Fast and Cheap Genome Wide Haplotype Construction via Optical Mapping

T.S. Anantharaman, V. Mysore, and B. Mishra

Pacific Symposium on Biocomputing 10:385-396(2005)

FAST AND CHEAP GENOME WIDE HAPLOTYPE CONSTRUCTION VIA OPTICAL MAPPING*

T.S. ANANTHARAMAN∗, V. MYSORE†, AND B. MISHRA†

∗*Wisconsin Biotech Center, Univ. Wisc., Madison WI, U.S.A*

† *Courant Institute of Mathematical Sciences, NYU, New York, NY, U.S.A.*

E-mail:tsa@biostat.wisc.edu; {vm40,mishra}@nyu.edu

We describe an efficient algorithm to construct genome wide haplotype restriction maps of an individual by aligning single molecule DNA fragments collected with Optical Mapping technology. Using this algorithm and small amount of genomic material, we can construct the parental haplotypes for each diploid chromosome for any individual. Since such haplotype maps reveal the polymorphisms due to single nucleotide differences (SNPs) and small insertions and deletions (RFLPs), they are useful in association studies, studies involving genomic instabilities in cancer, and genetics, and yet incur relatively low cost and provide high throughput. If the underlying problem is formulated as a combinatorial optimization problem, it can be shown to be NP-complete (a special case of K-population problem). But by effectively exploiting the structure of the underlying error processes and using a novel analog of the Baum-Welch algorithm for HMM models, we devise a probabilistic algorithm with a time complexity that is linear in the number of markers for an ϵ-approximate solution. The algorithms were tested by constructing the first genome wide haplotype restriction map of the microbe *T. pseudoana*, as well as constructing a haplotype restriction map of a 120 Mb region of Human chromosome 4. The frequency of false positives and false negatives was estimated using simulated data. The empirical results were found very promising.

1. Introduction

Diploid organisms, such as humans, carry two mostly similar copies of each chromosome, referred to as haplotypes. Variations in a large population of haplotypes at specific loci are called polymorphisms. The co-associations

*The work reported in this paper was supported by grants from NSF's Qubic program, NSF's ITR program, Defense Advanced Research Projects Agency (DARPA), Howard Hughes Medical Institute (HHMI) biomedical support research grant, the US Department of Energy (DOE), the US air force (AFRL), National Institutes of Health (NIH) and New York State Office of Science, Technology & Academic Research (NYSTAR).

of these variations across the loci indices are of intense interest in disease research.

The main limitation of most SNP based approaches is that each SNP is assayed separately without the related phasing information. Instead, the phase is inferred statistically from a large population of SNP data and employ certain simplifying assumptions such as: parsimony in the total number of different haplotypes in the population, the Hardy-Weinberg equilibrium, perfect phylogeny to combinatorially constrain the possible haplotypes. See the full paper [4] for a detailed survey of the literature.

For a genotyping method to be able to correctly determine the phasing between neighboring polymorphic markers in every individual haplotype map, it must ultimately be able to test single DNA fragments containing 2 or more heterozygous polymorphic markers in a single test. It is possible, of course, to assemble individual haplotype maps by sequencing the individual's entire genome using a modified sequence assembly algorithm [7,9] but the cost of doing this is prohibitive[a].

Here, we propose a direct and more cost-effective approach using the fairly well developed single molecule technology of Optical Mapping.

Each individual haplotype map of restriction sites will only detect a small fraction of all polymorphisms in the human genome, but using a commonly accepted linkage disequilibrium assumption (see[4]), approximately 8 individual haplotype restriction maps will contain more than the 300,000 SNPs required to infer all other known polymorphisms in the individual genome. Even with 50 fold data redundancy required, all date required for 8 individual haplotype restriction maps can be collected for under $1000.

