POPULATION SEQUENCING USING SHORT READS: HIV AS A CASE STUDY

VLADIMIR JOJIC, TOMER HERTZ AND NEBOJSA JOJIC*

Microsoft Research, Redmond, WA 98052

**E-mail: jojic@microsoft.com*

Despite many drawbacks, traditional sequencing technologies have proven to be invaluable in modern medical research, even when the targeted genomes are highly variable. While it is often known in such cases that multiple slightly different sequences are present in the analyzed sample in concentrations that vary dramatically, the traditional techniques typically allow only the most dominant strain to be extracted from a single chromatogram. These limitations made some research directions rather difficult to pursue. For example, the analysis of HIV evolution (including the emergence of drug resistance) in a single patient is expected to benefit from a comprehensive catalog of the patient's HIV population. In this paper, we show how the new generation of sequencing technologies, based on high throughput of short reads, can be used to link site variants and reconstruct multiple full strains of the targeted gene, including those of low concentration in the sample. Our algorithm is based on a generative model of the sequencing process, and uses a tailored probabilistic inference and learning procedure to fit the model to the obtained reads.

Keywords: sequence assembly, population, HIV, epitome,rare variants, multiple strains, variant linkage

1. Introduction

Sequencing multiple different strains from a mixed sample in order to study sequence variation is often of great importance. For example, it is well known that even single mutations can sometimes lead to various diseases[1]. On the other hand, mutations in pathogen sequences such as the highly variable HIV[14] may lead to drug resistance[12,19]. At any given time, an HIV positive individual typically carries a large mixture of strains, each with a different relative frequency[21], and some over a hundred times less abundant than the dominant strains, and any one of them can become dominant if others are under greater drug pressure. The emergence of drug resistant HIV strains has lead to assembling a large list of associated single

mutations[a]. However, new studies are showing that there are important linkage effects among some of these mutations[18] and that the linkage may be missed by current sequencing techniques[17].

When processing mixed samples by traditional methods[20], only a single strain can be sequenced in each sequencing attempt. Multiple DNA purifications may be costly and will usually provide accurate reconstruction only of several dominant strains. Picking the less abundant strains from the mixture is a harder problem. Recent computational approaches which infer a mixture of strains directly from the ambiguous raw chromatograms of mixed samples can deconvolve strains reliably only when their relative concentrations are higher than 20%, as the rarer variants get masked[6]. Note that unlike the problem of metagenome sequencing, where multiple species are simultaneously sequenced, the goal of multiple strain sequencing is to recover a mixture of different full sequence variants of the same species, which is complicated by the high similarity among them.

Recently, a number of alternative sequencing technologies have enabled high-throughput genome sequencing. For example, 454 sequencing[13] is based on an adaptation of the pyrosequencing procedure. Several studies have demonstrated its use for sequencing small microbial genomes, and even some larger scale genomes. One of the major advantages of pyrosequencing is that it has been shown to capture low frequency mutations. Tsibris et. al have shown that they can accurately detect low frequency mutations in the HIV env V3 loop[22]. A more recent work used pyrosequencing to detect over 50 minor variants in HIV-1 protease[2]. However, these technologies also have two important limitations. First, current sequencers can only read sequences of about 200 base pairs (and some even less). Second, sequencing errors, especially in homopolymeric regions, are high, making it potentially difficult to reconstruct multiple full sequences and estimate their frequencies. In this paper, we suggest a novel method for reconstructing the full strains from mixed samples utilizing technologies akin to 454. We formulate a statistical model of short reads and an inference algorithm which can be used to jointly reconstruct sequences from the reads and infer their frequencies. We validate our method on simulated 454 reads from the HIV sequences.

[a]see http://hivdb.stanford.edu/index.html

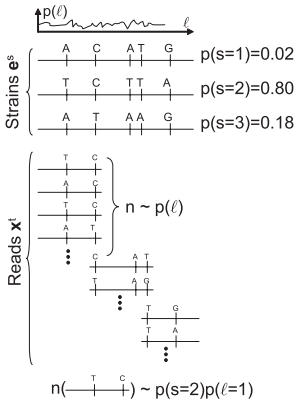

Figure 1. An illustration of population sequencing using short reads. In this toy example, three strains with five polymorphic sites are present in the sample. Short reads from various locations are taken. As the coverage depth depends on sequence content, the coverage depth will be proportional to the distribution $p(\ell)$ over the sequence location (the strains are assumed to differ little enough so that the depth of coverage of polymorphic variants of the same sequence patch are similar). The number of copies of a particular read (e.g., the TC variant shown at the bottom) depends both on the strain concentrations $p(s)$ and the depth distribution $p(\ell)$. See Section 2 for more details on notation and the full statistical model.

