

# Mixture Sequencing from Chromatogram Data

Delbert Dueck<sup>1</sup>, Chopra Abha<sup>2</sup>, Corey Moore<sup>2</sup>, Jules Sarich<sup>2</sup>, Damian Goodridge<sup>3</sup>, David Heckerman<sup>4</sup>, David Sayer<sup>3</sup>, Simon Mallal<sup>2</sup>, Nebojsa Jojic<sup>4</sup>

<sup>1</sup> University of Toronto, 10 King's College Road, Toronto, Canada M5S 3G4

<sup>2</sup> Centre for Clinical Immunology and Biomedical Statistics, Royal Perth Hospital, Wellington Street, Perth, Western Australia

<sup>3</sup> Conexio 4, East Fremantle 6158 Western Australia

<sup>4</sup> Microsoft Research, One Microsoft Way, Redmond WA 98052 USA  
jojic@microsoft.com

**Abstract.** One of the key components of sequencing technologies is proper separation of a single species/strain/allele of the targeted sequence from a sample. In traditional techniques, this has been achieved chemically (*e.g.*, using specific primer sequences), however, multiple different but related species can still possibly be picked up with the same primer. This is especially problematic in sequencing RNA or proviral DNA, when the virus in question is highly variable and each individual is infected with a different swarm of viral strains. In case of HIV, for example, when the dominant sequences in the population differ by one or more insertions and deletions, the standard sequencing techniques fail to recover any of the component strains sufficiently well. We show that the chromatograms of mixed sequences can be used to accurately infer the individual strains, removing the need for additional sequencing steps, *e.g.* new primer synthesis or cloning of individual viral variants. To this purpose, we have developed a statistical generative model of raw chromatogram data and an appropriate inference algorithm based on maximizing the likelihood of an observed chromatogram. To illustrate this technique, we used an automated ABI 3730XL sequencer to capture mixed samples of pro-viral DNA of HIV-infected patients. The chromatograms of the mixed samples were analyzed by the presented algorithm, providing the inferred individual strains for each mixture as output. The mixture components were then compared with the sequences of the original clones. In many cases, the separated components had fewer than 1% differences to the ground truth which compares favorably to the output of the basic sequencer, whose errors went up to 40%.

## 1 Introduction

In immunology research, obtaining as complete as possible a population of viral strains by sequencing a single sample is highly desirable, but difficult due to drawbacks of current sequencing technology. Sequencing highly polymorphic organisms, such as viruses, is complicated by diversity present in a sample. Current Sanger sequencing typically provides only the readout of a single dominant strain in the mix with some detectable ambiguities corresponding to individual site polymorphisms. However, linking polymorphic variants at different sites and

disambiguating complete multiple strains is not possible using standard software. Furthermore, signals from strains with very low concentrations are obliterated by more dominant strains.

The field of gene sequencing is at the crossroads. On the one hand, new techniques such as those proposed by 454 Life Sciences [1], allow sequencing of a large number of short (up to about 250 nucleotides) subsequences, possibly allowing detection of segments of strains that are present in fairly low concentrations. On the other hand, assembly of short segments acquired this way into full strains is an area of active research. The more traditional Sanger sequencing provides much longer readouts, but the problem of extraction of multiple strains from these signals is unsolved. Previous work has focussed only on *detecting* single-nucleotide polymorphisms (SNPs) in chromatograms ([2], [3], [4], [5], [6]), or insertions/deletions ([7]) without providing deconvolved multiple-strain outputs. Recently, an analysis of ambiguous decoding has also been proposed [18]. In this paper, we provide a new algorithm for analysis of low-level cues in chromatogram data that can help resolve the problem of disambiguating multiple strains from a single chromatogram signal, which on the one hand, makes Sanger technology more directly useful for population sequencing, and on the other provides a way for improving new generation of sequencing solutions, by providing a multi-strain skeleton for mapping short readouts. The importance of disambiguation of chromatograms cannot be overstated. In addition to fundamental technological problems that still need to be resolved for the next generation of sequencing solutions to be used for population sequencing of longer gene segments, the current Sanger technology is also more cost-effective, and large HIV sequencing studies are underway based on it. It is also likely that the new and old technologies can be combined for efficient solution for population sequencing.

