Supplementary Material

1 Phasing Algorithm

A Hidden Markov Model (HMM) is defined by a state space, transition probabilities, emission probabilities, and a set of observations. In our case, the state space is the set of ways that the children can inherit the two maternal copies of the chromosome (m_1 and m_2) and the two paternal copies (p_1 and p_2). For example, for a family with two children (m_1p_1, m_2p_1) represents a state where the first child inherits parental copies m_1 and p_1 and the second child inherits parental copies m_2 and p_1. Since we are working with whole-genome sequencing data, we also include a hard-to-sequence region flag in our state space in order to detect and flag regions with many sequencing errors. The transmission probabilities in our model represent recombination events where the chromosome inherited by a child switches from one parental copy to the other. We use estimates of $1.39e^{-8}$ and $9.23e^{-9}$ for the probability of maternal and paternal recombination per base-pair [4].

The emission probabilities in our model represent the probability of sequencing errors. We estimate these probabilities directly from the genotype data using a family-based method [5]. Finally, the observations are the variant calls for each family. We use the Viterbi algorithm [2] to identify the sequence of states that best explains the observed variant calls. The result is a fully phased family - which copy of parental chromosomes each child inherited at each region of the genome.

2 K-mer Length Choice

We chose a *k*-mer length of 100 using the following logic. We first estimated the average number of reads beginning at each loci by dividing the total number of reads that were properly paired and mapped to GRCh38 for each sample by the length of the diploid human genome ($6.27\text{x}10^{10}$) and taking the average. This gave us an average of 0.12 read starts per loci (and an average 17.61x coverage per loci per genome copy). A *k*-mer originating from position g in the genome will be present in a 150 bp read if the read starts in the interval $[g + k - 150, g]$. A *k*-mer from the genome will also not be present in a read if that read has a sequencing error within that *k*-mer region. Several studies have estimated Illumina short-read sequencing errors to be between ϵ=.01% and .05% per base [8, 6, 7]. Assuming reads are equally distributed in the genome, the probability that that a *k*-mer originating from a random location in the genome is contained in N total reads for a given sample can be described by a Poisson distribution $P(X = N; \mu = 0.12(151 - k)(1 - \epsilon)^k)$.

We limit our *k*-mer search for *k*-mers that occur at least twice in an individual to improve computation time by ignoring singletons. Using the Poisson distribution described, Fig. S1 shows the probability that a *k*-mer will occur at least N times in an individual's set of reads. For $N = 2$, across various error rates, we see that the elbow in the curve, the location where this probability begins to drop sharply off, is a little after k=100. Therefore, we chose k=100 as a good length to balance *k*-mer uniqueness and the probability the *k*-mer occurs in the raw reads. The *k*-mer extraction and localization pipelines are fairly memory and computationally expensive, so we did not experiment with *k*-mers of different lengths.

3 Maximum Likelihood Model

We previous developed and validated a proof-of-concept algorithm to localize 100-bp *k*-mers extracted from 150 bp reads [1]. We review the mathematics of this maximum likelihood model, and discuss the modifications that we added in order to allow localization of tandem repeats and for sequence originating from the sex chromosomes.

The goal of the maximum likelihood model is as follows: For each *k*-mer, we wish to find it's corresponding location (region r) in the genome that best explain the distribution of the *k*-mer counts in all of the families. We define the distribution of a given *k*-mer in all samples as $\mathbf{K}$, and the distribution of a given *k*-mer in family f as $\mathbf{K_f}$. Therefore, we want to find the region r that maximizes the likelihood of observing $\mathbf{k}$. We can rewrite this likelihood in log-likelihood form:

$$\ell(r; \mathbf{K}) = \sum_f log(P(\mathbf{K_f}|r)) \tag{1}$$

To compute this likelihood, we must compute $P(\mathbf{K_f}|r)$, the probability of a family's *k*-mer counts given the *k*-mer's hypothetical region. We assume that children inherit the given *k*-mer in a Mendelian fashion: from each parent, they receive either 0 or 1 copy of the *k*-mer. A parent with the *k*-mer present on both copies of their chromosome is guaranteed to pass down the *k*-mer to their child, a parent without the *k*-mer present on either copy of their chromosome will never pass the *k*-mer down to their offspring, and a parent who is heterozygous for the *k*-mer has a 50% chance of passing the