2. A statistical model of short sequence readouts from multiple related strains

In this section, we follow the known properties of high throughput, short read technologies, as well as the properties of populations of related sequences, e.g., a single patient's HIV population, to describe a hierarchical statistical process that leads to creation of a large number of short reads (Fig. 1). Such a generative modeling approach is natural in this case, as the process is indeed statistical and hierarchical. For example, the reads will be sampled from different strains depending on the strain concentrations in the sample, but the sampling process will include other hidden variables, such as the random insertions and deletions when the reads contain homopolymers. The statistical model will then define the optimization criterion in the form of the likelihood of the observed reads. Likelihood optimization ends up depending on two cues in the data to perform multi-strain assembly: a) different strain concentrations which lead to more frequently seen strains being responsible for more frequent reads, and b) quilting of overlapping reads to infer mutation linkage over long stretches of DNA.

We assume that the sample contains S strains $\mathbf{e}^s$ indexed by $s \in [1..S]$ with (unknown) relative concentrations $p(s)$. A single short read from the sequencer is a patch $\mathbf{x} = \{x_i\}_{i=1}^{N}$, with $N \approx 100$ and x_i denoting the $i-th$ nucleotide, taken from one of these strains starting from a random location ℓ. It has been shown that in 454 sequencing, a patch depth may be dependent on the patch content. We assume that different strains have highly related content in segments starting at the same location ℓ, and thus capture the expected relative concentrations of observed patches by a probability distribution $p(\ell)$, shared across the strains. This distribution will

also be unknown and will be estimated from the data. Under these assumptions, a simple model of the short reads obtained by the new sequencing technologies such as 454 sequencing is described by the following sampling process:

- Sample strain s from the distribution $p(s)$
- Sample location ℓ from the distribution $p(\ell)$
- Set $x_i = e^s_{i+\ell-1}$, for $i \in [1..N]$

Here we assume that the strains $\mathbf{e}^s$ are defined as nucleotide sequences $\mathbf{e}^s = \{e^s_i\}$. However, since we will be interested in the inverse process – assembling the observed patches $\mathbf{x}^t$ into multiple strains, we make the definition of $\mathbf{e}$ softer in order to facilitate smoother inference of patch mapping in early phases of the assembly when the information necessary for this mapping is uncertain. In particular, as in our previous work concerning diversity modeling and vaccine immunogen assembly[7], we assume that each site e^s_i is a *distribution* over the letters from the alphabet (in this case the four nucleotides). Thus, we denote by $e^s_i(x_j)$ the probability of the nucleotide x_j under the distribution at coordinates (s, i) of the strain description $\mathbf{e}$. We have previously dubbed the models of this nature *epitomes* as they are a statistical model of patches contained in larger sequences. Our generative model of the patches $\mathbf{x}$ is therefore refined into:

- Sample strain s from the distribution $p(s)$
- Sample location ℓ from the distribution $p(\ell)$
- Sample x by sampling for each $i \in [1..N]$ the nucleotide x_i from the distribution $e^s_{i+\ell-1}(x)$

While the epitome distributions capture both the uncertainty about reconstructed strains and the point-wise sequencing errors, in order to model possible insertions and deletions in the patch, which are important because of the assumed strain alignment (shared ℓ), we also add another variable into the process which we call 'transformation' τ, describing the finite set of possible *minor* insertions or deletions. The insertions and deletions come from two sources: a)homopolymer issues in sequencing and b) insertions and deletions among strains. The first set of issues arise when a sequence of several nucleotides of the same kind, e.g., $AAAA$ is present in the patch. In 454 sequencing, there is a chance that the number of sequenced letters in the obtained patch is not equal to the true number present in the sequence. As opposed to the indels among strains, which are usually multiples of three nucleotides to preserve translation into aminoacids, as well as consistent across the reads; the homopolymer indels are not limited in this way.

