HapTree-X: An integrative Bayesian framework for haplotype reconstruction from transcriptome and genome sequencing data

Emily Berger[1,2,3,*], Deniz Yorukoglu[1,*], Bonnie Berger[1,2,**]

1 Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

2 Department of Mathematics, MIT, Cambridge, MA, USA

3 Department of Mathematics, UC Berkeley, Berkeley, CA, USA

∗ **These authors contributed equally.**

∗∗ **Corresponding author: bab@csail.mit.edu**

Abstract

Identifying phase information is biomedically important due to the association of complex haplotype effects, such as compound heterozygosity, with disease. As recent next-generation sequencing (NGS) technologies provide more read sequences, the use of diverse sequencing datasets for haplotype phasing is now possible, allowing haplotype reconstruction of a single sequenced individual using NGS data. Previous haplotype reconstruction studies have ignored differential allele-specific expression in whole transcriptome sequencing (RNA-seq) data; however, intuition suggests that the asymmetry in this data (i.e. maternal and paternal haplotypes of a gene are differentially expressed) can be exploited to improve phasing power. In this paper, we describe a novel integrative maximum-likelihood estimation framework, HapTree-X, for efficient, scalable haplotype assembly of an individual genome using transcriptomic and genomic NGS read datasets, which makes use of differential allele-specific expression. Our advance includes the first method for haplotype assembly that uses differential expression, newly allowing the use of reads that cover only one SNP. We evaluate the performance of HapTree-X on real sequencing read data, both transcriptomic and genomic, from NA12878 (1000 Genomes Project and Gencode) and demonstrate that HapTree-X increases the number of SNPs that can be phased and sizes of phased-haplotype blocks, without compromising accuracy. We prove theoretical bounds on the precise improvement of accuracy as a function of coverage which can be achieved from differential expression-based methods alone. Thus, the advantage of our integrative approach substantially grows as the amount of RNA-seq data increases.

Introduction

By running standard genotype calling tools, it is possible to accurately identify the number of "wild type" and "mutant" alleles for each single-nucleotide polymorphism (SNP) site [1]. However, in the case of two heterozygous SNP sites, genotype calling tools cannot determine whether "mutant" alleles from different SNP loci are on the same chromosome or on different homologous chromosomes (i.e. compound heterozygote). In many cases, the latter can cause loss of function while the former is healthy; therefore, it is necessary to identify the phase (or diplotype) — the copies of a chromosome that the mutant alleles occur — in addition to the genotype. Identifying phase information of an individual is important in biomedical studies due to disease association of complex haplotype effects such as compound heterozygosity [2], as well as matching the donor and the host in organ transplantation [3, 4].

As more sequencing data becomes available [5], we seek to design efficient algorithms to obtain accurate and comprehensive phase information directly from transcriptomic, as well as the commonly-used genomic, NGS read data. Transcriptome sequencing data differs from genomic read data in that genes often have differential haplotypic expression [6]. We wish to leverage this asymmetry to increase the number of SNPs of an individual that can be phased.

Methods have been proposed for the computational identification of an individuals diplotype using pedigree (e.g. trio-based phasing) [7, 8], population structure of variants (e.g. phasing by linkage disequilibrium) [9–12] and more recently by identity-by-descent in unrelated individuals [13, 14], as well as various types of sequencing read datasets: DNA-seq [15–19], RNA-seq [20]; proximity-ligation [21]; and haploid sperm sequencing [22]. Population-based, IBD and pedigree-based methods require data from a group of individuals to perform phasing of a individual. For solving the single-individual haplotype reconstruction problem, using RNA-seq or DNA-seq data are the most viable approaches as they are widely-available and inexpensive.

Long-range diplotyping is important because it gives more statistical power for downstream analyses [21]. Compared to DNA-seq, RNA-seq allows for longer-range phasing due to RNA splicing in the transcriptome. To date, approaches that utilize RNA-seq data for phasing (e.g., [20]) can only make use of reads covering 2 or more heterozygous SNPs, as they repurpose existing genome phasing approaches which are based on sequence contiguity. However, only 10% of reads that overlap a heterozygous SNP fall into this category (Table 1). Thus, current methods are discarding 90% of potentially useful information. Though these reads do not overlap multiple SNPs, as do those conventionally used for phasing, they provide insight into differential haplotypic expression within genes. An advantage of using reads covering only a single SNP is that phasing is not limited by the length of the read or fragment, nor the transcriptomic or genomic distance between SNPs.

In this paper, we develop the first method for solving the haplotype reconstruction problem using differential allele-specific expression (DASE) information within RNA-seq data. We follow the intuition that DASE in the transcriptome can be exploited to improve phasing power because SNP alleles within maternal and paternal haplotypes of a gene are present in the read data at (different) frequencies corresponding to the differential haplotypic expression [6]. To solve this haplotype reconstruction problem, we introduce a new maximum-likelihood formulation, generalizing that from HapTree [19], which considers DASE and is thus able to exploit reads covering only one SNP. This formulation results in an integrative algorithm, HapTree-X, which determines a haplotype of maximal likelihood based on both RNA-seq and DNA-seq read data. Our method does not require fragments to cover at least two SNPs and therefore for the first time can leverage the large number of RNA-seq fragments that cover only one SNP to produce more accurate phasing, as well as both novel and larger blocks of phased SNPs.

