Regularized sequence-context mutational trees capture variation in mutation rates across the human genome

Christopher J. Adams[1], Mitchell Conery[1], Benjamin J. Auerbach[1], Shane T. Jensen[2], Iain Mathieson[3], Benjamin F. Voight[3,4,5]

Author affiliations:

1. Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2. Department of Statistics and Data Science, The Wharton School at the University of Pennsylvania, Philadelphia, USA
3. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
4. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
5. Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Correspondence to:

Benjamin F. Voight, PhD
Associate Professor of Systems Pharmacology and Translational Therapeutics
Associate Professor of Genetics
University of Pennsylvania - Perelman School of Medicine
3400 Civic Center Boulevard
10-126 Smilow Center for Translational Research
Philadelphia, PA 19104
Email: bvoight@pennmedicine.upenn.edu

ABSTRACT

Germline mutation is the mechanism by which genetic variation in a population is created. Inferences derived from mutation rate models are fundamental to many population genetics inference methods. Previous models have demonstrated that nucleotides flanking polymorphic sites – the local sequence context – explain variation in the probability that a site is polymorphic. However, limitations to these models exist as the size of the local sequence context window expands. These include a lack of robustness to data sparsity at typical sample sizes, lack of regularization to generate parsimonious models and lack of quantified uncertainty in estimated rates to facilitate comparison between models. To address these limitations, we developed Baymer, a regularized Bayesian hierarchical tree model that captures the heterogeneous effect of sequence contexts on polymorphism probabilities. Baymer implements an adaptive Metropolis-within-Gibbs Markov Chain Monte Carlo sampling scheme to estimate the posterior distributions of sequence-context based probabilities that a site is polymorphic. We show that Baymer accurately infers polymorphism probabilities and well-calibrated posterior distributions, robustly handles data sparsity, appropriately regularizes to return parsimonious models, and scales computationally at least up to 9-mer context windows. We demonstrate application of Baymer in two ways – first, identifying differences in polymorphism probabilities between continental populations in the 1000 Genomes Phase 3 dataset, and second, in a sparse data setting to examine the use of polymorphism models as a proxy for *de novo* mutation probabilities as a function of variant age, sequence context window size, and demographic history. We find a shared context-dependent mutation rate architecture underlying our models, enabling a transfer-learning inspired strategy for modeling germline mutations. In summary, Baymer is an accurate polymorphism probability estimation algorithm that automatically adapts to data sparsity at different sequence context levels, thereby making efficient use of the available data.

INTRODUCTION

Germline mutations are the primary source of genetic variation between and within species. Quantifying where, what type, and how frequently mutations arise is therefore of fundamental importance to population genetic inference and complex trait studies. Better estimates of mutation rates improve tools designed to quantify population divergence times[1], demographic history[2], and the effects of background selection[3]. Moreover, models for the underlying *de novo* mutation rate from which burden of mutations can be statistically assessed have enabled discovery of genes[4,5] and non-coding sequences[6,7] contributing to complex disease[4,5,8,9].

Our working hypothesis is that there exists an underlying structure to the context-dependent effects that shape the mutation rate. Here we focus on polymorphism probabilities as a proxy for the mutation rate that we hypothesize share the same context-dependent architecture subject to genetic drift, demography, selection, biased gene conversion, or additional phenomenon that operate across population history. The frequency of polymorphisms varies widely across the genome[10] and correlates with several genomic features[11–13], with new mutations caused by both exogenous and endogenous sources[14]. There is considerable evidence to suggest that local nucleotide context directly relates to the probability that a nucleotide mutates. A classic example of this is the ~14-fold higher rate of C>T transitions at methylated CpG sites, owing to spontaneous deamination of 5-methylcytosine[15–17]. Long tracts of low-complexity DNA have higher mutation rates, which is hypothesized to be the result of slippage of DNA polymerase during replication[18]. This prior work suggests that local sequence context is integral to understanding variation in polymorphism rates across the genome, and that the most predictive models will be best positioned to guide elucidation of the underlying mutational mechanisms.

Our previous work demonstrated that a sequence context window of seven nucleotides (i.e., ‘7-mer’) provided a superior model to explain patterns of genetic variation relative to smaller windows that are commonly used (e.g., 3-mers)[19]. While an advance, this model was fundamentally limited for three reasons: *scalability, regularization,* and *uncertainty*. First, the size of the model – which increases by a factor of four for each additional nucleotide added – presents intrinsic limits both computationally and in terms of statistical power. Second, while it is straightforward to assume that every sequence context is meaningful, a more parsimonious model – informed by biological intuition – might be that only a subset of contexts contributes meaningfully to the observed variation in data. This is particularly important for inference of somatic and *de novo* mutation rates or in other data-sparse situations. Finally, while our previous model provided a point estimate of the mean polymorphism probability, it did not immediately emit uncertainty resulting from multinomial variance and heterogeneity in larger sequence contexts. As sequence context sizes are expanded, there is functionally less data and thus more uncertainty in estimates, making point estimates even more unreliable. Quantifying uncertainty is also required for detecting differences in probabilities across models, for example when comparing differences in rates across populations[20–22] or at functional genomic features[23]. Ideally, a method should scale the inferred context length proportional to the amount of data and the biological signal that may be present within that data while providing uncertainty in estimated parameters and underlying probabilities.

Previous work has sought to address these challenges, though methods introduced to date do not address all limitations simultaneously. Sparsity and scalability have been tackled through a deep-learning framework[24] as well as an IUPAC-motif-based clustering approach[25] which modeled polymorphism probabilities up through 9-mers. Another method explored polymorphism probabilities up through 7-mers using DNA shape covariates to reduce the parameter space[26]. All three methods are robust and effective at measuring point estimates of polymorphism probabilities in expanded sequence contexts, however none explicitly estimate the uncertainty of these parameters. Finally, the CIPI model[27] is a Bayesian method that addresses these issues, but focuses on applications with smaller context-window motifs (5-mer) in variant settings with fewer mutation events (e.g., somatic mutations in cancer or mutations in viral genomes) and is not obviously scalable computationally to larger size context windows and sizes of contemporary population genomics data sets in humans (e.g., hundreds of millions of polymorphic sites).

Here, we develop a method that addresses all three limitations in the original model. We construct a Bayesian tree-based method that integrates sequence context window size, handles sparse data, and captures uncertainty in estimates of mutation probability via the posterior distribution. We apply our approach in two ways. First, we quantify differences in polymorphism probabilities between continental populations and place bounds on the effect sizes of potential undescribed context-dependent differences in the 1000 Genomes dataset[28]. Second, we explore the use of polymorphism datasets to predict *de novo* mutations. We measure the effect of population history, variant age, and sequence context size on model performance with the aim of generating a meaningful proxy to estimate the germline mutation rate.

RESULTS

A tree-based sequence-context model captures variation in polymorphism probabilities

We began by developing a model to describe the hierarchical relationship of sequence context dependencies over increasing window sizes. We structured this as a rooted, tree-based graph, where each type of substitution class is represented distinctly (**Fig. 1A**). Each level of the tree represents an increasing window size of sequence considered, alternating between incorporating nucleotides to the window on the 3′ end for even-sized contexts and on the 5′ end for odd-sized contexts. We fold over reverse complementary contexts to reduce the parameter count (**Methods**). To ease readability, we denote each mutation with the sequence context, the nucleotide in scope bolded, and the polymorphism indicated with an arrow (e.g T**C**C>T represents the polymorphism where the bolded cytosine has become a thymine). Each non-root edge represents the log-transformed, multiplicative shift in polymorphism probability captured by expanding sequence context. The root edge corresponds to an estimated base polymorphism probability for a given mutation type. For a given sequence context, each node in the tree represents the probability of observing a polymorphic site in the central nucleotide (referred to hereafter as polymorphism probability), and is the product of all edges, starting from the root that leads to the node (**Fig. 1B**). As our previous work has shown for a specific level of sequence context, the distribution of observed counts for each sequence context can be modelled via independent multinomial distributions[19] facilitating likelihood calculation. The resulting multinomial probability vector corresponds to the combination of individual polymorphism probability estimates across each mutation type tree for each sequence context (**Methods**).

Within the model, we incorporate two features essential for downstream applications when comparing the outputs of competing models. First, we employ a Bayesian formulation which generates posterior distributions for polymorphism probabilities (**Methods**). This approach naturally estimates parameter uncertainty which is essential for comparison of rates across different tabulated models. Second, we incorporate regularization in the parameter estimation procedure for tree edges. Previous sequence context models estimated parameters for all edges of the tree (ϕ), meaning that all values of were effectively non-zero. However, our previous work suggested that perhaps only a fraction of edges meaningfully contribute information[19]. Hypothesizing that only a subset of edges is informative for the polymorphism probability shifts, we regularize our tree model by incorporating a *spike-and-slab* prior on the ϕ parameters[29]. We tune the model such that the slab is favored when the evidence suggests a shift greater than 10% for a given context level (**Fig. 1C**). This value was choosen weighing the stability of model convergence with the goal of inferring the largest possible effects.

Because the posterior distribution is not analytically tractable, we implemented an adaptive Metropolis-within-Gibbs Markov Chain Monte Carlo (MCMC) sampling scheme[30] to sample from and thereby estimate the posterior distribution of this model. To further aid in convergence and enforce intermediate nodes to have informative polymorphism probabilities, we estimated parameters of the model level-by-level rather than all simultaneously, leveraging the conditional dependency structure of the hierarchical tree. Under this set-up, the unseen higher-order layers are assigned $\phi_{a,b} = 0$ shifts until their level has been sampled. We embedded this model and sampling scheme into software (named Baymer) for further testing and applications.