k-mer down to an offspring. The possible phased genotypes denoted in the order of (maternal allele, paternal allele) is $\mathbf{G} = \{\text{“}0/0\text{”}, \text{“}1/0\text{”}, \text{“}0/1\text{”}, \text{“}1/1\text{”}\}$ where 1 denotes that a person has the *k*-mer on the given copy of their chromosomes and 0 indicates it is absent. We will call mother's genotype g_m, father's genotype g_p (p for paternal), and a child's genotype g_c. We can also define mother's, father's, and child's *k*-mer counts as k_m, k_p, and k_c respectively. Using this notation and the law of total probability, we can rewrite our family-wise probability of observing the data as:

$$P(\mathbf{K_f}|r) = \sum_{g_m, g_p \epsilon G} P(\mathbf{K_f}|r, g_m, g_p) P(g_m, g_p) = \sum_{g_m, g_p \epsilon G} \left(P(k_p|g_p) P(k_m|g_m) \prod_c P(k_c|g_c) \right) P(g_m, g_p) \tag{2}$$

We iterate over all possible genotypes for the mother and the father but not for the children because from our phasing algorithm, we already know which copy of mom's DNA a child inherited, and which copy of dad's DNA a child inherited at any given region. We can therefore compute the child's genotype for a *k*-mer, given the region, and the parent's phased genotypes:

$$g_c = \textit{phase}(c, r, g_m, g_p) \tag{3}$$

The *phase*() function queries the inheritance pattern of a child c at a region r in the phasing dictionary. Using the phasing information, it then combines the maternal haploid genotype on the appropriate copy of g_m with the paternal haploid genotype on the appropriate copy of g_p to infer the child's diploid genotype.

Now let's compute $P(k, g)$, the probability of a *k*-mer count in a person, given a person's genotype. We assume that the sequencing pipeline, which used random-PCR targeted every region of the genome at an equal read depth, or at least that the PCR amplification bias has similar profiles across samples (the same regions of the genome are consistently under or over amplified). In our original algorithm, we assumed that *k*-mers were unique to a single location in the genome, with no copy number variation. *k*-mer depth then follows a Poisson distribution, dependent on genotype heterozygosity and average *k*-mer depth μ_k. A theoretical μ_k can be derived for each person using the total number of sequencing reads, the length of the person's genome (which differ slightly between males and females), and the length of the *k*-mer. The average μ_k of a 100-bp *k*-mer for our samples was 5.83. Using the syntax where $\sum(\text{“}0/0\text{”}) = 0$, $\sum(\text{“}0/1\text{”}) = \sum(\text{“}0/1\text{”}) = 1$, and $\sum(\text{“}1/1\text{”}) = 2$, a *k*-mer distribution can be summarized as follows:

$$P(k|g, \mu_k) = P_{\text{poisson}}(k; \mu_k \sum g) \tag{4}$$

However, we modified our original algorithm to account for tandem repeats, identical sequences that occur next to or near each other in tandem. We first normalize very high values of k to be between 0 and 10 by performing the following operation. For each family, using $max(k_f)$ as the maximum counts of a *k*-mer any family member, we convert each family member's k to:

$$k^* = \frac{k}{10^{\lfloor log_{10}(max(k_f)) \rfloor}} \tag{5}$$

We then only need to iterate over small numbers of possible repeats. To compute the probability of k* occurrences of a subsequence, ASLAN iterates not only over possible genotypes, but also over possible numbers of repeats. Rather than combinations of 0 and 1, the set of genotypes G in Eq. 2 then becomes the number of repeats a person has on each copy of their chromosome:

$$G = \{\text{“}c_1/c_2\text{”} : c_1 \epsilon C, c_2 \epsilon C\} \tag{6}$$

where c is the number of copies of a repeat on each copy of a genome. Since we have transformed k to be a single digit number, we iterate over C=$\{0, 1, 2, 3, 4, 5\}$ to reduce the number of iterations we must compute a probability for.

