Double trouble: Predicting new variant counts across two heterogeneous populations

Yunyi Shen
EECS, MIT
yshen99@mit.edu

Lorenzo Masoero*
Amazon.com
masoerl@amazon.com

Joshua Schraiber
Genome Interpretation Group, Illumina Inc
jschraiber@illumina.com

Tamara Broderick
EECS, MIT
tbroderick@mit.edu

1 Introduction

Collecting genomics data across multiple heterogeneous populations has the potential to improve our understanding of disease. For instance, we might compare cases and controls to discover biological underpinnings of disease and ultimately develop treatments. Or we might compare multiple cancer types to understand etiology and develop more broadly-applicable therapies. Despite sequencing advances, though, resources remain a constraint in how much data can be gathered. If we could predict the amount of new information to be gained from different potential amounts of data, that could guide experimental design to make optimal use of the resources at hand. And while many authors have developed prediction methods for the single-population case [Masoero et al., 2022, Chakraborty et al., 2019, Zou et al., 2016, Gravel and National Heart Lung & Blood Institute (NHLBI) GO Exome Sequencing Project, 2014, Ionita-Laza et al., 2009], that work will generally provide poor-quality predictions across multiple heterogeneous populations when these populations can share information. We develop a new methodology to provide accurate predictions in the face of multiple populations.

More precisely, we predict the number of (genetic) variants, i.e., differences in an observed genome relative to some reference genome. Variants have the potential to severely disrupt gene function and are implicated in many diseases [Tennessen et al., 2012, Cargill et al., 1999, 1000 Genomes Project Consortium, 2015]. We consider the case where the experimenter has collected (limited) pilot data on two populations.[2] The experimenter is planning a follow-up experiment of a particular size and wishes to predict (1) the number of new variants shared across the two populations or (2) the total number of new variants. Here, "new" refers to variants not already discovered in either population in the pilot experiment, and size is determined by the number of samples in each population. For instance, the shared variants predicted in task (1) might be used to help elucidate a common genetic basis across cancers [Rafnar et al., 2009, Bojesen et al., 2013, Rashkin et al., 2020]. Or a biologist might be interested in task (2) if they plan to compare cases and controls in a rare variant association study [Zuk et al., 2014].

Related work. Many authors have addressed the problem of predicting the number of new variants in a follow-up experiment of a particular size given a pilot experiment – when samples in both cases arise from a single, homogeneous population [Masoero et al., 2021, Chakraborty et al., 2019, Zou et al., 2016, Gravel and National Heart Lung & Blood Institute (NHLBI) GO Exome Sequencing Project, 2014, Ionita-Laza et al., 2009]. One option for task (2) is to treat the two populations as a single population and apply existing techniques. But we will see in our experiments that, if the populations are heterogeneous, the relative ratios of the two populations in the follow-up affects

*Work not related to Amazon

[2]Our method naturally extends to more than two populations.

the number of expected variants. Similarly, while we might use existing methods for task (1) by treating shared variants as the only variants, doing so cannot account for the differential impact of more samples in one or the other population.

Our contribution. We provide a new methodology (Section 2) for predicting the number of shared and total variants in a follow-up experiment from a pilot experiment when there are two potentially-heterogeneous populations. We provide theory to show that a seemingly natural extension of a state-of-the-art technique [Masoero et al., 2022] to independently model the variant frequencies in two populations fails for surprisingly fundamental reasons. To still use ideas from Bayesian nonparametrics (BNP) as in [Masoero et al., 2022], we need to develop a new BNP model, not previously seen in any literature. We demonstrate in experiments on real cancer data and simulations that our method provides accurate predictions (Section 3).

2 Setup and our model

Setup. We consider two populations. Let N_i be the number of samples in population $i \in \{1, 2\}$ in the pilot experiment. Let $N = (N_1, N_2)$, and let J_N be the total number of unique variants across both sets of samples. Let ψ_j be a (unique) label for the jth unique variant. Let $x_{i,n,j}$ equal 1 if the nth sample in population i exhibits the jth variant and 0 otherwise. Let δ_ψ represent a discrete mass of size 1 at location ψ. So the measure $X_{i,n} = \sum_{j=1}^{J_N} x_{i,n,j}\delta_{\psi_j}$ collects the variant information for sample n in population i by pairing each variant indicator $x_{i,n,j}$ with its variant label ψ_j. Equivalently, we can write $X_{i,n} = \sum_{j=1}^{\infty} x_{i,n,j}\delta_{\psi_j}$ if we set $x_{i,n,j} = 0$ for an imagined countable infinity of possible future variants ψ_j.[3]

We consider experimenters planning a follow-up study that will collect M_i samples from population i; let $M = (M_1, M_2)$. We would like to predict the total number of new variants found in the follow-up given the pilot; call that number $U_N^{(M)}$:

$$U_N^{(M)} = \sum_{j=1}^{\infty} \left[\prod_{i=1}^{2} \mathbf{1}\left(\sum_{n=1}^{N_i} x_{i,n,j} = 0\right)\right] \mathbf{1}\left(\sum_{i=1}^{2}\sum_{m=1}^{M_i} x_{i,N_i+m,j} > 0\right).$$

We would also like to predict the number of shared variants across the populations. In fact, we predict the number of new variants appearing k_1 times in population 1 and k_2 times in population 2; call that number $U_N^{(M,k)}$ where $k = (k_1, k_2)$:

$$U_N^{(M,k)} = \sum_{j=1}^{\infty} \left[\prod_{i=1}^{2} \mathbf{1}\left(\sum_{n=1}^{N_i} x_{i,n,j} = 0\right)\right] \left[\prod_{i=1}^{2} \mathbf{1}\left(\sum_{m=1}^{M_i} x_{i,N_i+m,j} = k_i\right)\right].$$

We can sum up across values of $k_1 > 0$ and $k_2 > 0$ to find the total number of shared variants.

Our model and theoretical results. Like all single-population methods [Masoero et al., 2022, Chakraborty et al., 2019, Zou et al., 2016, Gravel and National Heart Lung & Blood Institute (NHLBI) GO Exome Sequencing Project, 2014, Ionita-Laza et al., 2009], we ignore linkage disequilibrium. We assume that samples within a single population are exchangeable. It is natural then to assume the existence of an unknown $\theta_{i,j}$ that represents the frequency of variant j in population i. Let $\boldsymbol{\theta}_j = (\theta_{1,j}, \theta_{2,j})$ be the vector collecting variant frequencies across populations for variant j. We can further collect the frequencies across variants in a random vector-valued measure $\Theta = \sum_{i=1}^{\infty} \boldsymbol{\theta}_j \delta_{\psi_j}$. We assume $x_{i,n,j} \sim \text{Bernoulli}(\theta_{i,j})$ independently across i, j, n.

We take a Bayesian approach, so it remains to put a prior on Θ. Analogous to the principles from [Masoero et al., 2022, James, 2017, Broderick et al., 2018] for one population, we have the following desiderata for our prior across two populations: (A) the number of variants observed in any single sample should be almost surely (a.s.) finite, and (B) the number of latent variants should be countably infinite (to ensure we never hit an upper bound during our analysis); more precisely for desideratum (B), we assume each of the two elements of each $\boldsymbol{\theta}_j$ in the countably infinite sum above is strictly between 0 and 1.

[3] Though there exists a fixed, finite upper bound on the number of variants, it is conceptually and computationally convenient to not model it directly.

In the one-population case, Masoero et al. [2022] satisfy similar desiderata with a three-parameter beta process (3BP) prior [Masoero et al., 2022, James, 2017, Broderick et al., 2012, Teh and Gorur, 2009] on the single-population analogue of Θ. One might think to put independent 3BP priors on the variant frequencies in the multiple population case. But, surprisingly, the coupling induced via the sharing of variant labels ψ_j makes it impossible to use *any* independent priors on the two sets of population frequencies (within a Poisson point process framework) while simultaneously meeting the desiderata. We provide a precise theorem statement and a proof for this new result in Appendix S8.