2. Problem Formulation

Our problem can be formulated mathematically as follows: We assume that all individual single molecule DNA fragments are derived from a diploid genome (ignoring the case of sex chromosomes) with two copies of homologous chromosomes. Each DNA fragment is further mapped by cleavage with a restriction enzyme of choice and imaged by an imaging algorithm to produce an ordered sequence of "restriction fragment lengths" or equivalently, "restriction sites." The variations in these restriction fragment lengths are primarily due to RFLPs as well as SNPs at the restriction sites. Additionally, there are further variations introduced by the experimental

[a]This cost has been estimated to be over $10 million per individual.

process and could be assumed due to: sizing errors, partial digestion, short missing restriction fragments, false cuts, ambiguities in the orientation, optical chimerisms, etc. Thus, the genomes may be represented as two haplotype restriction maps, H_1 and H_2, for the same individual which differ only slightly from a genotype restriction map H by a small number of short insertions, deletions and SNPs that coincide with restriction sites. All such maps, H, H_1 and H_2, are assumed to be representable as a sequence of restriction sites (e.g. $H_{2,i}$, with indices $0 \leq i \leq (N+1)$, where $H_{2,0}$ and $H_{2,N+1}$ represent the chromosome ends), but are unknown. However, short DNA fragments of around 500 Kb derived from such maps, and further corrupted by experimental noise processes can be readily generated at high throughput and very low cost using a technology like Optical Mapping (see the full paper [4], and additional references therein). These short DNA fragments will be written as D_k, with indices $1 \leq k \leq M$, where M is the number of data fragments and each data fragment is in turn represented as a sequence of restriction sites (e.g. $D_{k,j}$, $0 \leq j \leq m_k + 1$) and can be aligned globally to create an estimate of genotype map H using algorithms described previously [2].

The algorithmic problem, we wish to study, is to further separate H into two maps H_1 and H_2 in such a manner that each data fragment D_k is aligned well to one haplotype or other and that H_1 and H_2 differ from H only by modifications consistent with SNPs or RFLPs polymorphisms.

Thus, ultimately, this problem corresponds to a problem of refining a multiple map alignment into two families, starting with one global alignment. A combinatorial generalization, where the number of such families is arbitrarily large ($k > 1$) and the cost of each alignment is arbitrarily unconstrained, has been shown to lead to computationally infeasible problems. See [8] for the proof of NP-completeness as well as a probabilistic analysis to show conditions under which the problem can be solved efficiently with a probability close to one. The key to an effective solution of these problems relies on careful experiment design (e.g., choice of coverage, restriction enzyme, experimental conditions, etc.) to ensure conditions under which a polynomial time probabilistic algorithm will work with high probability in conjunction with a Bayesian error model that encodes the error processes properly.

To construct individual haplotype maps from Optical Mapping data we use a mixture hypothesis of pairs of maps H_1 and H_2 for each chromosome, corresponding to the correct restriction map of the two parental chromosomes. We first assemble the data into a regular map of the entire genome

and use this assembly to separate the data into distinct chromosome sets: all maps from the same chromosome belonging to a pair will be included in the same set. We then use a probabilistic model of the errors in the data to derive conditional probability density expressions $f(D_k|H_1)$ and $f(D_k|H_2)$, and apply Bayes rule to maximize a score for the best alignment with respect to proposed H_1 and H_2, Equation 1.:

$$f(H_1, H_2|D_1, \ldots, D_M) \propto f(H_1, H_2) f(D_1, \ldots, D_M|H_1, H_2) \tag{1}$$

The first term on the right side is the prior probability of H_1 and H_2 and we just use a low prior probability for each polymorphism (difference in H_1 vs. H_2). For the conditional probability term, we can assume each map is a statistically independent sample from the genome and that the mapping errors are drawn from i.i.d. distributions and hence write:

$$f(D_1, \ldots, D_M|H_1, H_2) = \prod_{k=1}^{M} \frac{[f(D_k|H_1) + f(D_k|H_2)]}{2} \tag{2}$$

The conditional terms of the form $f(D_k|H_i)$ above can be written as a summation over all possible (mutually exclusive) alignments between the particular D_k and H_i, and for each alignment the probability density is based on an enumeration of the map errors in the alignment and multiplying together the probability associated with each error under some suitable error model. The exact form of the error models suitable for Optical Mapping is described in the next section, but for almost any error models used the sum of the probability for all alignments can be computed effectively using dynamic programming.