Our original algorithm was written for autosomes, but we made some modifications to give ASLAN the ability to localize sequences originating from the X and Y chromosomes as well. Instead of iterating over $g_m, g_p \epsilon G$ in Eq. 2 we account for the non-Mendelian inheritance patterns of the sex chromosomes. For regions on the Y chromosome, we iterate over $g_m \epsilon$ ‘“0/0” and $g_p \epsilon \{\text{“}0/c\text{”} : c \epsilon C\}$. For regions on the X chromosome, we iterate over $g_p \epsilon \{\text{“}c_1/c_2\text{”} : c_1, c_2 \epsilon C\}$ and $g_p \epsilon \{\text{“}c/0\text{”} : c \epsilon C\}$.

Given the phasings and *k*-mer counts, for every family we can now compute the log-likelihood of a given *k*-mer belonging to each region of the genome $log(P(\mathbf{K_f}|r))$ and we can take the cumulative log-likelihoods to compute the total log-likelihood of a given *k*-mer belonging to each region of the genome $log(P(\mathbf{K}|r))$.

Rather than reporting only the region with maximum likelihood and be at the whim of statistical noise, we estimate a maximum likelihood interval [3]. From our cumulative likelihoods, we find all the neighboring regions on the graph whose relative likelihoods are within a certain threshold. That is, our maximum likelihood interval is:

$$\{r : \frac{L(r)}{L(\hat{r})} >= t\} \tag{7}$$

where t is a certain threshold. Because *k*-mers very in their allele frequencies and families in which they are present, and because families vary in their IBD patterns, *k*-mer likelihood profiles are susceptible to different amounts of statistical

noise. For that reason, we choose t as a function of the standard deviation (σ) in each *k*-mer's log-likelihood profile. Specifically, we define our maximum likelihood region to be:

$$\{r : log\frac{L(r)}{L(\hat{r})}) >= -\gamma\sigma(log(L(r)))\} \tag{8}$$

where γ is a hyperparameter that we must tune. We tried values of γ between .01 and 1, ultimately choosing .1 for a balance of sensitivity and specificity, localizing 96% of medium-prevalence (prevalence of 0.2 to 0.8) *k*-mers, with a 90% accuracy and median resolution of 870Kb. If regions from multiple different chromosomes fell into our maximum likelihood region for a given *k*-mer, we considered that *k*-mer unlocalized. A schematic of this pipeline is shown in Fig. 1 in the manuscript, with actual family-wise and cumulative log-likelihood graphs computed by our algorithm on a *k*-mer from the data.

The code for our localization algorithm can be accessed at
`https://github.com/briannachrisman/alt_haplotypes`.

4 Validation

4.1 Synthetic Dataset Generation

We generated synthetic datasets of theoretical *k*-mer distributions. We analyzed the performance of the algorithm on *k*-mers with different prevalences in a population and average number of tandem repeats. For each theoretical *k*-mer, we generated a location, a prevalence, and the average number of tandem repeats. For *k*-mer location, randomly sample uniformly across the 263,461 regions separated by our datasets recombination points identified in the phasing algorithm. Since the y-chromosome technically accounts for only one of these 263,461 regions, we also append many theoretical y-chromosome *k*-mers to our simulated dataset. To generate the population prevalence of the *k*-mer, the probability that the *k*-mer appears in one copy of the genome by random uniform sampling between 0 to 1. We performed the entire pipeline three different times, using average tandem repeat numbers of 1, 10, and 100. We generate a distribution of counts for a theoretical *k*-mer by doing the following for each family:

1. For each parent's two copies of the genome, we choose whether that copy contains the *k*-mer in each copy of their genome by comparing a randomly generated number to the prevalence of the *k*-mer.

2. For copies that do contain the *k*-mer, we simulate the number of tandem repeats of the *k*-mer they contain. We do this by sampling from a Poisson distribution, where the mean is equal to the given average number of tandem repeats in the population. For the simulated datasets where the average tandem repeat number is 1, we assume the *k*-mer is non-repetitive and only has 1 repeat in every individual with the *k*-mer.

3. We simulate the children's "true" counts of a *k*-mer by using the phasing information to find which copy of their parents' chromosomes they inherited from the chosen location in the genome, and taking the sum of the number of *k*-mers on the appropriate parent copies.

