Decoding Genetic Variations: Algorithms for Haplotype Assembly

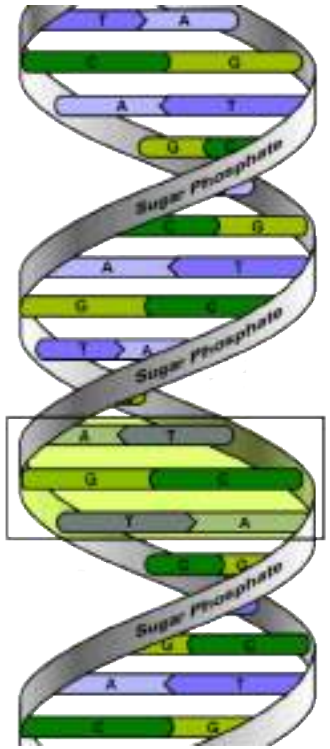
Haris Vikalo

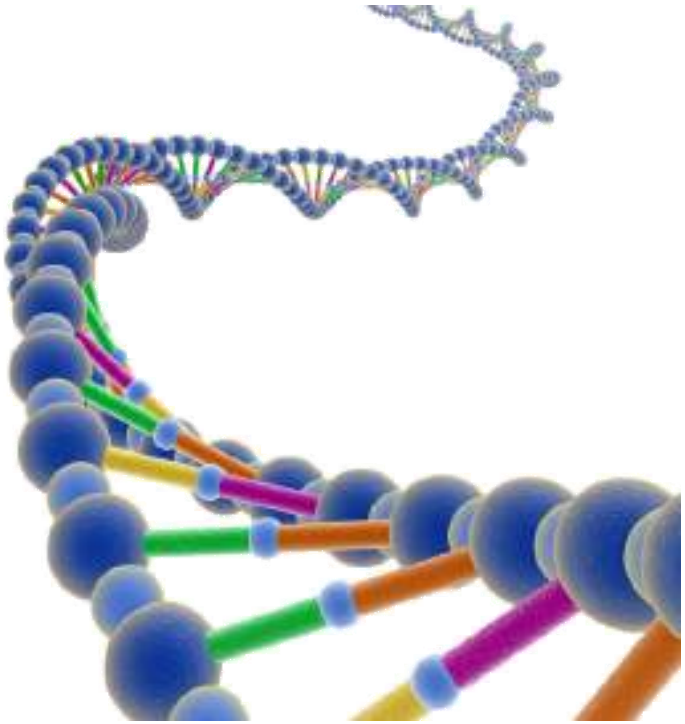
Electrical and Computer Engineering Department
The University of Texas at Austin