Intermediate output of sequencing technologies (e.g. smoothed chromatograms obtained by ABI sequencers) can often appear to be of low quality when in fact it should be appreciated as containing *more* usable data than a clean, high quality signal. Unreadable chromatograms are often caused by the presence of multiple strains or alleles of the targeted sequence region, especially when these strains differ from each other by one or more deletions and/or insertions. This situation will tend to appear in the most important sequencing tasks — the ones targeting variable regions of the sequence of interest. One such example is HIV sequencing, wherein a large fraction of chromatograms is discarded even though the very cause of the unreliable readout — the presence of multiple HIV strains — is of great interest to the research community.

In §3, we point out the amplitude and phase cues that could reveal the individual strains in the chromatograms of mixtures, then develop a statistical generative model that uses these cues as well as a diversity profile to assign likelihood to different chromatogram decodings, and the inference algorithm that separates the components in the mixture by maximizing this likelihood. We also present preliminary, but strongly indicative experimental validation of the algorithm: we show that in controlled mixing situations, where we can know the sequences for the mixed strains, we can infer the component mixtures with high accuracy (less than 1% of erroneously decoded nucleotides). The mixed components all have insertions and deletions, and they also differ by up to 10% of

point mutations, which makes the decoding especially error prone using standard sequencing methods (up to 40% error).

## 2 Chromatography Background

Chromatography is a fast and effective mode for sequencing DNA, and is most commonly performed with the *chain termination method* [8]. First, DNA molecules selected for sequencing will be amplified through rounds of PCR. Once the concentration is sufficiently high, a final annealing is performed. Previously, this would require the four deoxyribonucleotide triphosphates (dNTPs: dATP, dCTP, dGTP, dTTP) as nucleotide building blocks to help grow the strands, but this time a small amount of a synthetic dideoxynucleotides (ddNTPs: ddATP, ddCTP, ddGTP, ddTTP) are added instead. They differ by not having a hydroxide ion on their sugar component where the next nucleotide would attach; thus, the growing chain terminates. This process takes place stochastically at different locations in the strand for each molecule, thus the resulting mixture will contain some DNA fragments of each intermediate length.

Secondly, polyacrylamide gel electrophoresis [9] is applied to sort fragments by time taken to travel along gel-filled capillaries, which monotonically corresponds to fragment mass and length. Upon reaching the end of the capillary, fragments pass by a laser, and since ddNTPs are also labeled with different fluorescent compounds [10], they emit a color distinct to each nucleotide that can be detected and recorded. The ABI 3730 machine automates this process and includes four-channel scanning of the lanes, base-calling (interpreting the traces of the scanner), and output into a standard AB1 file.

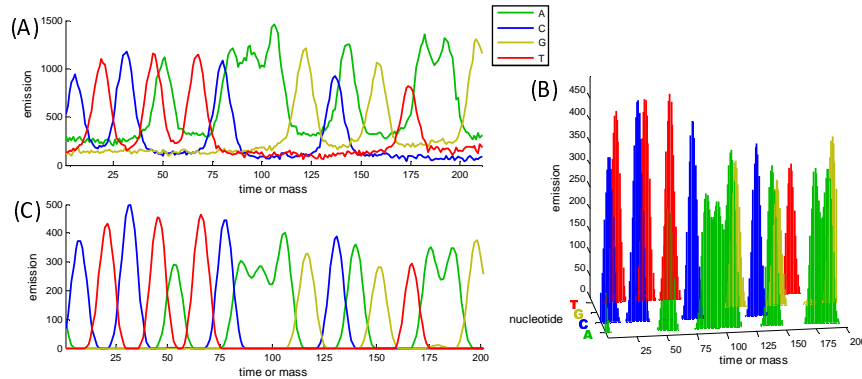
Because this process, like PCR, relies on primers to isolate DNA, it is susceptible to getting mixtures DNA sequences that differ in regions outside the primer. This is a common cause of noisy chromatogram trace files that confound standard base-callers looking for pure traces. It has been shown ([11]) that there are important cues in these chromatogram files, such as precise positioning and amplitude of peaks, that should enable decoding of chromatograms with more than one sequence.

## 3 The statistical generative model of chromatograms

The built-in ABI base-calling software is designed to operate on clean data and thus disregards much of the signal amplitude and position information present in chromatograms. More sophisticated base-callers, such as Phred [12] account for uncertainty and release confidence measures for each base they call, but they still only model and extract a single sequence from each chromatogram input.

In order to develop an algorithm for extracting more information from mixtures, we first develop a generative model that describes the variability typically observed in chromatograms of mixed signals through various involved variables such as component sequences, their amplitudes, and their phase offsets. The model is statistical in nature, and is described by conditional probability distributions which allow various types of noise and uncertainty in the model. The

product of these conditionals forms the joint probability distribution over all model variables, and given a noisy chromatogram, the estimate of the variables of interest (*e.g.*, the component sequences) can be obtained by statistical inference.



