes mean square error is the mean (Keener, 2011), we use the posterior predictive mean of the number of new variants as our predictor. In what follows let $U_N^{(M)}$ represent the number of new variants in a follow-up sample of size M after a preliminary study of size N, $U_N^{(M)} := \sum_{j=1}^{\infty} \mathbf{1}\left(\sum_{n=1}^{N} x_{n,j} = 0\right) \mathbf{1}\left(\sum_{m=1}^{M} x_{N+m,j} > 0\right)$.

Proposition 1 *Let $\Theta \sim \mathrm{3BP}(\alpha, \sigma, c)$, and let $X_n \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\Theta)$ for $n = 1, 2, \ldots$. Assume $\alpha > 0$, $c > -\sigma$ and $\sigma \in (0, 1)$. Then, for $(a)_{b\uparrow} := \Gamma(a + b)/\Gamma(a)$,*

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

In practice, sequencing a genome is complex and noisy; 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; Out of these, a number $D_{n,j,\mathrm{err}} \leq D_{n,j}$ reads give rise to an error, due to technological imperfections, and are discarded. The remaining $D_{n,j,\mathrm{noerr}} = D_{n,j} - D_{n,j,\mathrm{err}}$ reads are correctly processed, aligned to the reference genome, and recorded (Ionita-Laza and Laird, 2010). Let $C_{n,j} \leq D_{n,j,\mathrm{noerr}}$ denote the number of times that reads are correctly processed and we observe disagreement with the reference genome. A variant is "called" whenever some discrepancy criterion – the "variant calling rule" – is satisfied. Following Ionita-Laza and Laird (2010) we focus on simple threshold variant calling rules: given $T > 0$, variation is declared if the count $C_{n,j}$ exceeds T, i.e. $x_{n,j} = \mathbf{1}(C_{n,j} \geq T)$. We model the sequencing depth $D_{n,j}$ as a Poisson random variable with parameter $\lambda > 0$, i.i.d. across individuals and positions. The number of successful reads $D_{n,j,\mathrm{noerr}}$ is binomially distributed, with number of trials given by the sequencing depth $D_{n,j}$, and success probability given by $1 - p_{\mathrm{err}}$ – a fixed probability of reading error that depends on the sequencing technology. We write the probability, $\phi(\lambda, T, p_{\mathrm{err}})$, that at least T successful reads are obtained at any position j for any individual: $\phi(\lambda, T, p_{\mathrm{err}}) := \sum_{t=T}^{\infty} \frac{e^{-\lambda}\lambda^t}{t!} \sum_{i=T}^{t} \binom{t}{i} (1 - p_{\mathrm{err}})^i p_{\mathrm{err}}^{t-i}$. When planning the follow-up experiment under a fixed budget, there is a tradeoff in practice between sequencing depth λ and number of individuals sequenced, M. Let $\phi_{\mathrm{init}} = \phi_{\mathrm{init}}(\lambda_{\mathrm{init}}, T, p_{\mathrm{err}})$ and $\phi = \phi(\lambda_{\mathrm{follow}}, T, p_{\mathrm{err}})$ denote the value of $\phi(\lambda, T, p_{\mathrm{err}})$ for the pilot and the follow-up study respectively. We note that our prior over the variant frequencies has not changed: $\Theta \sim \mathrm{3BP}(\alpha, \sigma, c)$. But now we first draw whether organism n has variant with frequency θ_j according to Bernoulli(θ_j). If it does have the variant, we draw whether it is observed according to Bernoulli(ϕ_{init}) in the initial experiment and Bernoulli(ϕ_{follow}) in the follow-up.

Proposition 2 *Let $\Theta \sim \mathrm{3BP}(\alpha, \sigma, c)$. Let $X_n \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\phi_{\mathrm{init}}\Theta), n \in \{1, \ldots, N\}$. Let $X_{n+m} \mid \Theta \overset{iid}{\sim} \mathrm{BeP}(\phi_{\mathrm{follow}}\Theta), m \in \{1, \ldots, M\}$. Assume $\alpha > 0$, $c > -\sigma$ and $\sigma \in (0, 1)$.*

$$U_N^{(M)} \mid X_{1:N} \sim \mathrm{Poisson}\left(\hat{P}_N^{(M, \lambda_{\mathrm{follow}})}\right), \tag{2}$$

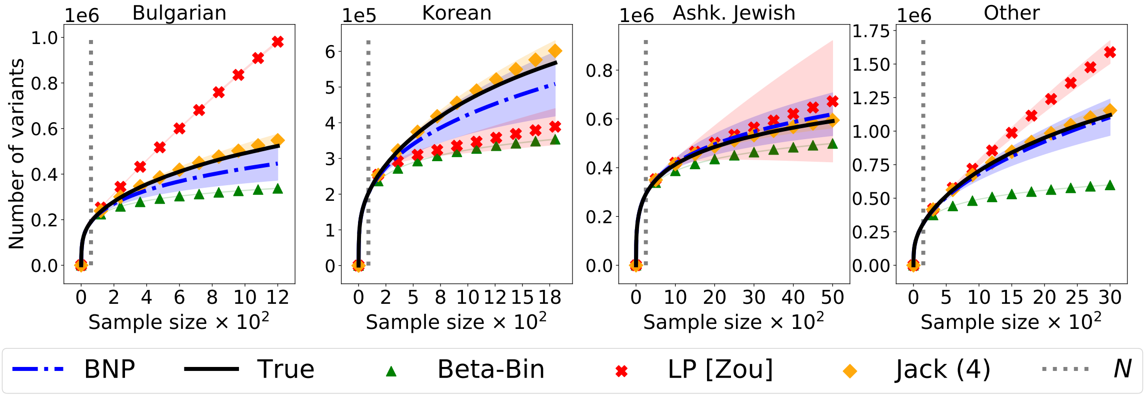

Figure 1: Prediction of the number of new variants. The dashed black line displays the true number of distinct variants (y-axis) as the sample size increases (x-axis). In every subpopulation, we divide the data in 33 samples of equal size N, and iteratively train our method (as well as competing methods), on one subset of the data. We report the mean prediction across subsets (blue: BNP, Eq. (1); green: Ionita-Laza et al. (2009); red: Zou et al. (2016); orange: Gravel (2014)). Shaded regions report an estimate of the prediction error, and cover one standard deviation.

with $\hat{P}_N^{(M,\lambda_{\text{follow}})} := \left(\alpha\phi_{\text{follow}} \sum_{m=1}^M \mathbb{E}_B\left\{(1-\phi_{\text{follow}}B)^{m-1}(1-\phi_{\text{init}}B)^N\right\}\right)$*, and where* $B \sim \text{Beta}(1-\sigma, c+\sigma)$*. Equation (2) can be used to optimally design follow-up experiments under a budget constraint, by finding*

$$(\lambda^\star, M^\star) \in \underset{M,\lambda_{\text{follow}}}{\arg\max}\, \hat{P}_N^{(M,\lambda_{\text{follow}})} \quad \textit{subject to} \quad f(M,\lambda_{\text{follow}}) \le D, \tag{3}$$

where D *is the available budget, and* $f(m,\lambda)$ *is a non-negative function, increasing in both arguments, encoding the cost of sampling* m *individual at sequencing depth* λ*.*