Other methods for assembling Optical Mapping data for relatively short clones into ordered restriction maps exist, as detailed in the full paper [4], but they only focus on genotyping, and with widely varying degrees of success. See the full paper [4] for a detailed survey of the literature. We believe that this paper describes the first published algorithm for assembling single molecule data into haplotype maps.

3. Algorithm

Following theorems form the basis for computing various conditional probabilities for a hypothesis.

Theorem 3.1. *Consider an arbitrary alignment between the data D and the hypothesis H, J^{th} restriction site of D matching the I^{th} restriction site of H. We will denote this aligned pair by $J \mapsto I$.*

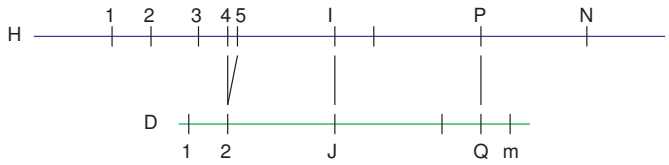

Figure 1. To define the notation required we consider a single arbitrary alignment between a particular data D and hypothesis H. Recall that N is the number of restriction sites in H and m the number of restriction sites in D. Any arbitrary alignment between D and H can be described as a list of pairs of restriction sites from H and D that describes which restriction site from H is aligned with which restriction site from D. As an example, Here the alignment consists of 4 aligned pairs $(4, 2)$, $(5, 2)$, (I, J) and (P, Q). Notice that not all restriction sites in H or D need be aligned. For example between aligned pairs (I, J) and (P, Q) there is one misaligned site on H and D each, corresponding to a missing site (false-negative) and extra-site (false-positive) in D. In this alignment a true small fragment between sites 4 and 5 in H are missing from D, which is shown by aligning both sites 4 and 5 in H with the same site 2 in D. Note that if two or more consecutive fragments in H are all missing in D, this would be described by aligning all sites for the missing fragments in H with the same site in D (rather than showing only the outermost of this set of consecutive sites in H aligned with D, for example). The expression for the conditional probability density of any alignment, such as the one here, can be written as the product of a number of probability terms corresponding to the regions of alignment between each pair of aligned sites, plus one probability term for each unaligned region at the two ends of the alignment.

Let the probability density of the unaligned portion on the left and right end of such an alignment be denoted by $f_{ur}(I, J)$ on the right end if $J \mapsto I$ is the rightmost aligned pair, and $f_{ul}(I, J)$ on the left end if $J \mapsto I$ is the leftmost aligned pair.

In addition, the following probability density functions f_m and f_a denote the following:

$$f_m(I, P) = Pr[H[I..P] \text{ is missing in the observed data } D].$$
$$f_a(I, J, P, Q) = Pr[H[I..P] \text{ is an aligned region but not a missing fragment with respect to the observed data region } D[J..Q]].$$

We assume that $I < P$ and $J < Q$.

Then the following holds:

$$f(D|H) = \sum_{I=1}^{N} \sum_{J=0}^{m+1} f_{ul}(I, J) f(D[J..m+1]|H[I..N] \wedge J \mapsto I).$$

$$
\begin{aligned}
&f(D[J..m+1]|H[I..N] \wedge J \mapsto I) \\
&= f_{ur}(I,J) + f_m(I,I+1) f(D[J..m+1]|H[I+1..N] \wedge J \mapsto (I+1)) \\
&\quad + \sum_{P=I+1}^{N} \sum_{Q=J+1}^{m+1} f_a(I,J,P,Q) f([Q..m+1]|H[P..N] \wedge Q \mapsto P)
\end{aligned}
$$

In particular, if the intermediate values are kept in a DP table $\mathsf{A}_{\mathrm{suf}}[I,J]$

$$\mathsf{A}_{\mathrm{suf}}[I,J] = f(D[J..m+1]|H[I..N] \wedge J \mapsto I)$$

then it is easily seen that $f(D|H)$ *can be computed exactly in* $O(m^2N^2)$ *time and* $O(mN)$ *space, assuming that* f_m *and* f_a *are* $O(1)$ *time functions and* f_{ul} *and* f_{ur} *are* $O(N)$ *time functions.* □