Reported RNA-seq phasing results using HapTree-X for a well annotated human lymphoblas-

toid cell line (GM12878) provide strong evidence for long-distance haplotype phasing capability of paired-end RNA-seq read alignments as well as the use of differential allele-specific expression as a practical haplotype reconstruction tool. Used jointly with genome reads in genotyping studies, RNA-seq reads can provide long distance scaffolds in order to be used for extending and merging haplotypes inferred from genome reads as well as introducing new long-distance phasing instances not possible to attain using short genome sequencing reads. We observe that compared to a state-of-the-art sequence-based haplotype reconstruction method, HapCut [15], HapTree-X, increases the total number of SNPs phased along with the sizes of phased haplotype blocks with improved accuracy, leveraging RNA-seq reads that only cover a single heterozygous-SNP in the transcriptome.

Method

Definitions and Notation

In this section we provide technical notation and discuss several basic assumptions we make.

The goal of phasing is to recover the unknown haplotypes (haploid genotypes), $H = (H_0, H_1)$, which contain the sequence of variant alleles inherited from each parent of the individual. As homozygous SNPs are irrelevant for phasing, we restrict ourselves to heterozygous SNPs (from now on referred to simply as a 'SNP') and we denote the set of these SNPs as S. We assume these SNPs to be biallelic, and because of these restrictions, H_0 and H_1 are complements. Let $H[s] = (H_0[s], H_1[s])$ denote the alleles present at s, for $s \in S$.

We denote the sequence of observed nucleotides of a fragment simply as a "read" (independent from single/paired-end reads). We assume each read is mapped accurately and uniquely to the reference genome, and moreover that each read is sampled independently. The set of all reads is denoted as R. Given a set of SNP loci S, we define a read $r \in R$ as a vector with entries $r[s] \in \{0, 1, -\}$, for $s \in S$, where a 0 denotes the reference allele, a 1 the alternative allele, and $-$ that the read does not overlap s or that it contains false allele information at s. We say a read $r \in R$ *contains* a SNP s if $r[s] \neq -$ and we let *size* of a read r, $|r|$, refer to the number of SNPs it contains. For each read r and for each SNP locus s, we assume a probability of opposite allele (reference if the true allele is the alternative, and vice versa) information $r[s]$ equal to $\varepsilon_{r,s}$ and represent these error probabilities as a matrix ε. We assume these errors to be independent from one another.

In genomic read data, all $r \in R$ are equally likely to be sampled from the maternal or paternal chromosomes. In RNA-seq data however, this may not always be the case. In this paper, we define the differential haplotypic expression (DHE) to represent the underlying *expression bias* between the maternal and paternal chromosomes of a particular gene. Throughout, we will refer to the probability of sampling from the higher frequency haplotype of a gene as β. We assume two genes g, g' have independent expression biases β, β'. Differential allele-specific expression (DASE) we define as the *observed* bias in the alleles at a particular SNP locus present in R. We define *concordant expression* as when the DASE of a SNP agrees with the DHE of the gene to which the SNP belongs; that is when the majority allele (allele occurring with higher frequency) occurring within the reads at a particular SNP locus is in agreement with the expected majority allele as determined by the DHE.

To perform phasing using the sequence contiguity within reads (contig-based phasing), upon the set of SNP loci S and read set R, we define a *Read Graph* such that there is a vertex for each SNP locus $s \in S$ and an edge between any two vertices s, s' if there exists some read r containing both s and s'. These connected components correspond to the haplotype blocks to be phased.

To phase using differential expression (DASE-based phasing), we assume the existence of some