3. Experiments

We validate our methods on data from the Genome Aggregation Database (gnomAD) (Karczewski et al., 2019), a recent extension of the Exome Aggregation Consortium (ExAC) data set (Lek et al., 2016) and the largest publicly available human genomic dataset. GnomAD contains genetic information from 125,748 individual exome sequences recorded at 1,195,872 genomic loci. Samples in gnomAD are arranged into eight subpopulations according to geographic origin, where one of the eight subpopulations is a catch-all category called "Other". These subpopulations vary in size from 1,335 Bulgarian samples to 17,720 African-American samples. See Appendix for details.

In Figure 1, we show that our method's performance on pure prediction of new variants in a follow-up study is competitive with the state-of-the-art when the pilot and follow-up studies are collected under the same experimental conditions. We compare to three recent alternatives proposed in the literature (Ionita-Laza et al., 2009; Gravel, 2014; Zou

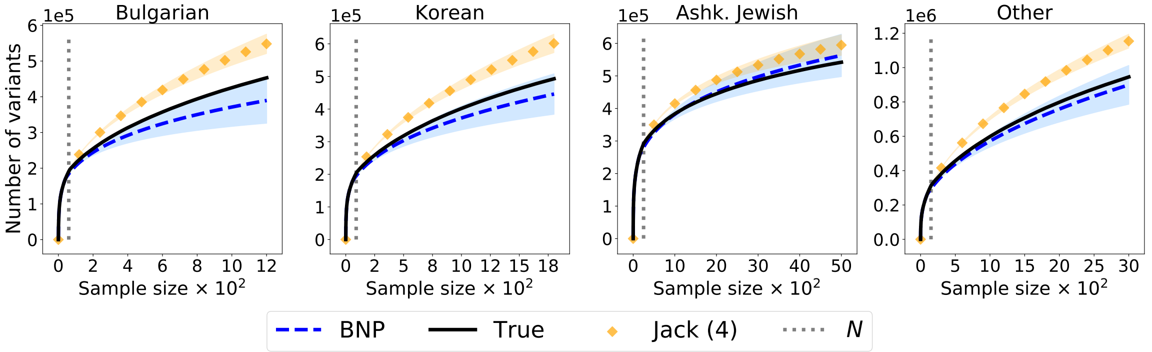

Figure 2: Predicting the number of new variants under different experimental conditions between the pilot and follow-up. Same four subpopulations (gnomAD). Pilot sequencing quality is $\lambda_{\text{init}} = 45$. Follow-up sequencing quality is $\lambda_{\text{follow}} = 32$. Horizontal axis is the number of samples. Vertical axis is the number of total observed variants across both pilot and follow-up. The threshold $T = 30$.

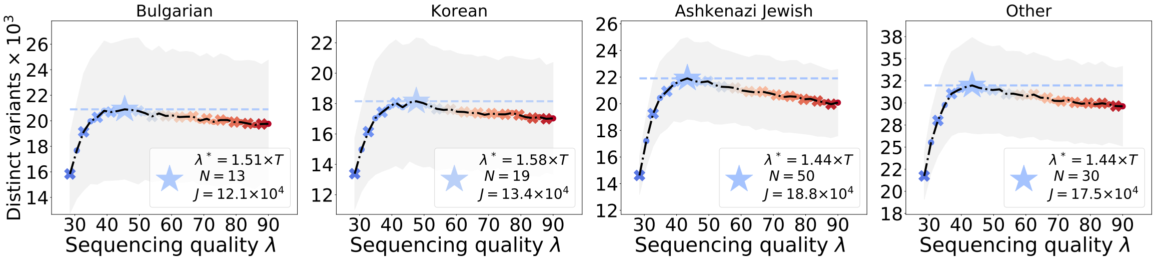

Figure 3: Optimal experimental design. For each subpopulation, we assume to have access to a pilot sample $X_{1:N}$ with $\lambda = 40$, $p_{\text{err}} = .01$. Sample sizes are as in Figure 1. We assume $\text{cost}(m, \lambda) = m \log \lambda$, $D = 3000$, $T = 30$. For every choice of feasible (λ, m), with $\lambda \in [.9 \times T, 3 \times T]$ (x-axis) we compute and plot the expected number of distinct variants that are going to be observed for the largest feasible extrapolation size M (y-axis). Shaded regions cover one standard deviation.

et al., 2016). In Figure 2 we show how, when experimental conditions change, the BNP method keeps on producing useful predictions, by adapting to the changing conditions, while competing methods fail to do so. Last, in Figure 3 we sketch how our predictors can be used to inform optimal experimental design of a follow up study, by leveraging the fomulation of Equation (3).

References

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

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

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

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

Trevor Campbell, Jonathan H Huggins, Jonathan P How, and Tamara Broderick. Truncated random measures. *Bernoulli*, In press.

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.

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

Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, and David L Altshuler. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

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

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

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019.

Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, and Beryl B Cummings. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536 (7616):285, 2016.

John Paisley, David Blei, and Michael Jordan. Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics*, pages 850–858, 2012.

Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

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

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

James Zou, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7:13293, 2016.

Appendix A. Additional details on the model

In Section 1 we introduced the beta-Bernoulli process model we use to describe the observed data. This is obtained by combining a 3BP process prior, with a collection of conditionally i.i.d. Bernoulli processes likelihoods. The parametric beta distribution prior forms a convenient conjugate prior to the parametric Bernoulli distribution likelihood. Analogously, the Bernoulli process likelihood has a convenient conjugate prior in the *beta process*. In the Bayesian nonparametric case, in contrast to the parametric case, conjugacy is all the more important due to the even greater potential difficulty of computation in infinite dimensions; in particular, conjugacy allows posterior computation via simple finite arithmetic operations, where the number of required operations is on the order of the (finite) number of observed variants. Recall that above we assumed independence of the $\{x_{nj}\}_{n=1}^{\infty}$ across j; this independence implies the independence of the θ_j across j. In agreement with this assumption, the beta process prior on Θ can be interpreted as a sequence of independent priors on the θ_j. In essence, the beta process prior provides control over these independent priors in a way that satisfies our goals: (A) a finite number of observed variants in any finite sample and (B) a number of observed variants that is unbounded as the number of samples grows. Since it is common for physical processes to exhibit power laws, we choose the *three-parameter beta process* (3BP) model for its ability to capture power laws. In particular, its three parameters are (1) a *mass parameter* α that scales the total number of variants observed, (2) a *discount parameter* σ that controls the power law of the growth in observed variant cardinality, and (3) a *concentration parameter* c that modulates the frequency of more widespread variants.