**Fig. 1.** Chromatography data from the ABI 3730, with one trace for each nucleotide. (A) shows raw data, before correcting for gel and mobility effects and smoothing, as shown in (C). (B) shows how the data can be seen as a four-channel histogram of DNA sequence fragments terminated by ddNTPs.

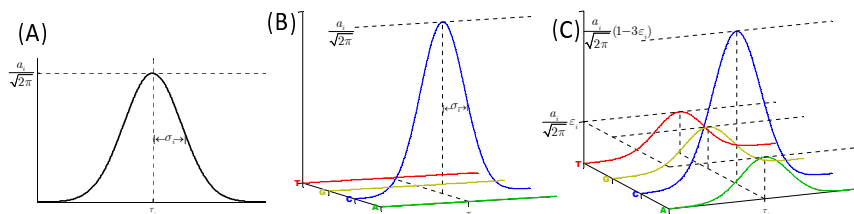
We will describe the model from bottom up, starting with a simple model of a chromatogram trace obtained from the ABI sequencer. ABI sequencer performs low-level signal processing on raw channel traces as shown in Fig. 1A, correcting for nonlinear gel mobility effects, and other effects which we are currently not modeling. Beginning with the signal for one nucleotide,  $x(t)$ , we can consider the signal as a histogram of terminating ddNTP appearance frequencies plotted against time (or mass),  $t$ , as illustrated in Fig. 1B. We construct a probability model of the data and express the likelihood of the data given the model as:

$$P(x | \text{model}) = \prod_t P[\text{ddNTP appearance at } t]^{x(t)} \quad (1)$$

If we treat ddNTP as being normally-distributed about discrete positions in the DNA chain, the probability model becomes a familiar mixture of Gaussians with a latent class variable  $\mathbf{c} = \{c_1, c_2, \dots, c_t, \dots\}$  assigning a class label to each time value:

$$P[\text{ddNTP appearance at } t] = a_{c_t} \frac{1}{\sqrt{2\pi}\sigma_{c_t}} e^{-(t-\tau_{c_t})^2/2\sigma_{c_t}^2} \quad (2)$$

where the model parameters,  $\Theta$ , are as follows:  $a_{c_t}$  represents the amplitude (mixture proportion),  $\tau_{c_t}$  the peak position (mean of the Gaussian), and  $\sigma_{c_t}$  the peak width (standard deviation of the Gaussian). One such peak (with index  $c_t = i$ ) is shown in Fig. 2A. This leads to the following complete log-likelihood



**Fig. 2.** Peaks in a chromatogram treated as components in a Gaussian mixture model. (A) shows the basic model with peak amplitude ( $a_i$ ), position ( $\tau_i$ ), and width ( $\sigma_i$ ) parameters. (B) generalizes to a four-channel signal with the peak being located in exclusively in one nucleotide’s channel. (C) further generalizes this by modelling uncertainty (allowing ‘leakage’) of proportion  $\varepsilon_i$  to each of the other nucleotide channels.

of the data and class labels:

$$\log P(X, \mathbf{c}|\Theta) = \sum_t x(t) \log \left\{ \frac{a_{c_t}}{\sqrt{2\pi}\sigma_{c_t}} e^{-(t-\tau_{c_t})^2/2\sigma_{c_t}^2} \right\} \quad (3)$$

The EM algorithm could be used to infer class labels and learn model parameters separately for each of the four data channels,  $\{x^A(t), x^C(t), x^G(t), x^T(t)\}$ , but we find it advantageous to consider the channels jointly, storing a class label for each (channel, time value) pair,  $c_{lt}$ . We also introduce a nucleotide parameter,  $\ell_{c_{lt}}$ , for each class:

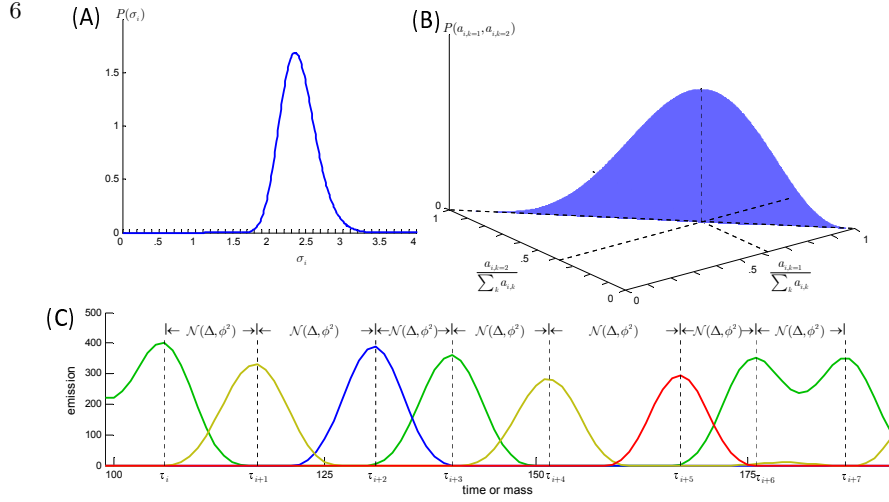
$$\log P(X, \mathbf{c}|\Theta) = \sum_{l \in \{A, C, G, T\}} \sum_t x^l(t) \log \left\{ [l = \ell_{c_{lt}}] \frac{a_{c_{lt}}}{\sqrt{2\pi}\sigma_{c_{lt}}} e^{-(t-\tau_{c_{lt}})^2/2\sigma_{c_{lt}}^2} \right\} \quad (4)$$

Here the Iverson notation is used where [true] = 1 and [false] = 0; this peak-to-nucleotide assignment is shown in Fig. 2B. It is important to keep class labels consistent with nucleotide parameters *i.e.* the label of channel A at time  $t$ ,  $c_{At}$ , should index a Gaussian whose nucleotide label is also A, which means  $\ell_{At} = A$  or else (4) will give a log-likelihood of  $-\infty$ . This leads to optimization difficulties in the EM algorithm — nucleotide labels will never greedily optimize beyond initialization — so we introduce a nucleotide ‘uncertainty’ or ‘leakage’ parameter,  $\varepsilon_{lt}$ , (see Fig. 2C) representing the proportion of the Gaussian placed on all other nucleotides. The complete log-likelihood then becomes:

$$\log P(X, \mathbf{c}|\Theta) = \sum_{t,l} x^l(t) \log \left\{ a_{c_{lt}} \frac{\varepsilon_{c_{lt}}^{[l \neq \ell_{c_{lt}}]} (1-3\varepsilon_{c_{lt}})^{[l = \ell_{c_{lt}}]}}{\sqrt{2\pi}\sigma_{c_{lt}}} \exp \left( \frac{-(t-\tau_{c_{lt}})^2}{2\sigma_{c_{lt}}^2} \right) \right\} \quad (5)$$

Next, we label each Gaussian with an ordered pair  $(i, k)$  where  $i$  indexes the base pair (there are typically several hundred in a chromatogram) and  $k = \{1, \dots, K\}$  indexes the strain number. We confine our experimental results to chromatograms containing  $K = 2$  distinct strains.

We use additional information in the form of conjugate prior distributions over parameters to help with sequencing. First, we put a gamma prior on  $\sigma_{ik}^{-2}$



**Fig. 3.** Conjugate priors on latent model variables. Because peak widths are well-known to fall within a certain range, it is given a tight gamma prior distribution as illustrated in (A). Mixing proportions are assumed to be a shared property of the entire mixture (one concentration per strain), and this is controlled with a Dirichlet prior as illustrated in (B) in two dimensions. Peaks spacing is fairly constant (C) – in increments equivalent to the incremental mass of each DNA fragment – so they are given a Gaussian prior around the previous peak location plus a global constant,  $\Delta$ .

reflecting the knowledge that peak width (i.e. ddNTP mass) is relatively constant between molecules:

$$P(\sigma_{ik}^{-2}) \propto (\sigma_{ik}^{-2})^{\gamma-1} e^{-\sigma_{ik}^{-2}/\beta} \quad (6)$$

where  $\gamma$  and  $\beta$  are shared shape and scale hyperparameters, respectively, dependant on signal resolution only. This is shown in Fig. 3A.

Chromatography emission levels *i.e.* amplitudes on chromatogram traces are proportional to the concentration of DNA fragments with a particular terminating ddNTP nucleotide. These tend to decrease as fragment mass increases (strains lengthen) beyond a certain point and signal quality degrades. In small neighborhoods, however, relative amplitudes of peaks should be approximately proportional to the relative concentrations of strains and thus can provide a clue for decoding. We use a Dirichlet prior with one hyperparameter ( $\mu_1, \dots, \mu_K$ ) per strain so  $P(a_{ik}) \propto (a_{ik})^{\mu_k}$ , where  $\mu_k$  is a constant reflecting the relative abundance of the strain in the mixture. See Fig. 3B for a multidimensional illustration of this distribution.