However, we show that an alternative approach can work. Namely, we propose that the $\boldsymbol{\theta}_j$ be generated from a Poisson point process with rate measure that does *not* factorize across the two populations: $\nu(d\boldsymbol{\theta}) = \gamma(\theta_1 + \theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta}$ with hyperparameters $\gamma, \alpha, \sigma_1, \sigma_2, c_1, c_2$. We show that this prior satisfies our desiderata when $\gamma, c_1, c_2 > 0$, $\sigma_1, \sigma_2 > 0$, and $\alpha \in [\sigma_1 + \sigma_2, \sigma_1 + \sigma_2 + 1)$; for details, see Proposition S4 in Appendix S7. The 3BP enjoys useful conjugacy properties; in particular, when the hyperparameters are known, its convenient form allows a closed-form prediction for the number of new variants in a follow-up. Our prior is similar in structure; it resembles a beta distribution in each dimension except for the leading factor $(\theta_1 + \theta_2)^{-\alpha}$. But due to this leading factor, there is no longer a convenient closed form for our prediction. Nonetheless, we can use numerical approximations, which we detail in Appendix S4.

Our method for fitting and making predictions. Whether we are predicting the number of new variants or new shared variants in a follow-up, we start by learning the hyperparameters of our model by using an empirical Bayes technique with the pilot data; see Appendix S2 for details. The hyperparameters control the behavior of the variant frequencies in Θ and hence the overall growth rate of variants and shared variants. With these hyperparameters in hand, we can then make predictions for the follow-up; see Appendix S1 for details of how we perform prediction in each task.

3 Synthetic and real data experiments

We next demonstrate empirically that our method is able to predict the total number of new variants, or particular combinations of new variants, across two populations in a follow-up study – given data from a pilot study across the same two populations.

Data. We simulate data according to our own model and check our recovery in Appendix S6. For real data, we use the MSK-IMPACT dataset [Cheng et al., 2015]. While the full data provides samples from patients with 341 different cancer types, we here restrict our attention to the two most prevalent types in the data to form our two populations: Lung Adenocarcinoma (749 individuals) and Breast Invasive Ductal Carcinoma (1,153 individuals). See Appendix S3 for more details.

Comparisons. We compare our new approach to two baselines that we derive from a state-of-the-art single-population approach [Masoero et al., 2022], which uses a three-parameter beta process (3BP) model. In particular, we let d3BP (for "(fully) dependent 3BP") represent the prediction we get by treating the two populations as a single population. And we let i3BP (for "independent 3BP") represent the case where we treat the two populations separately and assume no shared variants.

Experimental setup. For each i, we choose N_i points uniformly at random from the full available data set for the ith population. Together, these samples form the pilot study. We consider predicting the number of new variants in a follow-up study with M_i variants in the ith follow-up population. We take the follow-up samples chosen uniformly at random from the remaining samples not already chosen for the pilot. We describe the details of how we fit the hyperparameters of each model, including the comparison models, in Appendix S2.

Predicting the total number of new variants. In Figure 1A, we consider a pilot of size $(N_1, N_2) = (100, 100)$. In Appendix S6, we perform the analogous experiment but with different sizes (N_1, N_2) for the pilot – including varying proportions across the two populations. The horizontal axis of Figure 1A shows the total number of samples across both pilot and follow-up. Everything before 200 samples represents the (known) pilot data (shaded gray). Beyond 200 samples, we vary both the number of samples in the follow-up as well as the proportion from each population. Concretely, we start by introducing samples entirely from population 1 (Lung Adenocarcinoma). So our follow-up study has size (M_1, M_2) with $M_2 = 0$ and increasing M_1 until we reach $M_1 = 649$. At that point, we fix $M_1 = 649$ and add samples only from population 2 (Breast Invasive Ductal Carcinoma) until we reach 1,702 total follow-up samples. Each successive follow-up study contains the previous

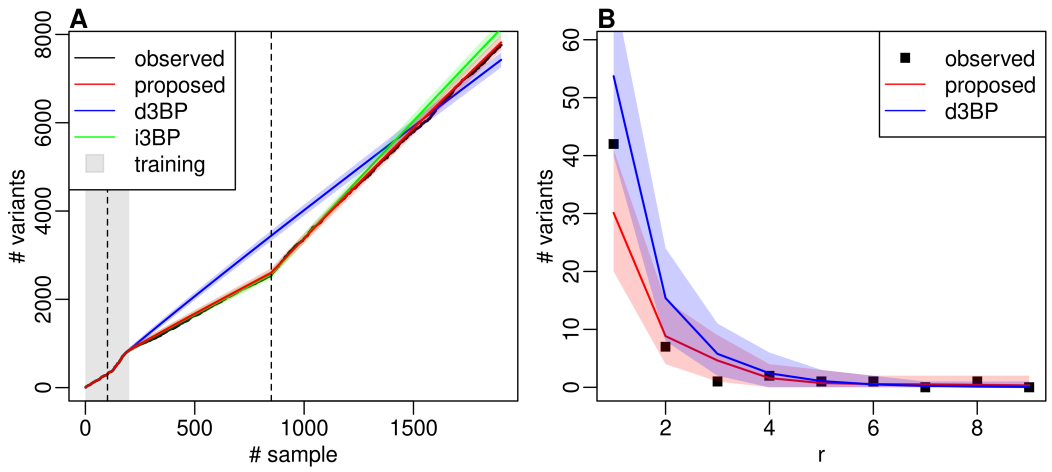

Figure 1: (A) Predicting the total number of new variants in a follow-up given data from a pilot. (B) Predicting the number of new variants appearing $(k_1, k_2) = (1, r)$ times in the two populations.

follow-up samples. The vertical dashed lines indicate where we switch from adding samples from population 1 to samples from population 2 in the pilot (left) and in the follow-up (right). The vertical axis shows the total number of unique variants observed so far – across both pilot and any follow-up samples observed at that point. Since we have access to the follow-up samples in this case, we can plot the ground truth (solid black line). We can see that population 1 has a noticeably smaller rate of new variants per sample than population 2.

Our method (red) tracks the ground truth closely. The d3BP (blue) is often far from the ground truth since it is not able to model the two populations separately; thus it performs particularly poorly when the overall ratio of populations in the full observed data does not closely match the ratio in the pilot data – e.g., when the follow-up is dominated by population 1 near the righthand vertical dashed line. The i3BP is able to capture the two different growth rates in the two populations; it likely performs well since there are few shared variants across these two populations relative to the total number of variants. For each prediction method, the shaded area represents a 95% credible interval.

Predicting the number of new shared variants. Next we consider a pilot of size $(N_1, N_2) = (100, 100)$ and a follow-up of size $(M_1, M_2) = (649, 1053)$ chosen uniformly at random from the two cancer populations, as above. We are interested in predicting how many new variants appear (k_1, k_2) times in the follow-up; this number of new variants forms the vertical axis in Figure 1B. We set $k_1 = 1$ and vary k_2 across the horizontal axis of Figure 1B. Our method's prediction appears in red, and the prediction of the d3BP appears in blue; in both cases, the shaded area represents a 95% credible interval. The i3BP does not model sharing of variants, so its predictions are effectively 0 throughout. We consider different combinations of (k_1, k_2) in Appendix S6.

Summary. We have seen that the i3BP can perform well in predicting the total number of new variants (in particular, when there are relatively few shared variants across populations). And the d3BP can perform well at predicting the number of new shared variants. But our new method is the only one that consistently performs well across all tasks and in the presence of shared variants.

Discussion. While our model naturally extends to more than two populations, computation may pose a practical challenge when the number of populations grows. This challenge arises since our method requires numerical integration both when fitting hyperparameters and when making predictions. We hope to explore ideas to speed up these computations in future work.