Evaluation of the model demonstrates robust inference of the underlying rates with uncertainty

A key feature of Baymer is that it estimates posterior distributions for each parameter, allowing for uncertainty in the probabilities of polymorphism at each sequence context. To evaluate the coverage of the estimated posterior probabilities, we used simulations to assess how often our posterior distribution captures simulated values. Using a pre-specified polymorphism probability table, we tested how frequently polymorphism probabilities estimated by Baymer captured the true value for each sequence context (**Methods**). We found that across all sequence context sizes, 89%, 93%, and 97% of context simulations contained the true polymorphism probability in the 90%, 95%, and 99% credible intervals, respectively (**Methods, Supplementary Table 1**).

A second important feature is that regularization is embedded into the method, allowing for the creation of parsimonious models that capture most of the information with the fewest non-zero parameters. This part is critical to address cases where the amount of data is not large and limits power, or when considering larger windows of sequence context that are rare and/or uninformative. If robustly calibrated, we would expect probabilities inferred in a holdout set to strongly correlate with those estimated during a test phase (i.e., minimal overfitting). To evaluate the robustness of the inferred rates, we partitioned the human genome reference into two sets - even and odd base-pairs - and used SNPs of allele count 2 or greater observed in the gnomAD[31] non-Finnish European (NFE) collection to independently train models (**Methods**). We compared the concordance of probabilities for models with sequence context windows up to 4 flanking nucleotides on either side (i.e., a 9-mer model) using the maximum likelihood estimate approach[19] and Baymer (**Supplementary Fig. 1**). For each comparison, in addition to the Spearman correlation, we also calculated the root mean squared perpendicular error (RMSPE) from each point to the x-y axis, as a measure of the tightness of the distribution from the true, shared value (**Methods**). The maximum likelihood estimates of polymorphism probabilities (**Fig. 2A,** Spearman correlation ρ = 0.915; RMSPE = 0.117) were less correlated and considerably less tightly distributed than those for Baymer-derived models (**Fig. 2B,** ρ = 0.990; RMSPE = 0.035). This result occurred even after omitting ~16,000 sequence contexts with zero mutations in either dataset (odd and even base pairs) from the maximum likelihood model comparison, rendering practical use of large swaths of the model useless due to substantial overfitting at the 9-mer level. If zero-mutation contexts omitted from the maximum likelihood model were included, the correlations would perform considerably worse (**Methods, Supplementary Fig. 1D,** ρ = 0.876; RMSPE = 0.744).

We next sought to evaluate the transferability of inferred models between experimental collections; while internally consistent, the above procedure could simply reflect data set specific biases[32]. For this, we compared non-admixed, non-Finnish European (EUR) samples obtained from the 1000 Genomes (1KG) Project (re-sequenced by the New York Genome Center)[33] with the gnomAD NFE sample described above. As before, we split the data into even and odd base pairs but also applied a variant down-sampling procedure to match total variant count and site-frequency spectrum between both sets **(Methods)**. By comparing variants found in the even base-pair genome of gnomAD with the odd base-pair genome of 1KG, this strategy ensures no variation overlapped between data sets. We observed that the probabilities estimated from both sample sets were strongly correlated (ρ = 0.981; RMSPE = 0.064; **Fig. 2C**) though were slightly weaker than the correlations from each internal comparison and fit less tightly (gnomAD ρ = 0.990; RMSPE = 0.035; **Fig. 2B**; 1KG ρ = 0.986; RMSPE = 0.042; **Supplementary Fig. 2**). This result demonstrates that some additional between-sample variation may exist, but that Baymer infers probabilities of polymorphism that are broadly consistent with one another, supporting the notion of model transferability.

We next aimed to quantify how well the model selects meaningful context features. We expected more proximal bases to the focal site to have a greater impact on polymorphism probabilities for two reasons, (i) due to data richness, and (ii) that proximity to the polymorphic site would suggest more direct impacts on mutability, e.g., the CpG context. Baymer estimates the fraction of posterior samples in the slab, implying a non-zero effect on polymorphism probabilities, and in the spike, which implies no effect. Thus, the probability of an edge being included in the slab is the equivalent of the posterior inclusion probability (PIP) for our model. Consistent with expectation, the fraction of sequence contexts with a PIP > 0.95 monotonically decreases as the sequence context size is increased (**Fig. 2D**).

Larger contexts best explain patterns of variation genome-wide

We note that over 61% of all sequence contexts with a PIP > 0.95 are found in the 8-mer and 9-mer levels of our model of polymorphism observed in the gnomAD NFE data. While fewer than 2% of 9-mer sequence contexts meaningfully impact the final estimates, they still account for the most total absolute contexts (7189 total contexts > 0.95 PIP). This observation holds even after filters for data sparsity (**Methods, Fig. 2E**). This implies a considerable impact on polymorphism probabilities in extended sequence contexts, consistent with previous work[19,23–25]. This general trend is similarly consistent across mutation types (**Fig. 2F**). We thus evaluated the overall improvement in likelihood by expanding window sizes up to 9-mers. Compared to lower context models (e.g., 3-mer, 5-mer, or 7-mer) on holdout data, 9-mer Baymer models substantially improved the likelihood and best fit to the data **(Methods, Supplementary Table 2)**.

Frequency of polymorphism across populations do not differ substantially across levels of sequence context

Prior work has centered around evaluating whether mutation rates have changed over evolutionary time by evaluating differences in the proportions of sequence-context-dependent polymorphism between human populations[21,22,34–36]. To determine whether polymorphism probabilities differ across human populations, we analyzed individuals from the NYGC resequencing of 1KG Phase III representing continental European, African, East Asian, and South Asian groups. We extracted variants private to these continental groups, down-sampling to match site-frequency spectra bins and overall sample sizes (**Methods**). We then applied Baymer to each individual dataset to model probabilities up to a 9-mer window of sequence context. We compared estimates of polymorphism probabilities in each population by assessing the degree to which the posterior distribution of each population's model parameters overlapped. The fraction overlap of each distribution is a proxy for the probability that the underlying polymorphism probabilities are the same. Due to the implicit tree structure of sequence context models, polymorphism probability shifts in edges will affect all edges downstream of the context in question. Therefore, we identified edges where both the estimated polymorphism probability and the immediate shift, $\phi_{a,b}^{m}$, were both considered very likely to be different.

Specifically, we identified contexts whose polymorphism probabilities and shifts both overlapped less than 1% in pairwise comparisons between the four populations (**Supplementary Table 3**). This included all the most notable previously reported 3-mer shifts across continental groups, including the increase in T**C**C>C mutations found in European relative to Non-European ancestry populations[20–22,34,36]. We also discovered a nested context within the classic T**C**C>T context, namely **C**C>T, as being very likely to differ between populations. This could simply be a trickledown signal from the T**C**C>C, A**C**C>C, and C**C**C>C effects implicated by Harris[21]. However, all four contexts from this 3-mer family have evidence of elevated polymorphisms probabilities in Europeans vs Africans, which might suggest a more parsimonious explanation of a second contributing signal, possibly with the same underlying mechanism.

We next focused on the remainder of 3-mer and wider extended sequence contexts (**Table 1**). While a handful of such sequence contexts have been implicated[34], these results are confounded by batch effects in the original 1KG sequencing data[37]. In our results, we observed the presence of nucleotide repeats, e.g., TA / CG dinucleotides; poly-C / poly-A in several of the divergent contexts, which could be explained by polymerase slippage[18].

While the population-specific polymorphism probabilities estimated and polymorphism counts are identical between each pairwise comparison and thus correlated, we still note that 15/28 pairwise differences are specific to a single continental group. Of these, only the two canonical European context mutation differences (T**C**C>T and T**C**T>T) are in 3-mer contexts, otherwise all are found in 5-mer and greater mer-levels. In South Asian samples, we find that the mean CT**A**TA>T polymorphism probabilities are approximately 1.6 times higher than the remaining populations and in Africans TAT**A**TATC>G is approximately 1.9 times higher. The largest population-specific effect was discovered in East Asians where ATA**C**CTC>A polymorphism probabilities are roughly 2.7 times higher than in European, African, or South Asian models. None of these effects have been explicitly documented before.

Taken collectively, we observed relatively few instances of shifts that were quantifiably different across continental groups, and those that were observed were largely confined to relatively small windows of context where we might have anticipated well powered tests (e.g., 3- and 5-mers). To quantify the power of our procedure and the sample size necessary to identify true shifts in polymorphism probabilities, we performed simulations where true effect differences were ‘spiked-in’ between two populations over a range of weak to stronger effects and across a sampling of different sequence contexts (**Methods**). Shifts for this experiment are defined as the natural log of the polymorphism probabilities ratio (NLPPR) between each simulated population. This allowed us to construct credible sets of effects that we were reasonably well powered (>80%) to discover (**Table 2**). Unsurprisingly, the power scaled proportional to the number of context instances, simulated mutations in the dataset, and the size of the spiked-in differences (**Supplementary Figure 3**). Notably, extremely subtle shifts (NLPPR <= 0.01; 0.99 – 1.01 fold change) were not detectable at any sequence context size. On the opposite side of the spectrum, we found that we were reasonably powered to identify shift differences where NLPPR > 1.0 (fold decrease <= 0.37 or fold increase >= 2.72) up through 5-mers and in 6-mers with large sample sizes. For reference, the T**C**C>T polymorphism has an NLPPR = 0.291 (~1.34 fold increase) – the largest difference of any 3-mer by our calculation.