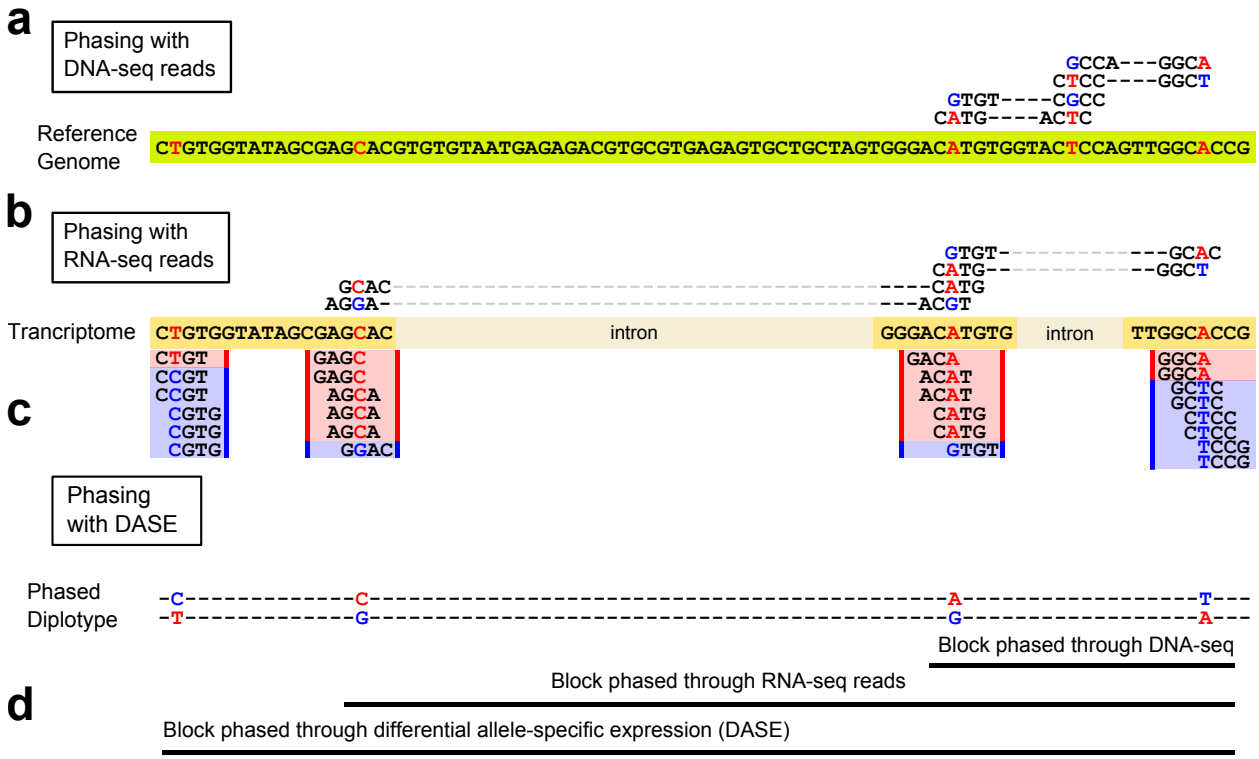

Figure 1: A toy example demonstrating the haplotype phasing capabilities of and differences between single-individual haplotype reconstruction methods using genome sequencing (DNA-seq) reads (a), transcriptome sequencing (RNA-seq) reads (b), and differential allele-specific expression (DASE) information that can be inferred from RNA-seq data (c). Green and orange blocks respectively represent reference genome and the transcriptome sequence, which contains only the exons in a gene separated by introns. Positions marked in red denote heterozygous-SNP loci. Paired-end sequencing reads are of length 2x4bp and have 3-4 bp insert lengths; reference alleles overlapping SNP loci are marked with red and alternative alleles are marked with blue. (a) Phasing using DNA-seq reads can be performed by looking at reads that overlap multiple heterozygous-SNP loci and observing the alleles that are connected through reads. Phasing distance is limited by maximum fragment length (12bp in the example). Multiple SNP loci can be chained together for phasing, but the probability of a switch error increases with the length of the chain. (b) Though limited to only the SNPs within the transcriptome, RNA-seq reads have longer distance phasing capability than DNA-seq reads due to long introns in the genome that are spliced-out in the sequenced transcript fragments. RNA-seq reads also provide higher accuracy phasing of SNPs within the transcriptome compared to DNA-seq, since DNA-seq phasing needs to chain through intron SNPs to connect the exons. (c) Differential allele-specific expression (DASE) at transcriptomic SNP loci is available within RNA-seq datasets in the form of allele-specific coverage ratios. For genes that display differential haplotypic expression (DHE), the majority of alleles can be phased together to obtain a single haplotype block for the entire gene. Depending on the DHE and depth-coverage, DASE-based phasing can perform accurate haplotype reconstruction, independent of gene/exon lengths, without requiring paired-end or long reads. (d) Phasing capabilities of DNA-seq, RNA-seq and DASE based phasing methods are demonstrated on the given toy example. The genome sequencing based approach is only able to provide haplotype blocks for the exons close together. The RNA-seq read based approach is able to reconstruct a longer haplotype block, phasing through the introns as well, but failing to phase far apart SNPs within the first exon. Whereas DASE-based phasing is able to reconstruct the complete gene haplotype by leveraging differential expression at SNP loci.

gene model G that specifies the genes (and their exons) within the genome. For each $g \in G$, we assume that the haplotypes (H_0, H_1) restricted to g are expressed at rates β_0, β_1 respectively due to DHE. The phasing blocks correspond to the SNPs in genes $g \in G$, though we will see that some SNPs are not phased due to insufficient probability of concordant expression. Two distinct genes g, g' may not be DASE-phased due to lack of correlation between their expression biases β, β'.

In the remainder of this paper, when DASE-phasing a particular gene, by H we mean the gene haplotype, that is H restricted to the SNPs within g.

The blocks which are able to be phased by HapTree-X integrating both contig and DASE-based phasing are defined as the connected components of a *Joint Read Graph*. In the Joint Read Graph, each vertex corresponds to a SNP phased by either method, and there is an edge between any two vertices (SNPs) s, s' if there exists some block that was phased by either method containing both s, s'.

Building on the definitions above, we describe the mathematical underpinnings of the haplotype reconstruction problem that assume the existence of DHE. We include an overview of our algorithm in **HapTree-X Framework**.