Acknowledgments and Disclosure of Funding

YS and TB were supported in part by the DARPA I2O LwLL program, an NSF Career Award, and an ONR Early Career Grant.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

S. E. Bojesen, K. A. Pooley, S. E. Johnatty, J. Beesley, K. Michailidou, J. P. Tyrer, S. L. Edwards, H. A. Pickett, H. C. Shen, C. E. Smart, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature Genetics*, 45(4):371–384, 2013.

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

T. Broderick, A. C. Wilson, and M. I. Jordan. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 24(4B):3181–3221, 2018. ISSN 13507265.

M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22 (3):231–238, 1999.

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(1):1–9, 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, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, and M. F. Berger. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of Molecular Diagnostics*, 17(3):251–264, 2015.

S. Gravel and National Heart Lung & Blood Institute (NHLBI) GO Exome Sequencing Project. Predicting discovery rates of genomic features. *Genetics*, 197(2):601–610, 2014.

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.

L. Masoero, J. Schraiber, and T. Broderick. Bayesian nonparametric strategies for power maximization in rare variants association studies. In *NeurIPS 2021 Workshop on Learning Meaningful Representations of Life*, 2021. URL `http://arxiv.org/abs/2112.02032`.

L. Masoero, F. Camerlenghi, S. Favaro, and T. Broderick. More for less: Predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika*, 109(1):17–32, 2022. ISSN 14643510. doi: 10.1093/biomet/asab012.

T. Rafnar, P. Sulem, S. N. Stacey, F. Geller, J. Gudmundsson, A. Sigurdsson, M. Jakobsdottir, H. Helgadottir, S. Thorlacius, K. K. H. Aben, T. Blöndal, T. E. Thorgeirsson, G. Thorleifsson, K. Kristjansson, K. Thorisdottir, R. Ragnarsson, B. Sigurgeirsson, H. Skuladottir, T. Gudbjartsson, H. J. Isaksson, G. V. Einarsson, K. R. Benediktsdottir, B. A. Agnarsson, K. Olafsson, A. Salvarsdottir, H. Bjarnason, M. Asgeirsdottir, K. T. Kristinsson, S. Matthiasdottir, S. G. Sveinsdottir, S. Polidoro, V. Höiom, R. Botella-Estrada, K. Hemminki, P. Rudnai, D. T. Bishop, M. Campagna, E. Kellen, M. P. Zeegers, P. de Verdier, A. Ferrer, D. Isla, M. J. Vidal, R. Andres, B. Saez, P. Juberias, J. Banzo, S. Navarrete, A. Tres, D. Kan, A. Lindblom, E. Gurzau, K. Koppova, F. de Vegt, J. A. Schalken, H. F. M. van der Heijden, H. J. Smit, R. A. Termeer, E. Oosterwijk, O. van Hooij, E. Nagore, S. Porru, G. Steineck, J. Hansson, F. Buntinx, W. J. Catalona, G. Matullo, P. Vineis, A. E. Kiltie, J. I. Mayordomo, R. Kumar, L. A. Kiemeney, M. L. Frigge, T. Jonsson, H. Saemundsson, R. B. Barkardottir, E. Jonsson, S. Jonsson, J. H. Olafsson, J. R. Gulcher, G. Masson, D. F. Gudbjartsson, A. Kong, U. Thorsteinsdottir, and K. Stefansson. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nature Genetics*, 41(2):221–227, 2009.

S. R. Rashkin, R. E. Graff, L. Kachuri, K. K. Thai, S. E. Alexeeff, M. A. Blatchins, T. B. Cavazos, D. A. Corley, N. C. Emami, J. D. Hoffman, E. Jorgenson, K. H. Kushi, T. J. Meyers, S. K. Van Den Eeden, E. Ziv, L. A. Habel, T. J. Hoffmann, L. C. Sakoda, and J. S. Witte. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nature Communications*, 11(1):1–14, 2020.

Y. Teh and D. Gorur. Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad Go, Seattle Go, and On behalf of the NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, 2012.

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(1):1–5, 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), 2014.

Appendix

S1 Method for prediction

In the center of our method is to predict $U_N^{M,k}$ for $k = (k_1, k_2)$ and $M = (M_1, M_2)$ while $N = (N_1, N_2)$, that is, number of new variants exist exactly k_1 times in population 1 and k_2 times in population 2 (in a follow up study of size M). Shared variants are those with $k_1, k_2 > 0$. Predicting number of new variants in either population thus the discover curve can be done via the predicting of $U_N^{M,k}$ and sum over $k > 0$ (see Corollary 1).

S1.1 Near conjugate prior

We assume hyperparameters for the near-conjugate prior is known. We describe fitting of those hyperparameters in Appendix S2. Consider the variants exists k_1 times in population 1 and k_2 times in population 2 in the follow up sequencing of size M_1 and N_2 while did not appear in the pilot study with size $N = (N_1, N_2)$. Formally, we consider

$$U_N^{(M,k)} = \sum_{j=1}^{\infty} \left[\prod_{i=1}^{2} \mathbf{1}\left(\sum_{n=1}^{N_i} x_{i,n,j} = 0 \right) \right] \left[\prod_{i=1}^{2} \mathbf{1}\left(\sum_{m=1}^{M_i} x_{i,N+m,j} = k_i \right) \right]. \tag{S1}$$

Proposition S1 (predicting $U_N^{(M,k)}$ under near conjugate prior)**.** *With prior being the near-conjugate prior, the number of follow-up k-r tons*

$$U_N^{(M,k)} | X_{1,1:N_1}, X_{2,1:N_2} \sim \text{Poisson}(\lambda_N^{(M,k)})$$

where

$$\begin{aligned} \lambda_N^{(M,k)} =& \gamma \binom{M_1}{k_1} \binom{M_2}{k_2} B(\sigma_1 + k_1, c_1 + N_1 + M_1 - k_1) B(\sigma_2 + k_2, c_2 + N_2 + M_2 - k_2) \\ & \mathbb{E}_{X,Y} (X + Y)^{-\alpha} \end{aligned} \tag{S2}$$

Where $B(a, b)$ is the beta function, and $\mathbb{E}_{X,Y}$ is the expectation with

$$\begin{aligned} X &\sim \text{Beta}(\sigma_1 + k_1, c_1 + N_1 + M_1 - k_1) \\ Y &\sim \text{Beta}(\sigma_2 + k_2, c_2 + N_2 + M_2 - k_2) \end{aligned}$$

independently.

We give the proof of this result in Appendix S4.

The number of *new* variants in in the follow-up given the pilot

$$U_N^{(M)} = \sum_{j=1}^{\infty} \left[\prod_{i=1}^{2} \mathbf{1}\left(\sum_{n=1}^{N_i} x_{i,n,j} = 0 \right) \right] \mathbf{1}\left(\sum_{i=1}^{2} \sum_{m=1}^{M_i} x_{i,N+m,j} > 0 \right).$$

Corollary 1 (Predicting number of new variants)**.** *With the prior being the near-conjugate model, the number of new variants seen with in the follow up study*

$$U_N^{(M)} | X_{1,1:L}, X_{2,1:D} \sim \sum_{k_1=1}^{M_1} \sum_{k_2=0}^{M_2} \lambda_N^{(M,k)}$$

Where $\lambda_N^{(M,k)}$'s are given in Equation (S2)*.*

Proof. Use the definition of $U_N^{(M)}$ and results on sum of independent Poisson random variables. □

S1.2 d3BP

Under d3BP, we essentially ignore the population label and use a 3BP to model the combined population following Masoero et al. [2021]. We use the same definition of kr-tons in Equation (S1).