Next, we will detail two equivalent representations for drawing $\Theta \sim 3\mathrm{BP}(\alpha, \sigma, c)$. In what follows, our focus will be on generating the θ_j. For our purposes here, the ψ_j serve merely to distinguish the variants (with no other significance to the labels), so it is enough to ensure that they are all almost surely distinct. To that end, in what follows, we always take $\psi_j \overset{iid}{\sim} \mathrm{Unif}[0,1]$, independently and identically distributed across j. For our first representation, we can write $\Theta \sim 3\mathrm{BP}(\alpha, \sigma, c)$ if

$$\Theta = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} B_{i,j}^{(i)} \prod_{\ell=1}^{i-1} \left(1 - B_{i,j}^{(\ell)}\right) \delta_{\psi_{i,j}},$$

where $C_i \overset{iid}{\sim} \mathrm{Poisson}(\alpha)$ and $B_{i,j}^{(\ell)} \overset{indep}{\sim} \mathrm{Beta}(1-c, \sigma + \ell c)$, independently across i, j, and l. This size-biased representation (Broderick et al., 2012; Paisley et al., 2012; Campbell et al., In press) demonstrates that the three-parameter beta process can be interpreted as a sequence of independent frequencies. In mathematical manipulations, though, the Poisson process representation of the beta process is often much more convenient. We therefore note that the representation above is equivalent (Broderick et al., 2012; Paisley et al., 2012) to drawing the $\{\theta_j\}$ from a Poisson point process with rate measure

$$\nu(\mathrm{d}\theta) = \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta^{-1-\sigma} (1-\theta)^{c+\sigma-1} \mathbf{1}_{[0,1]}(\theta) \mathrm{d}\theta. \tag{4}$$

To ensure that goals (A) and (B) above are satisfied, we must have the following restrictions on the 3BP hyperparameters: $\alpha > 0$, $c > -\sigma$ and $\sigma \in (0,1)$ (Broderick et al., 2018).

Appendix B. Proofs

B.1. Proof of Proposition 1

Proof By construction, the variant frequencies $\{\theta_j\}$ are formed from a Poisson point process with rate measure ν given in Equation (4). 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 Poisson 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

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

as was to be shown. ■

B.2. Proof of Proposition 2

Proof We start by showing that $U_N^{(M)}$ is almost surely finite, for any choice of N, M. To see the almost sure finiteness of the Poisson parameter and hence of $U_N^{(M)}$, 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. The θ factor will arise from $\mathrm{Bernoulli}(1 \mid \phi_{\mathrm{follow}}\theta)$.

The exact form of the Poisson parameter in Equation (2) arises by following the same thinning argument as in the proof of Theorem 1. To see the beta representation,

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

■

Appendix C. Empirics for the prediction and prediction uncertainty

C.1. Choosing the beta process hyperparameters

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 of 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 predictor, as proved in Section 2. So we will induce a dependency on the observed pilot data by fitting these hyperparameter values to the pilot data. Our approach may be seen to fit into the framework of empirical Bayes.

One common option in empirical Bayes is to maximize the probability of the data given the hyperparameters:

$$\arg\max_{\alpha,\sigma,c} \mathrm{pr}(X_{1:N}|\alpha, \sigma, c),$$

with $\mathrm{pr}(X_{1:N}|\alpha, \sigma, c) = \int_{\Theta} \mathrm{pr}(X_{1:N}|\Theta)\mathrm{pr}(d\Theta|\alpha, \sigma, c)$. 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 EFPF 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 may suffer from numerical instability in this optimization problem.

A much easier choice is to treat the prediction from our model above 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 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. Similarly, by our definition in Section 2, $U_n^{(m)} \mid X_{1:N}$ is 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 the error minimization problem:

$$\hat{\alpha}, \hat{\sigma}, \hat{c} := \underset{\substack{\alpha,\sigma,c: \\ \alpha>0,\ \sigma\in[0,1), c>-\sigma}}{\arg\min} \sum_{m=1}^{N-n} \left(P_n^{(m)}(\alpha,\sigma,c) - \left(U_n^{(m)} \mid X_{1:N} \right) \right)^2 . \tag{5}$$

In practice, we choose $n = \lfloor 2/3 \times N \rfloor$, and we use numerical solvers to estimate $\hat{\alpha}, \hat{\sigma}, \hat{c}$.[1] We choose to minimize the 2-norm (sum of squared differences), but different norms could be employed. In our experiments, we also tested the 1-norm (sum of squared differences), but observed empirically a slightly worse performance. Finally, we use the quantity $\hat{P}_N^{(M)} := P_N^{(M)}(\hat{\alpha}, \hat{\sigma}, \hat{c})$ as our predictor for the number of new variants in the follow-up study of size M after the full pilot study of size N.

C.2. Accounting for overdispersion

We expect our real data to be *overdispersed* relative to our putative posterior predictive distribution Equation (1) since the Poisson distribution has variance constrained to be equal to its mean. In theory, a non-Poisson model would allow improved uncertainty quantification; in practice, the conjugacy of the Poisson is what led to our easy-to-use predictive formula in Equation (1). To enable a speedy procedure, we instead propose a bootstrapped measure of uncertainty for our predictor. In more detail, let $S > 1$ be the number of bootstrap resamples. For each $s \in [S]$, we draw a bootstrap resample, without replacement, of size $n_{\text{boot}} = N(1-(1-1/N)^N) \approx 0.6323 \times N$. We choose this value because n_{boot} is the expected number of *distinct* values that we would expect to sample in a bootstrap resample - with replacement - of size N. We call this bootrapped subset $\hat{X}^{(s)}_{1:n_{\text{boot}}}$. For each bootstrap resample, we resolve Equation (5) for the model hyperparameters, which in turn induce a new prediction for the number of new variants in the follow-up study of size M. We report uncertainties by calculating standard deviations across these predictions.

Appendix D. Description of competing methods

In the following subsections we provide additional details about the competing methods considered. All the methods 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. Moreover, all these methods make use of the site-frequency-spectrum (SFS) or fingerprint of the sample $X_{1:N}$. This is nothing but a vector counting the frequencies of frequencies observed in the sample. Namely, given a pilot study $X = X_{1:N}$ with J distinct variants, the SFS of X is given by

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

[1] In our experiments, we tested various choices of n, and found $n = \lfloor 2/3 \times N \rfloor$ to be a good value across all the sample sizes considered. Different choices might be considered, e.g. when the sample size N is very small, or, conversely, very large. For the purposes of minimizing Equation (5), we used the differential evolution algorithm (Storn and Price, 1997), as implemented in the SciPy library (Jones et al., 2001–).