Likelihood of a Phase

We formulate the haplotype reconstruction problem as identifying the most likely phase(s) of set of SNPs S, given the read data R, and sequencing error rates ε. Furthermore, suppose we knew for each read r, the likelihood that r was sampled from H_i (denote this as β_i^r); we represent these probabilities as a matrix $\mathcal{B}$. While $\mathcal{B}$ is not given to us, we may estimate $\mathcal{B}$ from R (see **Maximum Likelihood Estimate of Differential Haplotypic Expression**). We derive a likelihood equation for H, conditional on $R, \mathcal{B}$ and ε.

Given a haplotype H, reads R, error rates ε, and $\mathcal{B}$, the likelihood of H being the true phase is

$$\mathrm{P}[H \mid R, \mathcal{B}, \varepsilon\,] = \frac{\mathrm{P}[R \mid H, \mathcal{B}, \varepsilon\,]\,\mathrm{P}[H \mid \mathcal{B}, \varepsilon]}{\mathrm{P}[R \mid \mathcal{B}, \varepsilon\,]}. \tag{1}$$

Since $\mathrm{P}[R \mid \mathcal{B}, \varepsilon]$ does not depend on H, we may define a *relative likelihood* measure, RL. Note that $\mathrm{P}[H|\mathcal{B}, \varepsilon] = \mathrm{P}[H]$ as the priors on the haplotypes are independent of the errors in R, and of $\mathcal{B}$.

$$\mathrm{RL}[H \mid R, \mathcal{B}, \varepsilon\,] = \mathrm{P}[R \mid H, \mathcal{B}, \varepsilon\,]\,\mathrm{P}[H]. \tag{2}$$

For the prior $\mathrm{P}[H]$, we assume a potential *parallel bias*, $\rho \geq .5$, (the prior probability of adjacent SNPs being phased in parallel as opposed to switched) which results in a distribution on H such that adjacent SNPs are independently believed to be phased in parallel $\binom{00}{11}$ with probability ρ and switched $\binom{01}{10}$ with probability $1 - \rho$. When $\rho = .5$ we have the uniform distribution on H. The general prior distribution on H in terms of ρ is

$$\mathrm{P}[H] = \rho^{P(H)}(1-\rho)^{S(H)} \tag{3}$$

where $P(H)$ and $S(H)$ denote the number of adjacent SNPs that are parallel and switched in H, respectively. Given the above model, as each $r \in R$ independent, we may expand $\mathrm{P}[R \mid H, \mathcal{B}, \varepsilon\,]$ as a product:

$$\mathrm{P}[R \mid H, \mathcal{B}, \varepsilon] = \prod_{r \in R} \mathrm{P}[r \mid H, \mathcal{B}, \varepsilon] \tag{4}$$

In the setting of RNA-seq, reads are not sampled uniformly across homologous chromosomes, but rather according to the DHE (*expression bias*) of the gene from which they are transcribed. We see in (5) how this asymmetry allows us to incorporate reads which contain only one SNP. Let $A(r, H_i), D(r, H_i)$ denote the SNP loci where r and H_i agree and disagree respectively, then

$$\mathrm{P}[r \mid H, \mathcal{B}, \varepsilon] = \sum_{i \in [0,1]} \left(\beta_i^r \prod_{s \in A(r,H_i)} (1 - \varepsilon_{r,s}) \prod_{s \in D(r,H_i)} \varepsilon_{r,s} \right). \tag{5}$$

When there is uniform expression $\beta_0^r = \beta_1^r$ (no bias) and if $|r| = 1$, then $\mathrm{P}[r \mid H, \mathcal{B}, e]$ is constant across all H. This is not the case when the expression bias is present however, and therefore reads covering only one SNP affect the likelihood of H.

If we knew the matrix $\mathcal{B}$, we could apply HapTree [19] to find H of maximal likelihood; the matrix $\mathcal{B}$, however, is unknown. Throughout this paper we provide methods for determining a maximum likelihood $\mathcal{B}$, and for which reads r we are sufficiently confident there this is in fact non-uniform expression, that is $\beta_0^r \neq \beta_1^r$. Moreover, we determine for which SNPs $s \in S$ (contained only by reads of size one), we have sufficient coverage and expression bias to determine with high accuracy the phase $H[s]$.

Maximum Likelihood Estimate of Differential Haplotypic Expression

For a fixed gene g, containing SNPs S_g, the corresponding reads R_g have expression biases β_0^r, β_1^r which are constant across $r \in R_g$. Let $\beta = \beta_0^r$ refer to this common expression; we wish to determine the maximum likelihood underlying expression bias β of g responsible for producing R_g. To do so, we formulate a Hidden Markov Model (HMM) and use the forward algorithm to compute relative likelihoods of R given β, ε.

To achieve the conditional independence required in a HMM, we define R'_g, a modification of R_g, containing only reads of size one, so that $R'_{g,s}$ (the reads $r \in R'_g$ which cover s) are independent from $R'_{g,s'}$ ($\forall s \neq s' \in S_g$). We restrict each $r \in R_g$ to a uniformly random SNP s, and include this restricted read of size one ($r|_s$) in R'_g (we note that if $|r| = 1$, then $r = r|_s$, by definition.) Therefore, $R'_{g,s}$ and $R'_{g,s'}$ are independent as all $r \in R'_g$ are of size one.