In contrast, our experiment had essentially no power to discover 9-mer shifts and extremely limited power for 8-mers, even for large shifts. Thus, there may exist large shifts at these sizes that we could not reliably capture. These results are consistent with our comparisons in the real data (**Table 1**), as only differences within the detectable range at each mer-level were implicated.

These power calculations suggest that, given the experiment we performed grouping all mutations together (agnostic to allele frequency or age, see **Discussion**), if any 3-mer differences greater than the T**C**C>T shift exist, we would have discovered these effects for a broad range of modest to very strong effects across a range of sequence contexts window sizes. This effectively sets bounds on the differences possible for this analysis scheme in this data.

A sequence context model that captures variability in *de novo* mutational rates

Given its formulation in handling data sparsity, we next sought to apply Baymer to develop a model that best captures rates of *de novo* mutations across the genome. We took advantage of a recent collection of 2,976 WGS Icelandic trios that identified 200,435 *de novo* events[38] and, analogous to the above, we partitioned *de novo* variants into even (for training) and odd (for testing) base pairs. We observed substantial improvement in the overall likelihood in the testing set for 5-mer size windows compared to 3-mers (3-mer vs 5-mer, delta-LL = 2,144), but only minimal improvement for increasing windows sizes further (5-mer vs 9-mer, delta-LL = 265). Indeed, Baymer did not select any sequence context feature beyond the 5-mer level with PIP > 0.95. This is not unexpected given our approach to regularization, as the number of events in larger sequence contexts is increasingly sparse, it is desirable to only include informative contexts to avoid overfitting.

We next used Baymer to improve upon this baseline model. Previous work has demonstrated that inference of *de novo* mutational probabilities can be captured via rare variant polymorphism data obtained from population sets as a proxy[23]. We hypothesized that a partitioned set of polymorphism data based on: (i) larger sample sizes that (ii) closely matched the ancestry of the *de novo* set and (iii) focused on rare variants as a proxy to capture the most recent mutation events would generate the most transferrable model and robust rate estimates. To build variant partitions, we used variant call set data from gnomAD, focused on either a population-matched proxy (i.e., NFE, the non-Finnish European subset) or variant calls from all samples in gnomAD regardless of ancestry (i.e., ALL). For each of these, we created three partitions focused (i) exclusively on variants with one allele count (i.e., singletons; labeled POP-1), (ii) exclusively on variants with two allele counts (i.e., doubletons; labeled POP-2), and (iii) variants with allele count of two or greater (labeled POP-2+). Beyond this, we also identified a set of putatively derived substitutions in the human lineage by comparing the GRCh38 human reference genome with ancestral sequences obtained from primates[39].

We applied Baymer to each variant set independently, comparing the likelihoods of each model to explain rates of *de novo* mutation in the test set after downscaling probabilities proportional to the sample size. First, we observed that for 3-mer sequence context models, the set of variants obtained from the *de novo* training set outperformed all other models despite there being 102 to 1,377 times fewer variants contributing to them than the polymorphism datasets (**Fig. 3A, Supplementary Table 4**). In contrast, for larger windows of context (i.e., 7-mer and 9-mer), several of the polymorphism partitions explained the data better than one trained directly from *de novo* events. This result indicates that increased sample size is required to detect meaningful shifts in polymorphism probabilities in larger sequence context windows.

Despite evidence to suggest singleton datasets should best recapitulate *de novo* variation[4,23,31], we were surprised to observe that models that trained exclusively on singletons and ALL-2 performed considerably worse than the rest across all windows of sequence context (**Fig. 3A, Supplementary Table 4**). This is particularly surprising for larger windows of sequence context, given the prior intuition that larger numbers of variants would have provided better rate estimates. Although we only used variants that passed gnomAD quality control checks, this filter still included a large proportion of variants with a negative log-odds ratio of being a true variant (AS_VQSLOD < 0; **Supplementary Fig. 4**). This pattern was also evident for other variant allele counts but were most striking in singletons and the ALL-2 variant groups. Stricter quality filters (AS_VQSLOD > 5-10) considerably improved model performance, but still did not surpass the *de novo* training model at the 3-mer level (**Supplementary Table 4**). Our NFE singleton Baymer model trained on the strictest quality filter tested (AS_VQSLOD > 10) nearly equaled our best performing model, NFE-2+, with ~ 1/30th the number of variants, but came up just short. In summary, we observed that training from a population matched sample which excluded singletons, NFE-2+, best predicted rates of *de novo* mutations in 5-mer or larger contexts, better than training on *de novo* events directly.

Next, we sought to determine which sample set best modelled the *de novo* test set adjusting for the total number of variants within the partition. To control for sample size differences, we down-sampled each partition to match the number of variants observed in the *de novo* training set (n=70,364) five times. After down-sampling and when considering 9-mer context models, we observed that the partitions which included NFE exclusively (noted in green, **Fig. 3B**) performed on average better than using the entirety of gnomAD, “ALL” (noted in orange in **Fig. 3B**), which included a more diverse panel of individuals within Europe (e.g., Finnish) but also beyond Europe (e.g., East and South Asian, African and African American). This is consistent with prior belief that, after controlling for the total sample size, variants that derive from samples where ancestries more closely match are the most informative.

A grafted tree approach provides superior estimates of *de novo* mutational probabilities

Given the observations that *de novo* models only outperform polymorphism-based models when either small sequence contexts are used (**Fig. 3A**) or the sample size is controlled (**Fig. 3B**), we next sought to explore a transfer learning-inspired[40] strategy to improve upon our model performance. Transfer learning has previously been employed in a sequence context modelling setting[24]. We hypothesized that regularization means that *de novo* models have reduced performance with expanded sequence contexts due to low sample sizes. Indeed, our *de novo* model did not have the power necessary to confidently (PIP > 0.95) include any non-zero shifts in sequence contexts larger than 5-mers in the model (**Fig. 4A**). The larger polymorphism datasets, however, were well-powered to detect shifts in every level of the tree **(Fig. 4A)**.

The nested tree structure of our polymorphism probability models provides a natural strategy where specific branches of the estimated trees can be interchanged, i.e., a “grafted” tree. We asked how similar estimates for edges in expanded sequence contexts are between our *de novo* model and the best-performing polymorphism model, NFE-2+. In edges in 2-mer and greater levels where the *de novo* training model is powered enough to detect shifts (PIP > 0.95), the mean posterior estimates of shifts are highly correlated **(Fig. 4B)**. This suggests a grafted tree approach is feasible, leveraging the polymorphism datasets for those edges the *de novo* model is incapable of estimating properly due to sparsity **(Fig. 4C)**. Therefore, we built a grafted tree model using 1- to 3-mer edges estimated in the *de novo* training data model, and 4- to 9-mer edges estimated using the NFE-2+ data model. The resulting combined model had a greater fit to the holdout *de novo* data than either the NFE-2+ model or *de novo* model alone **(Fig. 4D, Methods)**.

DISCUSSION

Here, we present Baymer, a Bayesian method to model mutation rate variation that computationally scales to large windows of nucleotide sequence context, robustly manages sparse data through an efficient regularization strategy, and emits posterior probabilities that capture uncertainty in estimated probabilities. Consistent with previous studies[24–26], we show that expanded sequence context models in most current human datasets are overfit with classic empirical methods but considerably improve model performance when properly regularized. As a result, this method allows for renewed evaluation of experiments that originally were statistically limited to polymorphism probability models with small sequence context windows.

We examined differences in polymorphism probabilities between the continental populations in the 1KG project. While differences in 3-mer polymorphism probabilities have been well-documented[20–22] and expansions up to 7-mers have been tested[34], both methods rely on empirical models with frequentist measures of uncertainty. Here, we expanded the search space out to 9-mer windows and leverage the uncertainty estimated in the model to directly quantify differences in these populations. We note that many of the differences discovered contain poly-nucleotide repeats. There is some prior literature on the mechanism of slippage in polymerases during replication of such sequences[18], so differential efficiencies of these enzymes across populations could conceivably result in these patterns. However, it is also very possible that artifacts from sequencing errors with differential effects across populations could explain the differences.

Despite being well-powered to identify a large range of differences in 3-mer and smaller contexts we identified very few contexts that differ with high probability between the populations tested. This implies that if large-scale population differences in the mutation spectrum do exist at these window context sizes, they are most likely comprised of numerous subtle shifts rather than a few large changes, in agreement with conclusions from prior work[22].

We also explicitly placed bounds on the magnitude of differences that could possibly exist in this dataset without being detected, quantifying what differences we can expect to be discovered given the way variants are grouped in this experiment. Even though the 1KG project is relatively small compared to current datasets, the number of sequence contexts available for modeling is dataset-independent and inherently limited by the sequence diversity of the human genome. Thus, while more polymorphism data could lead to the discovery of additional smaller shifts in the future, bigger datasets will not improve the power to detect larger shifts in this allele frequency agnostic setting. In fact, for very large samples, polymorphisms in some contexts can become saturated,[41] reducing the information content in a similar manner as overly sparse data. Thus, both to increase power and to improve modeling resolution, it will become necessary to partition the data (e.g., by allele frequency or variant age[35], or other genomic features).