2015 MBMC Workshop, December 3-4, 2015

DNA Sequencing: History and State-of-the-Art

- Determine the order of nucleotides in a DNA sequence

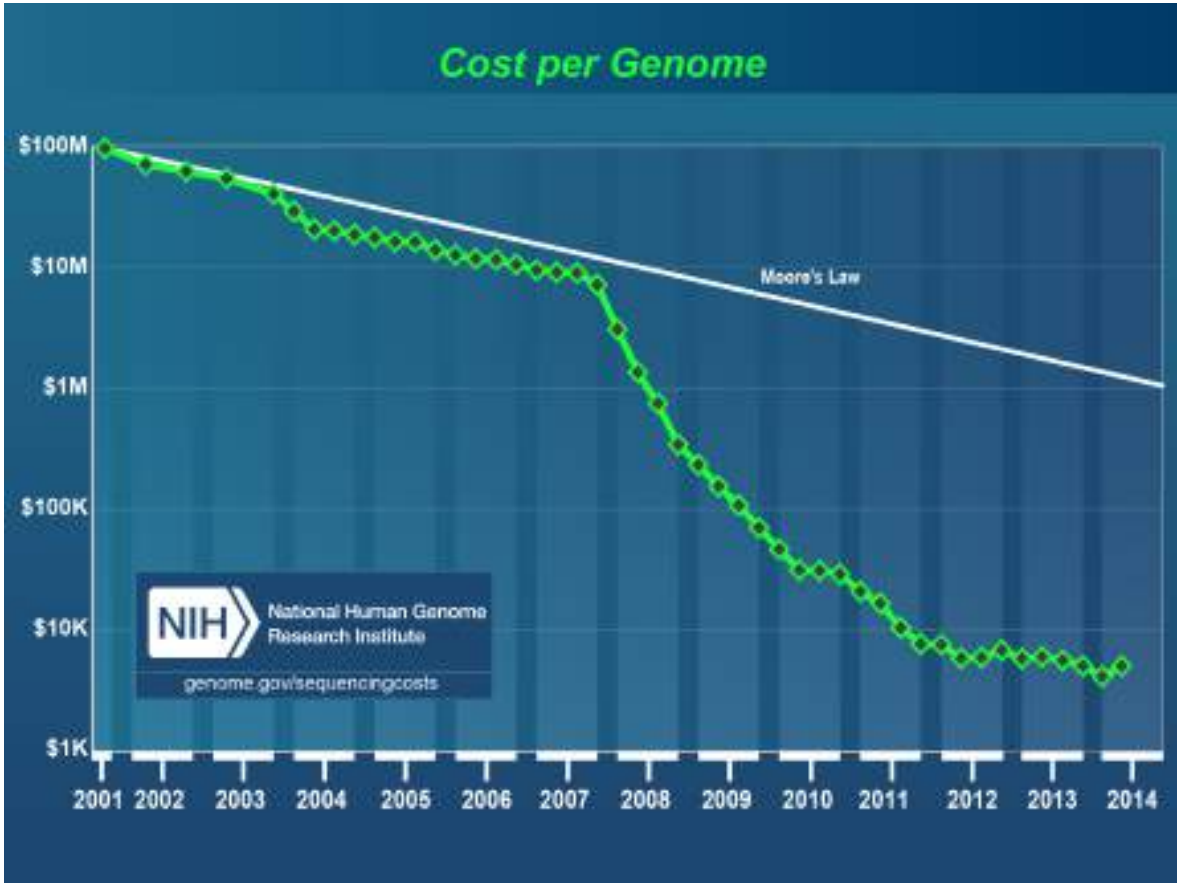




- Human Genome Project: mapping the genetic blueprint
  - followed by rapid advancements in sequencing technology

Advancements in Sequencing Technology

- Fast, accurate and affordable whole-genome sequencing



- Methodology: sequencing-by-synthesis, shotgun sequencing

Genetic Variations

- Routine sequencing enables studies of genetic variations...

~1.5% DNA difference

... but finding such rare events remains challenging.

Genetic Variations in Humans

- Genetic variations between individuals are very rare
  - 1 single nucleotide polymorphism (SNP) in 1000 nucleotides

- Major effects on human health and medical treatments
  - hereditary diseases (Huntington's, cystic fibroses, sickle cell anemia)
  - complex diseases, determine how we metabolize medications

Genetic Variations in Humans

- Human genome $\sim$ 3 bil. bases (A, C, G, T), 23 chromosomes
  - diploids: chromosomes come in pairs (non-identical/homologous)
  - each chromosome inherited from on of the parents
- Characterizing genetic variations
  - SNP calling determines locations and type of polymorphisms
  - example: A/G, C/G, A/G

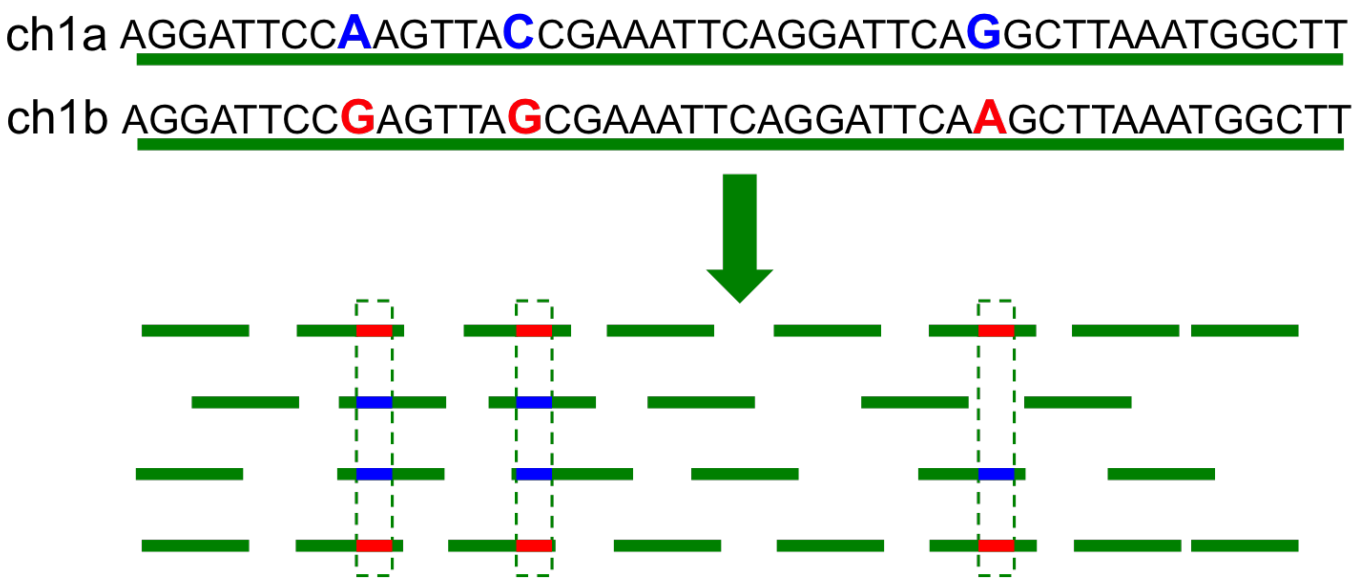
AGGATTCC**A**AGTTA**C**CGAAATTCAGGATTCA**G**GCTTAAATGGCTT