Our goal is the determine the maximum likelihood β, given R'_g. We assume a uniform prior on β, and therefore $\mathrm{P}[\beta \mid R'_g, \varepsilon]$ is proportional to $\mathrm{P}[R'_g \mid \beta, \varepsilon]$ (immediate from Bayes theorem). We may theoretically compute $\mathrm{P}[R'_g \mid \beta, \varepsilon]$ by conditioning H (which is independent from β, ε)

$$\mathrm{P}[R'_g \mid \beta, \varepsilon] = \sum_H \mathrm{P}[R'_g \mid H, \beta, \varepsilon] \, \mathrm{P}[H]$$

and expand $\mathrm{P}[R'_g \mid H, \beta, \varepsilon]$ as a product over $r \in R'_g$ as in (4) and (5). This method, however, requires enumerating all H; since $|H| = 2^{|S_g|}$ we seek different approach. Indeed, we translate this process into the framework of a Hidden Markov Model, apply the forward algorithm to compute $f(\beta) := \mathrm{P}[R'_g \mid \beta, \varepsilon]$ exactly for any β, and since f has a unique local maxima for $\beta \in [.5, 1]$, we can apply Newton-Rhapson method to determine β of maximum likelihood.

To set this problem in the framework of a Hidden Markov Model, we let the haplotypes H correspond to the hidden states, R'_g to the observations, and let the time evolution be the ordering of the SNPs S_g. The observation at time s in this context is $R'_{g,s}$, the reads covering SNP s. The emission distributions are as follows:

$$\mathrm{P}[R'_{g,s} \mid H[s], \beta, \varepsilon] = \prod_{r \in R'_{g,s}} \mathrm{P}[r \mid H[s], \beta, \varepsilon]$$

$$\mathrm{P}[r \mid H[s], \beta, \varepsilon] = \begin{cases} \beta_0(1 - \varepsilon_{r,s}) + (1 - \beta_0)\varepsilon_{r,s} & \text{if } r[s] = H_0[s] \\ \beta_1(1 - \varepsilon_{r,s}) + (1 - \beta_1)\varepsilon_{r,s} & \text{if } r[s] = H_1[s] \end{cases} \tag{6}$$

where $H[s]$ is H restricted to s.

To determine the hidden state transition probabilities, recall our prior on H in (3). We may equivalently model this distribution H as a Markov chain, with transition probabilities:

$$\mathrm{P}[H[s_{i+1}] \mid H[s_i]] = \begin{cases} \rho & \text{if } H_0[s_i] = H_0[s_{i+1}] \\ 1-\rho & \text{if } H_0[s_i] \neq H_0[s_{i+1}] \end{cases}$$

These emission probabilities and hidden state transition probabilities are all that are needed to apply the forward algorithm and determine the β of maximum likelihood.

Likelihood of Concordant Expression

A Solution of Maximum Likelihood

In this section we prove that the intuitively correct solution (under mild conditions CND1, CND2, and CND3) is that of maximal likelihood. In doing so, we see the role played by concordant expression, and motivate its use as a probabilistic measure for determining which SNPs we believe we may phase with high accuracy.

We derive H^+, a haplotype solution of a gene g, of maximum likelihood given R'_g, β and ε and conditions CND1, CND2, and CND3. Let C_s^v denote the number of reads $r \in R'_{g,s}$ such that $r[s] = v$ where $v \in \{0, 1\}$. Provided error rates are constant (CND1) (say ϵ) and $\epsilon < .5$ (CND2), and assuming a uniform prior distribution ($\rho = .5$) (CND3), we can show a solution of maximum likelihood is $H^+ = (H_0^+, H_1^+)$, where $H_0^+[s] = v$ such that $C_s^v \geq C_s^{1-v}$. In words, H_0^+ and H_1^+ contain the alleles that are expressed the majority and minority of the time (respectively) at each SNP locus; given sufficient expression bias and coverage, intuitively, H^+ ought to correctly recover the true haplotypes. It is easy to show that CND1 and CND3 can be removed if one is willing to specify a minimum coverage; we do not show this here. Intuitively, CND2 must not be removed.

To prove H^+ is of maximal likelihood, we introduce the terms *concordant expression* and *discordant expression*. We say R and H have *concordant expression* at s if $C_s^{H_0[s]} > C_s^{H_1[s]}$, *discordant expression* if $C_s^{H_0[s]} < C_s^{H_1[s]}$, and *equal expression* otherwise. In words, since we assume $\beta_0 > \beta_1$, we *expect* to see the allele $H_0[s]$ expressed more than the allele $H_1[s]$ in $R_{g,s}$ (concordant expression.)