It remains a challenge to disentangle the contribution of demography[20,36,42] versus changes in the underlying mutation rate on the mutation spectrum. Here, we control for the site frequency spectrum of variants included, but the next stage of this model will need to incorporate more sophisticated demographic features. Integrating Baymer-derived trees with a joint mutation spectrum and demographic history method, such as mushi[36], is a promising future direction.

Next, we asked to what degree polymorphism datasets could be used to approximate the *de novo* mutation rate. Currently, true *de novo* mutation datasets are limited in size, which place bounds on the scope of inference for adequate sequence context modeling. We demonstrate that polymorphism datasets are accurate proxies for *de novo* mutation models and largely share the same context-dependent mutability shifts, though in contrast to reports in the literature[4,23,31], the focus exclusively on singleton variants (at least, using gnomAD calls) performed poorly relative to all other considered models. Indeed, our experiment indicates that it is preferable to use germline mutation models based on large polymorphism datasets that can estimate shifts through the 9-mer level than it is to use the largest 3-mer de novo dataset, as is frequently the norm[4,5,31]. Including exclusively variants from either polymorphism data or *de novo* data was also suboptimal, however, as the best possible model we built for estimating *de novo* mutation rates used *de novo* mutations in concert with polymorphism datasets. The success of this experiment implies a general context-dependent mutability shift structure that underlies the human mutation spectrum. The similarity of the derived dataset, which in theory represents the oldest subset of variants tested, to the *de novo* variation further strengthens this argument and suggests that although there have been some well-documented small changes in context-dependent mutation rates, the general architecture remains largely conserved during modern human history.

One limitation of the model is the treatment of multi-allelic sites. Currently, multi-allelic sites are treated as separate polymorphisms which violates assumptions of the multinomial model, where only one outcome is possible for each locus. When we excluded multi-allelic sites, we observed biases in the rates of **C**pG>A and **C**pG>G mutations, which are disproportionately filtered as a side-effect of sharing the same sequence contexts with **C**pG>T mutations. A more nuanced approach that models multiallelic and biallelic sites separately and then integrates jointly would deal with this issue, though multiple mutations at the same nucleotide position with the same allele change would require additional effort[43].

Finally, although we can identify regions of the tree where polymorphism probabilities diverge and thus infer critical points in the tree, this model is tailored towards polymorphism probability estimation rather than explicitly for motif discovery[27]. Our objective is to estimate polymorphism probabilities rather than finding those contexts with the largest effect sizes. Adding one nucleotide at a time pseudo-symmetrically for tree generation reduces the computational sampling load but makes for more awkward interpretation of the resulting mono-nucleotide impacts.

In all of our experiments, we focused on the entirety of the accessible, non-coding genome. That said, Baymer can easily be applied to any genomic features of interest for both polymorphism probability estimates and comparisons of feature-dependent sequence context shifts. Our approach does not currently incorporate genomic features in the model, but given genomic area bounds, polymorphism probabilities can be tailored to a biological question of interest. Addressing questions regarding the impact of genomic features on observed polymorphisms will be enhanced with well-regularized models, as smaller genomic areas or specific variant conditions can induce considerable data sparsity by reducing the number of contexts and/or polymorphisms available. Therefore, Baymer paves the way for exciting possibilities to study the effects of genomic features, variant age, and smaller subpopulations on sequence context-dependent mutation rate variation.

METHODS

Sample Data Sources

We sourced samples from the 1KG Phase III New York Genome Center resequencing project[33], gnomADv3.0[31], and trios from Halldorsson et al[36]. The genomic area for all sample sources was condensed to only include coordinates included within the 1KG accessibility mask[28] and outside of RefSeq coding regions to approximate the mappable non-coding genome. Only non-indel SNVs designated as "PASS" by the data source were retained. Based on confidence calls within the FASTA sequence files, high-confidence ancestral states (designated as those sites where all sequences agree on ancestral state) were inferred for all variants and contexts within the genomic area specified, where data allowed. Otherwise, variants and sites were omitted[39]. Ancestral allele counts were used for partitioning variants into different count brackets. Variants with allele frequency greater than 0.85 were removed to control for ancestral state misidentification[44]. We also compiled all sites where the high-confidence ancestral state and GRCh38 reference genome disagree, treating this collection as a call-set of derived variants. See Data Accessibility section for URLs for all data sources.

Baymer Model Description

In Baymer, increasing windows of sequence context are modeled as nested trees where each sequence context has 4 children – one for each of the four nucleotides added to expand the window size. For even-sized contexts, nucleotides are added to the 5′ end, and for odd-sized contexts, to the 3′ end. In this way, sequence context trees can be iteratively constructed to a given window size. We build one such tree for every reverse-complement folded 1-mer mutation type (i.e. **A**>C, **A**>G, **A**>T, **C**>A, **C**>G, **C**>T). Note that we designate the polymorphic nucleotide in focus in bold. For a given mutation type tree, *m*, let every edge be parameterized by $\phi_{a,b}^{m}$ where *a* denotes the edge's tree level and *b* the edge index. Edges in the first level of the tree represent the baseline **A**>* and **C**>* polymorphism probabilities (i.e., '1-mer') and center the polymorphism probabilities. These edges can take any value between zero and one and are given uninformative priors $\phi_{1,0}^{m} \sim$ *Uniform*(0,1). All edges beyond the first levels represent the log-transformed multiplicative shifts in polymorphism probability from their respective parent nodes. The polymorphism probability for any node is therefore given by the product of the edge log-transformed multiplicative shifts leading to that node and the root node in the tree corresponding to mutation type *m*.

$$p_{a,b}^{m} = \phi_{1,0}^{m} \prod_{a*,b*} \exp\left(\phi_{a*,b*}^{m}\right) \tag{1}$$

where *a** and *b** represent the level and index of exclusively those edges leading to the context in question. For every leaf context, *i,* where the mer-level, *a*, is equal to the maximum sequence context size considered, we let $\boldsymbol{p_i}$ denote the multinomial probabilities. Stated more explicitly:

$$\boldsymbol{p}_i = \left[p_{a,b}^{m_1}, p_{a,b}^{m_2}, p_{a,b}^{m_3}, 1 - \sum_{m*} p_{a,b}^{m*}\right] \tag{2}$$

where $m_{1\text{-}3}$ denote the three mutation types possible for this context. The corresponding outcomes, $\boldsymbol{x_i}$, for these probabilities is a length four vector for each of the three mutation types and the number of non-polymorphic context sites. We let n_i denote the total number of occurrences of leaf context *i* in the genomic area specified. Over *k* leaf nodes, the likelihood for the model can be calculated as:

$$p(y|\boldsymbol{\phi}) = \prod_i^k Multinom(n_i, \boldsymbol{p_i}, \boldsymbol{x_i}) \tag{3}$$

To provide regularization for the edges that are included in the model, we placed a spike-and-slab[29] prior on ϕ[22]:

$$\phi_{a,b}^{m} \sim \begin{cases} N(0, c^2\sigma_a^2) & w.p.\ 1-\alpha_a \\ N(0, \sigma_a^2) & w.p.\ \quad \alpha_a \end{cases} \tag{4}$$

where α_a is the mixture probability that a given edge in mer level *a* belongs to the spike or slab. We use an uninformative prior for α_a ~ *Uniform*(0, 1). Both the slab and spike distribution are specified to be Gaussian with a hyperparameter, *c*, representing the ratio between each distribution's standard deviation. The variance of the slab distribution for each level, σ_a^2, is a prespecified hyperparameter. For our models, we set this variance to ensure that the slab is favored when the evidence suggests a shift greater than 10% for a given context level (c = 500; σ_a^2 = 0.729). These chosen hyperparameters were informed by our prior biological intuition for meaningful effect sizes and a balanced ratio between the spike and slab distributions. These hyperparameters are at the discretion of the user, but a value of *c* less than or equal to 10000 is recommended[45].

Finally, we define a latent variable, *I*, that specifies whether a given edge belongs to the spike (I=0) or slab distribution (I=1). This yields the joint posterior distribution of the model:

$$p\left(\boldsymbol{\phi}, \boldsymbol{I}, \boldsymbol{\alpha}, \boldsymbol{\sigma^2}\middle|y\right) \propto p(y|\phi)p(\phi|I,\sigma^2)p(I|\alpha)p(\alpha)p(\sigma^2) \tag{5}$$

To estimate the posterior distribution above, we use an adaptive Metropolis-within-Gibbs MCMC sampling scheme[30]. Every level of the tree is estimated in ascending order, setting higher-order levels (i.e., larger windows of sequence context) to have uninformative shifts to aid convergence and enforce intermediate nodes to have informative polymorphism probabilities.

Our MCMC sampling scheme follows this approach. For the level-by-level sampling scheme, edges in levels higher, *a*′, than the level currently being sampled, *a*, are set to have no impact on the ultimate probabilities estimated, i.e., $\phi^m_{a',*}$= 0.

For the first layer of the tree:

1. Initialize all $\phi^m_{1,0}$ with a random value drawn from $Uniform(0,1)$ for iteration x = 0.
2. Sample new values of each $\phi^m_{1,0,x}$ for this iteration *x*, from $Normal(\phi^m_{1,0,x-1}, \tau^m_{1,0,x-1})$ using a Metropolis step[46], where $\tau^m_{1,0,x-1}$ represents the variance of the normal proposal density for $\phi^m_{1,0,x-1}$ at the previous iteration *x-1*.
3. Repeat step 2 until algorithm convergence.