In a later section we will see how to reduce the complexity to linear time when we only require an ϵ-approximate value $\tilde{f}$

$$f(D|H) - \epsilon < \tilde{f}(D|H) < f(D|H) - \epsilon,$$

for the probability density function arising in the context of optical mapping as follows:

$$
\begin{aligned}
f_m(I,I+1) &= P_\nu^{H_{I+1}-H_I} \\
f_a(I,J,P,Q) &= \lambda^{Q-J-1} P_d (1-P_d)^{P-I-1} \\
&\qquad (1-P_\nu)^{H_P-H_I}\, G_{(H_P-H_I),\sigma^2(H_P-H_I)}(D_Q - D_J),
\end{aligned}
$$

where P_d = the digest rate, λ = the false-positive site rate, $\sigma^2 h$ = the Gaussian sizing error variance for a fragment of size h, P_ν = the probability of missing a fragment of unit size, and R_e = the breakage rate of DNA (the inverse of the expected fragment size). For a random variable x following a Gaussian distribution $\mathbf{N}(\mu, \sigma^{\mathbf{2}})$, the probability density value at d is $G_{\mu,\sigma^2}(d) = \exp[-(d-\mu)^2/2\sigma^2]/(\sqrt{2\pi}\sigma)$.

The exact form of the functions for f_{ul} and f_{ur} for Optical Mapping are complicated, but do not affect the complexity of the algorithm; thus a detailed discussion is omitted here, but can be seen in the full paper [4]. The key assumption required is that f_{ul} and f_{ur} permit $O(1)$ ϵ-approximation.

As it has been shown elsewhere [1], a good approximate location of the best alignment between D and H can be determined in $O(1)$ expected time, if the conditional probability density has been previously evaluated for a similar H or alternatively, through a geometric hashing algorithms. Only

a $O(1)$-width band of the DP table needs to be evaluated to compute an ϵ-approximation $\tilde{f}(D|H)$. In particular, the band width of the DP table used in practice is usually about $\Delta = 8$; more generally for Optical mapping Δ is bounded by

$$(1 - P_d)^{\Delta-1} = \epsilon, \quad \text{or} \quad \Delta = 1 + \frac{\ln(\epsilon)}{\ln(1 - P_d)}.$$

With this approach we achieve a reduced time complexity of $O(\min(m, N))$ (more explicitly, $O(\min(m\Delta^3, N))$).

Now we show how we may recompute conditional probabilities for a modification to hypothesis: How can one re-evaluate the conditional probability distribution function, $f(D|H' = p(H))$ when the new hypothesis, H', has been obtained by locally changing H in just one place (corresponding to a polymorphism). There are three cases to consider. We study one of the three cases here in detail and refer the reader to [4] for the remaining cases. The omitted cases are similar but tedious.

We may obtain H' by

(1) Deleting one of the existing restriction sites in H, as the site may contain a heterozygous SNP;
(2) Adding a new restriction site at a specified location in H, symmetrical to the previous case;
(3) Increasing or decreasing a restriction fragment length in H, an RFLP;

Consequently, we may also need to compute the first and second derivative of $f(D|H)$ relative to the change in any fragment size in H.

Theorem 3.2. *Consider an arbitrary alignment between the data D and the hypothesis H, J^{th} restriction site of D matching the I^{th} restriction site of H. Using the notations of the previous discussion, we write:*

$$\mathsf{A}_{\text{suf}}[I, J] = f(D[J..m + 1]|H[I..N] \wedge J \mapsto I),$$

and

$$\mathsf{A}_{\text{pref}}[I, J] = f(D[0..J]|H[1..I] \wedge J \mapsto I).$$

Then

$$\mathsf{A}_{\text{suf}}[I, J] = f_{ur}(I, J) + f_m(I, I + 1)\mathsf{A}_{\text{suf}}[I + 1, J] + \sum_{P=I+1}^{N} \sum_{Q=J+1}^{m+1} f_a(I, J, P, Q)\mathsf{A}_{\text{suf}}[P, Q],$$