The transformation τ describes a mini alignment between the read and the epitome segment describing the appropriate strain s starting at a given location ℓ. We assume that the transformation τ will affect epitome segment just before the patch is generated by sampling from it. Thus, the statistical generative model that we assume for the rest of the paper consists of the following steps:

- Sample strain s from the distribution $p(s)$
- Sample location ℓ from the distribution $p(\ell)$
- Sample patch transformation τ from $p(\tau)$ and transform the epitome segment $\{\mathbf{e}_i^s\}_{i=\ell}^{\ell+N+\Delta}$, with Δ allowing all types of indels we want to model. This transformation provides the new set of distributions $e^s_{\tau(k)}$, where we use operator notation for τ to denote the mapping of locations.
- Sample $\mathbf{x}$ from $p(\mathbf{x}|s,\ell,\tau,\mathbf{e}) = \prod_i e^s_{\tau(i+\ell-1)}(x_i)$ by sampling for each $i \in [1..N]$ the nucleotide x_i from the distribution $e^s_{\tau(i+\ell-1)}(x)$

Each read $\mathbf{x}^t$ has a triplet of hidden variables s^t, ℓ^t, τ^t describing its unknown mapping to the catalog of probabilistic strains (epitome). In addition to hidden variables, the model has a number of parameters, including relative concentrations of the strains $p(s)$, the variable depth of coverage for different locations in the genome $p(\ell)$, and the uncertainty over the nucleotide x present at any given site i in the strain s, as captured by the distribution $e_i^s(x)$ in the epitome $\mathbf{e}$ describing the S strains. If the model is fit to the data well, the uncertainty in the epitome distributions e_i^s should contract to reflect the measurement noise (around 1%). But, if an iterative algorithm (e.g., EM) is used to jointly estimate the mapping of all reads $\mathbf{x}^t$ and the (uncertain) strains $\mathbf{e}^s$, then the uncertainty in these distribution also serves to smooth out the learning process and avoid hard decisions that are known to lead to local minima. Thus, these distributions will be uncertain early in such learning procedures and contract as the mappings become more and more consistent. In the end, each of the distributions $\mathbf{e}_i^s$ should focus most of the mass on a single letter, and the epitome $\mathbf{e}$ will simply become a catalog of the top S strains present in the sampled population. If more than S strains are present, this may be reflected by polymorphism in some of the distributions $\mathbf{e}_i^s$.

3. Strain reconstruction as probabilistic inference and learning

We now derive a simple inference algorithm consisting of the following intuitive steps:

- Initialize distributions e_i^s, strain concentrations $p(s)$ and coverage depth $p(\ell)$. More on initialization in the next section.
- Map all reads to $\mathbf{e}$ by finding the best strain s^t, location in the strain ℓ^t and the mini alignment τ that considers indels.
- Re-estimate model parameters by (appropriately) counting how many reads map to different locations ℓ and different strains s. Also count how many times each nucleotide ended up mapped to each location (s, i) in the strain reconstruction $\mathbf{e}$ and update the distributions e_i^s to reflect the relative counts.
- Iterate until convergence.

We can show that this meta algorithm corresponds to an expectation-maximization algorithm that optimizes the likelihood of obtaining the given set of reads $\mathbf{x}^t$ from the statistical generative model described in the previous section. The log likelihood of observing a given set of patches (reads) is

$$\mathcal{L} = \sum_t \log p(\mathbf{x}^t) = \sum_t \log \sum_{s^t,\ell^t,\tau^t} p(s^t)p(\ell^t)p(\tau^t)p(\mathbf{x}^t|s^t,\ell^t,\tau^t). \quad (1)$$

We note that $\mathcal{L}$ is a function of model parameters $\mathbf{e}$, $p(s)$, $p(\ell)$ and $p(\tau)$, and our goal is to maximize this likelihood wrt $\mathbf{e}$ as well as $p(s)$, as our output should be the catalog of strains, or epitome $\mathbf{e}$, present with the component concentrations $p(s)$. It is also beneficial to to maximize the log likelihood wrt other parameters, i.e. estimate the varying coverage depth for different parts of the strains as well as distribution over typical indels. Not only do these parameters may be of interest in their own right, but an appropriate fitting of these parameters increases the accuracy of the estimates of strains and their frequencies.