For each subsequent level, *a* > 1:

1. Draw initial values (x=0) for parameters $\boldsymbol{\phi^m_{a,b}}$, $\boldsymbol{I^m_{a,b}}$, α_a
   a. $\boldsymbol{\phi^m_{a,b}}$ is drawn from $Uniform(-0.7,0.7)$, such that the total multinomial probabilities sum to 1
   b. $\boldsymbol{I^m_{a,b}}$ is drawn from $Bernoulli(0.5)$
   c. $\alpha_a$is drawn from $Uniform(0,1)$
2. Sample new values of $\phi^m_{a,b,x}$ from $Normal(\phi^m_{a,b,x-1}, \tau^m_{a,b,x-1})$ using a Metropolis step
3. Sample new values of $I^m_{a,b,x}$ using a Gibbs sampling step:

$$I^m_{a,b,x} \sim Bernoulli\left(\frac{p(I=1|\phi^m_{a,b,x},\sigma_{a,x},\alpha_{a,x})}{p(I=1|\phi^m_{a,b,x},\sigma_{a,x},\alpha_{a,x})+p(I=0|\phi^m_{a,b,x},\sigma_{a,x},\alpha_{a,x})}\right) \tag{6}$$

4. Sample new values of α_a using a Gibbs sampling step,

$$\alpha_{a,x} \sim Beta(1+\textstyle\sum_{m,i=1}^{j} I^m_{a,b,x}\, , 1+j-\sum_{m,i=1}^{j} I^m_{a,b,x}) \tag{7}$$

   where *j* represents the total number of edges in the current level.
5. Repeat steps 2-4 until algorithm convergence.

Posterior coverage estimation simulations

Polymorphism probabilities for our simulations were set using the mean of the posterior distribution estimated with Baymer when applied to private European variant data with minimal jitter added to avoid over-regularized estimates while still maintaining realistic human context-dependent polymorphism probability patterns. Jitter was added by sampling every 9-mer polymorphism probability, $p^m_{a,b}$, from $Normal(p^m_{a,b}, (p^m_{a,b})^{1.5})$, where the variance was set to scale to the underlying polymorphism probability. This dataset was chosen as it had the property of reaching sparsity limits at the 7-mer level and beyond. Thus, simulations evaluated up to 7-mers would provide a mixture of sparse and data-rich sequence contexts, providing a representative proxy for larger datasets run up through the 9-mer level. Using these polymorphism probabilities, new datasets were simulated by sampling from the multinomial distribution for each 9-mer sequence context. After applying Baymer to each individual dataset, we calculated the frequency that the true polymorphism probabilities were included in different sized credible sets. 2000 simulations were run for every sequence context up until 7-mers. Equal-tailed intervals were used to assign the credible intervals. Note that to aid computational tractability of this number of contexts and simulations, the alpha mixing parameter was sampled by using the posterior distributions for each level of the underlying base probability model used to generate simulated data.

Model comparisons for even/odd base-pair subsets

All non-Finnish European (NFE) variants with a derived allele count greater than or equal to 2 in the filtered gnomAD dataset were collected. Variants were next partitioned according to genomic coordinate parity (even/odd base pairs) to evenly divide the two groups as randomly as possible. Baymer was run on even and odd sets independently and the mean posterior estimates of polymorphism probability parameters were returned.

The root mean squared perpendicular error (RMSPE) was calculated by measuring the perpendicular distance between each point (estimated polymorphism probability) and the x=y line, that assumes each estimate is identical between models.

For transferability experiments, all European samples, excluding Finnish samples, from the 1KG Phase III designated as non-admixed[28] were aggregated and trimmed to only include sites with a minimum of 2 derived alleles and again partitioned according to genomic position parity. Opposite parities between 1KG and gnomAD datasets were grouped together. For each dataset, 100 equally-sized allele frequency bins between the minimum allele frequency in the two datasets and 1.0 were set. Each dataset was randomly down-sampled to ensure the same number of variants in each allele frequency bin. Baymer was applied to each down-sampled dataset and mean posterior estimates were compared.

Extended Sequence Context Likelihood Estimation

The gnomAD NFE data was partitioned into even and odd base pairs as described above. For each split, models were estimated using Baymer up through 9-mers. Smaller models correspond to the Baymer tree with all edges in larger sequence contexts not being considered assigned uninformative shifts ($\phi_{a,b}^{m} = 0$). We calculated likelihoods using the mean posterior probability estimate at the 9-mer level on the opposite parity polymorphism count data.

Data Sparsity Filters

To distinguish the degree to which estimates of PIP are simply a byproduct of data sparsity, we filtered out all sequence contexts with fewer than 50,000 total instances or fewer than 50 mutations in the non-coding genomic area considered.

Private Variant Analyses

All continental populations without substantial recent admixture (African, European, South Asian, East Asian) from the NYGC 1KG phase III resequencing dataset were filtered to only include variants private to each continental group. Each population was trimmed to only include variants with a minimum allele count of 2 and then down-sampled and site frequency spectra-matched to match the smallest variant counts across the four continental groups. Baymer was then applied to each resulting dataset. The resulting posterior distributions of the polymorphism probabilities and ϕ shifts of each model were then pairwise compared by calculating the fraction overlap of the distributions, as a proxy for the probability they are the same. Distributions are parameterized using a Gaussian kernel density estimate on the posterior samples.

Power Estimates

Truth polymorphism probabilities used in our simulations to estimate power were set using the same model as the variance calibration experiments. For a given sequence context mutation, we tested the discoverability of a spectrum of deviations from the "truth" model. We simulated 1000 9-mer count tables using polymorphism probabilities from both the "truth" model and the deviated model. Both count tables were modeled using Baymer and the resulting posterior distributions used to assess the fraction overlap for the context mutation in focus. A shift is considered discovered if the degree of fraction overlap is less than 1%. As running this experiment for all context mutations was intractable, we tested at most 100 CpG and 100 non-CpG contexts at each mer-level. Contexts were chosen to give an even spread across the sample size spectrum, as dictated by total contexts.

Grafted Tree Scheme

Baymer models were built independently on *de novo* even data and gnomAD NFE polymorphism data with allele count greater than or equal to 2. The *de novo* model parameter estimates were used up through 3-mers. For the remaining levels (for 5-mers and larger windows), NFE-2+ parameter mean point estimates were used in place of the equivalent de novo edges. Thus, the grafted tree polymorphism parameters were the product of the point estimates for each branch of the tree, given the data source described above. The multinomial likelihood of the resulting model was calculated on the odd de novo holdout data, as before.

Data and Code Accessibility

All data analyzed here are publicly available at the following websites:

NYGC resequencing of 1KG Phase III data:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/

gnomADv3.0:
https://gnomad.broadinstitute.org/downloads

Halldorsson et al. trio data:
https://science.sciencemag.org/highwire/filestream/721792/field_highwire_adjunct_files/7/aau1043_DataS5_revision1.tsv

1KG accessibility mask:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/PilotMask/20160622.allChr.pilot_mask.bed

RefSeq coding regions:
http://www.ensembl.org/biomart/

Ancestral FASTA:

ftp://ftp.ensembl.org/pub/release97/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38.tar.gz

Code Accessibility

We have implemented our Baymer method into software that is freely available as a python package. This can be accessed on the Voight Lab GitHub repository: https://github.com/bvoightlab/Baymer

Acknowledgements

B.F.V. acknowledges support for this work from the NIH/NIDDK (DK101478 and DK126194). We would also like to thank Dr. Ziyue Gao for her very helpful feedback on the manuscript and in the development process.

Author Contributions

The experiments were conceived and designed by C.J.A., S.T.J., I.M., and B.F.V. C.J.A. and B.F.V. performed statistical analyses. C.J.A., M.C., B.J.A., and B.F.V. analyzed the data. C.J.A. and B.F.V. drafted the initial manuscript. All authors contributed and edited the final manuscript. The work was supervised by B.F.V.

REFERENCES

1. Wang, Y. & Nielsen, R. Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias. *Mol Ecol* (2012) doi:10.1111/j.1365-294X.2011.05413.x.
2. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* (2009) doi:10.1371/journal.pgen.1000695.
3. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* (2009) doi:10.1371/journal.pgen.1000471.
4. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* (2016) doi:10.1038/nature19057.
5. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat Genet* (2019) doi:10.1038/s41588-018-0294-6.
6. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* 2022.03.20.485034 (2022) doi:10.1101/2022.03.20.485034.
7. Petrovski, S. *et al.* The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet* **11**, e1005492 (2015).
8. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9**, e1003671 (2013).
9. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat Genet* (2018) doi:10.1038/s41588-018-0062-7.
10. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* Preprint at https://doi.org/10.1038/nrg3098 (2011).
11. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* (2009) doi:10.1038/ng.363.
12. Fryxell, K. J. & Moon, W. J. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol* (2005) doi:10.1093/molbev/msi043.
13. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* (2012) doi:10.1038/nature11273.
14. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* Preprint at https://doi.org/10.1016/j.cell.2019.02.051 (2019).
15. Holliday, R. & Grigg, G. W. DNA methylation and mutation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **285**, 61–67 (1993).
16. Sung, W. *et al.* Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Mol Biol Evol* (2015) doi:10.1093/molbev/msv055.
17. Lujan, S. A. *et al.* Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res* (2014) doi:10.1101/gr.178335.114.
18. Bzymek, M. & Lovett, S. T. Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proc Natl Acad Sci U S A* (2001) doi:10.1073/pnas.111008398.

19. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* (2016) doi:10.1038/ng.3511.

20. Mathieson, I. & Reich, D. Differences in the rare variant spectrum among human populations. *PLoS Genet* (2017) doi:10.1371/journal.pgen.1006581.

21. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences* (2015) doi:10.1073/pnas.1418652112.

22. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *Elife* (2017) doi:10.7554/elife.24284.

23. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* (2018) doi:10.1038/s41467-018-05936-5.

24. Fang, Y., Deng, S. & Li, C. A deep learning-based framework for estimating fine-scale germline mutation rates. *bioRxiv* (2021).

25. Bethune, J., Kleppe, A. S. & Besenbacher, S. A method to build extended sequence context models of point mutations and indels. *bioRxiv* (2021).

26. Liu, Z. & Samee, M. A. H. Mutation rate variations in the human genome are encoded in DNA shape. *BioRxiv* (2021).

27. Ling, G., Miller, D., Nielsen, R. & Stern, A. A Bayesian Framework for Inferring the Influence of Sequence Context on Point Mutations. *Mol Biol Evol* (2020) doi:10.1093/molbev/msz248.

28. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* (2015) doi:10.1038/nature15393.

29. George, E. I. & McCulloch, R. E. Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**, 881–889 (1993).

30. Roberts, G. O. & Rosenthal, J. S. Examples of adaptive MCMC. *Journal of computational and graphical statistics* **18**, 349–367 (2009).

31. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020) doi:10.1038/s41586-020-2308-7.

32. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol* **14**, 1–20 (2013).

33. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* 2021.02.06.430068 (2021) doi:10.1101/2021.02.06.430068.

34. Aikens, R. C., Johnson, K. E. & Voight, B. F. Signals of Variation in Human Mutation Rate at Multiple Levels of Sequence Context. *Mol Biol Evol* (2019) doi:10.1093/molbev/msz023.

35. Gao, Z., Zhang, Y., Przeworski, M. & Moorjani, P. Timing and causes of the evolution of the germline mutation spectrum in humans. *bioRxiv* 2022.06.17.496622 (2022) doi:10.1101/2022.06.17.496622.

36. DeWitt, W. S., Harris, K. D., Ragsdale, A. P. & Harris, K. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences* **118**, e2013798118 (2021).

37. Anderson-Trocmé, L. *et al.* Legacy Data Confound Genomics Studies. *Mol Biol Evol* **37**, 2–10 (2020).

38. Halldorsson, B. v *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science (1979)* **363**, eaau1043 (2019).

39. Ensembl. Ensembl, Data from “homo_sapiens_ancestor_GRCh38.” *http://ftp.ensembl.org/pub/release-97/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38.tar.gz*.

40. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J Big Data* **3**, 1–40 (2016).

41. Agarwal, I. & Przeworski, M. Mutation saturation for fitness effects at human CpG sites. *Elife* **10**, e71513 (2021).

42. Gao, Z. *et al.* Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proceedings of the National Academy of Sciences* **116**, 9491–9500 (2019).

43. Johnson, K. E. & Voight, B. F. Identifying non-identical-by-descent rare variants in population-scale whole genome sequencing data. *bioRxiv* 2020.05.26.117358 (2020) doi:10.1101/2020.05.26.117358.

44. Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Context Dependence, Ancestral Misidentification, and Spurious Signatures of Natural Selection. *Mol Biol Evol* **24**, 1792–1800 (2007).

45. George, E. I. & McCulloch, R. E. Approaches for Bayesian variable selection. *Stat Sin* 339–373 (1997).

46. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J Chem Phys* **21**, 1087–1092 (1953).

Table 1. Baymer modeled 1KG private continental context mutations with extreme polymorphism probability differences

Population Comparison	Context Mutation	log(Poly. Prob. Fraction)	Poly. Prob. Fraction Overlap	Shift Difference	Shift Fraction Overlap	Population Specificity
European v African	T**C**C>T	0.291	0	-0.174	1.4E-157	European
	T**C**T>T	0.136	1.6E-18	-0.116	8.5E-16	European
	GCA**A**TTA>G	0.569	4.7E-03	-0.668	2.4E-03	
	TAT**A**TATC>G	-0.660	7.2E-03	0.730	5.6E-03	African
European v South Asian	T**C**C>T	0.112	1.2E-09	-0.059	2.7E-03	European
	T**C**T>T	0.063	5.0E-03	-0.066	2.9E-03	European
	CT**A**TA>T	-0.587	2.9E-03	0.493	7.3E-03	South Asian
	AT**C**TTC>G	-0.606	7.6E-03	0.668	5.4E-03	
European v East Asian	C**C**C>T	0.081	1.4E-03	0.075	6.6E-04	
	T**C**C>T	0.312	0	-0.156	2.4E-97	European
	G**C**T>T	-0.064	5.7E-03	0.095	6.1E-05	
	T**C**T>T	0.133	3.0E-19	-0.102	9.6E-06	European
	GCA**A**CCA>G	1.056	5.3E-03	-1.104	5.0E-03	
	ATA**C**CTC>A	-1.029	4.2E-03	0.830	5.0E-03	East Asian
African v South Asian	T**C**C>T	-0.179	1.7E-118	0.115	3.4E-12	
	CT**A**TA>T	-0.507	6.1E-03	0.482	7.4E-03	South Asian
	CCC**C**CAG>G	-0.818	2.6E-03	0.767	2.7E-03	
	TAT**A**TATC>G	0.668	3.3E-03	-0.738	2.2E-03	African
African v East Asian	G**C**T>T	-0.063	9.1E-03	0.074	2.2E-03	
	CT**C**GCG>T	1.240	2.8E-03	-1.243	3.6E-03	
	TAA**A**ATA>T	-1.160	3.9E-03	1.135	4.8E-03	
	ATA**C**CTC>A	-1.061	4.6E-03	0.829	5.7E-03	East Asian
	TAT**A**TATC>G	0.712	3.9E-04	-0.748	1.3E-04	African
East Asian v South Asian	T**C**C>T	-0.200	2.4E-155	0.097	5.4E-05	
	CT**A**TA>T	-0.519	5.3E-03	0.479	7.8E-03	South Asian
	CT**C**GCG>T	-1.244	2.0E-03	1.247	2.7E-03	
	ATA**C**CTC>A	0.906	8.5E-03	-0.819	9.1E-03	East Asia
	CCC**C**CAG>G	-0.819	3.8E-03	0.764	4.4E-03	

Table 2. Power estimates for 1KG continental private polymorphism probabilities.

abs(log(adj. poly. prob / null poly. prob.))	# contexts sample size percentile	fraction of contexts with >80% power at each mer level						
		3mers	4mers	5mers	6mers	7mers	8mers	9mers
0.01	0-25%	0	0	0	0	0	0	0
	26-50%	0	0	0	0	0	0	0
	51-75%	0	0	0	0	0	0	0
	76-100%	0	0	0	0	0	0	0
0.1	0-25%	0.44	0.11	0	0	0	0	0
	26-50%	0.63	0.04	0	0	0	0	0
	51-75%	0.73	0.03	0	0	0	0	0
	76-100%	0.58	0.10	0	0	0	0	0
0.5	0-25%	1.00	0.92	0.30	0.21	0.01	0	0
	26-50%	1.00	1.00	0.68	0.15	0.01	0	0
	51-75%	1.00	1.00	0.76	0.27	0.02	0	0
	76-100%	1.00	1.00	0.87	0.20	0.03	0	0
1	0-25%	1.00	0.99	0.81	0.34	0.20	0.02	0
	26-50%	1.00	1.00	1.00	0.73	0.23	0.06	0
	51-75%	1.00	1.00	1.00	0.87	0.24	0.04	0
	76-100%	1.00	1.00	1.00	0.87	0.37	0.08	0
1.5	0-25%	1.00	1.00	0.96	0.61	0.25	0.08	0
	26-50%	1.00	1.00	1.00	0.91	0.39	0.18	0.02
	51-75%	1.00	1.00	1.00	0.99	0.59	0.20	0
	76-100%	1.00	1.00	1.00	0.99	0.69	0.26	0