We may now equivalently define H^+ as a solution which assumes concordant or equal expression at every SNP s. Because we assume uniform priors, $\mathrm{P}[H \mid R'_g, \beta, \epsilon]$ is proportional $\mathrm{P}[R'_g \mid H, \beta, \epsilon]$ (see (1)), and since each read is of size one, we can factor across S_g in the following way:

$$\mathrm{P}[R \mid H, \beta, \epsilon] = \prod_{s \in S_g} \mathrm{P}[R_{g,s} \mid H[s], \beta, \epsilon]$$

Therefore, to show H^+ is of maximal likelihood, it only remains to show that concordant expression is at least as likely as discordant expression, as intuition suggests. Let $\gamma_i = \beta_i(1-\epsilon) + (1-\beta_i)\epsilon$, then as in (6) we may deduce

$$\mathrm{P}[R_{g,s} \mid H[s], \beta, \epsilon] = \prod_{i \in \{0,1\}} \gamma_i^{C_s^{H_i[s]}}$$

Let $H^- = (H_1^+, H_0^+)$, the opposite of H^+. We can now compare the likelihood of concordant (or equal) expression at s ($H^+[s]$) with that of discordant (or equal) expression at s ($H^-[s]$.) For ease of notation, let $v_i = H_i^+[s]$ and $w_i = H_i^-[s]$.

$$\frac{\mathrm{P}[R_{g,s} \mid H^+[s], \beta, \epsilon]}{\mathrm{P}[R_{g,s} \mid H^-[s], \beta, \epsilon]} = \frac{\prod_{i \in \{0,1\}} \gamma_i^{C_s^{v_i}}}{\prod_{i \in \{0,1\}} \gamma_i^{C_s^{w_i}}} = \frac{\gamma_0^{C_s^{v_0}} - C_s^{w_0}}{\gamma_1^{C_s^{w_1}} - C_s^{v_1}} = \left(\frac{\gamma_0}{\gamma_1}\right)^{C_s^{v_0} - C_s^{v_1}} \geq 1 \tag{7}$$

The rightmost equality results from the fact that $H_i^+ = H_{1-i}^-$, and hence $v_i = w_{1-i}$. Since $\epsilon < .5$, we have $\gamma_0 > \gamma_1$; $C_s^{v_0} - C_s^{v_1} \geq 0$ by the definition of H^+, which proves the inequality.

Computing Likelihood of Concordant Expression

We just showed that under mild conditions, the solution of maximal likelihood is, intuitively, that which has concordant expression at each SNP locus s. Therefore, to determine which SNPs we believe we can phase with high accuracy, we measure the probability of concordant expression at that SNP, and only phase when that probability is sufficiently high.

These probability of concordant expression can be immediately derived from (7). We assume a uniform error rate of ϵ for ease of notation, though is not required. Let $\mathrm{CE}(R_{g,s}, H[s])$ denote the event of concordant expression at s, then

$$\mathrm{P}[\mathrm{CE}(R_{g,s}, H[s]) \mid \beta, \epsilon] = \frac{\mathrm{P}[R_{g,s} \mid H^+[s], \beta, \epsilon]}{\mathrm{P}[R_{g,s} \mid H^+[s], \beta, \epsilon] + \mathrm{P}[R_{g,s} \mid H^-[s], \beta, \epsilon]} = \frac{1}{1 + \left(\frac{\gamma_1}{\gamma_0}\right)^{|C_s^0 - C_s^1|}} \tag{8}$$

Furthermore, given N reads, an expression bias β, and a constant error rate ϵ, we compute likelihood of concordant expression using the standard binomial distribution $B(N, \gamma_0)$ by equating 'successes' in the binomial model to observations of the majority allele, expressed with bias γ_0 (recall γ_i takes errors into account):

$$\mathrm{P}[\mathrm{CE} \mid N, \beta, \epsilon] = \sum_{i=\lceil \frac{N+1}{2} \rceil}^{N} \binom{N}{i} \gamma_0^i \gamma_1^{N-i} \;\geq\; 1 - e^{-N \frac{1}{2\gamma_0} (\gamma_0 - \frac{1}{2})^2} \tag{9}$$

To obtain the bound on the right hand side, apply the Chernoff bound $\mathrm{P}[X < (1-\lambda)\mu)] \leq e^{-\frac{\lambda^2 \mu}{2}}$ where X corresponds to the number of 'successes' and $\mu = \mathrm{E}[X] = N\beta$. This bound shows that the probability of concordant expression increases exponentially with the coverage (N).

We remark for large N, the Binomial Distribution $\mathrm{B}(n, \beta)$ converges to the normal distribution $\mathcal{N}(N\beta, N\beta(1-\beta))$, and therefore this probability can always be easily computed. See Figures 2 and 3 for a sense of these likelihoods.

HapTree-X Framework

HapTree-X is a novel Bayesian haplotype reconstruction framework, tailored to RNA-seq read datasets, which employs simultaneous contig-based and DASE-based haplotype phasing.

HapTree-X outputs phased haplotype blocks, given an input of RNA-seq (and optionally in addition DNA-seq) read alignment files (bam/sam), a VCF file containing the individual's genotype, and a gene model which specifies the genes (and their exons) within the genome.