To express the expectation-maximization (EM)[5] algorithm for this purpose, we introduce the auxiliary distributions $q(s^t, \ell^t)$ and $q(\tau^t|s^t, \ell^t)$ that describe the posterior distribution over the hidden variables for each read $\mathbf{x}^t$, and use Jensen's inequality to bound the log likelihood:

$$\begin{aligned}\mathcal{L}(\mathbf{e}) &= \sum_t \log \sum_{s^t,\ell^t,\tau^t} q(s^t,\ell^t)q(\tau^t|s^t,\ell^t)\frac{p(s^t)p(\ell^t)p(\tau^t)p(\mathbf{x}^t|s^t,\ell^t,\tau^t,\mathbf{e})}{q(s^t,\ell^t)q(\tau^t|s^t,\ell^t)},\\ &\geq \sum_t \sum_{s^t,\ell^t,\tau^t} q(s^t,\ell^t)q(\tau^t|s^t,\ell^t)\log\frac{p(s^t)p(\ell^t)p(\tau^t)p(\mathbf{x}^t|s^t,\ell^t,\tau^t,\mathbf{e})}{q(s^t,\ell^t)q(\tau^t|s^t,\ell^t)}\end{aligned}$$

The bound is tight when the q distribution captures the true posterior distribution $p(s^t, \ell^t, \tau^t|\mathbf{x}^t)$[16], thus the reference to q as a posterior distribution. By optimizing the bound with respect to the q distribution parameters (under the constraint that the appropriate probabilities add up to one), we can

derive the E step as

$$q(\tau^t|s^t,\ell^t) \propto p(\tau^t)p(\mathbf{x}^t|s^t,\ell^t,\tau^t,\mathbf{e}) \quad (2)$$

$$q(s^t,\ell^t) \propto p(s^t)p(\ell^t)e^{\sum_{\tau^t} q(\tau^t|s^t,\ell^t)\log p(\tau^t)p(\mathbf{x}^t|s^t,\ell^t,\tau^t\mathbf{e})} \quad (3)$$

where both the computation of $q(\tau^t|s^t,\ell^t)$ and the summation over τ in the second equation are performed efficiently by dynamic programming. These operations reduce to well known HMM alignment of two sequences (in this case, one probabilistic sequence, $\{e_i^s\}_{i=\ell}^{\ell+N+\Delta}$, and one deterministic sequence, $\mathbf{x}^t$), because they estimate optimal alignment (and the distribution over alignments, and an expectation under it), in the presence of indels. In our experiments, we make an additional assumption that $q(\tau^t|s^t,\ell^t)$ puts all probability mass on one, best, alignment.

The bound simplifies the estimation of model parameters under the assumption that the q distribution is fixed. For example, the estimate of the (relative) strain concentrations and the spatially varying (relative) depth coverage is performed by

$$p(s) = \frac{\sum_t \sum_{\ell^t} q(s^t=s,\ell^t)p(\ell^t)}{\sum_t \sum_{\ell^t}\sum_{s^t} q(s^t,\ell^t)p(\ell^t)} \quad p(\ell) = \frac{\sum_t \sum_{s^t} q(s^t,\ell^t=\ell)p(s^t)}{\sum_t \sum_{\ell^t}\sum_{s^t} q(s^t,\ell^t)p(s^t)} \quad (4)$$

The estimate for the epitome probability distributions describing (with uncertainty) the strains present in the population is

$$e_i^s(x) = \frac{\sum_t \sum_{\ell^t,\tau,j|\tau(j+\ell-1)=i}[x_j=x]q(s^t=s,\ell^t)q(\tau|s^t=s,\ell^t)}{\sum_t \sum_{\ell^t,\tau,j|\tau(j+\ell-1)=i} q(s^t=s,\ell^t)q(\tau|s^t=s,\ell^t)}, \quad (5)$$

where [] denotes the indicator function. This equation simply counts how many times each nucleotide mapped to site s,i, using probabilistic counts expressed in q expectations under the possible patch alignments described by τ are again computed efficiently using dynamic programming, or, as in our experiments, they can be simplified by using the most likely alignment.