and similarly,

$$\mathsf{A}_{\text{pref}}[I,J] = f_{ul}(I,J) + \mathsf{A}_{\text{pref}}[I-1,J]f_m(I-1,I) + \sum_{P=1}^{I-1}\sum_{Q=0}^{J-1}\mathsf{A}_{\text{pref}}[P,Q]f_a(I,J,P,Q),$$

If $H \setminus \{H_K\}$ is obtained from H by deleting the site H_K, then

$$\begin{aligned}
&f(D|H\setminus\{H_K\})\\
&= Pr[\textit{Alignments with rightmost aligned } I < K]\\
&\quad + Pr[\textit{Alignments with leftmost aligned } J > K]\\
&\quad + Pr[\textit{Alignments with a fragment spanning } H[K-1..K+1]]\\
&= \sum_{I=1}^{K-1}\sum_{J=0}^{m+1}\mathsf{A}_{\text{pref}}[I,J]f_{ur}^{(-H_k)}(I,J) + \sum_{I=K+1}^{N}\sum_{J=0}^{m+1}f_{ul}^{(-H_k)}(I,J)\mathsf{A}_{\text{suf}}[I,J]\\
&\quad + \mathbb{I}_{K<N}\sum_{J=0}^{m+1}\mathsf{A}_{\text{pref}}[K-1,J]f_m(K-1,K+1)\mathsf{A}_{\text{suf}}[K+1,J]\\
&\quad + \sum_{J=0}^{m+1}\sum_{I=1}^{K-1}\sum_{P=K+1}^{N}\sum_{Q=J+1}^{m+1}\mathsf{A}_{\text{pref}}[I,J]\frac{f_a(I,J,P,Q)}{1-P_d}\mathsf{A}_{\text{suf}}[P,Q],
\end{aligned}$$

where $f_{ul}^{(-H_k)}$ and $f_{ur}^{(-H_k)}$ are computed respectively from f_{ul} and f_{ur} by suitable simple modifications.

Then it is seen that $f(D|H\setminus\{H_K\})$, $\forall K\ 1 \le K \le N$, can be computed exactly in $O(m^2N^2)$ time and $O(mN)$ space, assuming that f_m and f_a are $O(1)$ time functions and f_{ul} and f_{ur} are $O(N)$ time functions.

If we only wish to compute an ϵ-approximation $\tilde{f}$, for some consecutive range of m different K values, one can compute these m probabilities $f(D|H\setminus\{H_K\})$ for each kind of modification in $O(min(m,N))$ time [4].

3.1. *Search Algorithm for Haplotypes*

The recurrence equations of the previous subsections and the dynamic programming algorithms based on those allow us to efficiently compute the posterior probability for a single possible pair of maps H_1 and H_2 and their modifications

$$\begin{bmatrix} H_1^{(0)} \\ H_2^{(0)} \end{bmatrix} \Rightarrow \begin{bmatrix} H_1^{(1)} \\ H_2^{(1)} \end{bmatrix} \Rightarrow \begin{bmatrix} H_1^{(2)} \\ H_2^{(2)} \end{bmatrix} \Rightarrow \cdots$$

The computationally expensive part of computing the haplotype map algorithm is the search over possible maps H_1 and H_2 in order to find the one with the highest posterior probability.

Initially, we assume that a single genotype map hypothesis H has been computed and it has been determined that H best matches all data. The algorithms to compute such maps have been developed [3,2] and have been in use for more than five years. The speed of the main algorithm, GenTig, has been improved through an important heuristic stage that relies on geometric hashing to quickly identify the maps that overlap, and can also be used in the context of haplotyping. The time complexity of this geometric-hashing-stage is super-linear and is given as

$$T_H = O(N + M_D^{4/3}), \quad \text{where } M_D = \sum_{j=1}^{M} m_j + 1,$$

i.e., M_D is the total number of fragments in the optical mapping data. We will see that the actual time for this stage T_H is dominated by the remaining computation involving search over possible haplotype pairs H_1 and H_2, unless the genome we are dealing with is much larger than the human genome; see next subsection.