The HapTree-X pipeline is initiated by determining which genes are expressed using the gene model and RNA-seq data. For each of these genes, a maximum likelihood expression bias (DHE) is computed (see **Maximum Likelihood Estimate of Differential Haplotypic Expression**). Furthermore, we determine which SNPs within those genes have high likelihood of concordant expression (see **Computing Likelihood of Concordant Expression**); we phase only those SNPs.

For reads containing only such SNPs, we assign to them the computed expression bias of the gene they cover; for all other reads, we assign a non-biased expression. Finally, applying a generalized version of HapTree [19], we determine a haplotype of maximal likelihood (as defined in **Likelihood of a Phase**) which depends on the DASE present in the RNA-seq data, as well as the sequence contiguity information within the reads.

HapTree-X is available at `http://groups.csail.mit.edu/cb/haptreex/`.

Results

Theoretical Performance of HapTree-X Framework

We demonstrate in **Computing Likelihood of Concordant Expression** the differential haplotypic expression level of a gene, β, and its coverage determine likelihood of concordant expression. We show this relationship below for varying β and levels of coverage. While these functions are derived from an idealized model of the data (for genes without alternative splicing and no amplification bias), this relationship suggests that as the depth-coverage of a dataset increases, so does the likelihood of concordant expression, and hence the accuracy of HapTree-X. Figure 1 displays the theoretical curves depicting the exponential growth of likelihood of concordant expression as a function of coverage and β, as described in (9). We infer from this theoretical result that requiring a lower bound of DHE is beneficial for reliable DASE-based phasing given moderate coverage (30-50+). Furthermore, we present a table including minimum coverage required to obtain a probability of at least $1 - 10^{-\alpha}$ of concordant expression, given β.

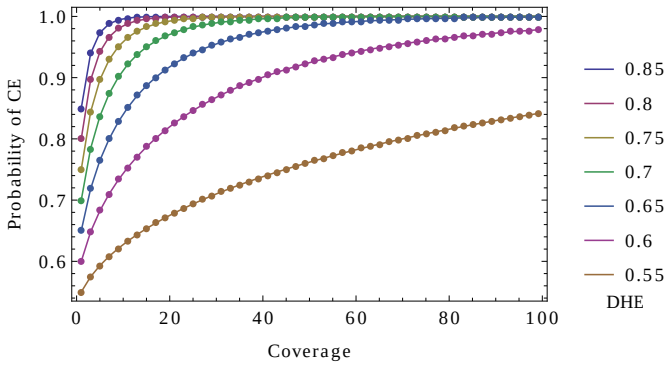

Figure 2: Likelihood of concordant expression (CE) as a function of coverage and differential haplotypic expression $\beta \in [.55, .6, .65, .7, .75, .8, .85]$

$\beta \setminus \alpha$	2	3	4	5
.85	9	15	21	27
.8	13	21	31	41
.75	19	33	49	63
.7	31	55	79	105
.65	57	101	147	193
.6	133	235	339	447
.55	539	951	1377	1811

Figure 3: Coverage needed to obtain likelihood $1 - 10^{-\alpha}$ of concordant expression given a differential haplotypic expression of β and an assumed opposite allele error rate of 2%.

Experimental Results

Datasets and Experimental Setup

We evaluate haplotype reconstruction performance of HapTree-X on diploid RNA-seq and DNA-seq read datasets from GM12878, a well-studied lymphoblastoid cell line from a human female individual with European Ancestry (1000 Genomes Project [23]).

To assess the accuracy of phased haplotype blocks generated by HapTree-X, we compare our phasing results to a high-quality trio-phased SNP annotation of GM12878 (1000 Genomes Project Phase I), the gold-standard phasing reference. RNA-seq raw read datasets of GM12878 are obtained from ENCODE CSHL Long RNA-seq (wgEncodeCshlLongRnaSeq) [24] track with average sequencing depth of 100 million mate-pairs (2x76bp), transcriptome fragments sequenced from the nucleus with Poly-A^+ and Poly-A^- profiling.

For each RNA-seq dataset, we performed 2-pass alignments using STAR aligner v2.4.0d [25] by initially aligning raw reads to hg19 reference genome and then realigning reads to a second index generated from the splice junctions inferred from the first alignment.

We restricted DASE-based phasing within HapTree-X only to the SNPs that are located within the same gene in the GENCODE gene annotation v19 (wgEncodeGencodeCompV19) [26]. For joint

DNA-seq and RNA-seq phasing experiments, we obtained genome sequencing reads of NA12878 from 1000 Genomes Project Pilot 2 release aligned to the hg19 reference genome using bwa aligner [27] and input both genome and transcriptome reads to the HapTree-X haplotype reconstruction framework.

We compared our results to those of HapCUT v0.7 sequence-based haplotype reconstruction tool [15]. To accommodate for long range splicing-junctions within RNA-seq read alignments, we defined maximum insert-size (maxIS) parameter to be longer than each chromosome's length.

Results from GM12878 data sets

In the RNA-seq read datasets from GM12878 (PolyA+ and PolyA- together), we observe that majority of the reads ($\sim 89\%$) only cover a single heterozygous SNP in the genome. The distribution of read sizes are given in Table 1. Of the 19782889 reads containing one SNP, we are able to confidently assign expression biases to 675892 of them; we use these reads to phase, increasing the total number of reads to be used in phasing by 28%.