As discussed previously, the position of each peak in a chromatogram trace is proportional to the time a DNA fragment terminating at that location takes to reach the sensor. This time may depend on the mass, orientation, and shape of each fragment, with fragments from the same strain being correlated by these properties. For example, the mass of fragments in each strain grows in discrete steps corresponding to the mass of each dNTP, and so after enough differences between strains, the mass difference may become sufficient to incur a delay of one strain with respect to the other as they travel through the gel. Thus, it is important to capture statistical dependencies among the assigned peak positions

by modeling the  $i^{\text{th}}$  peak position in the  $k^{\text{th}}$  strain as being normally-distributed about the previous peak position in the same strain, plus a constant,  $\Delta$ :

$$P(\tau_{ik}|\tau_{i-1,k}) = \frac{1}{\sqrt{2\pi}\phi_k} e^{-(\tau_{ik}-\tau_{i-1,k}-\Delta)^2/2\phi^2}$$

Here,  $\Delta$  is assumed to be a constant dependant on the sample resolution of the chromatogram, and  $\phi^2$  a shared variance representing the importance placed on this prior. This is illustrated in Fig. 3C.

The model can specialize on a certain species by involving its diversity profile. For our purposes, we use an HIV profile,  $H$ , computed from the data in [13]. This profile consists of position-independent multinomial distributions (over four nucleotides) for the 9700 HIV base positions. (Alternatively, a mixture model based on the same data can be used). The assigned peaks describe a decoding of each strain  $k$ , and this decoding is expected to be in accordance with this profile, given some alignment  $\mathbf{b}_k$ . While this alignment is most generally another mapping of the base pairs, in our implementation of the model we consider the chromatograms in shorter overlapping windows and so it is sufficient to describe the alignment  $\mathbf{b}_k$  as a single specific location in the profile. We use this information as a prior on the nucleotide assignments of each peak as follows:

$$P(\ell_{ik}|b_{ik}) = (1 - 3\varepsilon_{ik}) \cdot h_{b_{ik}}(\ell_{ik}) + \varepsilon_{ik} \cdot (1 - h_{b_{ik}}(\ell_{ik}))$$

where  $b_{ik}$  is a base position pointer into the profile for the  $i^{\text{th}}$  peak in the  $k^{\text{th}}$  strain and  $H = \{h_1(l), h_2(l), \dots, h_{9700}(l)\}$  with  $l \in \{\text{A, C, G, T}\}$  is the HIV diversity profile. For example, the probability that the nucleotide at position 6500 in the envelope region of the profile is ‘G’ is  $h_{6500}(\text{G})$ .

To compute the most likely decodings for all strains, in our preliminary experiments we used a fast approximate variational inference technique [14]. We approximate the product of the likelihood,  $P(X, C|\Theta)$ , and the priors described above, with another distribution,  $Q(C)$ :

$$P(X, \mathbf{c}|\Theta)P(\sigma_{ik}^{-2})P(a_{ik})P(\tau_{ik})P(\ell_{ik}|b_{ik}) \approx Q(\mathbf{c}) \quad (7)$$

where

$$Q(\mathbf{c}) = \prod_i \prod_k \prod_l \prod_t q_{iklt}^{[c_{lt}=(i,k)] \cdot x^l(t)} \text{ s.t. } \sum_{i,k} q_{iklt} = 1 \quad \forall l, t \quad (8)$$

There will inevitably be a considerable amount of uncertainty in some strain assignments where amplitude, profile, and spacing cues for peaks are inconsistent. In these cases, reversing a close decision involves switching peaks between strains which does not happen easily if the Gaussians are already mostly-fit and the free energy is near a local minimum. For this reason, we introduce a final latent variable,  $\mathbf{r}$ , associated with each peak position,  $i$ , and use it to enumerate all possible strain permutations. For two-strain case, there are  $2! = 2$  possible permutations, namely  $(1, 2) \rightarrow (1, 2)$  and  $(1, 2) \rightarrow (2, 1)$ . The  $Q$ -distribution is modified accordingly:

$$Q(\mathbf{c}, \mathbf{r}) = Q(\mathbf{c}) \cdot \prod_i \prod_{\kappa=1}^{K!} (\rho_{i\kappa})^{[r_i=\kappa^{\text{th}} \text{ permutation}]} \text{ s.t. } \sum_{\kappa=1}^{K!} \rho_{i\kappa} = 1 \quad \forall i \quad (9)$$