EM algorithm for our model should iterate equations (2- 5). These equations are a more precise version of the algorithm description from the beginning of the section. The iterative nature of the algorithm allows a refinement in one set of parameters to aid in refining other parameters. For example, iterating the two equations in (4) lead to estimates of strain frequency and variability in read coverage that are compatible with each other - the first equation takes into account the fact that some regions of the genome are under represented when assigning a frequency to strains based on the read counts; and the second equation discounts the effect of strain frequency on read counts in order to compute the read content dependent (approximated as genome position dependent) variability in coverage. On

the other hand, the estimate of the epitome (i.e., the catalog of strains) and the strain frequency estimates are coupled through the posterior distribution q - a change in either one of these model parameters will affect the posterior distribution (2) which assigns reads to different strains, and this will in turn affect these same model parameters in the next iteration.

4. Computational cost and local minima issues

A good boost to the algorithm's performance is achieved by its hierarchical application. The epitome $\mathbf{e}$ is best initialized by an epitome $\mathbf{e}$ consisting of a smaller number learned in a previous run of the same algorithm, e.g., by repeating each of the original S strains K times and then adding small perturbations to form an initial epitome with SK strains. If the first number of strains S was insufficient, this new initial catalog of strains contains rather uncertain sites wherever the population is polymorphic, but the alignments of variables ℓ from the previous run for all patches are likely to stay the same, so that part of each distribution $q(s, \ell)$ is transferred from the previous run, and does not change much, thus making it possible to avoid search over this variable and reduce complexity. An extreme application of this recipe, that according to our experiments seems to suit HIV population sequencing, is to run the algorithm first with $S = 1$, which essentially reduces to consensus strain assembly in noisy conditions, and then increase the catalog $\mathbf{e}$ to the desired size. For a further speed up, a known consensus sequence (or a profile) can be used to initialize all strains in the epitome.

The simple inference technique described above still suffers from two setbacks. One problem is computational complexity. The number of reads can be very large, although these reads may be highly redundant at least for all practical purposes in the early iterations of the algorithm. Another, more subtle problem is the weakness of the concentration cues in inference using our model, which may cause local maxima problems. Our generative model mirrors the true data generation process closely, and thus the correct concentrations in conjunction with properly inferred strains correspond to the best likelihood. But if pure EM learning is applied, the concentration cue can be too weak to avoid local minima in $\mathbf{e}$. Fortunately, a simple technique can be used to address both of these two issues. Reads are clustered using agglomerative clustering and the initial q distributions are estimated by mapping the cluster representations rather than all reads.

The ℓ mapping is considered reliable and fixed after that point as the described initialization makes all strains similar enough to the true solution for the purposes of ℓ mapping (but not for inferring strain index s). In the first

few iterations after that, clusters are mapped to different strains, but the epitome distributions are not considered in this mapping - the assumption is made that the final set of parameters will map clusters so that all strains in the epitome are used. Each cluster mapping is iterated with updates of concentrations of $p(s)$. This results in loosely assigning read clusters with similar frequencies to the same strain. After 2-3 such iterations, epitome distributions are inferred based on the resulting q distribution, and then the full EM algorithm, over all patches, is continued. This is necessary as the agglomerative clusters may not be sufficient to infer precisely the content of all sites until individual reads are considered. It should be noted that due to the high number and overlap of reads, it is in principle possible to have a substantially lower reconstruction error than the measurement error(1%).

In our implementation, the computational cost is quadratic in the number of patches associated with particular offset in the strains, due to the agglomerative clustering step. The cost of an EM iteration is proportional to the product of the number of patches (reads) and the total length of the epitome (strain catalog).

5. Experimental validation

We assessed performance of our method on sequence data for nef and env regions of HIV. Starting with these sequences, we simulated 454 reads as 80-120 nucleotide long patches $\mathbf{x}^t$ generated by the statistical generative model described in Section 2. The generated reads, without the model parameters or results of the intermediate steps, were then analyzed using the inference technique in Section 3 to reconstruct the hidden variables, such as read-to-genome alignments ℓ^t and read-to-strain assignments s^t, and estimate the model parameters, most importantly the epitome, or strain catalog, $\mathbf{e}$, and the strain frequencies $p(s)$. These were then compared to the ground truth.