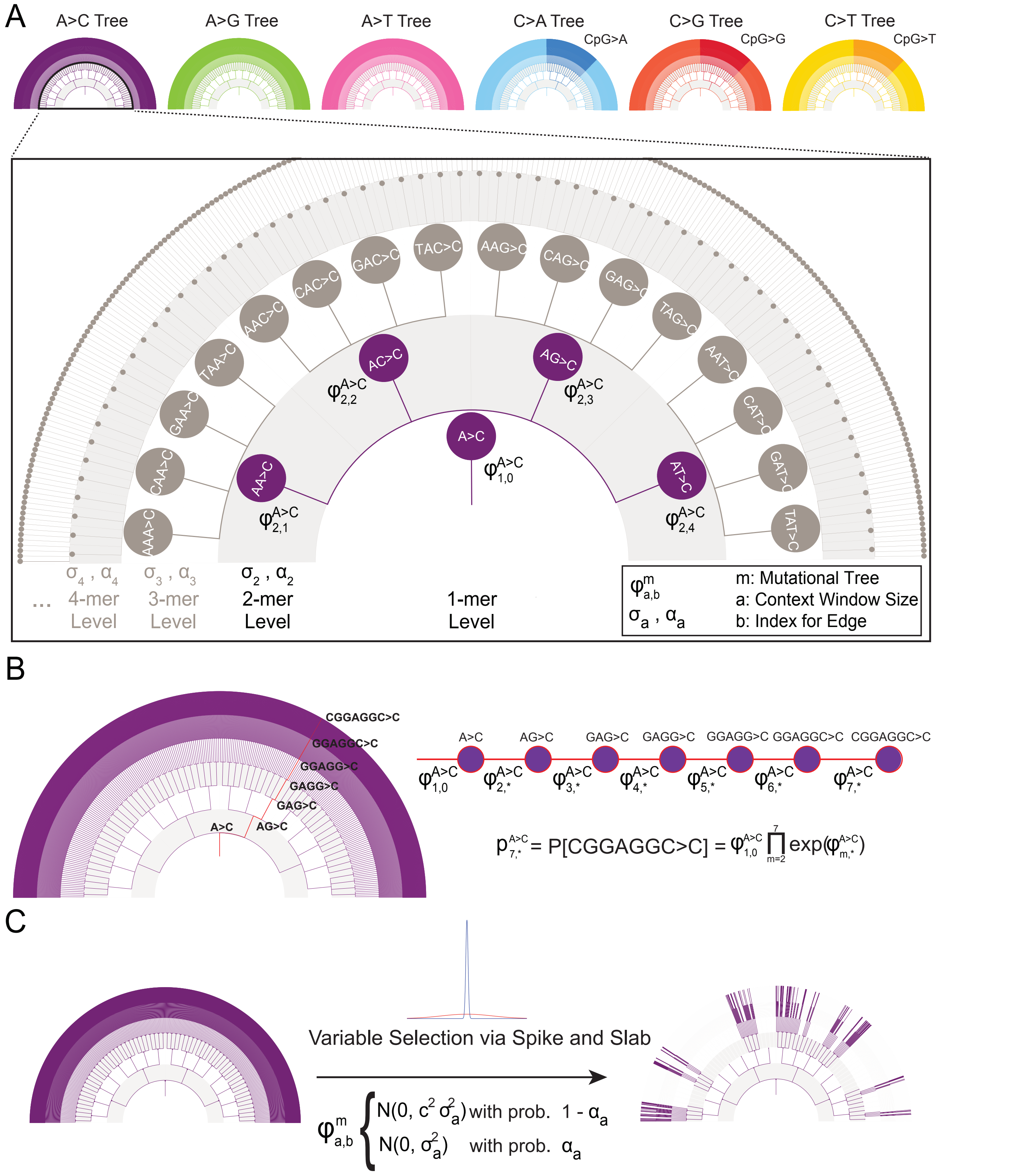
A
A>C Tree
A>G Tree
A>T Tree
C>A Tree
CpG>A
C>G Tree
CpG>G
C>T Tree
CpG>T
A>C
AA>C
AC>C
AG>C
AT>C
AAA>C
CAA>C
GAA>C
TAA>C
AAC>C
CAC>C
GAC>C
TAC>C
AAG>C
CAG>C
GAG>C
TAG>C
AAT>C
CAT>C
GAT>C
TAT>C
$\varphi^{A>C}_{1,0}$
$\varphi^{A>C}_{2,1}$
$\varphi^{A>C}_{2,2}$
$\varphi^{A>C}_{2,3}$
$\varphi^{A>C}_{2,4}$
... σ_4 , α_4 4-mer Level
σ_3 , α_3 3-mer Level
σ_2 , α_2 2-mer Level
1-mer Level
$\varphi^{m}_{a,b}$ m: Mutational Tree
σ_a , α_a a: Context Window Size
b: Index for Edge
B
CGGAGGC>C
GGAGGC>C
GGAGG>C
GAGG>C
GAG>C
A>C
AG>C
A>C AG>C GAG>C GAGG>C GGAGG>C GGAGGC>C CGGAGGC>C
$\varphi^{A>C}_{1,0}$ $\varphi^{A>C}_{2,*}$ $\varphi^{A>C}_{3,*}$ $\varphi^{A>C}_{4,*}$ $\varphi^{A>C}_{5,*}$ $\varphi^{A>C}_{6,*}$ $\varphi^{A>C}_{7,*}$
$p^{A>C}_{7,*} = P[CGGAGGC>C] = \varphi^{A>C}_{1,0} \prod_{m=2}^{7} \exp(\varphi^{A>C}_{m,*})$
C
Variable Selection via Spike and Slab
$\varphi^{m}_{a,b} \begin{cases} N(0, c^2 \sigma^2_a) \text{ with prob. } 1 - \alpha_a \\ N(0, \sigma^2_a) \text{ with prob. } \alpha_a \end{cases}$

Figure 1. Hierarchical relationship of sequence contexts and key algorithmic elements of Baymer. (A) Each mutation type is represented by a separate sequence context tree, related by the shared mer level parameters and joint multinomial likelihood distribution. Each sequence context tree has a nested structure where information is partially pooled across each shared parent. (B) Polymorphism probabilities are parameterized as the product of the series of edges that lead to the sequence context of interest. (C) Sequence context trees are regularized using a spike-and-slab prior distribution.

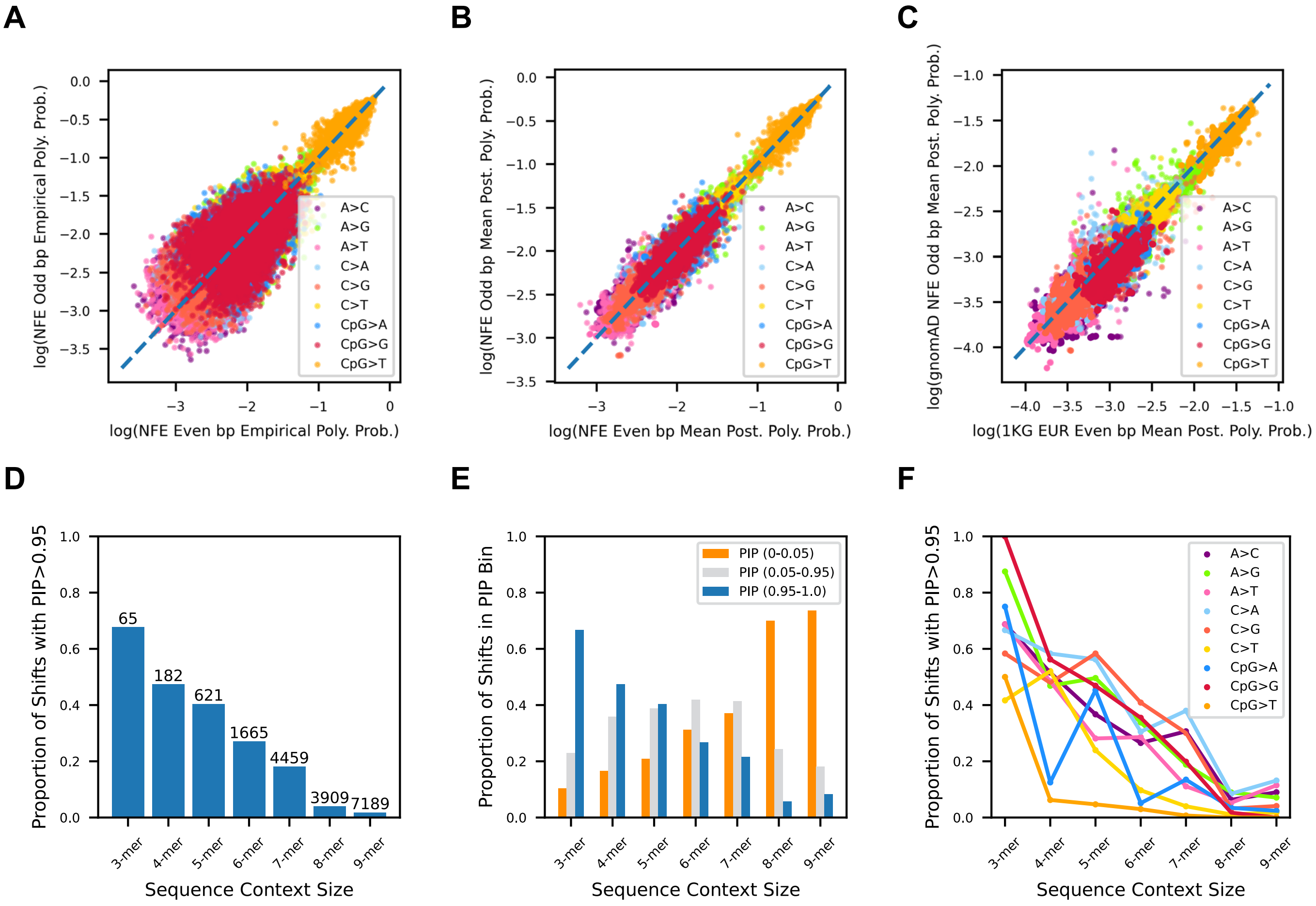

Figure 2. Baymer model validation, transferability, and regularization in gnomAD non-Finnish European (NFE) polymorphisms with derived allele count greater than or equal to 2 in non-coding accessible regions. (A) Empirical 9mer polymorphism probabilities for context mutations with at least 1 occurrence in both datasets (15910 omitted context mutations) are plotted against one another (Spearman correlation 0.915; $p < 10^{-100}$; RMSPE = 0.12). (B) Baymer mean posterior estimates for 9mer polymorphism estimates in even and odd bp datasets (Spearman correlation 0.990; $p < 10^{-100}$; RMSPE = 0.035). (C) Baymer mean posterior estimates for 9mer polymorphism estimates in odd bp non-Finnish European gnomAD data and even bp NYGC 1KGPIII data, down-sampled to match total number of polymorphisms and site frequency spectrum (Spearman correlation 0.981; $p < 10^{-100}$; RMSPE = 0.063). (D) Fraction of edges in the NFE model with a PIP > 0.95 in each sequence context window layer. Absolute count of edges above bars. (E) For high-data contexts with at least 100,000 total instances in the non-coding genome and 50 total mutations, fraction of edges at each sequence context window size across PIP bins. (F) Proportion of high-data contexts within each mutation type at each sequence context window size with PIP>0.95.