Proposition S2 (predicting number of k-r tons under d3BP)**.**

$$U_N^{(M,k)}|X_{1,1:L}, X_{2,1:D} \sim \text{Poisson}(\lambda_{dep,N}^{(M,k)})$$

Where

$$\lambda_{dep,N}^{(M,k)} = \alpha \frac{(c+\sigma)_{(N_1+N_2+M_1+M_2-k_1-k_2)\uparrow}(1-\sigma)_{(k_1+k_2-1)\uparrow}}{(c+1)_{(N_1+N_2+M_1+M_2-1)\uparrow}} \times \binom{M_1}{k_1}\binom{M_2}{k_2}$$

For number of variants, we follow Masoero et al. [2021], which is equivalent as using Corollary 1.

S1.3 i3BP

Th i3BP model is not capable for predicting shared variants. The number of new variants in each population can be predicted follow Masoero et al. [2021] and we add up the new variants as a postprocessing.

S2 Method for fitting the hyperparameter

Regardless of predicting task, we took a empirical Bayes approach to find the hyperparameters $(\gamma, \alpha, \sigma_1, \sigma_2, c_1, c_2)$ for the near-conjugate prior, (α, σ, c) for d3BP, $(\alpha_1, \alpha_2, \sigma_1, \sigma_2, c_1, c_2)$ for i3BP. We used likelihood function on prediction of kr tons.

S2.1 Objective function on trimmed ℓ-tons

From the training set of size (N_1, N_2), we can calculate the observed number of variants exist $\ell = (\ell_1, \ell_2)$ times $u_0^{(N,\ell)}$. We can treat these as the prediction when there is no pilot data. From Proposition S1, we have the predictive distribution of$u_0^{(N,\ell)}$ of for different ℓ (and $\ell_1 + \ell_2 > 0$) they are independent Poisson with parameters $\lambda_0^{(N,\ell)}$ (Equation (S2)). To reduce computation cost we trim ℓ with some upper bound v such that $\ell_1, \ell_2 \leq v$ (we set v=10 in the real data analysis). Denote $\text{pois}(x|\lambda)$ as the pmf of Poisson distribution with parameter λ The objective function is given by the log likelihood

$$\mathcal{L}(\gamma, \alpha, \sigma_1, \sigma_2, c_1, c_2) = \sum_{\ell_1=1}^{v}\sum_{\ell_2=0}^{v} \log\left[\text{pois}\left(u_0^{(N,\ell)}|\lambda_0^{(N,\ell)}\right)\right] + \sum_{\ell_2=1}^{v} \log\left[\text{pois}\left(u_0^{(N,\ell)}|\lambda_0^{(N,(0,\ell_2))}\right)\right] \quad \text{(S3)}$$

where $\lambda_0^{(N,\ell)}$ is given in Equation (S2)

Then we set the hyperparameter to be

$$(\hat{\gamma}, \hat{\alpha}, \hat{\sigma}_1, \hat{\sigma}_2, \hat{c}_1, \hat{c}_2) = \arg\max \mathcal{L}(\gamma, \alpha, \sigma_1, \sigma_2, c_1, c_2)$$

We optimize the objective function using double annealing from `scipy` with default parameters.

S2.2 Fitting i3BP and d3BP with trimmed ℓ-ton's

The d3BP essentially treat the two population as one. We used the result from Masoero et al. [2021] on the predictive distribution of ℓ-ton, i.e. the number of variants exist exactly ℓ_s in a study with $N_s = N_1 + N_2$ samples times being independent Poisson distributed with mean $\lambda_{N_s}^{\ell_s} = \alpha\binom{N_s}{\ell_s}\frac{(1-\sigma)_{(\ell_s-1)\uparrow}(c+\sigma)_{(N_s-\ell_s)\uparrow}}{(c+1)_{(N_s-1)\uparrow}}$. For d3BP, we can calculate the number ℓ_s-tons $u_{N_s}^{\ell_s}$ up to an upper bound $\ell \leq v$ by ignoring the label of two populations. The objective function for d3BP is then

$$\mathcal{L}(\alpha, \sigma, c) = \sum_{\ell_s=1}^{v} \log\left[\text{pois}\left(u_{N_s}^{\ell_s}|\lambda_{N_s}^{\ell_s}\right)\right]$$

For i3BP, we treat the two population independently and count $u_{1,N_1}^{\ell_1}$, i.e. the ℓ_1-tons in population 1 and $u_{2,N_2}^{\ell_2}$ ℓ_2-tons in population 2 up to an upper bound v such that $\ell_1, \ell_2 \leq v$. The predictive distribution for them are Poisson with expectation $\lambda_{1,N_1}^{\ell_1} = \alpha_1\binom{N_1}{\ell_1}\frac{(1-\sigma_1)_{(\ell_1-1)\uparrow}(c_1+\sigma_1)_{(N_1-\ell_1)\uparrow}}{(c_1+1)_{(N_1-1)\uparrow}}$ and $\lambda_{2,N_2}^{\ell_2} =$

$\alpha_2\binom{N_2}{\ell_2}\frac{(1-\sigma_2)_{(\ell_2-1)\uparrow}(c_2+\sigma_2)_{(N_2-\ell_2)\uparrow}}{(c_2+1)_{(N_2-1)\uparrow}}$ respectively. The objective function for i3BP is

$$\mathcal{L}(\alpha_1,\sigma_1,c_1,\alpha_2,\sigma_2,c_2)=\sum_{\ell_1=1}^{v}\log\left[\text{pois}\left(u^{\ell_1}_{1,N_1}|\lambda^{\ell_1}_{1,N_1}\right)\right]+\sum_{\ell_2=1}^{v}\log\left[\text{pois}\left(u^{\ell_2}_{2,N_2}|\lambda^{\ell_2}_{2,N_2}\right)\right]$$

We optimize the objective function using double annealing from `scipy` with default parameters.

S3 Details on experiments

S3.1 Synthetic example

In the synthetic example we sample the product Bernoulli process whose rate coming from d3BP and near conjugate prior. We simulate 300 individuals for each population and take $(D,L)=(30,30),(30,50),(50,30),(50,50)$ from two populations as training and the rest as test sets. Simulation methods was described in Appendix.S5. We set the near conjugate prior's hyperparameters to be $\gamma=150,\alpha=1,\sigma_1=0.2,\sigma_2=0.8,c_1=c_2=20$ and the hyperparameters for the d3BP to be $\alpha=20,\sigma=0.5,c=10$ to generate the synthetic dataset.

We fit the hyperparameters for the near-conjugate prior, d3BP and i3BP prior using the empirical Bayes method described in Appendix S2. We then make predictions based on the posterior distribution of new variants. We calculate the posterior distribution of new variants in each follow-up size following the method described in Appendix S1 and add them to the variants number in the training sets.

S3.2 MSK-IMPACT example

In real data example, we first take $(D,L)=(50,100),(100,50),(100,100)$ samples from the first and second cancer types. With this "pilot" sample, we 1) find the hyperparameters for near-conjugate, d3BP and i3BP via empirical Bayes method described in Appendix S2 and 2) condition on the sample, we predict new variants and new shared variants seen in the follow up samples.

In fit the hyperparameter for the near-conjugate prior, d3BP and i3BP prior using the empirical Bayes method described in Appendix S2.

We then make predictions based on the posterior distribution of new variants. We calculate the posterior distribution of new variants in each follow-up size following the method described in Appendix S1 and add them to the variants number in the training sets. In predicting the new shared variants we calculate the posterior follow (S2) (for near conugate) and Masoero et al. [2021] using different k_1,k_2 values.

S4 Details on predicting method

S4.1 Near conjugate prior

Here we give the proof of Proposition S1.