AGGATTCC**G**AGTTA**G**CGAAATTCAGGATTCA**A**GCTTAAATGGCTT

- The complete information is provided by haplotypes
  - the ordered list of SNPs on a chromosome
  - example: (A,C,G) and (G,G,A)

Shotgun Sequencing and Haplotype Assembly

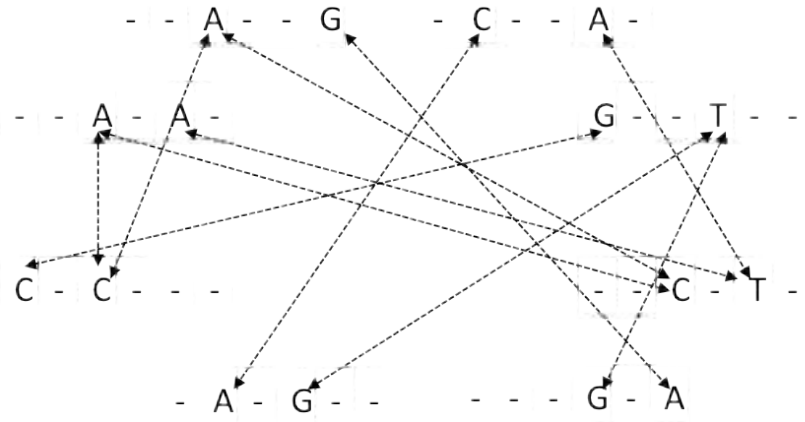
- Sequencing chromosome pairs:



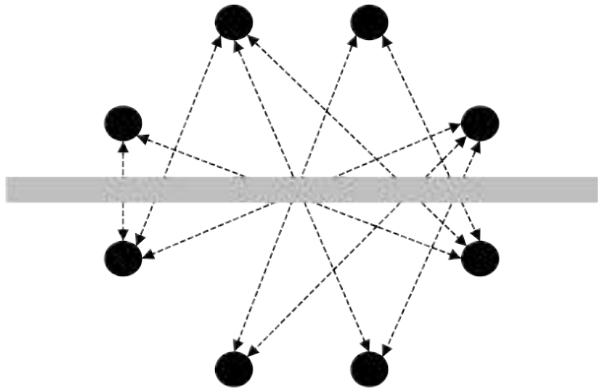
- Sample the chromosomes with short reads, some 'cover' SNPs
  - read association to chromosomes not known a priori

A Fragment Conflict Graph Interpretation

- Represent reads by nodes, conflicts by edges
  - fragments are in conflict if they cover a common SNP location but have different nucleotides there (so, different chromosomes)



(a)



(b)

- If data is error-free, conflict graph is bipartite
  - in this case, haplotype assembly is straightforward

Various Formulation of the Haplotype Assembly Problem

- If the conflict graph is not bipartite, assembly is non-trivial
- Approach: minimize the number of transformation steps needed to alter the graph so that it becomes bipartite
  - minimum fragment removal (MFR), minimum SNP removal (MSR)
- Minimum error correction (MEC): find the smallest number of nucleotides in reads whose flipping to a different value resolves conflicts among the fragments from the same chromosome
  - essentially, remove edges in the conflict graph by assuming the fewest possible sequencing errors; NP hard
- Existing methods: [Li, Kim, Waterman, 2004], HapCut [Bansal & Banfa, 2008], HapCompass [Aguiar & Istrail, 2013], HapTree [Berger et al., 2014]

Notation and Data Structure

- Assume we performed SNP/genotype calling
- Label SNPs in the i^{th} haplotype position as $h_i^1, h_i^2 \in \{1, -1\}$
  - define $\mathbf{h} = \mathbf{h}^1 = -\mathbf{h}^2 = [h_1^1 \;\; h_2^1 \;\; \ldots \;\; h_n^1]$
- Organize reads into a matrix $\mathbf{R}$, row $\mathbf{r}_i$ is the i^{th} read

