Genomics via Optical Mapping II(A): Restriction Maps from Partial Molecules and Variations

(Extended Abstract)

THOMAS ANANTHARAMAN and BUD MISHRA[1]

Abstract

In this paper, we extend an algorithmic approach to constructing ordered restriction maps from images of a population of individual DNA molecules (clones) digested by restriction enzymes. The original algorithm was capable of producing high-resolution, high-accuracy maps rapidly and in a scalable manner given a certain class of data errors, including contamination, sizing errors, false and missing restriction sites and unknown orientation. Here we extend this set of errors to include possibly broken molecules where the amount of breakage is not known beforehand, which is necessary for handling larger clones. In an earlier paper [AMS97], we had shown that the problem of making maps from molecules with end fragments missing as the only source of error is NP-complete. We also show how to handle multiple reliability levels in the input data when calling restriction sites.

1 Genomics and Optical Mapping

Optical mapping [CAH+95, CJI+96, HRL+95, JRH+96, MBC+95, SCH+95, SLH+93, WHS95] is a single molecule methodology for the rapid production of ordered restriction maps from individual DNA molecules. Restriction enzyme cleavage sites are visible as gaps that appear flanking the relaxed DNA fragments (pieces of molecules between two consecutive cleavages). Relative fluorescence intensity (measuring the amount of fluorochrome binding to the restriction fragment) or apparent length measurements (along a well-defined "backbone" spanning the restriction fragment) are used as size-estimates of the restriction fragment and used to construct the final restriction map using an algorithmic approach described in [AMS97]. This approach is based on Bayesian inference and is capable of recovering from a number of data errors of unknown magnitude including sizing errors, false and missing restriction sites and unknown orientation by exploiting the redundancy in the multiple DNA molecules. This is done by hypothesizing a probabilistic model of the data and the errors and looking for the hypothesis that best fits the observed data given a prior distribution of restriction maps. The best hypothesis is found using a heuristic global search combined with local function optimization. This is not guaranteed to find the correct solution, but the Bayesian approach is capable of providing a probabilistic confidence measure which signals when the solution is not reliable due to insufficient data or too much error in the data. Since the problem is shown to be NP-complete in the presence of each of many of these error sources [AMS97], this is the best one can expect.

In this paper we extend the error model to allow for molecules which have a piece at either end broken off and therefore missing from the data. This error term is important when handling large cloned DNA molecules (e.g., large BAC based clones roughly of length $150Kb$), since they are highly likely to have pieces broken off at either end during the handling. The entire model including this new error term is

[1] *Authors' Current Address: Courant Institute, New York University, 251 Mercer St, NYC, NY-10012.* *The research presented here was partly supported by an NSF Career grant: IRI-9702071 and an NIH Grant: NIH R01 HG0025-07.*

presented. In addition we show how to handle multiple reliability levels in the input data when calling restriction sites, where the exact reliability levels are not known and must be deduced from the data.

2 Maps by Bayesian Inference

The Bayesian approach to compute maps consists of:

- A Model or Hypothesis $\mathcal{H}$, of the map of restriction sites including the errors.
- A prior density distribution over the Hypothesis $f(\mathcal{H})$.
- A conditional density distribution of the data molecules D_j

$$f(D_j|\mathcal{H}),$$

- The conditional density distribution of the Hypothesis via Bayes' rule:

$$f(\mathcal{H}|\mathcal{D}) = \frac{f(\mathcal{H})\ \prod_{j=1}^{M} f(D_j|\mathcal{H})}{f(\mathcal{D})}.$$

Using this formulation, we search over the space of all hypotheses to find the most "plausible" hypothesis that maximizes $f(\mathcal{H}|\mathcal{D})$. There is no need to compute $f(\mathcal{D})$, the prior of the data, since it is the same for all $\mathcal{H}$. We use a simple prior $f(\mathcal{H})$ which is just a Poisson distribution of the number of cuts (restriction sites) reflecting our expectation of a certain number of cuts per unit length based on the enzyme used. The expression for $f(D_j|\mathcal{H})$ is based on our data and error model and is more complex.

2.1 The Data Model

Unless otherwise specified, the indices i, j, u and k are to have the following interpretation:

- The index i ranges from 1 to N and refers to cuts in the hypothesis.
- The index j ranges from 1 to M and refers to data items (i.e., molecules).
- The index u ranges from 1 to U_j and refers to a possible orientation and breakage hypothesis for molecule j.
- The index k ranges from 1 to K_{ju} and refers to a specific alignment of cuts in the hypothesis with data j and breakage/orientation u.