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.

D.1. Beta-Bernoulli product model

We start by reviewing the approach proposed by Ionita-Laza et al. (2009). The authors consider the same problem of genomic variation described in Section 1. 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 the j-th feature is displayed by any observation (row) with probability $\theta_j \in [0,1]$, where the frequencies θ_j, $j = 1, \ldots, K$ are distributed according to a beta distribution with parameters a, b, i.e.

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

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

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

so that the columns of the matrix $X_{1:N}$ are i.i.d., 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.

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

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

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

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

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

Because we can't observe more than N variants in N trials, and since the 0 class is never observed, these probabilities are then normalized as follows:

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

It follows that the log likelihood for the observed data $X_{1:N}$ is given by

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

Letting $M = tN$ be the number of additional samples to be observed, we can compute the expected number of new hitherto unseen variants in additional M samples after N samples have been observed as

$$\begin{aligned}
\Delta_N(M) &= \mathbb{E}\left[\sum_{j=1}^{K} \mathbf{1}\left(\sum_{m=1}^{M} x_{m,j} > 0\right) \mathbf{1}\left(\sum_{n=1}^{N} x_{n,j} = 0\right)\right] \\
&= \frac{K}{\boldsymbol{B}(a,b)} \int_{[0,1]} (1-(1-\theta)^{(t+1)N}) - (1-(1-\theta)^{N})\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta \\
&= \frac{K}{\boldsymbol{B}(a,b)} \int_{[0,1]} (1-\theta)^{N} - (1-\theta)^{(t+1)N}\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta \\
&= \frac{\eta_1}{a}\frac{N+b-1}{N}\left[1 - \frac{\boldsymbol{B}(a,(t+1)N+b)}{\boldsymbol{B}(a,N+b)}\right],
\end{aligned} \tag{6}$$

where $\eta_1 := \mathbb{E}[f_{N,1}]$ is the expected number of features which appear exactly one time in a sample of size N. To use the estimator $\Delta_N(M)$, Ionita-Laza et al. (2009) substitute η_1 with its empirical counterpart $_{N,1}$, the number of features which have been observed once in the sample $X_{1:N}$. Then, they find the parameters a, b via maximization the 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 3 *The estimator obtained in Equation (6) crucially relies on the empirical frequency of features 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 $M > 0$.*

D.2. Linear program to estimate the frequencies of frequencies

Zou et al. (2016) assume the same setting as Ionita-Laza et al. (2009) and formalize the problem of hitherto unseen variants prediction as that of recovering the distribution of frequencies of all the genetic variation in the population, including those features 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 i.i.d. 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 $z_{N,j} := \sum_{n=1}^{N} x_{n,j} \sim \text{Binomial}(N, \theta_j)$. 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 P to integers

$$h_P : (0,1] \to \mathbb{N} \cup \{0\} \tag{7}$$

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 $\boldsymbol{f}_N^{(\kappa)} = \{f_{N,j}/N : j \leq \kappa\}$. In their analysis, the authors only consider $\kappa = 0.01$, 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 (7), they impose 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:

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

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 $\text{Binomial}(N,\theta,j)$ 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.3. Jackknife estimators

Jackknife estimators for the problem of estimating the number of hitherto unseen species were first introduced by Burnham and Overton (1978) in the capture-recapture literature.

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

$$\mathbb{E}[\hat{\psi}_N] = \psi + \frac{a_1}{N} + \frac{a_2}{N^2} + \dots, \tag{8}$$

for fixed constants $a_1, a_2, \dots$. 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

$$\hat{\psi}_N^{(p)} = \binom{N}{p}^{-1} \sum_{\mathcal{I}:|\mathcal{I}|=p} \hat{\psi}_{N-p,\mathcal{I}} \tag{9}$$

The idea of the Jackknife estimator is that, if the assumption of Equation (8) holds, we can improve over $\hat{\psi}_N$ by using a correction originating from Equation (9). The p-th order Jackknife estimator is defined as

$$\hat{\psi}_N^{J_p} = \frac{1}{p} \sum_{\ell=0}^{p} \left((-1)^{\ell} \binom{p}{\ell} (N-\ell)^p \hat{\psi}_N^{(\ell)} \right). \tag{10}$$

Under the assumption of Equation (8), the estimator of Equation (10) has bias approaching zero polynomially fast in the correction order, $\mathrm{Bias}(\hat{\psi}_N^{J_p}) \sim N^{-p-1}$.

D.3.1. 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 $X_n = [x_{n,1}, \dots, x_{n,J}]$, with $x_{n,j} = 1$ if species j has been observed on the n-th experiment, and 0 otherwise. Moreover, assume that each species $j \in [K]$ – both observed and unobserved ones – is observed across all trials with a fixed, but unknown probability $\theta_j \in (0,1]$.

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 $\hat{K}\mathrm{MLE}(X_{1:N}) = \hat{K}_N^{\mathrm{MLE}} = J$. Clearly $J \leq K$, therefore J is a biased estimate for K. If one assumes, in a similar spirit to Equation (8), that

$$\mathbb{E}[\hat{K}_N^{\mathrm{MLE}}] = K + \frac{a_1}{N} + \frac{a_2}{N^2} + \dots, \tag{11}$$

then one could use the jackknife estimator of Equation (10) to estimate K. This requires computing $\hat{\psi}_N^{(\ell)}$ for $\ell = 1, \dots, p$, 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 on the n-th trial,

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

Then,

$$\hat{K}_N^{(1,\backslash n)} = J - q_{N,n} \quad \text{and} \quad \hat{K}_N^{(1)} = \frac{1}{N}\sum_{n=1}^{N} \hat{K}_N^{(1,\backslash n)} = J - \frac{f_{N,1}}{N}. \tag{12}$$

Therefore, the order 1 jackknife estimator for the total population size is obtained by plugging in $\hat{\psi}_N^{(J_0)} = J$ and $\hat{\psi}_N^{(J_1)} = J - \frac{f_{N,1}}{N}$ in Equation (10):

$$\hat{K}_N^{J_1} = J + \frac{N-1}{N} f_{N,1} \tag{13}$$

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

$$\hat{K}_N^{(p)} = J - \binom{N}{p}^{-1} \sum_{\ell=1}^{p} \binom{N-\ell}{p-\ell} f_{N,\ell} \tag{14}$$

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 $\boldsymbol{f}_N$,

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

D.3.2. 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 $\Delta(N+M,N) := H_{N+M-1} - H_{N-1} = \sum_{\ell=N}^{M+N-1}$, 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:

$$\hat{V}_N^{(M)} = V(N) + \sum_{\ell=1}^{p} a_{\ell,p} \Delta(N+M,N)^{\ell}, \tag{15}$$

where $\boldsymbol{a} = [a_1, \ldots, a_p]$ are constants which depend on the initial sample size n and on the fingerprint of the sample $\boldsymbol{f}$. 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:

$$\hat{V}_N^{(M)} = \hat{V}_{N-1}^{(M)} = \ldots = \hat{V}_{N-p}^{(M)}. \tag{16}$$