The overall error rate in 454 reads is estimated at 0.6%. For our generated reads, we set substitutions errors of 1.0%, and for homopolymers (of length at least 2 nucleotides) we set the rate of insertion at 2%, and deletion at 0.5%. The read selection probability – the probability of obtaining a read from a particular offset from a particular strain – is set to be proportional to the product of depth of coverage $p(\ell)$ at the offset ℓ and the frequency of the strain $p(s)$ (see also Fig. 1). The depth of coverage is randomly drawn from a preset range of values (and, as other parameters, it was not later provided to the inference engine, which had to reconstruct it to infer correct strain frequencies). We assume that overlap between reads is no less than 50 nucleotides.

Table 1. The fraction of nucleotides reconstructed correctly in the least frequent strain as a function of that strain's frequency and the minimum number of reads.

Min. reads \ Frequency	0.1%	0.5%	1%	2%
10	40.93%	92.59%	100%	100%
20	62.25%	95.10%	100%	100%
30	100%	100%	100%	100%

In order to assess ability of the method to reconstruct low frequency strains we first created a dataset of 10 nef strains[14]. The nef region is approximately 621 nucleotides long. We randomly picked one strain as the low frequency strain. For this lowest frequency we considered four possibilities: 0.1%, 0.5%, 1%, and 2%. For the other 9 sequences, we randomly chose frequencies between 2% and 100% and then normalized them so that the sum of frequencies is 100%, i.e., $\sum p(s) = 1$. Then, the short reads were generated from the mixture as described above. Though the depth of coverage $p(\ell)$ was randomly assigned across the region, we ensured, by scaling the total number of reads, that a minimum number of reads is guaranteed for each genome location. We experimented with three possibilities for this minimum number of reads: 10, 20, and 30. The Table 1 illustrates the impact of the minumum number of reads on our ability to reconstruct sequences with small concentrations. Even in case of the minor strain frequency of just 0.1% we were able to reconstruct all ten sequences as long as we had suitable number of reads available. Furthermore, all strain frequencies were recovered with negligible error.

We also assessed the impact of the density of viral mutations on our ability to reconstruct the full strains. We used 10 HIV env strains from MACS longitudinal study[9]. All sequences originated from the same patient and were obtained from samples collected at 10 different patient visits. The visits occurred approximately every 6 months. Whereas variable strain frequencies may help us disambiguate between frequent and infrequent strains, in case of comparable frequencies, it is the mutations which occur in the overlap between reads which enable linking of site variants and the reconstruction of full sequences. In order to assess the number and proximity of mutations in env, we analyzed sequences collected from a single patient over a number of visits spanning 8 years. These sequences contained 280 nucleotides of gp120, followed by V3 loop, followed by 330 nucleotides of gp41, total of 774 nucleotides. The entropy of these sequences at each site is shown in Figure 2. The positions with high entropy are spaced almost

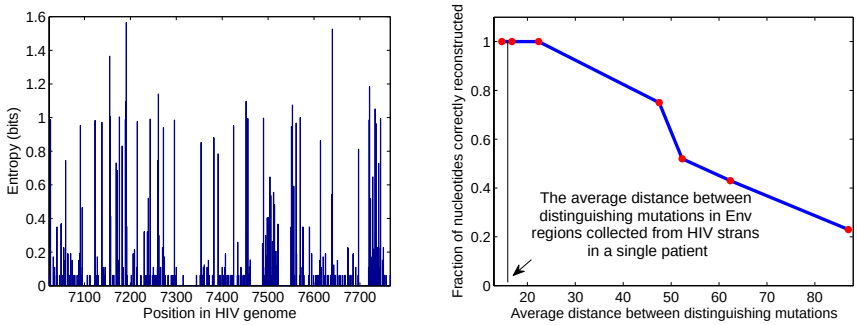

Figure 2. Left: Site entropy for an Env region, estimated over 137 sequences originated from the same patient. Note that the positions with high entropy are spaced almost uniformly throughout this region. The average distance between positions with entropy greater than 0.5 is 14.67. Right: From this dataset we selected 8 different sets of 10 sequences, each with different density of distinguishing mutable positions. We evaluated fraction of nucleotides correctly reconstructed for various densities of distinguishing mutations, represented as average distance between the distinguishing mutable positions. The vertical line traces the average distance between mutable positions in Env.

uniformly throughout this region, with separation between significantly mutable position (entropy greater than 0.5) reaching up to 57 nucleotides.