Our Bayesian model is specified as follows:

$N \equiv$ Number of cuts in the hypothesis $\mathcal{H}$.

$h_i \equiv$ The ith cut location on $\mathcal{H}$.

$m_j \equiv$ Number of cuts in the data D_j.

$U_j \equiv$ The number of predefined possible breakage values and orientations for data D_j. This is typically the same for all j.

$K_{ju} \equiv$ Number of possible alignments of the data D_j against the hypothesis $\mathcal{H}$ with breakage and orientation u.

$s_{ijuk} \equiv$ The cut location in D_j (suitably oriented) matching the cut h_i in $\mathcal{H}$, given the alignment A_{juk}. In case such a match occurs, this event is denoted by an indicator variable m_{ijuk} taking the value 1.

$m_{ijuk} \equiv$ An indicator variable, taking the value 1 iff the cut s_{ijuk} in D_j matches a cut h_i in the hypothesis $\mathcal{H}$, given the alignment A_{juk}. It takes the value 0, otherwise.

$F_{juk} \equiv$ Number of false (non-matching) cuts in the data D_j for alignment A_{juk}, that do not match any cut in the hypothesis $\mathcal{H}$. Thus $F_{juk} = m_j - \sum_{i=1}^{N} m_{ijuk}$

$p_{ci} =$ Probability that the ith sequence specific restriction site in the molecule will be visible as a cut.

$\sigma_i =$ Gaussian standard deviation of the observed position of the ith cut when present and depends on the accuracy with which a fragment can be sized.

$\lambda_f =$ Expected number of false-cuts per molecule observed. Since all sizes will be normalized by the molecule size, this will also be the false-cuts per unit length.

$p_b =$ Probability that the data is invalid ("bad"). In this case, the data item is assumed to have no relation to the hypothesis being tested, and could be an unrelated piece of DNA or a partial molecule with a significant fraction of the DNA missing. The cut-sites (all false) on this data item are assumed to have been generated by a Poisson process with the expected number of cuts $= \lambda_n$.

$\lambda_n =$ Expected number of cuts per "bad" molecule.

${Z^{\mathrm{L}}}_{ju}$, ${Z^{\mathrm{R}}}_{ju} \equiv$ The mean amounts of breakage on the left and right side of $\mathcal{H}$ respectively for breakage hypothesis and orientation u of data D_j.

$\sigma_e =$ The Gaussian standard deviation common to all breakage values with means ${Z^{\mathrm{L}}}_{ju}, {Z^{\mathrm{R}}}_{ju}$.

$\mathrm{pr}_{ju} \equiv$ The probability that any molecule j would break according to hypothesis $u : \sum_u \mathrm{pr}_{ju} = 1$ over all u corresponding to the same orientation.

$\mathrm{sz}_{ju} \equiv 1 - {Z^{\mathrm{L}}}_{ju} - {Z^{\mathrm{R}}}_{ju} =$ The rescaling factor for D_j.

$E_{iju} \equiv \frac{1}{\sqrt{2\pi}} \operatorname{erf}((h_i - {Z^{\mathrm{L}}}_{ju})/\sigma_e) + \frac{1}{\sqrt{2\pi}} \operatorname{erf}((1 - {Z^{\mathrm{R}}}_{ju} - h_i)/\sigma_e) - 1 =$ The probability that cut h_i is *not* in the broken off part of molecule D_j. Note that $\operatorname{erf}(x) = \int_{-\infty}^{x} e^{-x^2/2}\, dx$.

$S_{ijuk} \equiv s_{ijuk}\mathrm{sz}_{ju} + {Z^{\mathrm{L}}}_{ju}$ The expected rescaled value of s_{ijuk} given breakage hypothesis u. (This is a slight approximation since breakages cannot be negative)

We assume that a broken molecule is less likely to have false cuts compared to unbroken molecules in proportion to its DNA.