4. To generate the observed number of counts of a *k*-mer in each parent and child's WGS reads, we sample from a Poisson distribution. For the mean, we use the expected number of *k*-mer counts given the true counts of a *k*-mer, the individual's coverage metrics, and the sex-specific size of the genome.

We generate synthetic datasets using average tandem repeat counts of 1, 10, and 100. For each of these datasets, we generate distributions for 100,000 autosomal and X-chromosome *k*-mers and 1,000 y-chromosome *k*-mers.

4.2 Extracting Reads and Locations from Alternative Sequences from the Decoy Genome

While the synthetic dataset shows our algorithm worked in theory, the synthetic dataset was generated using many of the same assumptions that the algorithm was based on and therefore is somewhat of an unfair proof of accuracy. First of all, the synthetic dataset was generated assuming the phasings of the iHART dataset were correct. Secondly, to generate the number of *k*-mers observed for any individual, it uses a Poisson distribution assuming equally likely coverage of every location, the same distribution that the model assumes.

Therefore, we additionally validate our model using *k*-mers extracted from alternative sequences in the decoy genome. The decoy genome contains several hundred alternative sequences (identified by the "ALT" suffix in their contigs), corresponding to sequences that are not in the primary reference genome, but that have been found in enough individuals that the Human Genome Reference Consortia included them in GRCh38 [9]. While some of these sequences are unplaced scaffolds, many do have known locations with respect to the reference genome and are labelled accordingly. As localizing non-reference sequences to the human reference genome is the precise goal of our algorithm, these labelled ALT sequences serve as an excellent validation dataset. Using the ASLAN pipeline, we extracted all unique 100-mers (1,094,247 total *k*-mers) from each of the labelled ALT sequences (233 total contigs), counted their distributions and localized them.

4.3 Maximum Likelihood Interval Hyper-parameter Tuning

Our phasing and maximum likelihood algorithms rely mainly on distributions, parameters, and probabilities derived from basic statistics and biology. However, the final step of the localization process involves selecting a maximum likelihood interval that is within a certain range of the maximum likelihood interval. This range is based off of the standard deviation of the final likelihood distribution multiplied by a constant, λ. This constant is the only hyperparameter that we tune. Using a higher λ, a kmer's final selected regions comprise either longer localization region, or may include non-contiguous regions (upon which a k-mer would be considered unlocalized). Using a smaller λ, a k-mer will get localized to a tighter region, but it is more likely that k-mer is localized incorrectly. As shown in Figs. S2E and S3F, we tried 6 different values of λ between .01 and 1, and identified .1 to provide a good balance between localization success, accuracy, and region length in both the synthetic and decoy validation datasets. Subsequent results and figures that do not specify the parameter of the λ, can be assumed to use λ=0.1