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & -1 & 0 & -1 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & -1 \\ 0 & -1 & 0 & 0 & -1 & 0 \end{bmatrix}$$

Contributions

- Algorithms
  - communications-inspired haplotype assembly
    - haplotype assembly as the minimum distance decoding on a BSC [Z. Puljiz and HV, IEEE TCBB, 2015]
  - correlation clustering
    - SDP formulation, inherent sparsity of the solution enables fast heuristics [S. Das and HV, BMC Genomics, 2015]
  - structured low-rank matrix factorization [C. Cai, S. Sanghavi and HV, 2016]
- Analysis
  - an information-theoretic view: necessary/sufficient conditions for the number of reads for assembly [H. Si, S. Vishwanath and HV, 2015]

Structured Low-Rank Matrix Factorization

- The data matrix $\mathbf{R}$

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & -1 & 0 & -1 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & -1 \\ 0 & -1 & 0 & 0 & -1 & 0 \end{bmatrix}$$

Structured Low-Rank Matrix Factorization

- The underlying matrix $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} -1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

Structured Low-Rank Matrix Factorization

- The underlying matrix $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} -1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

- Define $P_\Omega(\mathbf{S})$ as

$$P_\Omega(\mathbf{S})_{ij} = \begin{cases} S_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise} \end{cases}$$

- Measurement model

$$R_{ij} = \begin{cases} S_{ij}, & \text{w.p.} \quad 1-p, \\ -S_{ij}, & \text{w.p.} \quad p. \end{cases}$$

Structured Low-Rank Matrix Factorization

- The underlying matrix $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 & -1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

- Define $P_\Omega(\mathbf{S})$ as

$$P_\Omega(\mathbf{S})_{ij} = \begin{cases} S_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise} \end{cases}$$

- Measurement model

$$R_{ij} = \begin{cases} S_{ij}, & \text{w.p.} \quad 1-p, \\ -S_{ij}, & \text{w.p.} \quad p. \end{cases}$$

Structured Low-Rank Matrix Factorization

- The underlying matrix $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \end{bmatrix}$$

- Define $P_\Omega(\mathbf{S})$ as

$$P_\Omega(\mathbf{S})_{ij} = \begin{cases} S_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise} \end{cases}$$

- Measurement model

$$R_{ij} = \begin{cases} S_{ij}, & \text{w.p.} \quad 1-p, \\ -S_{ij}, & \text{w.p.} \quad p. \end{cases}$$

Sampling a Low-Rank Matrix

- Matrix $\mathbf{S}$ has low rank, admits factorization

$$\mathbf{S} = \mathbf{U}\mathbf{V}^T$$

  - $\mathbf{U}$ and $\mathbf{V}$ are $n \times k$ and $m \times k$ matrices, respectively
  - k denotes the ploidy (the number of haplotypes)
- $\mathbf{U}$ consists of rows that indicate read origins
  - $\mathbf{u}_i = \mathbf{e}_j$ means the i^{th} read is obtained by sampling the j^{th} chromosome/haplotype
- The j^{th} column of $\mathbf{V}$, $\mathbf{v}_j$, is the sequence of the j^{th} haplotype
- The goal is to find $(\mathbf{U}, \mathbf{V})$ that minimize

$$f(\mathbf{U}, \mathbf{V}) = \|P_\Omega(\mathbf{R} - \mathbf{U}\mathbf{V}^T)\|_F^2$$

Structurally-Constrained Gradient Search

- Structurally-constrained gradient search

$$\mathbf{V}_{t+1} = \mathbf{V}_t - \alpha \nabla f(\mathbf{V}_t)$$

and

$$\mathbf{U}_{t+1} = \arg\min_{\mathbf{u}_i \in \mathbf{\Phi}} f(\mathbf{U}, \mathbf{V}_{t+1}),$$

- Convergence is guaranteed with the choice

$$\alpha = C \frac{\|\nabla f(V_t)^T\|_F^2}{\|P_\Omega(U_t \nabla f(V_t)^T)\|_F^2}, \quad C \in (0,1).$$

Structurally-Constrained Alternating Minimization

- Structurally-constrained alternating minimization

$$\mathbf{V}_{t+1} = \arg\min_{\mathbf{V}} \sum_{(i,j)\in\Omega} \left\| P_{\Omega}(\mathbf{R} - \mathbf{U}_t \mathbf{V}^T) \right\|_F^2$$

and

$$\mathbf{U}_{t+1} = \arg\min_{u_i \in \Phi} \sum_{(i,j)\in\Omega} \left\| P_{\Omega}(\mathbf{R} - \mathbf{U}\mathbf{V}_{t+1}^T) \right\|_F^2$$

until a termination criterion is met and then rounding the entries in $\mathbf{V}_{t_{max}}$ to ± 1.