We then proceed to minimize the free energy:

$$F = \int_{\mathbf{r}} \int_{\mathbf{c}} Q(\mathbf{c}, \mathbf{r}) \log \frac{Q(\mathbf{c}, \mathbf{r})}{P(\mathbf{X}, \mathbf{c}|\Theta) P(\sigma^{-2}) P(\mathbf{a}) P(\mathbf{t}) P(\ell|\mathbf{b})} \quad (10)$$

Minimization is efficiently performed by setting parameter derivatives to zero (i.e. co-ordinate descent). The one exception is the  $\mathbf{b}$  profile pointer parameter, which we learn in each iteration by maximizing the correlation between overlapping windows (e.g. 30 base pairs) of the decoded sequence and the profile.

## 4 Experimental Results: Identifying mixtures of sequences

We tested our algorithm on ten chromatograms obtained from HIV patients in the Centre for Clinical Immunology and Biomedical Statistics in Perth, Australia [13]. Four of the chromatograms, labeled ‘C1’, ‘C2’, ‘C3’, and ‘C4’, were obtained from ‘clean’ samples, *i.e.*, the chromatogram contained one unambiguous sequence. The remaining six (‘M1’, ‘M2’, ‘M3’, ‘M4’, ‘M5’, ‘M6’) were created by mixing clean samples in different proportions in the lab, and then performing chromatography on them.

After analyzing these six chromatograms with our algorithm, we aligned the inferred mixture strains with each of the ground truth sequences comprising the mixture, and obtained error rates shown in Table 4 (only the best matches are shown for each sequence). ABI basecalling refers to the decoding provided by the standard ABI software, designed with the assumption that the sample is clean. Since the samples were in fact mixed, ABI software must fail to decode one of the sequences, and so we report a single error rate for the one that was best decoded. To establish an error rate for ABI software, we compared its decoding with both mixture components as shown in Fig. 4.

Figure 4 shows alignment and error rates of each the six mixed chromatograms to the corresponding pair of mixture component chromatograms. As chromatogram quality declines after several hundred bases (and may be spotty at the beginning too), we compute error rates only in central regions (kept consistent across all algorithms for each sample) shown as green backgrounds. The component sequences were substantially different from each other: sequences C1 and C2 contained much of the gp120 envelope region in HXB2 and differed in 16% of the sites. In addition, while C1 and C2 have the same start, they differ by an insertion of three nucleotides followed later by a deletion of three nucleotides. Clean sequences C3 and C4 contained the gp41 envelope region and differed in 14% of sites, and also exhibit differences due to insertion/deletions. The sample mixes were created so that they lead to typical ambiguous chromatograms that HIV researchers encounter in practice, but in a controlled environment where the ground truth is known. It should be noted that ENV genes are very important for vaccine design, as the ENV protein is potentially exposed to antibodies. However, the immune pressure leads to high variability in this region of HIV, and so most of the ENV chromatograms are not single-strain. More often than not,

Error rates for identifying DNA mixtures by ABI base calls and our method, after alignment with ground truth. Some chromatograms deteriorate sooner than others, so we compute error rates for different regions of some samples as shown in Fig. 4’s green backgrounds. Mixture concentration estimates are also shown for our method.

Chromatogram label	Composition (ground truth)	Decoding Method	Mixture Estimate	Difference
<b>M1</b>	(50% C1+ 50% C2)	ABI basecaller	C2	42.54%
		decoded strain #1	59% C2+	0.55%
		decoded strain #2	41% C1	2.39%
<b>M2</b>	(62.5% C1+ 37.5% C2)	ABI basecaller	C1	30.59%
		decoded strain #1	60% C1+	0.34%
		decoded strain #2	40% C2	1.01%
<b>M3</b>	(83% C1+ 17% C2)	ABI basecaller	C1	0.75%
		decoded strain #1	75% C1+	0%
		decoded strain #2	25% C2	19.58%
<b>M4</b>	(50% C3+ 50% C4)	ABI basecaller	C4	38.35%
		decoded strain #1	59% C3+	9.02%
		decoded strain #2	41% C4	8.65%
<b>M5</b>	(62.5% C3+ 37.5% C4)	ABI basecaller	C3	9.86%
		decoded strain #1	64% C3+	9.02%
		decoded strain #2	36% C4	8.65%
<b>M6</b>	(83% C3+ 17% C4)	ABI basecaller	C3	1.62%
		decoded strain #1	74% C3+	2.83%
		decoded strain #2	26% C3	22.10%

they are similar to the mixtures we analyze here. In practice, however, a mixed sample is typically discarded and a new sample is taken.