5 Supplementary Figures

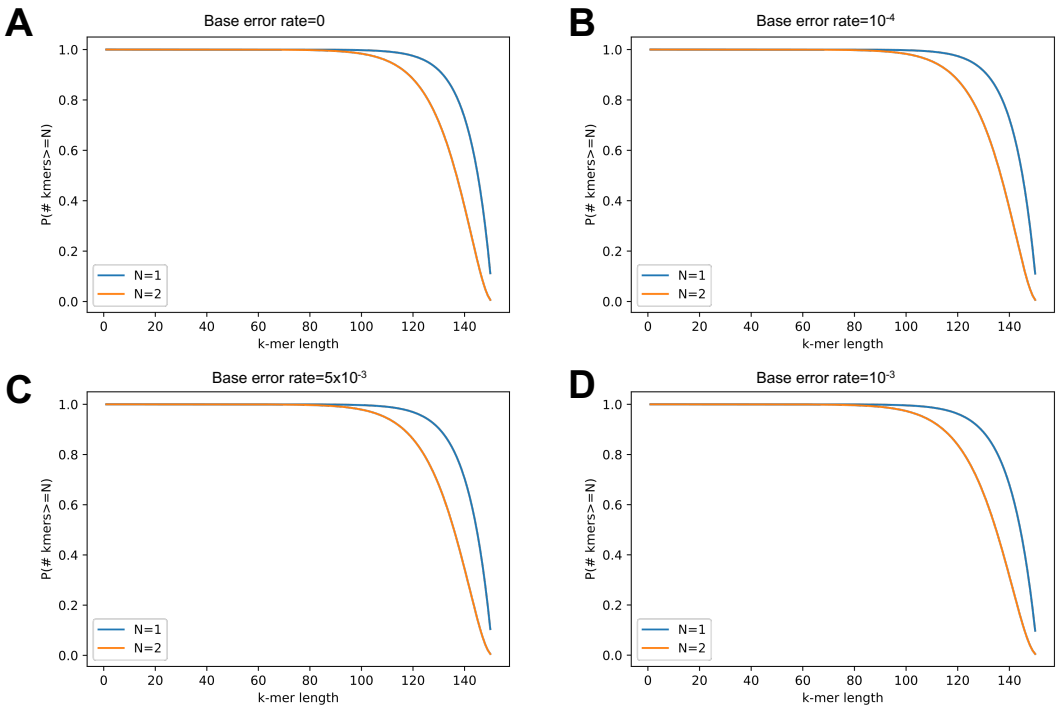
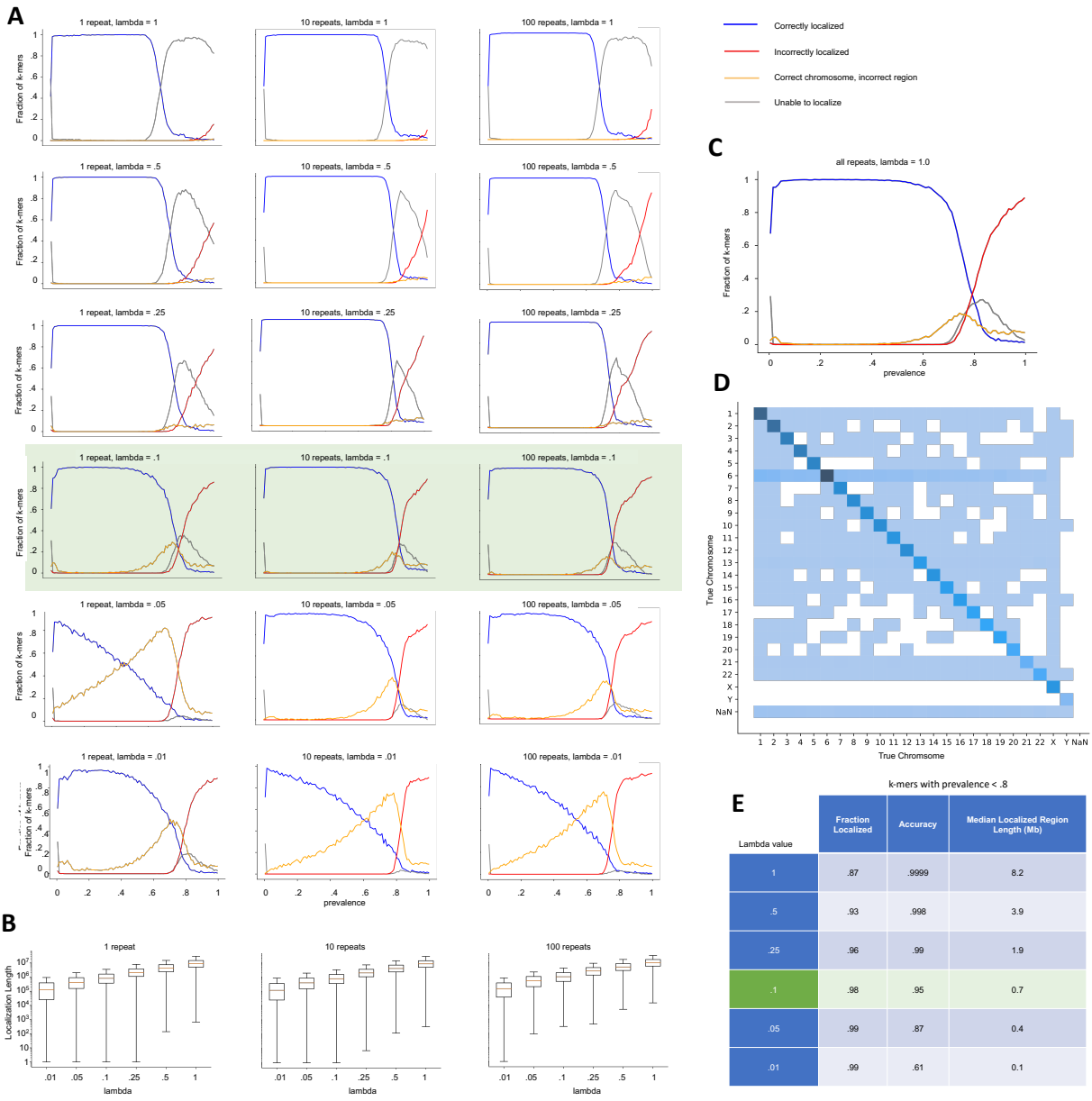