Equating $\hat{V}_N^{(M)}$ to $\hat{V}_{N-j}^{(M)}$ using Equation (15) for $j = 1, \dots, p$, we obtain a system of p equations of the form

$$V(N) - V(N-j) = \sum_{\ell=1}^{p} a_{\ell,p} (\Delta(N+M, N-j)^{\ell} - \Delta(N+M, N)^{\ell}). \tag{17}$$

Using the equality

$$V(N) - V(N-\ell) = \sum_{j=1}^{\ell} \frac{\binom{\ell}{j}}{\binom{N}{j}} f_{N,j}. \tag{18}$$

we can solve for $a_{\ell,p}$ 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.3.3. 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 on 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 $p = 1, 2, \dots$ one sequentially performs the following test:

$$H_{0,p}: \ \mathbb{E}(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p}) = 0 \quad \text{versus} \quad H_{a,p}: \ \mathbb{E}(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p}) \neq 0. \tag{19}$$

If $H_{0,p}$ is rejected, this has to be interpreted ad evidence of significant bias reduction relative even to the increased variance of $\hat{S}_{J_{p+1}}$ (Burnham and Overton, 1978). The first order p for which the test fails to reject the null hypothesis is picked as the jackknife order.

The test relies on the following observations:

$$\hat{K}_N^{(J_{p+1})} - \hat{K}_N^{J_p} = \sum_{\ell=1}^{p+1} \tilde{a}_p f_{N,p}, \tag{20}$$

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

$$\text{est var}(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p} \mid J) = \frac{J}{J-1} \left\{ \sum_{\ell=1}^{p} \tilde{a}_\ell^2 f_{N,\ell} \frac{(\hat{K}_N^{(J_{p+1})} - \hat{K}_N^{J_p})^2}{J} \right\}. \tag{21}$$

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

$$T_p = \frac{\hat{K}_N^{(J_{p+1})} - \hat{K}_N^{J_p}}{\sqrt{\text{est var}(\hat{K}_N^{J_{p+1}} - \hat{K}_N^{J_p} \mid J)}} \tag{22}$$

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.

Appendix E. Additional experimental results

E.1. The gnomAD dataset

For real data, we use the Genome Aggregation Database (gnomAD) (Karczewski et al., 2019), a recent extension of the Exome Aggregation Consortium (ExAC) data set (Lek et al., 2016). GnomAD contains genetic information from 125,748 individual exome sequences recorded at 1,195,872 genomic loci. Samples in gnomAD are arranged into eight subpopulations according to geographic origin, where one of the eight subpopulations is a catch-all category called "Other". These subpopulations vary in size from 5,040 Ashkenazi Jewish samples to 56,885 non-Finnish European samples.

In order to test the performance of our estimator we consider both synthetic and real data, with the goal of understanding how the predictive performance of the estimators changed across different data generating regimes. We compared the performance of the estimators derived in Section 2 to all the alternative estimators proposed in the literature – to the best of our knowledge – namely the ones proposed in Ionita-Laza et al. (2009); Gravel (2014); Zou et al. (2016), described in detail in Appendix D.1, Appendix D.2 and Appendix D.3.

First, as a sanity check, for those estimators for which an explicit model for the frequencies is available, we verified that the we were able to learn the true parameters of the data generating process and the proposed estimators performed well in practice. Across all the experiments performed, we found our estimators to perform comparably, if not better, to the best alternative method.

E.2. 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 2 to work particularly well. We tested 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 3 was able to recover the true growth rate of the distinct variants. Interestingly, in some instances, the optimization recovered parameters that differ from the true parameters that generated the process, but still have better empirical performance, due to the sampling variability associated to the process generation (see Figure 4).

E.3. Comparison with the method of Ionita-Laza et al. (2009)

In a nutshell, the method proposed by Ionita-Laza et al. (2009) is a parametric Bayesian counterpart to the nonparametric approach introduced in Appendix D.1 for a thorough discussion). Specifically, Ionita-Laza et al. (2009) assume that there exists a fixed, unknown number K of variants, whose frequencies are i.i.d. draws from a beta distribution. In our experiments, we find that this parametric approach is extremely effective when tested on synthetic data in which the variants' distribution follows an exact beta distribution with mean sufficiently larger than 0. The performance of the estimator rapidly worsens under misspecification, i.e. when the variants' distribution deviates from a beta distribution. In the appendix, we report experimental results for power law distributed frequencies, and

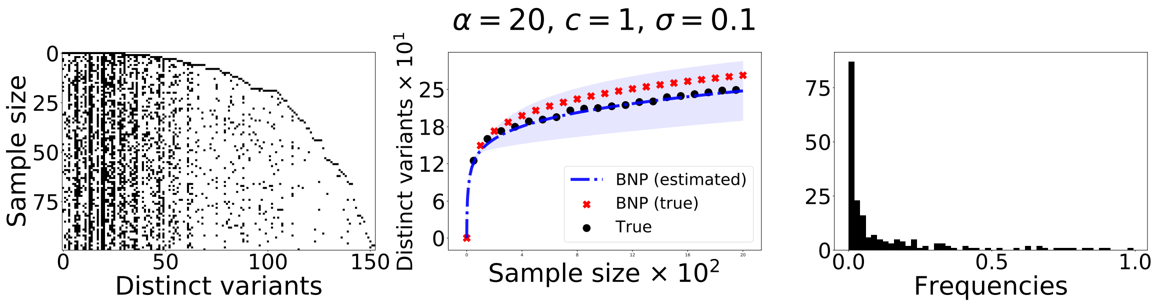

Figure 4: A draw from a three-parameter Indian buffet process. Here, $\alpha = 20$, $c = 1$, $\sigma = 0.1$. In the left panel, we see the binary matrix X containing the first $N = 100$ samples (x-axis) from the process, in its left-ordered-form (lof) – i.e. variants (y-axis) are sorted by the order 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.