The difficulty of disambiguating strains of comparable frequency is dependent on the maximal distance between pairs of adjacent mutations. In regions where two nearest mutable positions are separated by a conserved region longer than the read length, there will be no reads spanning both of those mutable positions, and we may not be able to tell whether mutations at the two sites are occurring in the same strain or not. In these cases, we should assume that linking of mutations is correct only in parts up to and after the conserved region, but not across this region, unless the strain frequencies are sufficiently different to allow our algorithm to correctly match the separated pieces based on the frequency of site variants. Therefore, the density of the distinguishing mutable positions is a measure of difficulty of disambiguating strains of comparable frequency.

We varied the average distance between adjacent mutations in a controlled manner. More specifically, we created 8 sets of 10 Env sequence mixtures, with average distances ranging from 10-80 bases apart, and computed the percentage of correct reconstructions for each set. Figure 2 shows reconstruction accuracy as a function of mutation density, defined as an average distance between the distinguishing mutable positions.

6. Conclusion

We introduced a population sequencing method which recovers full sequences and sequence frequencies. The method leverages inherent differences in the strain frequencies, as well as the sequence differences across the strains in order to achieve perfect reconstruction under a noise model mirroring the measurement error of the 454 sequencing method. We have shown that our method can reconstruct sequences with as small a frequency as 0.01%. While our experiments have been performed on simulated (but realistic) mixes of short segments of HIV, there is no technical reason why the technique would not work for longer genomes (e.g., entire HIV sequences or longer viral sequences). For most of HIV, the density of mutable positions is so high, that the technique should work with significantly shorter reads than 200. For more information, visit www.research.microsoft.com/∼jojic/popsequencing.html.

References

1. E. J. Baxter, et al. *Lancet*, 365(9464):1054–1061, Mar 2005.
2. C. Wang, et al. *Genome Res*, Jun 2007.
3. J. M. Coffin. *Science (New York, N.Y.)*, 267(5197).
4. DA. Lehman and C. Farquhar. *Rev Med Virol*, Jun 2007.
5. A. P. Dempster, N. M. Laird, et al. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
6. N. Jojic. Population sequencing from chromatogram data. In *ISMB, PLOS track*. 2006.
7. N. Jojic, et al. In Y. Weiss, B. Schölkopf, et al., eds., *Advances in Neural Information Processing Systems 18*, pp. 587–594. MIT Press, Cambridge, MA, 2006.
8. D. Jones, et al. *AIDS Res Hum Retroviruses*, 21(4):319–324, Apr 2005.
9. R. A. Kaslow, et al. *Am J Epidemiol*, 126(2):310–318, Aug 1987.
10. P. Kellam and B. A. Larder. *J Virol*, 69(2):669–674, Feb 1995.
11. B. Li, et al. *J Virol*, 81(1):193–201, Jan 2007.
12. S. Lockman, et al. *N Engl J Med*, 356(2):135–147, Jan 2007.
13. M. Margulies, et al. *Nature*, 437(7057):376–380, Sep 2005.
14. C. B. Moore, et al. *Science*, 296(5572):1439–1443, May 2002.
15. S. M. Mueller, et al. *J Virol*, 81(6):2887–2898, Mar 2007.
16. R. Neal and G. Hinton. In M. I. Jordan, ed., *Learning in Graphical Models*. Kluwer, 1998.
17. S. Palmer, et al. *J Clin Microbiol*, 43(1):406–413, Jan 2005.
18. S.-Y. Rhee, et al. *PLoS Comput Biol*, 3(5):e87, May 2007.
19. T. Ridky and J. Leis. *J Biol Chem*, 270(50):29621–29623, Dec 1995.
20. F. Sanger, et al. *Biotechnology*, 24:104–108, 1992.
21. T.-K. Seo, et al. *Genetics*, 160(4):1283–1293, Apr 2002.
22. A. Tsibris, et al. In *Antivir Ther.*, vol. 11:S74 (abstract no. 66). 2006.