Proof. The posterior followed a thinned Poisson process with a rate measure:

$$\begin{aligned}&\nu(d\boldsymbol{\theta})\text{Bernoulli}(0|\theta_1)^{N_1}\text{Bernoulli}(0|\theta_2)^{N_2}\\&\quad\binom{M_1}{k_1}\text{Bernoulli}(1|\theta_1)^{k_1}\text{Bernoulli}(0|\theta_1)^{M_1-k_1}\binom{M_2}{k_2}\text{Bernoulli}(1|\theta_2)^{k_2}\text{Bernoulli}(0|\theta_2)^{M_2-k_2}\\=&\gamma\binom{M_1}{k_1}\binom{M_2}{k_2}(\theta_2+\theta_1)^{-\alpha}\theta_2^{\sigma_2+k_2-1}(1-\theta_2)^{c_2+N_2+M_2-k_2-1}\theta_1^{\sigma_1+k_1-1}(1-\theta_1)^{c_1+N+M_1-k_1-1}\end{aligned}\tag{S4}$$

Thus the posterior distribution of k,rton's are Poisson with expectation:

$$\begin{aligned}\lambda_N^{(M,k)}=&\gamma\binom{M_1}{k_1}\binom{M_2}{k_2}B(\sigma_2+k_2,c_2+N_2+M_2-k_2)B(\sigma_1+k_1,c_1+N_1+M_1-k_1)\\&\mathbb{E}_{X,Y}(X+Y)^{-\alpha}\end{aligned}$$

Where $B(a,b)$ is the beta function, and $\mathbb{E}_{X,Y}$ is the expectation with

$$\begin{aligned}X&\sim\text{Beta}(\sigma_1+k_1,c_1+N_1+M_1-k_1)\\Y&\sim\text{Beta}(\sigma_2+k_2,c_2+N_2+M_2-k_2)\end{aligned}$$

independently.

□

The numerical double integral might be speed up via special function:

$$\int_0^1 (x+y)^{-\alpha} x^{\sigma_1-1} y^{\sigma_2-1} (1-x)^{c_1-1} (1-y)^{c_2-1} dx = B(b,d) y^{-\alpha+\sigma_2-1} (1-y)^{c_2-1} {}_2\mathrm{F}_1(\alpha, \sigma_1, \sigma_1+c_1, -1/y)$$

where ${}_2\mathrm{F}_1$ is the Gaussian hypergeometric function and $B()$ is the Beta function.

Further we can transform ${}_2\mathrm{F}_1$ using the Euler type transformation

$$\begin{aligned} {}_2\mathrm{F}_1(\alpha, \sigma_1, \sigma_1+c_1, -1/y) &= (1+1/y)^{-\alpha} {}_2\mathrm{F}_1(\alpha, c_1, \sigma_1+c_1, 1/(y+1)) \\ &= \left(\frac{y}{y+1}\right)^{\alpha} {}_2\mathrm{F}_1(\alpha, c_1, \sigma_1+c_1, 1/(y+1)) \end{aligned}$$

S4.2 d3BP

We give the proof of Proposition S2. The proof largely follow Masoero et al. [2021], by realizing that the total number follows a 3BP and allocation of the variants in two population follows a hypergeometric distribution.

Proof. The posterior followed a thinned Poisson process with a rate measure:

$$\begin{aligned} &\nu(d\theta)\mathrm{Bernoulli}(0|\theta)^{N_1}\mathrm{Bernoulli}(0|\theta)^{N_2} \\ &\quad \binom{M_1}{k_1}\mathrm{Bernoulli}(1|\theta)^{k_1}\mathrm{Bernoulli}(0|\theta)^{M_1-k_1}\binom{M_2}{k_2}\mathrm{Bernoulli}(1|\theta)^{k_2}\mathrm{Bernoulli}(0|\theta)^{M_2-k_2} \\ =&\binom{M_1}{k_1}\binom{M_2}{k_2}\nu(d\theta)\mathrm{Bernoulli}(1|\theta)^{k_1+k_2}\mathrm{Bernoulli}(0|\theta)^{N_1+N_2+M_1+M_2-k_1-k_2} \end{aligned} \tag{S5}$$

whose integral is the expectation, following Masoero et al. [2021], this is given by

$$\lambda_{dep,N}^{(M,k)} = \alpha \binom{M_1}{k_1}\binom{M_2}{k_2} \frac{(c+\sigma)_{(N_1+N_2+M_1+M_2-k_1-k_2)\uparrow}(1-\sigma)_{(k_1+k_2-1)\uparrow}}{(c+1)_{(N_1+N_2+M_1+M_2-1)\uparrow}}$$

□

S4.3 i3BP

The i3BP method cannot predict number of kr-tons, but we still could use the model to predict number of new variants. We treat two population independently model as 3BP's following Masoero et al. [2021] and add the number of new variants together to get the discover curve. That is, suppose we sample M_1 individual from population 1 and then M_2 from population 2 in the follow up study and N_1, N_2 individuals in the pilot, the predicted number of new variants at sample m_1 ($m_1 \leq M_1$) is Poisson with mean

$$\lambda_{m_1} = \sum_{k_1=1}^{m_1} \alpha_1 \frac{(c_1+\sigma_1)_{(M_1+k_1-1)\uparrow}}{(c_1+1)_{(M_1+k_1-1)\uparrow}}$$

and for all population 2 sample, the new variants seen after sampling population 1 is then added by the number seen in population 1, i.e. the predicted number of new variants at m_2 the sample of population 2 ($m_2 \leq M_2$) thus the $m_2 + M_1$th sample of the entire followup is Poisson with mean

$$\lambda_{m_2} = \sum_{k_1=1}^{M_1} \alpha_1 \frac{(c_1+\sigma_1)_{(M_1+k_1-1)\uparrow}}{(c_1+1)_{(M_1+k_1-1)\uparrow}} + \sum_{k_2=1}^{m_2} \alpha_2 \frac{(c_2+\sigma_2)_{(M_2+k_2-1)\uparrow}}{(c+1)_{(M_2+k_2-1)\uparrow}}$$

S5 Details on sampling

The sampling follows the general principle in Broderick et al. [2018]. That is, instead of sampling the Poisson process, we derive the marginal representation of the prior-product Bernoulli process.

S5.1 Near conjugate prior

Proposition S3 (Marginal representation of near-conjugate-Bernoulli process)**.** *Suppose we have* $x_{1,n,k}|\boldsymbol{\Theta} \sim \text{Bernoulli}(\theta_{1,k})$ *and* $x_{2,m,k}|\boldsymbol{\Theta} \sim \text{Bernoulli}(\theta_{2,k})$ *where* $\boldsymbol{\Theta}$ *come from a near conjugate prior, we have*

$$h_{cond}(x_{1,n,k}=1|x_{1,-n,k},x_{2,1:m,k}) = \frac{\mathbb{E}X_1(X_1+X_2)^{-\alpha}}{\mathbb{E}(X_1+X_2)^{-\alpha}} \tag{S6}$$

For $X_1 \sim \text{Beta}(\sigma_1+\sum_{i=1}^{n-1} x_{1,i,k}, c_1+n-1-\sum_{i=1}^{n-1} x_{1,i,k})$ *and* $X_2 \sim \text{Beta}(\sigma_2+\sum_{j=1}^{m} x_{2,j,k}, c_2+m-\sum_{j=1}^{m} x_{2,j,k})$ *independently.*