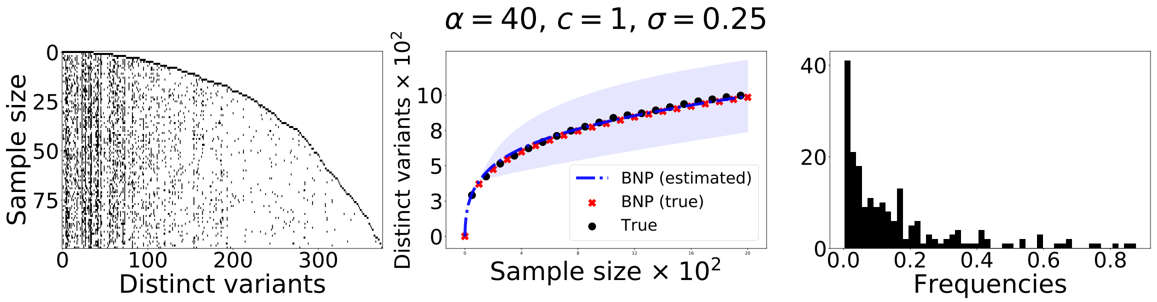

Figure 5: In this figure, we reproduce the visualizations explained in Figure 4 for a draw from a three-parameter Indian buffet process with parameters $\alpha = 40$, $c = 1$ and $\sigma = 0.25$

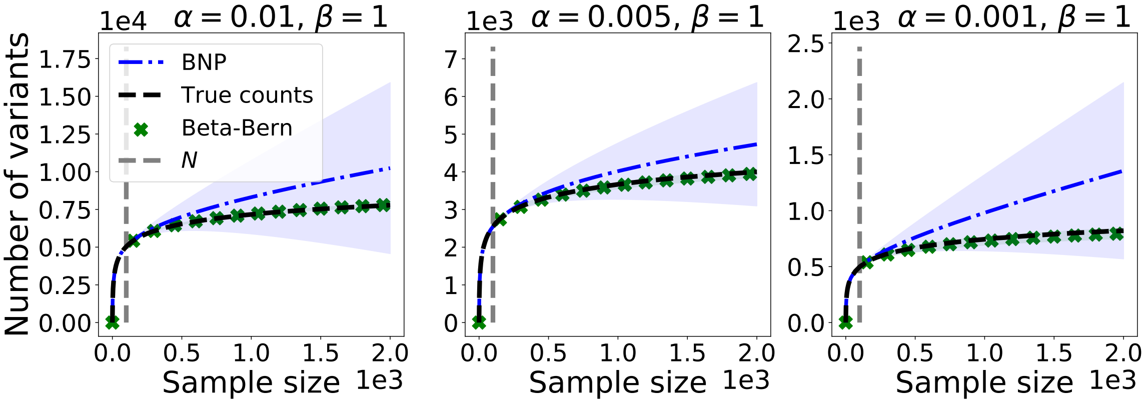

Figure 6: Performance of the beta-Bernoulli predictor (green crosses) proposed by Ionita-Laza et al. (2009) and of the nonparametric Bayesian predictor (dotted blue line, Theorem 1) on three different datasets (each panel represents a different dataset). Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimensions $K = 10^4$. The K coordinates are i.i.d. draws from a beta distribution. Conditionally on $\boldsymbol{\theta}$, we draw a random matrix X with $N = 2000$ rows and K columns. The (n, j)-th entry $x_{n,j}$ is Bernoulli distributed with parameter θ_j, so that the columns of X are i.i.d.. We retain the first $N = 200$ rows as training set and obtain the two estimators. We project up to $N + M = 2000$ observations. To produce estimates of the prediction error, we subdivide the whole data X in 10 distinct subsets of $N = 200$ rows, and iteratively train both models on each subset. We report the mean (solid lines, and crosses) and one standard deviation (shaded regions) of the predicted values at each extrapolation level $N+m$for each $m = 201, \ldots, 2000$ across the ten subsets. From left to right, we vary the first shape parameter of the beta distribution $\alpha \in \{10^{-2}, 5 \times 10^{-3}, 10^{-3}\}$, driving the mean of the distribution to zero, while keeping the second parameter $\beta = 1$ fixed.

show that even for moderate power law behavior the estimator severely underestimate the rate at which distinct variants appear. Because power laws arise in a vast number of natural phenomena, including the genomic application considered here, this represents a major limitation of the Bayesian parametric approach.

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 i.i.d. draws from a power law distribution, i.e. for some tails exponent $\xi \geq 0$

$$\theta_j \sim f(\theta) \propto \theta^{-\xi} \mathbf{1}_{[0,1]}(\theta). \tag{23}$$

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 small frequencies. Power laws arise in a vast number of natural phenomena, including ecology, biology, physical and social sciences. Therefore, having an estimator that is effective when

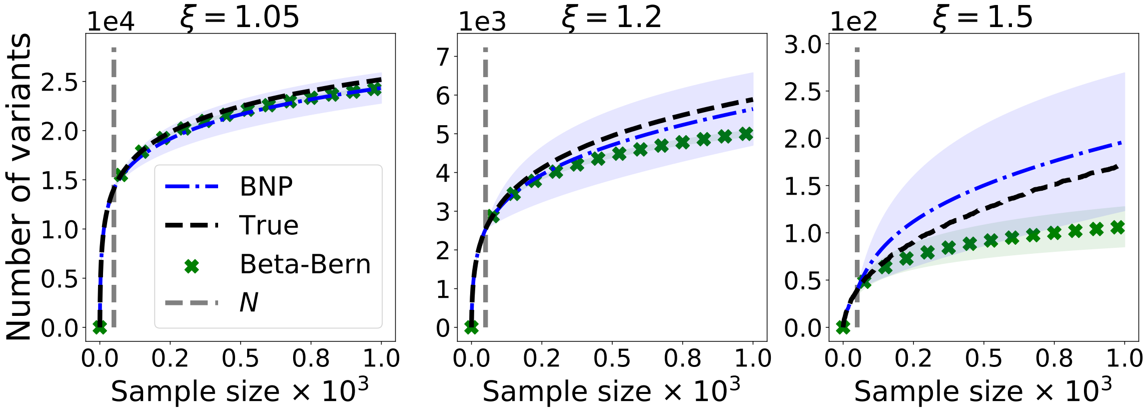

Figure 7: Performance of the beta-Bernoulli predictor (green solid line) proposed by Ionita-Laza et al. (2009) versus the nonparametric Bayesian predictor (dotted blue line) on three different datasets (each panel represents a different dataset). Each dataset is generated as follows: we first draw a random vector $\boldsymbol{\theta}$ of dimensions $K = 10^4$. The K coordinates are i.i.d. draws from a power law distribution as described in Equation (23). Conditionally on $\boldsymbol{\theta}$, we draw a random matrix 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 X are i.i.d.. We retain the first $N = 50$ rows as training set and obtain the two estimators. We project up to $N + M = 1000$ observations. We repeat the procedure over ten resamples of the same data. Uncertainty estimates are obtained by computing one empirical standard deviation across the ten predictors for each $\ell = 101, \ldots, 1000$. From left to right, we vary the exponent of the power law distribution (left, $\xi = 1.05$, center, $\xi = 1.2$, right $\xi = 1.5$).