Table 4 shows that in most cases our method dramatically outperforms the single sequence decoding provided by the ABI software, while, importantly, decoding both components from the mixture.

For example, we get good results for all mixture concentrations of C1 and C2. While we were able to extract both underlying clean sequences with low error rate, the traditional ABI base-caller’s 30.6% error rate in mixture M2 and 42.6% in mixture M1 are higher than the genetic diversity in the region, thus rendering sequencing output useless. This is the reason why mixed chromatograms with insertions or deletions between strains in ENV region are typically discarded in the HIV community. In this case, C1 and C2 sequences, though binding to the same primer, differ by deletions and insertions which create more ambiguities than there really are site mutations. We should note that even extending the sequencing further into a less reliable part of the chromatogram, we were able to decode over 500 nucleotides from both strains with similarly low error rates.

It is interesting that in case of 50/50 mixtures, which we had expected to be harder due to equal expected amplitudes of both components, were decoded well by our algorithm. This indicates the strength of the position ( $\tau_i - \tau_{i-1} \approx \Delta$ ) cue

for disambiguation. On the other hand, when one strain is present in very low concentrations, the algorithm found it difficult to lift it out of the noise, which explains the worse performance for weaker component of 83/17 mixes. In the case of chromatogram M6, if the algorithm decodes it assuming  $K = 1$  strains instead of  $K = 2$ , the error rate is reduced to 1.48%, beating the ABI basecaller.

We also found that chromatogram amplitude was a fairly good indicator of the ground truth mixture proportions of the clean sequences in samples other than 50/50 mixtures. This was estimated from the  $\mu_1$  and  $\mu_2$  amplitude hyper-parameters, which were optimized in the algorithm.

Finally, and probably most importantly, when the likelihood of the model is evaluated for the ground truth sequences, it was, in all cases, *higher* than the local maximum found by our algorithm. This means that the data does contain even more information than it might seem from our results, and that the model is correct. Thus, further improvements to inference algorithm are possible. Despite this fact, even our current approach is significantly better than existing solutions. In terms of dominant strain decoding, our approach provided an average error rate of 3.5% compared to the average error rate of 20.6% of the ABI caller (which does not return the secondary strain). The recently proposed approach [18] of simply matching two best templates from among the known strains to explain the ambiguous decoding of Phred is also inferior in case of HIV genes, esp. for Env, due to its high variability. The best matching two templates in the regions we sequenced will match a mixture of two strains taken from a database of 262 strains from [13] only in 89% of sites on average (std 4%), while the distribution of error for our method (over both strains) is significantly lower ( $p < 10^{-3}$ ), indicating that analysis at the signal level is important.

## 5 Conclusion

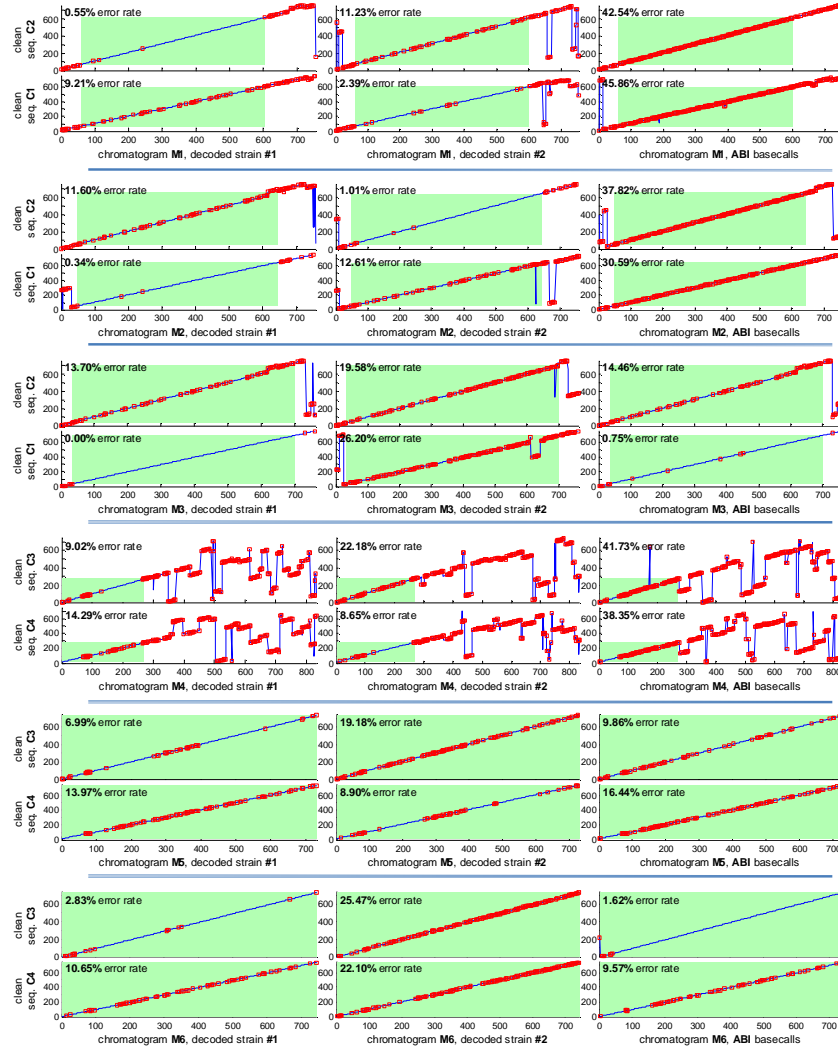
We have shown that for chromatograms of mixed sequences, probabilistic techniques can be employed to accurately infer the individual strains' sequences. This can be done by exploiting information overlooked by traditional base-callers, such as exact amplitude and position information.