If our initial hypothesis is H, then $H_1^{(0)} = H_2^{(0)} = H$, and at each stage $H_1^{(i)}$ and $H_2^{(i)}$ must then be refined by trying to add or delete restriction sites and by adjusting the distance between restriction sites by doing a gradient optimization of the probability density of all maps for each fragment size. The result is $H_1^{(i+1)}$ and $H_2^{(i+1)}$.

Note that at each hypothesis-recomputation step, trying each new restriction site polymorphism involves modifying H_1 or H_2 by adding or deleting a restriction site from H_1 (or H_2) only, while trying an RFLP involves modifying the same interval in both H_1 and H_2 by adding some δh to H_1 and subtracting the same δh from H_2. In each case both possible "phases" of each polymorphism is to be accounted for, reversing the use of H_1 and H_2 above. Since both phases must be tested and the better scoring one selected, except when adding the first polymorphism to H_1 and H_2, the search process can easily turn in to $2^{O(N)}$.

Note also that if the data cannot allow the phasing to be determined because there are no (or insufficient) data molecules spanning both polymorphisms, both phases (orientations) will score almost the same. This fact is also recorded since it marks a break in the phasing of polymorphisms.

Further note that RFLP polymorphisms are more expensive to score, since in addition to the phasing (whether H_1 or H_2 has the bigger fragment)

it is necessary to determine the amount of the fragment size difference for H_1 and H_2 (the δh value), which can be searched for in $O(1)$ expected time, and the constant is essentially logarithmic in the ratio of the expected fragment length to the resolution of optical mapping. More precisely, this step involves trying a number of different multiples of δh values that is logarithmic in the number of total possible values using the well known unimodal function maximization algorithm based on the golden mean ratio. As an example, the total number of δh values required for any fragment can be bounded by about 20 if the resolution of δh is set at 0.1Kb and the largest restriction fragment length is 50Kb; usually, this number is extremely small: just 1 or 2 small δh values are sufficient to verify that no polymorphism exists.

A purely greedy addition of polymorphisms to H_1 and H_2 is not sufficient to get the phases correct as the search can get stuck in local maxima when two or more polymorphisms are nearby. We avoid this problem by using a heuristic look ahead distance of w restriction sites, and scoring all combinations of polymorphisms in this window, before committing the best scoring set of polymorphisms in H_1 and H_2. With a sufficiently large window size w, the fraction of the polymorphic sites the algorithm misses or phases incorrectly can be made negligible. Since this heuristic can increase the worst case complexity of the algorithm exponentially with the window size w we heuristically determine the smallest possible window w by using simulated data and search the space of possible polymorphisms within a window by adding/deleting just one or two polymorphisms at a time until no further improvement in the probability density occurs.

The overall algorithm must try every possible restriction site and fragment as a possible polymorphic SNP or RFLP respectively using a rolling window of size w restriction sites. This process must be repeated a few times until no further polymorphisms are detected. Typically just two to three iterations of scanning all restriction sites suffice.

The overall complexity of the basic haplotype search algorithms described here, just using the basic DP algorithm from Theorem 3.1, is $O(M_D^2/C + N)$, where C = coverage and $C = (1/N)\sum_{j=1}^{M} m_j$. Several simple tricks to speed up the evaluation of conditional probabilities, coupled with a judicious applications of dual DP tables ultimately improves the asymptotic time complexity to $O(M_D)$. Detailed analysis is in the full-paper [4].

4. Empirical Results

In this section, we summarize experiments constructing a haplotype map of *T. pseudoana* and of a 120 Mb region of the human chromosome to demonstrate the feasibility and relevance of our approach. Finally, we also summarize results based on simulation to provide insight into the accuracy of our results. See the full-paper [4] for a detailed description of these studies.