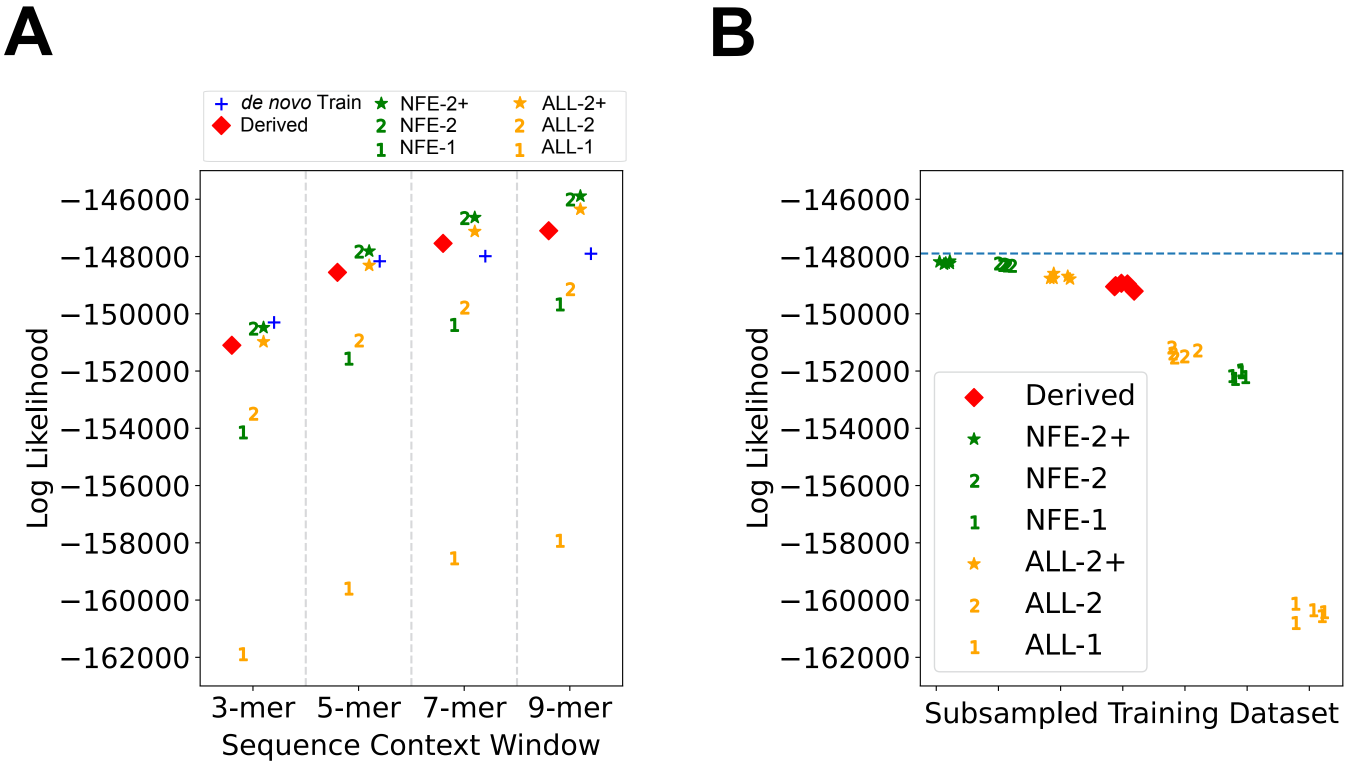

Figure 3. Modeling de novo mutation probabilities using polymorphism datasets. Even bp Halldorsson et al. *de novo* training data modeled by Baymer is compared to Baymer-modelled polymorphism datasets partitioned by allele count. (A) Multinomial likelihoods for each model are calculated on odd *de novo* bp test data at various sequence context sizes. Polymorphism probability estimates were linearly scaled to match the mean polymorphism probability of the holdout dataset. (B) Polymorphism datasets were down-sampled to match the size of the even bp de novo data (70,364 variants) and multinomial likelihoods were calculated on odd de novo bp data. Each dataset was down-sampled using 5 different random seeds. The LL of the 9mer *de novo* training model is indicated with the blue dotted line.

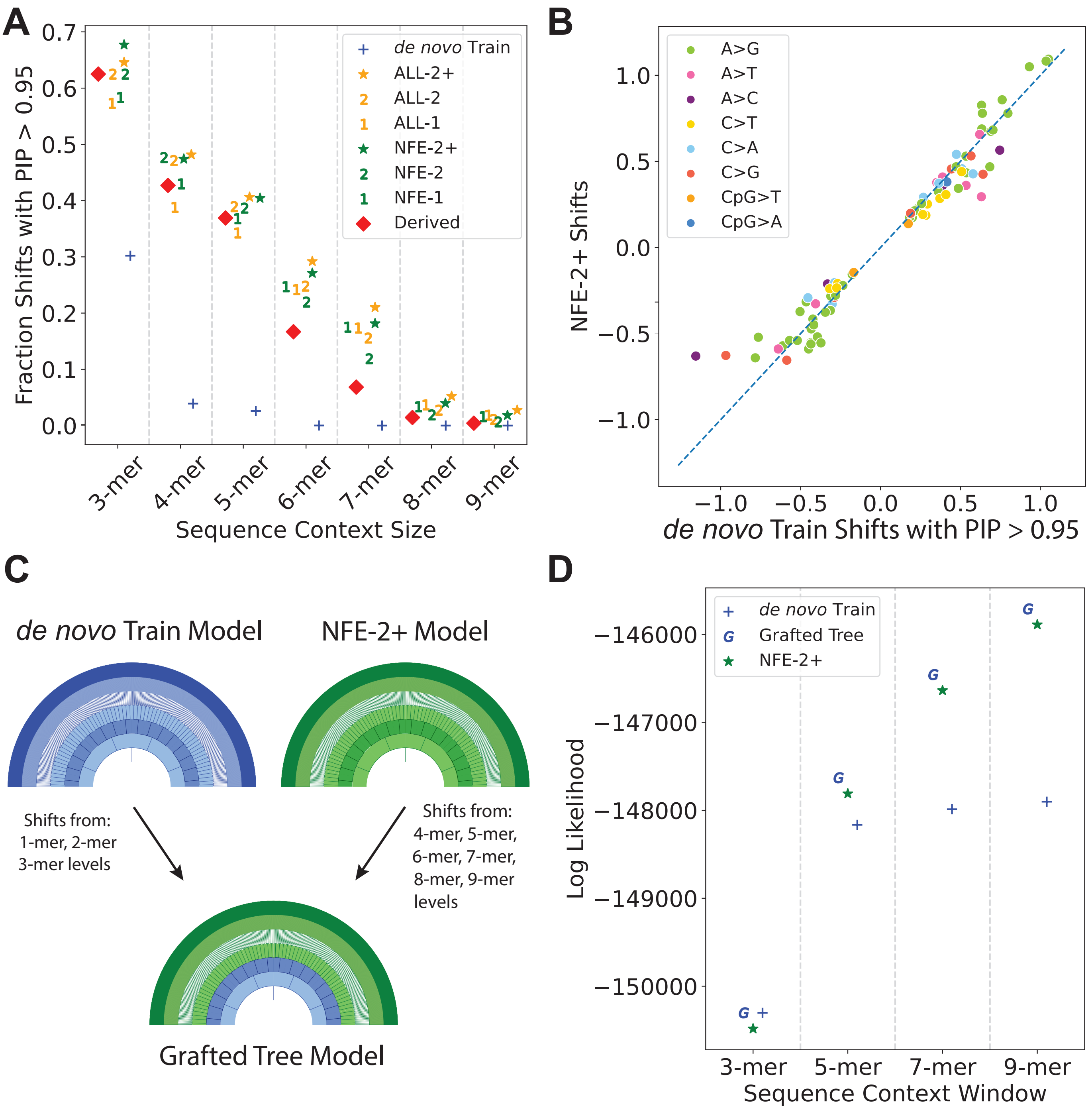

Figure 4. Tree grafting strategy to share information between Baymer models. (A) For each de novo proxy model, we calculated the fraction of context polymorphism probability shifts with a PIP > 0.95 in 2-mer – 9-mer mer-levels as a proxy for the degree of regularization in each model. (B) Polymorphism probability shifts in the de novo training model that are included with high-confidence (PIP>0.95) are very similar in magnitude and direction to their equivalents in the best-performing proxy model, NFE-2+, in 2-mer – 9-mer levels, implying a shared polymorphism probability shift structure. (C) Proposed tree-grafting schema for modeling de novo mutations that leverages mer-levels where de novo data is plentiful (1-mer – 3-mers) and uses polymorphism data to model the remainder of each model in larger mer-levels (4-mer – 9-mers) where the de novo model is underpowered. (D) The grafted tree method outperforms the previously best-performing model, NFE-2+.