Proof. Follow the result in Broderick et al. [2018], the marginal representation for the independent likelihood process with likelihood $h_1(x_{1,n,k}|\theta_1)$ and $h_2(x_{2,m,k}|\theta_2)$ is given by

$$h_{\text{cond}}(x_{1,n,k}=x|x_{1,-n,k},x_{2,1:m,k}) = \frac{\iint_{\mathbb{R}_+^2} \nu(d\boldsymbol{\theta}) h_1(x|\theta_1)\prod_{i=1}^{n-1} h_1(x_{1,i,k}|\theta_1)\prod_{j=1}^{m} h_2(x_{2,j,k}|\theta_2)}{\iint_{\mathbb{R}_+^2} \nu(d\boldsymbol{\theta})\prod_{i=1}^{n-1} h_1(x_{1,i,k}|\theta_1)\prod_{j=1}^{m} h_2(x_{2,j,k}|\theta_2)}$$

substitute $h_1(x_{1,n,k}|\theta_1)$ by θ_1 and $h_2(x_{2,m,k}|\theta_2)$ by θ_2 and $\nu(\boldsymbol{\theta})$, we have

$$\begin{aligned}
&h_{\text{cond}}(x_{1,n,k}=1|x_{1,-n,k},x_{2,1:m,k})\\
=&\frac{\iint_{\mathbb{R}_+^2} \nu(d\boldsymbol{\theta}) h_1(x|\theta_1)\prod_{i=1}^{n-1} h_1(x_{1,i,k}|\theta_1)\prod_{j=1}^{m} h_2(x_{2,j,k}|\theta_2)}{\iint_{\mathbb{R}_+^2} \nu(d\boldsymbol{\theta})\prod_{i=1}^{n-1} h_1(x_{1,i,k}|\theta_1)\prod_{j=1}^{m} h_2(x_{2,j,k}|\theta_2)}\\
=&\frac{\iint_{[0,1]^2} \theta_1(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1+\sum_{i=1}^{n-1}x_{1,i,k}}\theta_2^{\sigma_2-1+\sum_{j=1}^{m}x_{2,j,k}}(1-\theta_1)^{c_1-1+n-1-\sum_{i=1}^{n-1}x_{1,i,k}}(1-\theta_2)^{c_2-1+m-\sum_{j=1}^{m}x_{2,j,k}}}{\iint_{[0,1]^2}(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1+\sum_{i=1}^{n-1}x_{1,i,k}}\theta_2^{\sigma_2-1+\sum_{j=1}^{m}x_{2,j,k}}(1-\theta_1)^{c_1-1+n-1-\sum_{i=1}^{n-1}x_{1,i,k}}(1-\theta_2)^{c_2-1+m-\sum_{j=1}^{m}x_{2,j,k}}}\\
=&\frac{\mathbb{E}X_1(X_1+X_2)^{-\alpha}}{\mathbb{E}(X_1+X_2)^{-\alpha}}
\end{aligned}$$

For $X_1 \sim \text{Beta}(\sigma_1+\sum_{i=1}^{n-1} x_{1,i,k}, c_1+n-1-\sum_{i=1}^{n-1} x_{1,i,k})$ and $X_2 \sim \text{Beta}(\sigma_2+\sum_{j=1}^{m} x_{2,j,k}, c_2+m-\sum_{j=1}^{m} x_{2,j,k})$ independently. □

S5.2 d3BP

For a d3BP-Bernoulli process of size (M,N), we sample a 3BP-Bernoulli process following Masoero et al. [2021] with size $M+N$ and randomly split individual into population 1 of size M and population 2 of size N.

S6 More experiment results

S6.1 Synthetic experiments

We present the discover curve of fitting different generating process. When the data was sampled from d3BP-Bernoulli process, the near conjugate prior perform at least as well as the d3BP model while i3BP tend to over estimate due to double counting (Fig.S2). Similarly when the data was sampled from a near conjugate-Bernoulli process the near conjugate prior perform better than other two method we tested (Fig.S3).

S6.2 More sampling schemes in MSK-IMPACT data

When the training data has relatively small sample in population 2 all methods tend to over estimate the number of new variants at the latter stage but near-conjugate is close to the ground truth (Fig.S4). In this case the i3BP has good performance because the shared variants are rare in cancer dataset.

S7 View as multivariate complete random measure

In a multivariate case, we consider an atom has vector valued frequencies, and we collect the atom-frequencies pair in a vector valued random measure as simply an analogues of (scalar valued) measure, consider some

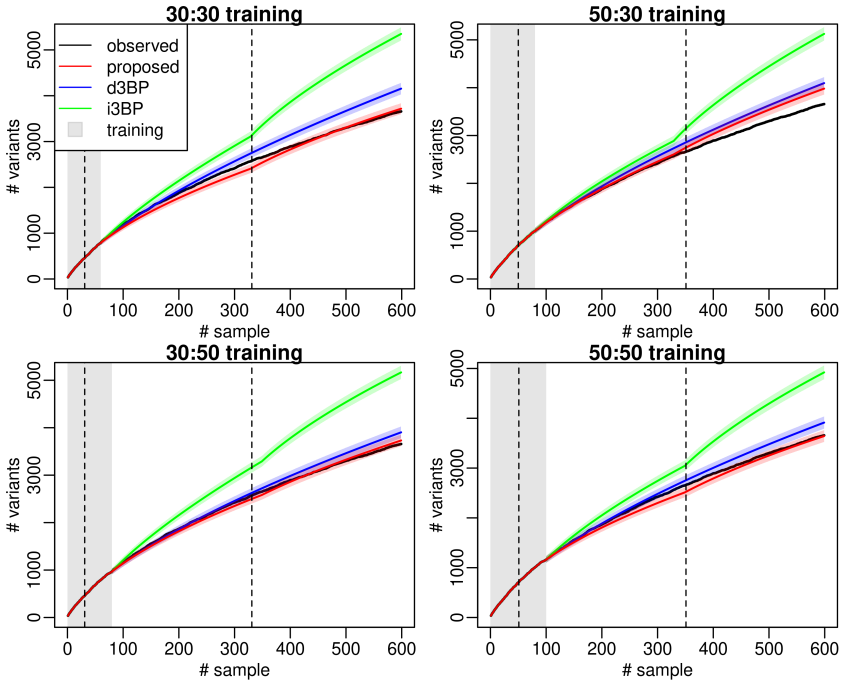

Figure S2: Discover curve under different sampling scheme in the training set. The data was sampled from a d3BP-Bernoulli process.

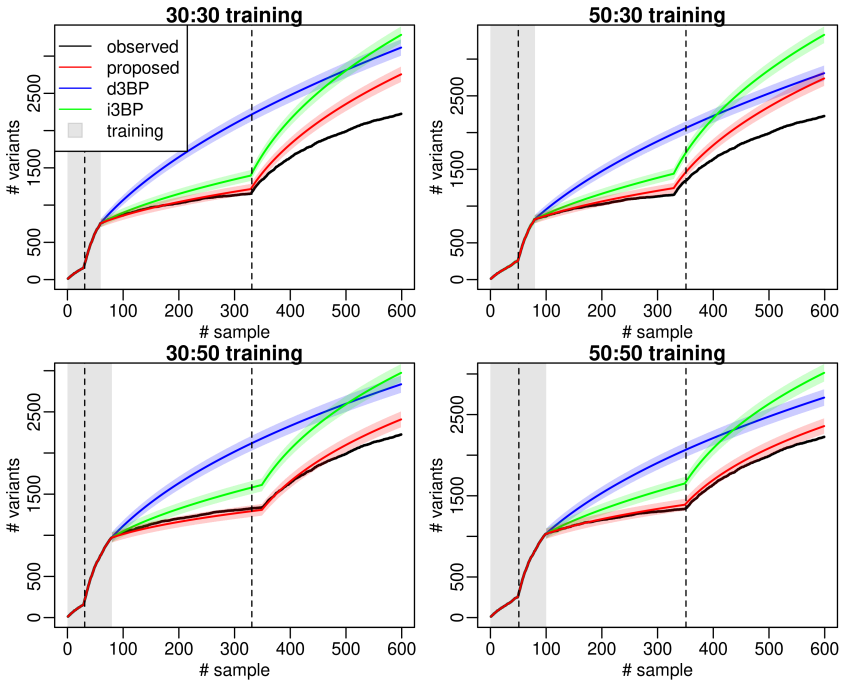

Figure S3: Discover curve under different sampling scheme in the training set. The data was sampled from the near-conjugate-Bernoulli process.