Figure S1: Determination of optimal k-mer length. We computed the probability that a k-mer of a given length would occur at least N times in an individual's set of reads, using different values of N sequencing error rate per base. We used error rates of (A) 0, (B) .0001, (C) .0004, and (D) .001. Even at the highest error rate, a k-mer length of 100 gives ¿ .95% probability that a k-mer occurs at least twice in an individual's reads if it occurs in their genome.



Lambda value	Fraction Localized	Accuracy	Median Localized Region Length (Mb)
1	.87	.9999	8.2
.5	.93	.998	3.9
.25	.96	.99	1.9
.1	.98	.95	0.7
.05	.99	.87	0.4
.01	.99	.61	0.1

Figure S2: Results and hyperparameter tuning of ASLAN on synthetic dataset. (A) Performance vs k-mer prevalence as various k-mer prevalences, tandem repeat numbers, and λ values. (B) Distribution of localized region length at varying λ values and tandem repeat numbers. (C) Summary of performance at chosen value of λ=0.1. (D) Confusion matrix of predicted chromosome vs simulated chromosome for aggregated repeat counts, using λ=.1. (E) Summary of ASLAN performance statistics across varying λ values.

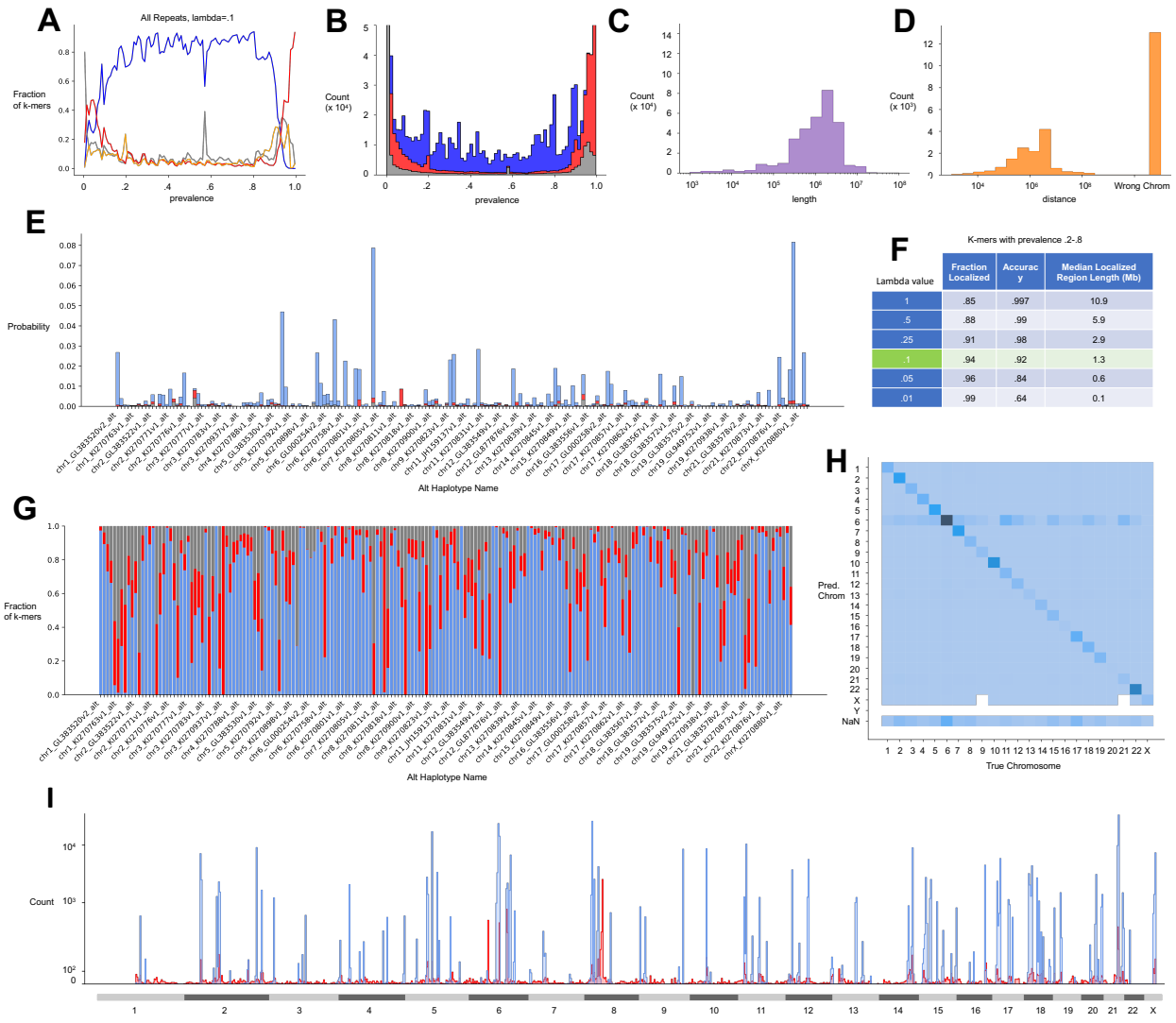


Figure S3: Results of running ASLAN on *k*-mers extracted from reads aligning to ALT sequences in the decoy genome. (A) Localization ability and accuracy for *k*-mers of various prevalences. (B) Absolute number of *k*-mers unlocalized, localized correctly, and localized incorrectly for varying prevalences. (C) Distribution of localized region lengths. (D) Distance from true location for incorrectly localized sequences. (E) Breakdown of correctly localized, incorrectly localized, and unable to localize in terms of absolute counts for each different contig. (F) Localization ability, accuracy, and median region lengths tuning λ to different values. (G) Breakdown of performance in terms of relative fraction for each contig. (H) Confusion matrix of chromosome localization. (I) Distribution of correct and incorrect localizations with respect to loci.