- For $k = 2$, explicit relation between $\|\mathbf{M} - \mathbf{U}_t\mathbf{V}_t\|_F^2$ and the number of iterations t (i.e., a bound on $\|\mathbf{M} - \mathbf{U}_t\mathbf{V}_t\|_F^2$).

Benchmarking Results

- Simulations on a benchmark database (diploids)

	Algorithms	$n = 350$			$n = 700$		
		$m = 5n$	$m = 8n$	$m = 10n$	$m = 5n$	$m = 8n$	$m = 10n$
Reconstr. rates	SHR-three	0.724	0.742	0.728	0.716	0.743	0.726
	MLF	0.858	0.933	0.962	0.809	0.863	0.884
	2d-mec	0.913	0.964	0.978	0.880	0.948	0.965
	HapCUT	0.913	0.896	0.888	0.916	0.896	0.889
	SpeedHap	0.959	0.984	0.984	0.947	0.985	0.986
	Fast Hare	0.945	0.985	0.995	0.949	0.986	0.995
	SC Gradient Descent	0.959	0.996	0.998	0.951	0.997	0.999

- Additional testing on experimental data (1000 Genomes Project, HuRef, fosmid)
  - ground truth rarely available, MEC score as a performance metric

Benchmarking Results

- Performance of haplotype assembly for triploids ($k = 3$)

Coverage	SC Gradient Descent			HapCompass		
	SWER	MEC	Time(s)	SWER	MEC	Time(s)
5	0.022	404	4.763	0.032	2259	1259.21
10	0.010	367	9.08	0.011	5093	1315.09
15	0.003	2064	22.07	0.023	7784	995.05

- Performance of haplotype assembly for tetraploids ($k = 4$)

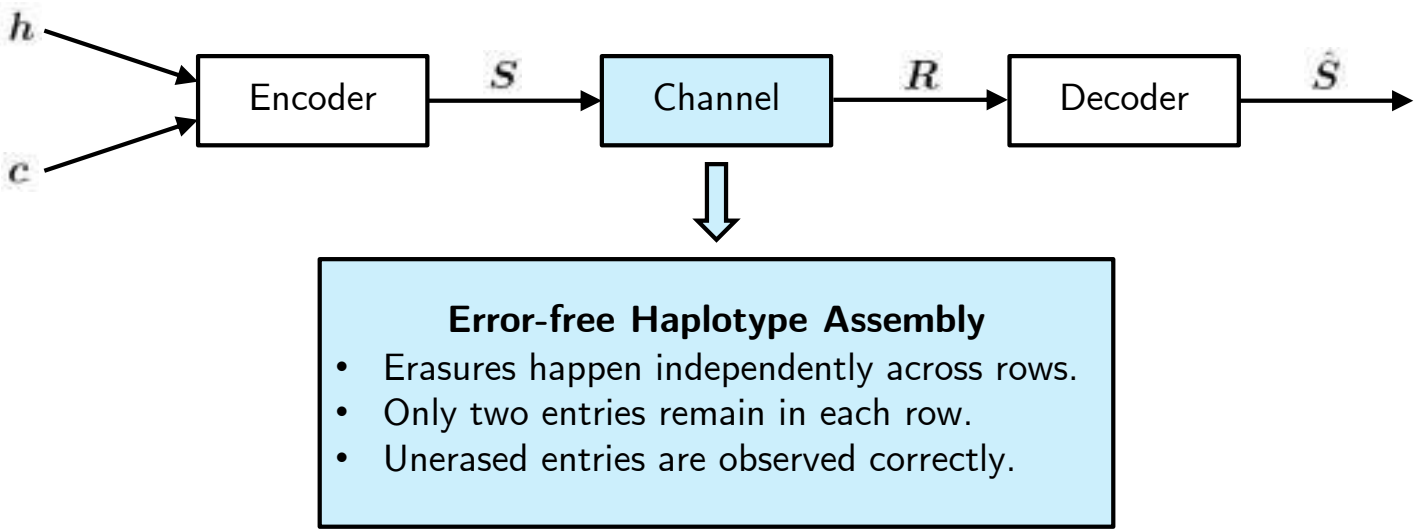
Coverage	SC Gradient Descent			HapCompass		
	SWER	MEC	Time(s)	SWER	MEC	Time(s)
5	0.083	1592	4.126	0.121	3510	1913.34
10	0.031	1962	14.529	0.114	7298	1329.46
15	0.027	2632	21.339	0.121	12239	2482.15