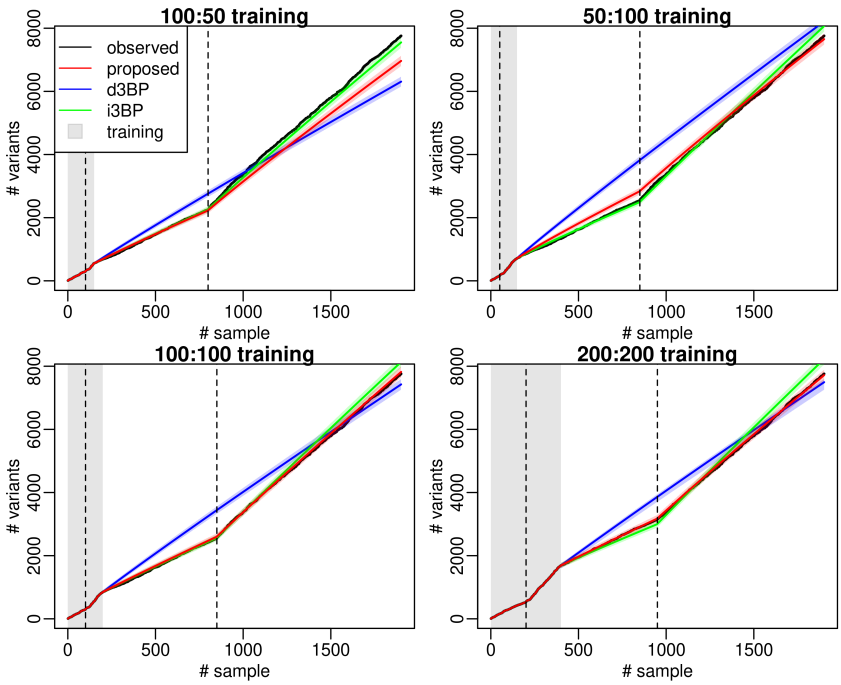

Figure S4: Discover curve under different sampling scheme in the training set in the real data example.

$\boldsymbol{\theta}_k \in \mathbb{R}_+^p$ and $x_{i,n,k} \in \{0, 1, \dots\}_{i=1}^q$ for a finite p, q, for K atoms that can be infinite:

$$\boldsymbol{\Theta} = \sum_{k=1}^{K} \boldsymbol{\theta}_k \delta_{\psi_k}$$

$$X_{i,n} = \sum_{k=1}^{K} x_{i,n,k} \delta_{\psi_k}$$

s.t. we have $\boldsymbol{x}_{n,k} \sim H(\Pi_i dx_i | \boldsymbol{\theta})$

Consider the ordinary components is a Poisson process with rate measure ν. There are some requirements for this construction, we wish to have first:

$$\nu(\mathbb{R}_+^p) = \infty$$

We also want at each coordinate of the feature vector, we see finitely many features when only finitely sampled.

For coordinate $1 \leq j \leq p$, we want:

$$\sum_{x_j=1}^{\infty} \sum_{\boldsymbol{x}_{-j}=0}^{\infty} \nu_{\boldsymbol{x}}(\mathbb{R}_+^p) < \infty \text{ for } \nu_x = \nu(d\boldsymbol{\theta}) h(\boldsymbol{x}|\theta)$$

Equivalently, we can denote $g(x_j|\boldsymbol{\theta})$ being the marginal likelihood of x_j,

$$\sum_{x_j=1}^{\infty} \nu_{x_j}(\mathbb{R}_+^p) \text{ for } \nu_{x_j} = \nu(d\boldsymbol{\theta}) g(x_j|\boldsymbol{\theta})$$

The several method for two independent Bernoulli process, i.e. $H(x_{i,k}|\boldsymbol{\theta}) = Ber(\theta_i)$ independent, can then be viewed in a similar way, i.e. a CRM whose ordinary component is a Poisson process on the unit square.

Example 1 (Single 3BP: diagonal model)**.** *The one 3BP ignoring label can be viewed as a "diagonal measure". I.e. we generate frequencies in population 1 follow a 1-d 3BP, then dictate the frequencies in the other population for the same variant to have the same value*

$$\nu(\boldsymbol{\theta}) = \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta_1^{-1-\sigma} (1-\theta_1)^{\sigma+c-1} \delta(\theta_1 - \theta_2)$$

Example 2 (Separate two 3BPs: axis model)**.** *In the model we have two separate 3BPs, we essentially assumed all variants are "private", i.e. will only appear in one population, we can take each 3BP and have a δ measure product*

$$\nu(\boldsymbol{\theta}) = \alpha_1 \frac{\Gamma(1+c_1)}{\Gamma(1-\sigma_1)\Gamma(c_1+\sigma_1)} \theta_1^{-1-\sigma_1}(1-\theta_1)^{\sigma_1+c_1-1}\delta(\theta_2) + \alpha_2 \frac{\Gamma(1+c_2)}{\Gamma(1-\sigma_2)\Gamma(c_2+\sigma_2)} \theta_2^{-1-\sigma_2}(1-\theta_2)^{\sigma_2+c_2-1}\delta(\theta_1)$$

Example 3 (Hierarchical-3BP by Masoero et al. [2021])**.** *This model was proposed by Masoero et al. [2021] as a prior of Bernoulli process. In their formulation, $f_j(\theta_j|\theta)$ takes a Beta form, that is*

$$\mu(d\boldsymbol{\theta}) = \int_\Theta \prod_{j=1}^2 \mathrm{Beta}(\theta_j|a_j\theta, b_j(1-\theta))\nu(d\theta)$$
$$\nu(d\boldsymbol{\theta}) \sim 3BP$$

The infinite mass can be shown by first applying Tonelli theorem and then observing that the integral on θ_j's are factorized and the overall integral is the integral of 3BP mass. The finite expectation can be shown by evaluating the expectations

$$\begin{aligned}
\iint \theta_1 \mu(d\boldsymbol{\theta}) &= \iint \theta_1 \int_\Theta \prod_{j=1}^2 \mathrm{Beta}(\theta_j|a_j\theta, b_j(1-\theta))\nu(d\theta)d\boldsymbol{\theta} \\
&= \int_\Theta \iint \theta_1 \prod_{j=1}^2 \mathrm{Beta}(\theta_j|a_j\theta, b_j(1-\theta))d\boldsymbol{\theta}\nu(d\theta) \\
&= \int_\Theta \int \theta_1 \mathrm{Beta}(\theta_1|a_1\theta, b_1(1-\theta))d\theta_1\nu(d\theta) \\
&= \int_0^1 \frac{a_1\theta}{b_1+(a_1-b_1)\theta}\nu(d\theta)
\end{aligned}$$

For any $a_1, b_1 > 0$ the integral is finite since $\theta\nu(d\theta)$ is proper Beta distribution up to a constant scaling while the denominator $b_1 + (a_1 - b_1)\theta$ is bounded.

Importantly we show our proposed method is proper

Proposition S4 (near conjugaet prior is proper)**.** *For the near conjugate prior with rage measure $\nu(\boldsymbol{\theta}) = \gamma(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}$. When $\gamma, c_1, c_2 > 0$, $\sigma_1, \sigma_2 > 0$ and $\alpha \in [\sigma_1+\sigma_2, \sigma_1+\sigma_2+1)$, the prior satisfies our desiderata of infinite mass and finite expectation on both axis, i.e. $\iint_{[0,1]^2}\nu(\boldsymbol{\theta}) = \infty$ while $\iint_{[0,1]^2}\theta_i\nu(d\boldsymbol{\theta}) < \infty$ for $i = 1, 2$.*

Proof. We take a small $\epsilon > 0$. Define the set $A = [0, \epsilon]^2$. Observe that we have $\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}$ being a (unormalized) independent Beta densities. We have

$$\begin{aligned}
\iint_{[0,1]^2}\nu(\boldsymbol{\theta}) &= \int_A \nu(\boldsymbol{\theta}) + \int_{A^c}\nu(\boldsymbol{\theta}) \\
&= \int_A \gamma(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta} \\
&+ \int I_{A^c}(\boldsymbol{\theta})\gamma(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta}
\end{aligned}$$

Consider the second term, we have $I_{A^c}(\boldsymbol{\theta})\gamma(\theta_1+\theta_2)^{-\alpha} \le \gamma\epsilon^{-\alpha} < \infty$, thus

$$\begin{aligned}
&\int I_{A^c}(\boldsymbol{\theta})\gamma(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta} \\
=&(B)(\sigma_1, c_1)(B)(\sigma_2, c_2)\mathbb{E}I_{A^c}(X+Y)^{-\alpha} \\
\le&\mathrm{B}(\sigma_1, c_1)\mathrm{B}(\sigma_2, c_2)\gamma\epsilon^{-\alpha} < \infty
\end{aligned}$$

where B is the beta function and the first equal is by writing the integral as an expectation of two independent Beta distributions.