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 7). Moreover, when a substantial portion of the variants have very low frequency, but the empirical distribution of the frequency has sufficiently fat right tails – as in the true data – the parametric approach is unreliable (see Figure 8). In all these cases instead, the Bayesian nonparametric estimator performs reasonably well.

E.4. Comparison with the method of Zou et al. (2016)

The method proposed by Zou et al. (2016) is frequentist, and nonparametric (see Appendix D.2). The goal of such method is to estimate the whole population histogram of variants' frequencies from observed samples by means of a linear program. Once the histogram has been estimated, quantities such as the number of hitherto unseen variants that will be discovered in M additional samples from the population are obtained from

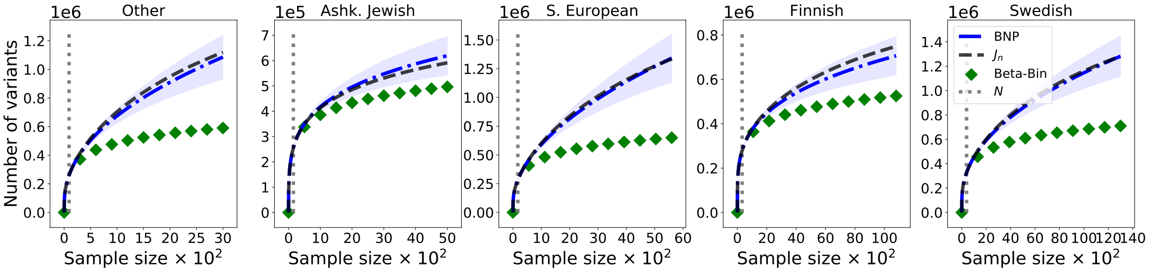

Figure 8: Results of the estimation of the number of new variants on some sub populations of the gnomAD dataset. We consider the Other Ashkenazi-Jewish, Southern European, Finnish and Swedish subpopulations. The x-axis displays the total number of samples collected. On the y-axis, we plot the number 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: $N \in \{93, 152, 173, 325, 393\}$). The blue line is the empirical mean of the predicted number of distinct variants observed according to the Bayesian nonparametric estimator across ten resamples samples of size N. The green crosses are the empirical means of the Bayesian parametric estimator of Ionita-Laza et al. (2009) across the same resamples. The shaded blue and green regions cover one standard empirical deviation for the predictors.

the population frequencies through a binomial sampling model (see Appendix D.2). The nonparametric nature of the method guarantees its robustness to various frequencies' distribution. However, the method suffers from other issues: in particular, the linear program is designed to only estimate *rare* variants frequencies, while frequencies of common variants are approximated using their empirical counterpart. Importantly, the algorithm requires to specify upfront a threshold $\kappa > 0$. The value of κ determines which frequencies are considered to be "rare" (those appearing less than κ% of the times in the sample), and which frequencies are instead "common" (those appearing at least κ% of the times in the sample). The input to the algorithm is the variants' frequencies empirical distribution up to κ%. The output of the linear program is an estimated histogram of rare frequencies, i.e. of the interval $[0, \kappa\%]$. In practice, we found that the optimization method is extremely sensitive to the choice of such parameter κ. The authors suggest picking $\kappa = 1$: this corresponds to learning only the histogram for frequencies $\theta \in [0, 10^{-2}]$, and approximate the remaining portion of the histogram using its empirical counterpart. In our experiments, we found that setting $\kappa = 1$ can lead to numerical instability and poor predictive performance. In particular, when the sample size N is small, it is often necessary to specify a much larger value of κ in order not to incur in numerical issues. Learning κ from the data is not straightforward, and, as we show in Figure 10 and Figure 11, different choices of κ yield dramatically different predictions for different frequencies distributions and different sample sizes. In the experiments on the gnomAD dataset reported in Figure 9 we train on 3% of the available data. We notice that for sample sizes considered, using small values of κ (e.g. $\kappa < 10$) can lead to numerical issues in the optimization. We show results using $\kappa = 20$,

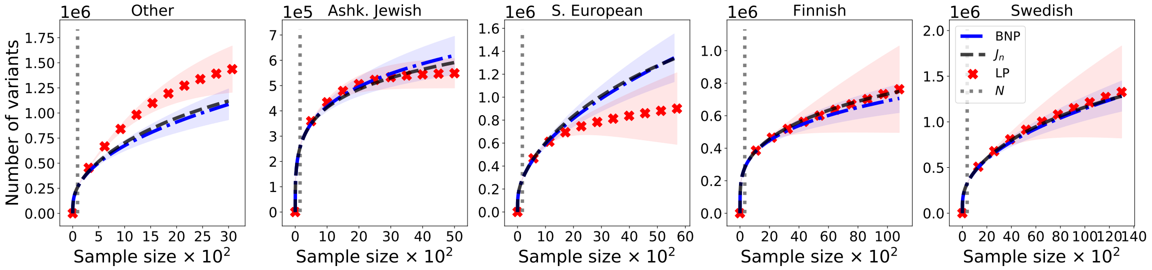

Figure 9: Results of the estimation of the number of new variants on some sub populations of the gnomAD dataset. The x-axis displays the total number of samples collected. On the y-axis, we plot the number of distinct genomic variants observed. The solid black line keeps track of the true number of distinct variants, the vertical grey line is placed in correspondence of the training sample size N. The blue dotted line is the empirical mean of the predicted number of distinct variants observed according to the Bayesian nonparametric estimator 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 cover two standard empirical deviations for the estimators. Here, we fix $\kappa = 1\%$, the value considered in Zou et al. (2016).

a value for which the optimization successfully converges for all datasets considered. We note that, especially when the sample size is small, the choice of κ has a relevant impact on the prediction quality, while its influence becomes negligible as the sample size N increases. However, we imagine these kinds of methods to be particularly interesting for the small N regime.

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 10 and Figure 11).

E.5. Comparison with the jackknife approach

The jackknife approach of Burnham and Overton (1978), recently proposed by Gravel et al. (2011); Gravel (2014) in a genomics context, was the best competitor of the Bayesian nonparametric estimator. The jackknife estimators require to specify and order $p \in \{1, 2, \dots\}$. Higher orders incur in less bias, at the cost of higher variance, while lower orders incur in less variance but have larger bias. Burnham and Overton (1978) proposed a procedure to