Roadmap

- Algorithms
  - communications-inspired haplotype assembly
    - haplotype assembly as the minimum distance decoding on a BSC
  - correlation clustering
    - SDP formulation, inherent sparsity of the solution enables fast heuristics
  - structured low-rank matrix factorization
- Analysis
  - an information-theoretic view: necessary/sufficient conditions for the number of reads for assembly [H. Si, S. Vishwanath and HV, 2015]

How Many Reads Are Needed: The Error-Free Case

- The joint source-channel coding formulation



- Q: What are the conditions on the number of reads (m) for accurate recovery?

How Many Reads Are Needed: The Error-Free Case

Theorem

Given the matrix $\mathbf{R}$ *with* 2 *reliable observations at arbitrary positions in each row, the original haplotype matrix* $\mathbf{S}$ *can be reconstructed if and only if the number of reads satisfies*

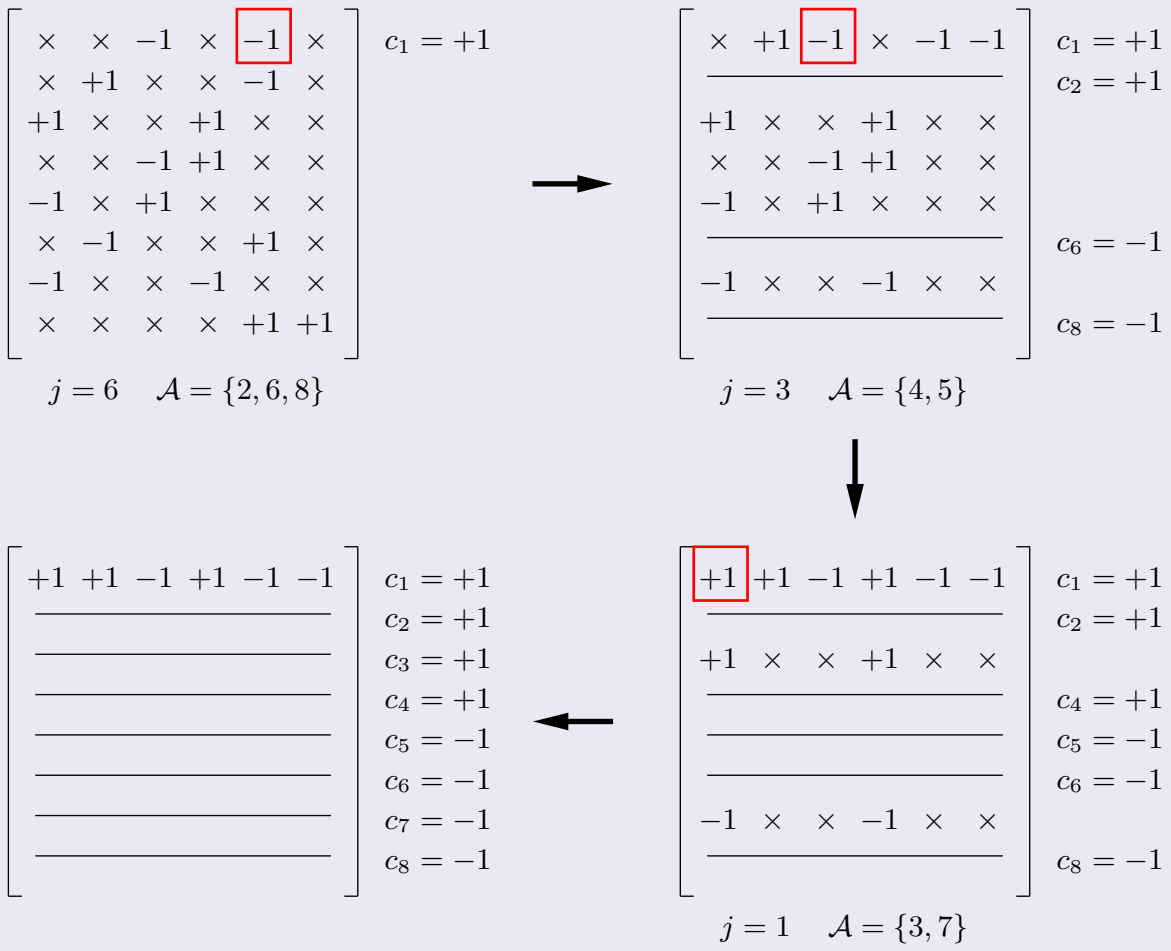
$$m = \Theta(n \ln n),$$

where n *is the length of the target haplotype, and the scaling factor can be chosen as* $1/2$.