Similarly for the expectation requirement, without lose of generality we consider θ_1, as our requirement and the rate measure is symmetric

$$\begin{aligned}
\iint_{[0,1]^2}\theta_1\nu(\boldsymbol{\theta}) &= \int_A \theta_1\nu(\boldsymbol{\theta}) + \int_{A^c}\theta_1\nu(\boldsymbol{\theta}) \\
&= \int_A \gamma\theta_1(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta} \\
&+ \int I_{A^c}(\boldsymbol{\theta})\gamma\theta_1(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta}
\end{aligned}$$

Consider the second term, we have $I_{A^c}(\boldsymbol{\theta})\gamma\theta_1(\theta_1+\theta_2)^{-\alpha} \leq \gamma\epsilon^{-\alpha} < \infty$ as well, thus use the same argument, we have

$$\int I_{A^c}(\boldsymbol{\theta})\gamma\theta_1(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}(1-\theta_1)^{c_1-1}(1-\theta_2)^{c_2-1}d\boldsymbol{\theta} < \infty$$

Thus, it suffice to check under the parameter range, we have $\int_A \nu(\boldsymbol{\theta}) = \infty$ and $\int_A \theta_i\nu(\boldsymbol{\theta}) < \infty$. Observe that for $c_i > 0$, we have $(1-\theta_1)^{c_1-1}, (1-\theta_2)^{c_2-1}$ being bounded on A, thus it suffice to check $\int_A(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}d\boldsymbol{\theta} = \infty$ and $\int_A \theta_1(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}d\boldsymbol{\theta} < \infty$. For the first requirement, consider the range where $\theta_1 \geq \theta_2$, we have

$$\begin{aligned}
\int_A (\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}d\boldsymbol{\theta} &\geq 2^{-\alpha}\int_{A\cap\{\theta_1\geq\theta_2\}} \theta_1^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}d\boldsymbol{\theta} \\
&\geq 2^{-\alpha}\int_0^{\epsilon}\theta_1^{-\sigma_2-1}\int_0^{\theta_1}\theta_2^{\sigma_2-1}d\theta_2 d\theta_1 \\
&= \frac{2^{-\alpha}}{\sigma_2}\int_0^{\epsilon}\theta_1^{-1}d\theta_1 \\
&= \infty
\end{aligned}$$

We used $\alpha \geq \sigma_1+\sigma_2$ in the first inequality. Suppose $\alpha \leq \sigma_1+\sigma_2+1-\delta$ for an arbitrary $\delta > 0$, consider the expectation

$$\begin{aligned}
\int_A \theta_1(\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1-1}\theta_2^{\sigma_2-1}d\boldsymbol{\theta} &= \int_A (\theta_1+\theta_2)^{-\alpha}\theta_1^{\sigma_1}\theta_2^{\sigma_2-1}d\boldsymbol{\theta} \\
&\leq \int_A (\theta_1+\theta_2)^{-\sigma_1-1+\delta/2}\theta_1^{\sigma_1}(\theta_1+\theta_2)^{-\sigma_2+\delta/2}\theta_2^{\sigma_2-1} \\
&\leq \int_A \theta_1^{-\sigma_1-1+\delta/2}\theta_1^{\sigma_1}\theta_2^{-\sigma_2+\delta/2}\theta_2^{\sigma_2-1} \\
&= \int_0^1 \theta_1^{\delta/2-1}d\theta_1\int_0^1\theta_2^{\delta/2-1}d\theta_2 \\
&< \infty
\end{aligned}$$

□

S8 Product measure in prior process

We can show that if the prior process is a product measure (not having δ measure) and the likelihood factorize in both data and parameters, then the prior cannot be properly normalized. This result tells us that, in the two population problem, atoms' behavior in the first population must contain information about their behavior in the other population, either through likelihood or through prior.

Theorem S1 (Incompatibility between product measure, infinite mass and finite expectations when likelihood factorizes)**.** *If the likelihood process are factorized for all dimensions (i.e. $g(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_j g_j(x_j|\theta_j)$), then the prior process (rate measure $\nu(d\boldsymbol{\theta})$) can not have all three properties: (1) being product measure (2) infinite mass (3) finite positive expectations on all coordinates*

Proof. By assumption, we have (3) means:

$$0 < \sum_{x_j=1}^{\infty}\nu_{x_j}(\mathbb{R}_+^p) \text{ for } \nu_{x_j} = \nu(d\boldsymbol{\theta})g(x_j|\theta_j) = \int \nu(d\boldsymbol{\theta})(1-g(0|\theta_j)) < \infty$$

Suppose we have (1) and (2). By (1) we have $\nu(d\boldsymbol{\theta}) = \prod_j \nu_j(d\theta_j)$, i.e. a product measure, then by (2), we have there exist at least one coordinate k, such that $\int \nu_k(d\theta_k) = \infty$ and all other coordinates are positive. We show (3) is not true for coordinate $j \neq k$:

$$\int \nu(d\boldsymbol{\theta})g(0|\theta_j) = \prod_{i\neq k,j}\int\nu_i(d\theta_i) \times \int\nu_k(d\theta_k) \times \int(1-g(0|\theta_j))\nu_j(d\theta_j)$$

The first factor is positive and second is ∞, while the last one is non-negative, thus the expectation can only be either 0 or ∞.

Suppose we have (1) and (3), take coordinate j, we have:

$$0 < \int \nu(d\boldsymbol{\theta})(1 - g(0|\theta_j)) = \prod_{i \neq j} \int \nu_i(d\theta_i) \times \int (1 - g(0|\theta_j))\nu_j(d\theta_j) < \infty$$

We must have for all $i \neq j$ $\int \nu_i(d\theta_i) < \infty$, then we can take $k \neq j$, which will imply for all $i \neq k$ $\int \nu_i(d\theta_i) < \infty$, i.e. for all coordinate we must have $\int \nu_i(d\theta_i) < \infty$. Since (1) implies $\nu(d\boldsymbol{\theta})$ is a product measure, we must have $\int \nu(d\boldsymbol{\theta}) < \infty$.

Suppose we have (2) and (3), use the above two results, by contradiction we cannot have (1). □

Note that the problem only happen when the likelihood factorize (in both data and parameters), which is stronger than independent in data. Independent is not enough to cause the problem. One simple counter example is when one coordinates actually contains information of the parameter that determines the behavior of the second coordinates

Example 4 (Independent $\mathrm{Bernoulli}(\theta_1\theta_2)$)**.** *Suppose we have* $g(\boldsymbol{x}|\boldsymbol{\theta}) = (\theta_1\theta_2)^{x_1+x_2}(1-\theta_1\theta_2)^{1-x_1+1-x_2}$*, i.e.* $x_1, x_2 \sim \mathrm{Bernoulli}(\theta_1\theta_2)$ *independently. The product measure of 3BP's (denote as* $\nu(\theta)$*) are properly normalized. Because both have finite expectations.*

$$\iint_{[0,1]^2} g(x_i = 1|\boldsymbol{\theta})\nu(d\theta_1)\nu(d\theta_2) = \iint_{[0,1]^2} \theta_1\theta_2\nu(d\theta_1)\nu(d\theta_2) = \int_0^1 \theta_1\nu(d\theta_1) \int_0^1 \theta_2\nu(d\theta_2) < \infty$$