- The optical mapping data for *T. Pseudoana* (Diatom) was analyzed by our algorithm; for all except chromosome 19, it successfully phased all polymorphisms and generated two separate maps.
- Our algorithm found 233 restriction site polymorphisms and 12 fragment length polymorphisms in the human chromosome 4 data, and was able to phase all polymorphisms into 2 contiguous regions. The nature of the polymorphisms detected was somewhat surprising and is discussed in details in the full paper [4].
- The simulated data, with statistical characteristics derived from human chromosome 21, was assembled using different data redundancy of 6×, 12×, 16×, 24×, 50× and 100× (per haplotype). The results are summarized in Table 1. From the simulated data we can infer that 16× redundancy is required to eliminate most errors in SNPs and about 50× redundancy is required to eliminate most errors in indels (RFLPs)

Redundancy	**fp SNPs**	**fn SNPs**	**fp RFLPs**	**fn RFLPs**	**Phase err**	**Molecules**
6x	5	5	1	18	7/26	30
12x	4	2	4	16	2/55	60
16x	2	1	0	12	2/71	80
24x	2	1	1	11	3/111	120
50x	0	1	1	5	4/228	250
100x	0	0	2	1	2/441	500

Figure 2. Haplotyping algorithm performance for 16 SNPs and 24 RFLPs.

5. Discussions and Future Work

Single molecule mapping technologies, such as Optical Mapping, are ideal for detecting genetic markers with phasing information and without population-based assumptions. It elegantly circumvents many problems

that have proven unsurmountable in all other population-based approaches (see discussion in [4]).

Furthermore, we estimate that our approach is currently the only approach that can produce a genome wide individual haplotype map for under $1000 (based on 8 restriction enzyme haplotype maps). The dominant SNP based approach requires testing of about 300,000 SNPs which costs at least ten times more per person. Our approach can be applied to other single molecule mapping technologies. When applied to single molecule technologies to map short 6–8bp LNA hybridization probes, it can be used to sequence the entire human genome: With 50× coverage the location of probes can be determined to within about 200bp. Hence well known error tolerant SBH (Sequencing by Hybridization) algorithms [6] can be used to determine the sequence within any 200bp window from maps of a universal set of about 2048 probes of 6bp, allowing a draft quality individual haplotype sequence to be assembled for about $20,000.

References

1. T. Anantharaman, B. Mishra, and D.C. Schwartz, "Genomics via Optical Mapping II: Ordered Restriction Maps," *J. of Comp. Bio.*, **4**(2):91–118, 1997.
2. T. Anantharaman, B. Mishra, and D.C. Schwartz, "Genomics via Optical Mapping III: Contiging Genomic DNA and Variations," ISMB '99 , 7:18–27, 1999
3. T. Anantharaman, and B. Mishra, "A Probabilistic Analysis of False Positives in Optical Map Alignment and Validation," WABI '01,**LNCS 2149**:27–40, 2001
4. T. Anantharaman, V. Mysore, and B. Mishra, *Fast and Cheap Genome wide Haplotype Construction via Optical Mapping*, NYU Tech. Report# TR 2004-852, 2004. `http://cs.nyu.edu/web/Research/technical_reports.html`.
5. R. Britten *et al.*, "Majority of Divergence between Closely Related DNA Samples is due to Indels," *PNAS*, **100**(8):4461–4465, 2003.
6. E. Halperin *et al.*, "Handling Long Targets and Errors in Sequencing by Hybridization," *J. of Comp. Bio.*, **10**:3–4, 2003.
7. G. Lancia *et al.*, "Practical Algorithms and Fixed-Parameter Tractability for the Single Individual SNP Haplotyping Problem," *WABI '02*: 29–43.
8. B. Mishra and L. Parida, "Partitioning Single-Molecule Maps into Multiple Populations: Algorithms And Probabilistic Analysis," *Discrete Applied Mathematics*, **104**(l-3):203-227, 2000.
9. M. Waterman *et al.*, "Haplotype Reconstruction from SNP Alignment," *RECOMB 03*: 207–216, 2003