- Proof by means of establishing conditions for connectivity of a random graph representing the problem + erasure decoding

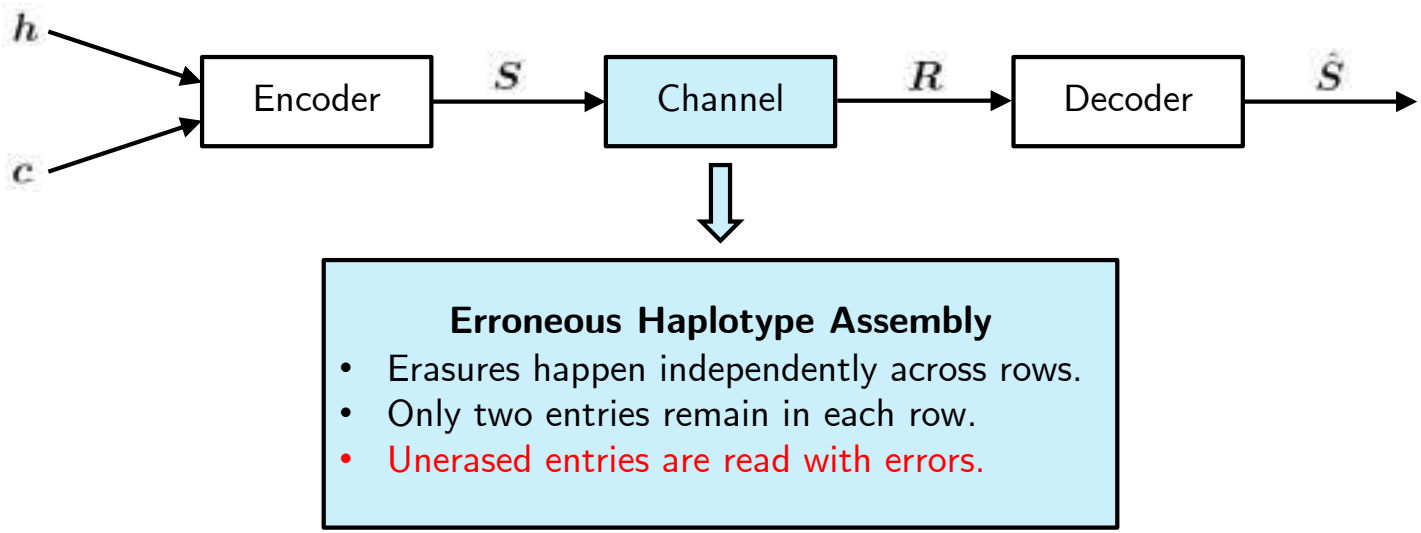
How Many Reads Are Needed: The Error-Free Case

- Sufficient condition: via erasure decoding



How Many Reads Are Needed: The Erroneous Case

- The joint source-channel coding formulation
  - every informative entry is "flipped" with probability p



- $\mathbf{R} = \mathcal{P}_{\Omega}(\mathbf{S} \oplus \mathbf{N})$, where $\mathbf{S} = \mathbf{c}^T \cdot \mathbf{h}$