This setup enabled us to identify and sequence mixtures of clean sequences in many cases. More sophisticated techniques, such as including raw (unsmoothed) trace data in the model, are currently being explored.

## References

1. Marcel Margulies, M., Egholm1, M., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**:376–380.
2. Bonfield, J., Rada, C., and Staden, R. (1998) Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Research*, **26**(14):3404–3409.
3. Mattocks, C., Tarpey, P., *et al.* (2000) Comparative sequence analysis (CSA): A new sequence-based method for the identification and characterization of mutations in DNA. *Human Mutation*, 16(5):437–443.

4. Crowe, M.L. (2005) SeqDoC: rapid SNP and mutation detection by direct comparison of DNA sequence chromatograms. *BMC Bioinformatics*, **6**:133.
5. Matukumalli, L.K., Grefenstette, J.J., *et al.* (2006) SNP-PHAGE – High Throughput SNP discovery pipeline. *BMC Bioinformatics*, **7**:468.
6. Chen K., McLellan, M.D., *et al.* (2007) PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Research*, **17**:659–666.
7. Bhangale, T.R., Stephens, M., and Nickerson, D.A. (2006) Automating resequencing-based detection of insertion-deletion polymorphisms. *Nature Genetics*, **38**:1457–1462.
8. Sanger, F., Nicklen, S. and Coulson, A. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.*, **74**:5463–5467.
9. Swerdlow, H. and Gesterland, R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucl. Acids Res.* **18**(6):1415–1419.
10. Prober, J.M. *et al.* (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*, **238**:336–341.
11. Thornley, D., and Petridis, S. (2006) Machine Learning in Basecalling - Decoding trace peak behaviour. *IEEE Symp. on Comp. Intel. in Bioinf. & Comp. Biol.*
12. Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, **8**:175–185.
13. Moore, Corey B. *et al.* (2002) Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science*, **296**:1439–1443.
14. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1998) An introduction to variational methods for graphical models. In M.I. Jordan (ed.), *Learning in Graphical Models*. Norwell, MA: Kluwer Academic Publishers.
15. Haan, N.M., and Godsill, S.J. (2002) Bayesian models for DNA sequencing. *Proc. IEEE Conf. on acoustics, speech, and signal processing*. IV: 4020–4023.
16. Jojic, N., *et al.* Learning MHC I - peptide binding. *Bioinformatics*, **22**(14): e227–e235.
17. Mallal, S. (1998) The Western Australian HIV Cohort Study, Perth, Australia. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*; **17** (Suppl 1): S23–S27.
18. Tenney, A.E., Wu, J.Q., *et al.* (2007) A tale of two templates: Automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices. *Genome Research*, **17**:212–218.



**Fig. 4.** Alignment and differences between clean sequences (C3, G4, A1, A2) and mixed sequences (F3, A4, D4, C1, E1, G1) with differences marked as a red  $\square$ . Error rates for the interpretable regions (shown with green background) appear in the upper-left for each algorithm. The multiple-strain decoding algorithm achieves much lower error rates than the ABI sequencer except in 83/17 mixtures with one extremely-dominant component.