From these definitions and assumptions it is not hard to write $\mathcal{L}$ the log of the probability density expression $f(\mathcal{D}|\mathcal{H})$ with respect to the (unscaled) cut locations in data D_j:

$$
\begin{aligned}
\mathcal{L} &= \sum_{j=1}^{M} \log f_j, \quad \text{where} \\
f_j &\equiv p_b e^{-\lambda_n} {\lambda_n}^{m_j} + \frac{(1-p_b)}{2} \sum_{u=1}^{U_j} \sum_{k=1}^{K_{ju}} f_{juk} \\
f_{juk} &= e^{-\lambda_f \mathrm{sz}_{ju}} {\lambda_f}^{F_{juk}} \mathrm{sz}_{ju}^{m_j} \times \prod_{i=1}^{N} \left[\left(p_{ci} E_{iju} \frac{e^{-(S_{ijuk}-h_i)^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i} \right)^{m_{ijuk}} (1 - p_{ci}E_{iju})^{(1-m_{ijuk})} \right].
\end{aligned}
$$

A detailed derivation can be found in [AMS97] for all terms not related to broken molecules. For broken molecules the preceding equations propose to consider a fixed number of breakage values at either end in different combinations ${Z^{\mathrm{L}}}_{ju}, {Z^{\mathrm{R}}}_{ju}$ with their probability distribution given by pr_{ju}. In this abstract the pr_{ju} values are assumed fixed, even though in the full paper we describe a parameterized form of pr_{ju} whose parameters are optimized along with the other Bayesian parameters.

2.2 Local Search Algorithm

In order to find the most plausible restriction map for any size N, we shall optimize $\mathcal{L}$ over the following parameters:

$$\begin{aligned} \text{Cut Sites} &= h_1, h_2, \ldots, h_N, \\ \text{Cut Rates} &= p_{c1}, p_{c2}, \ldots, p_{cN}, \\ \text{Std. Dev. of Cut Sites} &= \sigma_1, \sigma_2, \ldots, \sigma_N, \\ \text{Auxiliary Parameters} &= p_b, \lambda_f \text{ and } \lambda_n. \end{aligned}$$

Let us denote any of these parameters by θ. Thus, we need to solve the equation

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0,$$

to find an extremal point of $\mathcal{L}$ with respect to the parameter θ. Often the equation cannot be solved directly, but an approximate solution is used to get a better parameter estimate and the gradient is recomputed. We list the results below.

2.2.1 Case: $\theta \to p_b$

In this case it is computationally easy to compute both the first and second derivations of $\mathcal{L}$ hence we use Newton's equation:

$$\begin{aligned} p_b &:= p_b - \frac{\partial \mathcal{L}/\partial p_b}{\partial^2 \mathcal{L}/\partial {p_b}^2} \quad \text{where} \\ \frac{\partial \mathcal{L}}{\partial p_b} &= \sum_j \frac{(e_j - d_j)}{p_b e_j + (1-p_b) d_j}, \\ \frac{\partial^2 \mathcal{L}}{\partial {p_b}^2} &= -\sum_j \frac{(e_j - d_j)^2}{[p_b e_j + (1-p_b) d_j]^2}, \quad \text{and} \\ e_j &\equiv p_b e^{-\lambda_n} {\lambda_n}^{m_j}, \quad \text{and} \quad d_j \equiv \frac{(1-p_b)}{2} \sum_{u=1}^{U_j} \sum_{k=1}^{K_{ju}} f_{juk}. \end{aligned}$$

2.2.2 Case 2: $\theta \to h_i$, p_{c_i}, σ_i $(i = 1, \ldots, N)$, or λ_f

In this case,

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{j=1}^{M} \sum_{u=1}^{U_j} \mathrm{pr}_{ju} \sum_{k=1}^{K_{ju}} \pi_{juk} \chi_{juk},$$

where

$$\begin{aligned} \pi_{juk} &\equiv \left(\frac{1-p_b}{2}\right) \frac{f_{juk}}{f_j} \\ &\equiv \text{Relative probability of the alignment } A_{juk} \text{ for data item } D_j. \\ \chi_{juk}(\theta) &\equiv \left(\frac{F_{juk}}{\lambda_f} - \mathrm{sz}_{ju}\right) \frac{\partial \lambda_f}{\partial \theta} \\ &\quad + \sum_{i=1}^{N} \left[\frac{m_{ijuk}}{p_{ci}} \frac{\partial p_{ci}}{\partial \theta} - \frac{1 - m_{ijk}}{1 - p_{ci} E_{iju}} \frac{\partial p_{ci} E_{iju}}{\partial \theta}\right] \\ &\quad + \sum_{i=1}^{N} m_{ijuk} \left[\frac{\partial}{\partial \theta}\left(\frac{-(S_{ijuk} - h_i)^2}{2\sigma_i^2}\right) - \frac{1}{\sigma_i} \frac{\partial \sigma_i}{\partial \theta} + \frac{1}{E_{iju}} \frac{\partial E_{iju}}{\partial \theta}\right]. \end{aligned}$$