How Many Reads Are Needed: The Erroneous Case

Theorem

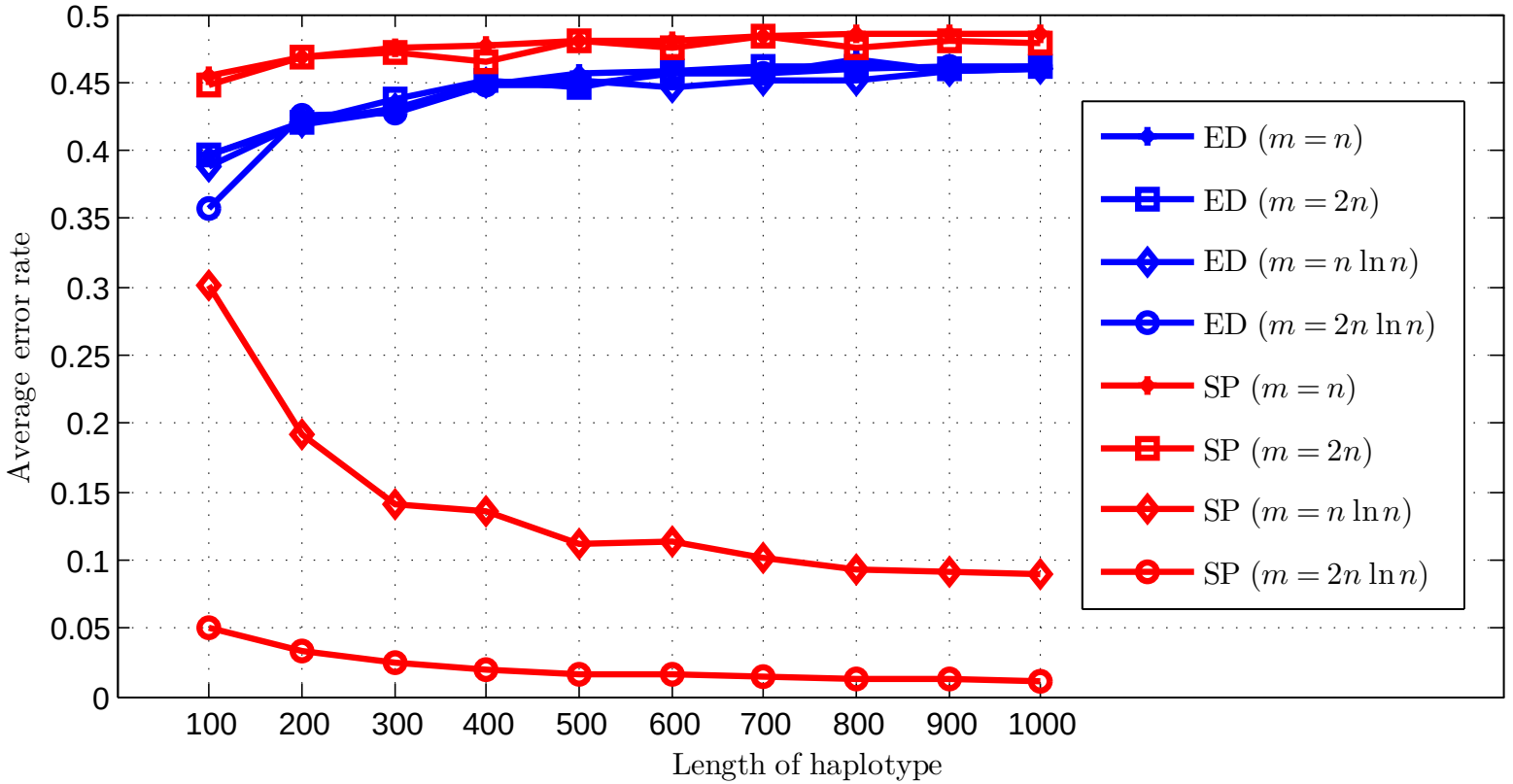
Given the matrix $\mathbf{R}$ *with* 2 *reliable observations at arbitrary positions in each row, the original haplotype matrix* $\mathbf{S}$ *can be reconstructed if and only if the number of reads satisfies*

$$m = \Theta(n \ln n),$$

where n *is the length of the target haplotype.*

- Yet again, proof via establishing conditions for connectivity of a random graph representing the problem
  - sufficient condition via spectral partitioning: construct adjacent matrix, perform SVD, clustering

Assembly Error Rate vs. Haplotype Length



Set $p = 0.1$, varied m

Summary and Future Work

- Haplotype assembly algorithms and analysis
  - low-rank matrix factorization significantly outperforms state-of-the-art methods
  - information-theoretic analysis provides conditions for recovery, recommendations on experimental specs
- Future work
  - performance analysis
  - applications
    - characterizing spectrum of viral quasi species, metagenomics, immunology

Backup Slide: Analysis of Alternating Minimization

Theorem

Let $\mathbf{M} = \mathbf{U}^\star \mathbf{\Sigma} \mathbf{V}^{\star \mathbf{T}} = \hat{\mathbf{U}}^\star \hat{\mathbf{V}}^{\star \mathbf{T}}$ *denote a rank*-1 *SNP fragment matrix with* μ*-incoherent rows and columns, and let* $\mathbf{N}$ *be the noise matrix. If the entries of* $\mathbf{R} = \mathbf{M} + \mathbf{N}$ *are observed uniformly and independently with probability*

$$p > C \frac{\kappa^4 \mu^4 \log n \log \frac{\|M\|_F}{\epsilon}}{m \delta_2^2},$$

where κ *is condition number of* $\mathbf{M}$, $\delta_2 \leq \frac{1}{64\kappa}$, $C > 0$, *then with high probability, after* $t = C' \frac{\log \|M\|_F}{\epsilon}$ *iterations it holds*

$$\|M - \hat{\mathbf{U}}^{(t)} [\hat{\mathbf{V}}^{(t)}]^T \|_F \leq \epsilon + C' \mu \kappa^2 \|N\|_F.$$