References

References

[1] Brianna Sierra Chrisman et al. "A method for localizing non-reference sequences to the human genome". In: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022*. World Scientific. 2021, pp. 313–324.

[2] G. David Forney. "The Viterbi Algorithm". In: *Proceedings of the IEEE* (1973). ISSN: 15582256. DOI: 10.1109/PROC.1973.9030.

[3] Peter Hall and Barbara La Scala. "Methodology and Algorithms of Empirical Likelihood". In: *International Statistical Review / Revue Internationale de Statistique* (1990). ISSN: 03067734. DOI: 10.2307/1403462.

[4] Julie Hussin et al. "Age-dependent recombination rates in human pedigrees". In: *PLoS Genetics* 7.9 (2011). ISSN: 15537390. DOI: 10.1371/journal.pgen.1002251.

[5] Kelley Paskov et al. "Estimating sequencing error rates using families". In: *BioData Mining* 14.1 (2021), pp. 1–19.

[6] Franziska Pfeiffer et al. "Systematic evaluation of error rates and causes in short samples in next-generation sequencing". In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-29325-6.

[7] Melanie Schirmer et al. "Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data". In: *BMC Bioinformatics* (2016). ISSN: 14712105. DOI: 10.1186/s12859-016-0976-y.

[8] Nicholas Stoler and Anton Nekrutenko. "Sequencing error profiles of Illumina sequencing instruments". In: *NAR genomics and bioinformatics* 3.1 (2021), lqab019.

[9] Xiangqun Zheng-Bradley et al. *Alignment of 1000 Genomes Project reads to reference assembly GRCh38*. 2017. DOI: 10.1093/gigascience/gix038.