Before, examining the updating formula for each parameter optimization, we shall introduce the following notations for future use. The quantities defined below can be efficiently accumulated for a fixed value of the set of model parameters.

$$
\begin{array}{llll}
\Psi_{0i} & \equiv \sum_j \sum_u \mathrm{pr}_{ju} \sum_k \pi_{juk} m_{ijuk} & \equiv & \text{Number of cuts matching } h_i \\
\Psi_{1i} & \equiv \sum_j \sum_u \mathrm{pr}_{ju} \sum_k \pi_{juk} m_{ijuk} S_{ijuk} & \equiv & \text{Sum of scaled cut locations matching } h_i \\
\Psi_{2i} & \equiv \sum_j \sum_u \mathrm{pr}_{ju} \sum_k \pi_{juk} m_{ijuk} S^2_{ijuk} & \equiv & \text{Sum of square of scaled cut locations matching } h_i \\
\mu_g & \equiv \sum_j \sum_u \mathrm{pr}_{ju} \sum_k \pi_{juk} & \equiv & \text{Number of ``good'' molecules} \\
\gamma_g & \equiv \sum_j \sum_u \mathrm{pr}_{ju} \sum_k \pi_{juk} m_j & \equiv & \text{Total cuts in ``good'' molecules} \\
\mathrm{BG} & \equiv \sum_j \sum_u \mathrm{pr}_{ju} \sum_k \pi_{juk} \mathrm{sz}_{ju} & \equiv & \text{Sum of broken molecule sizes} \\
\mathrm{FE}_i & \equiv \sum_j \sum_u \mathrm{pr}_{ju} E_{iju} \sum_k \pi_{juk} & \equiv & \text{Number of ``good'' molecules spanning } h_i \\
\mathrm{ZE}_i & \equiv \sum_j \sum_u \frac{\mathrm{pr}_{ju} G_{iju}}{E_{iju}} \sum_k \pi_{juk} m_{ijuk} & \equiv & \text{Weighted } \Psi_{0i} \\
\mathrm{ZEP}_i & \equiv \sum_j \sum_u \frac{\mathrm{pr}_{ju} G_{iju}}{1-p_{c_i} E_{iju}} \sum_k \pi_{juk} (1 - m_{ijuk}) & \equiv & \text{Weighted } \mu_g - \Psi_{0i} \\
\mathrm{FEP}_i & \equiv \sum_j \sum_u \frac{\mathrm{pr}_{ju} E_{iju}}{1-p_{c_i} E_{iju}} \sum_k \pi_{juk} (1 - m_{ijuk}) & \equiv & \text{Weighted } \mu_g - \Psi_{0i},
\end{array}
$$

where

$$G_{iju} \equiv \frac{\partial E_{iju}}{\partial h_i} = G\left(\frac{h_i - Z^{\mathrm{L}}{}_{ju}}{\sigma_e}\right) - G\left(\frac{1 - Z^{\mathrm{R}}{}_{ju} - h_i}{\sigma_e}\right).$$

We note here that f_j and the Ψ's can all be computed efficiently using a simple updating rule that modifies the values with one data item D_j (molecule) at a time. This rule can then be implemented using the *dynamic programming* recurrence equations described in [AMS97]. All other quantities can be computed with negligible additional overhead since they involve the same summations over index k.