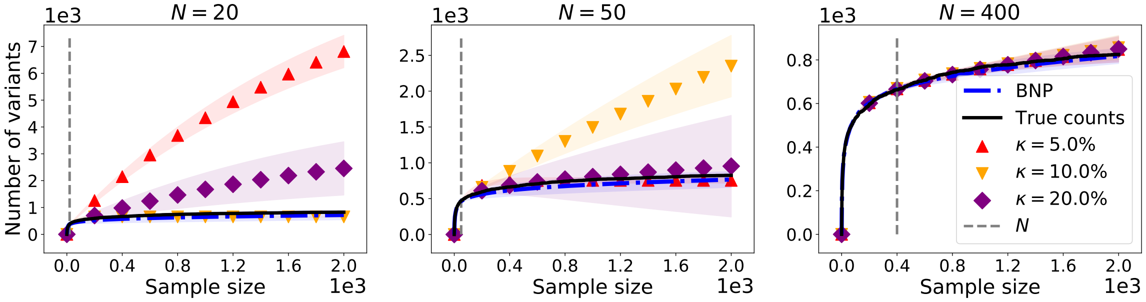

Figure 10: Comparison of the Bayesian nonparametric estimator (blue dotted line) to the frequentist nonparametric estimator proposed by Zou et al. (2016). We generate one synthetic dataset as follows: we first draw a random vector $\boldsymbol{\theta}$ of $K = 10^4$ i.i.d. beta random variables with parameters $\alpha = 0.001$ and $\beta = 1$. Conditionally on $\boldsymbol{\theta}$, we draw a random matrix X with $N = 2000$ rows and k columns. In each subplot, we retain a different fraction of rows of 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, purple) 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.

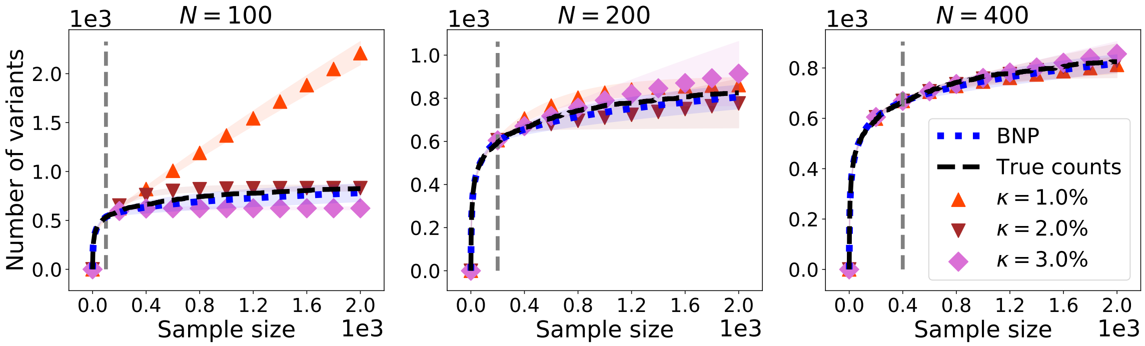

Figure 11: Comparison of the Bayesian nonparametric estimator (blue dotted line) to the frequentist nonparametric estimator of Zou et al. (2016). We use the same data showed in Figure 10 but using much smaller values of $\kappa \in \{1\%, 2\%, 3\%\}$. Trying to run the linear program for these values 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.

decide the optimal value of the estimator for the data under considerations. This consists in sequentially increasing the order of the jackknife estimator until the gain in bias reduction is exceeded by the variance increase (see Appendix D.3). All the estimators are obtained by modeling the number of missed variants using a parametric form derived from consistency requirements. In particular, the jackknife estimator of the p-th order, which estimates the number of missed variants which are going to be observed in M additional samples after initial N samples displaying k_N unique variants have been collected, is obtained by computing a linear equation of the first p frequencies of frequencies of the observed samples (see Appendix D.3 for more details). On the real data, we found order 3 or order 4 estimators to be optimal according to the sequential procedure. The results were extremely precise and the only competitive with the Bayesian nonparametric approach,

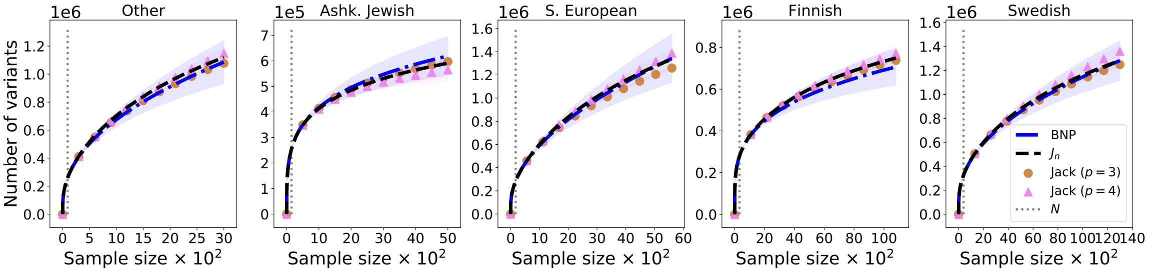

Figure 12: Again for the same sub populations considered in Figure 9, we compare the Bayesian nonparametric estimator to the Jackknife estimator proposed in Gravel (2014), for the third and fourth orders. Lower order consistently underestimate the number of distinct variants.

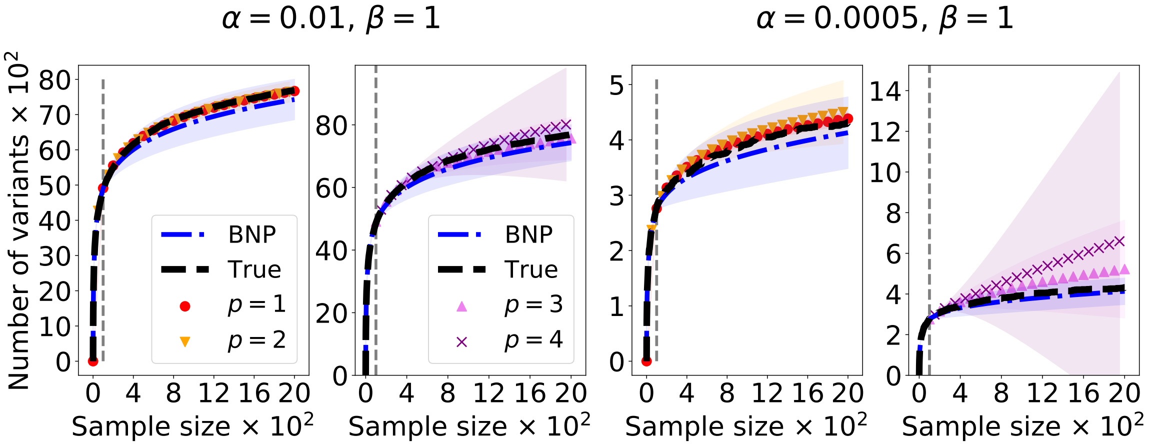

Figure 13: Comparison of the Bayesian nonparametric estimator (blue dotted line) to the jackknife estimator of Gravel (2014) for different choices of the order p. We generate two datasets as follows: for $\alpha \in \{0.01, 0.0005\}$ and $\beta = 1$, we generate two sets of $K = 10^4$ i.i.d. beta distributed draws $\boldsymbol{\theta}$ with parameters α, β. We then draw a random matrix 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.