Using the above notation we can compute the first derivatives of $\mathcal{L}$ and hence the update equations for the parameters in this case as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial h_i} &= \frac{1}{\sigma_i^2}\left(\Psi_{1i} - h_i \Psi_{0i}\right) - p_{ci}\mathrm{ZEP}_i - \mathrm{ZE}_i \\
&\Rightarrow \quad h_i := \frac{\Psi_{1i}}{\Psi_{0i}} - \frac{\sigma_i^2}{\Psi_{0i}}\left(p_{ci}\mathrm{ZEP}_i + \mathrm{ZE}_i\right) \\
\frac{\partial \mathcal{L}}{\partial p_{ci}} &= \frac{\Psi_{0i}}{p_{ci}} - \mathrm{FEP}_i \\
&\qquad \text{with } \mathrm{FEP}_i \propto \frac{1}{1 - p_{ci}\mathrm{FE}_i/\mu_g} \\
&\Rightarrow \quad p_{ci} := \Psi_{0i}(\Psi_{0i}\mathrm{FE}_i/\mu_g) + \mathrm{FEP}_i(1 - p_{ci}\mathrm{FE}_i/\mu_g) \\
\frac{\partial \mathcal{L}}{\partial \lambda_f} &= \frac{\gamma_g - \sum_i \Psi_{0i}}{\lambda_f} - \mathrm{BG} \\
&\Rightarrow \quad \lambda_f := \frac{\gamma_g - \sum_i \Psi_{0i}}{\mathrm{BG}} \\
\frac{\partial \mathcal{L}}{\partial \sigma_i} &= \frac{1}{\sigma_i^3}\left(\Psi_{2i} - 2h_i\Psi_{1i} + h_i^2\Psi_{0i}\right) - \frac{\Psi_{0i}}{\sigma_i} \\
&\Rightarrow \quad \sigma_i^2 := \frac{\left(\Psi_{2i} - 2h_i\Psi_{1i} + h_i^2\Psi_{0i}\right)}{\Psi_{0i}} \\
\frac{\partial \mathcal{L}}{\partial \lambda_n} &= \frac{1}{\lambda_n}\left(\sum_j m_j - \gamma_g\right) - (M - \mu_g)
\end{aligned}
$$

$$\Rightarrow \qquad \lambda_n := \frac{\sum_j m_j - \gamma_g}{M - \mu_g}.$$

Frequently it is possible to constrain $p_c = p_{c1} = \cdots = p_{cN}$ and $\sigma = \sigma_1 = \cdots = \sigma_N$. In that case the corresponding gradients and update equation for p_c and σ become:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma} &= \sum_i \left[\frac{1}{\sigma_i^3} \left(\Psi_{2i} - 2h_i \Psi_{1i} + h_i^2 \Psi_{0i} \right) - \frac{\Psi_{0i}}{\sigma_i} \right] \\
&\Rightarrow \qquad \sigma^2 := \frac{\sum_i (\Psi_{2i} - \Psi_{1i}^2 / \Psi_{0i})}{\sum_i \Psi_{0i}} \\
\frac{\partial \mathcal{L}}{\partial p_c} &= \sum_i \left(\frac{\Psi_{0i}}{p_c} - \mathrm{FEP}_i \right) \\
&\qquad \text{with } \left(\sum_i \mathrm{FEP}_i \right) \propto \frac{1}{1 - p_c \left(\frac{1}{N} \sum_i \mathrm{FE}_i / \mu_g \right)} \\
&\Rightarrow \qquad p_c := \frac{\sum_i \Psi_{0i}}{(\sum_i \Psi_{0i})(\frac{1}{N} \sum_i \mathrm{FE}_i / \mu_g) + (\sum_i \mathrm{FEP}_i)(1 - \frac{p_c}{N} \sum_i \mathrm{FE}_i / \mu_g)}.
\end{aligned}$$

2.3 Efficient Selection of Breakage Hypothesis

In practice it would be prohibitively expensive to actually compute the previous equations for all U_j possible breakage hypothesis for each molecule j. For example considering just breakages of up to 40% (say, in steps of 1%) of the molecule would require a total of $41 \times 42/2 = 861$ hypothesis per molecule, which would make the total run time 861 times more expensive just to consider broken molecules. Fortunately for each j, only a small number of all U_j breakage hypotheses actually make a significant contribution to the total probability and gradients. These can be selected by performing an approximate match computation of each breakage hypothesis against the hypothesis $\mathcal{H}$ and selecting only those that have a significant effect. Since U_j covers both orientations of the molecule, we take advantage of the fact that usually in one orientation all hypothesis score uniformly poorly. Making the number of hypotheses selected for complete evaluation adaptive to each j results in the most efficient code with an average of only 12 hypotheses needing to be evaluated. Compared to the effectively two orientations in [AMS97] this is still about 6 times slower, but is essential for handling data where 20—80% of the molecules are broken in varying amounts.

3 Input Cut Site Data with Multiple Confidence Levels

The image processing software is often not able to tell with certainty if a location along a molecule corresponds to a cut or not. In such cases more information is preserved if the image processing software passes along whatever information it had on the certainty with which it was able to determine the presence of a cut. Unfortunately the heuristics used in the image processing software do not allow this certainty to be expressed as a probability and the best that can be expected is to assign a score along some scale say a number v in $[1, W]$.

We can use this information optimally as part of our Bayesian inference mechanism by using a new index v ranging from 1 to W corresponding to the possible input confidence scores on cuts and the following new parameters:

$p_{f_v} \equiv$ False cut probability for cuts with confidence v in "good" molecules.

$F_v \equiv$ Fraction of cuts with confidence v in "good" molecules.

$F'_v \equiv$ Fraction of cuts with confidence v in "bad" molecules.

We show in the full paper how the log probability density expression $\mathcal{L}$ can be extended to include these new parameters and describe the gradients and update equations for these parameters. Only one new expensive term Ψ_v corresponding to "the number of cuts with confidence v matching any model cut" needs to be computed increasing the total run time by about another 30%.

4 Conclusion

In this paper, we make two contributions toward the construction of restriction maps with optical mapping data.

1. We extend the model in [AMS97] to include broken molecules.
2. We extend the model to allow for different levels of reliability in the input data, where the exact reliability levels are not known and must be inferred from the data itself.

Acknowledgment. Our thanks go to David Schwartz, Chris Aston, Joe Giacalone, Ed Huff and our many other colleagues in the laboratory for their invaluable help and advice with our research and for making the experimental data available to us.

References

[AMS97] T.S. Anantharaman, B. Mishra and D.C. Schwartz. "Genomics via Optical Mapping II: Ordered Restriction Maps," *Journal of Computational Biology*, **4**(2):91–118, 1997.

[CAH+95] W. Cai, H. Aburatani, D. Housman, Y. Wang, and D.C. Schwartz. "Ordered Restriction Endonuclease Maps of Yeast Artificial Chromosomes Created by Optical Mapping on Surfaces," *Proc. Natl. Acad. Sci., USA*, **92**:5164–5168, 1995.

[CJI+96] W. Cai, J. Jing, B. Irvine, L. Ohler, E. Rose, U. Kim, Shizuya, M. Simon, T. Anantharaman, B. Mishra and D.C. Schwartz. "High Resolution Restriction Maps of Bacterial Artificial Chromosomes Constructed by Optical Mapping," Submitted to *Genomics*, 1997.

[HRL+95] E.J. Huff, J. Reed, I. Lisanskiy, J.-S. Lo, B. Porter, T. Anantharaman, B. Mishra, D. Geiger and D.C. Schwartz. "Automatic Image Analysis for Optical Mapping," In *1995 Genome Mapping and Sequencing Conference*, Cold Spring Harbor, New York, May 10–14, 1995.

[JRH+96] J. Jing, J. Reed, J. Huang, X. Hu, V. Clarke, D. Housman, T. Anantharaman, E. Huff, B. Mishra, B. Porter, A. Shenker, E. Wolfson, C. Hiort, R. Cantor and D.C. Schwartz. "Automated High Resolution Optical Mapping Using Arrayed, Fluid Fixed DNA Molecules," Submitted to *Science*, 1996.

[MBC+95] X. Meng, K. Benson, K. Chada, E. Huff and D.C. Schwartz. "Optical Mapping of Lambda Bacteriophage Clones Using Restriction Endonuclease," *Nature Genetics*, **9**:432–438, 1995.

[SCH+95] A. Samad, W.W. Cai, X. Hu, B. Irvin, J. Jing, J. Reed, X. Meng, J. Huang, E. Huff, B. Porter, A. Shenker, T. Anantharaman, B. Mishra, V. Clarke, E. Dimalanta, J. Edington, C. Hiort, R. Rabbah, J. Skiadas, and D.C. Schwartz. "Mapping the Genome One Molecule At a Time—Optical Mapping," *Nature*, **378**:516–517, 1995.

[SLH+93] D.C. Schwartz, X. Li, L.I. Hernandez, S.P. Ramnarain, E.J. Huff and Y.K. Wang. "Ordered Restriction Maps of *Saccharomyces cerevisiae* Chromosomes Constructed by Optical Mapping," *Sciences*, **262**:110–114, 1993.

[WHS95] Y.K. Wang, E.J. Huff and D.C. Schwartz. "Optical Mapping of the Site-directed Cleavages on Single DNA Molecules by the RecA-assisted Restriction Endonuclease Technique," In *Proc. Natl. Acad. Sci. USA*, **92**:165–169, 1995.