Read Size	1	2	3	4	5	6	7	$8-13$
Count	19782889	2027207	290489	47424	17176	11941	10623	9119
%	89.12	9.133	1.311	.2137	.0774	.0538	.0479	.0411

Table 1: Distribution of read sizes (#heterozygous-SNPs covered) in GM12878 RNA-seq data (PolyA+ and PolyA-).

Table 2 summarizes the haplotype reconstruction performance of HapTree-X in comparison to a contig-based algorithm, HapCut. Running HapTree-X without any DASE-based phasing (using only reads covering at least two SNPs) yields identical statistics (besides switch error) to HapCut, as both employ the ReadGraph structure to determine the SNPs and blocks to be phased. The switch error rate of HapTree-X without DASE-based phasing is consistent with that from with DASE-based phasing.

Datasets \ Stats	SNPs	Switch Errors	Blocks	Edges	SNP Pairs
HapTree-X (DNA-seq & RNA-seq)	979181	3767	298637	680544	5121692
HapCut (DNA-seq & RNA-seq)	978811	5718	298710	680101	5101488
HapTree-X (RNA-seq)	220849	641	88355	132494	412534
HapCut (RNA-seq)	220386	669	88403	131985	380718
HapTree-X (DASE only)	1580	6	435	1145	4884

Table 2: Haplotype reconstruction results from HapTree-X and HapCut using DNA-seq and RNA-seq datasets from NA12878. Both HapCut and HapTree-X results are reported on RNA-seq read datasets as well as DNA-seq and RNA-seq merged datasets. DASE-based phasing only results from HapTree-X are also reported. For each dataset we report total number of phased SNPs, switch errors, haplotype blocks, edges and SNP pairs.

Results indicate that incorporating differential allele-specific expression in haplotype phasing increases the total number of SNPs phased, without increasing the switch error rate (with respect to the trio-phased gold-standard annotation). Furthermore, HapTree-X reduces the total number of blocks while increasing their overall sizes. We represent this by #Edges = #SNPs − #Blocks, equivalently the total number of pairs of adjacent (within a block) phased heterozygous-SNPs. This is also demonstrated by the large increase of total phased SNP pairs (any two SNPs within the same block). This indicates that HapTree-X produces longer haplotype blocks as a result of DASE-based phasing, as desired.

As discussed in **Likelihood of Concordant Expression**, the solution of maximum likelihood (for any gene g) corresponds to that with concordant expression at all SNP loci within g. HapTree-X therefore uses a threshold λ (negative log-likelihood of concordant expression) which requires any SNP to be concordantly expressed with probability at least $1 - e^{-\lambda}$, in order to be phased. We run HapTree-X while varying this threshold λ; we compute the percentage of concordantly expressed SNPs and the total phased SNPs as we increase this threshold. As the threshold increases, HapTree-X demands any SNP to be phased to have a correspondingly high likelihood of concordant expression; as a result, the phasing accuracy of HapTree-X increases. The cost paid for this increase in accuracy is a decrease in the total number of SNPs phased, as seen in Figure 4.

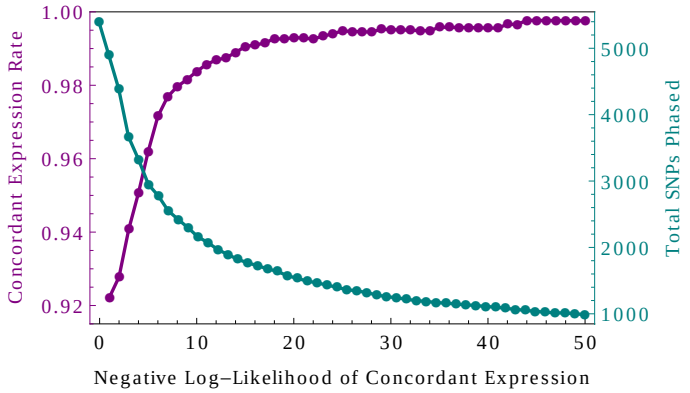

Figure 4: Rate of concordantly expressed SNPs (purple) and total number of SNPs phased (green) by HapTree-X, as a function of λ, the negative log-likelihood of concordant expression.

For the results reported in Table 2, we used a threshold value of 20. In theory, this threshold value λ would produce a percentage of concordantly expressed SNPs equal to $1 - e^{-\lambda}$; however because of the structural noise commonly observed in aligned RNA-seq data due to false mapping, RNA-editing, as well complex alternative splicing events, we require a $\lambda' > \lambda$ to meet desired accuracy levels. Additionally, we require an estimated $\beta \geq .6$ for any gene to be phased using DASE, for motivation see Figure 2. Finally, we have several methods for managing alternative splicing events. HapTree-X can (1) avoid all genes with alternative splicing, (2) phase s, s' only if the set of isoforms containing s, s' are equal, and (3) phase independent of isoforms but require s, s' to have coverage and DASE that are sufficiently similar. (3) was used in Table 2; (2), and especially (1), result in higher accuracy for lower λ, but of course phase fewer SNPs.

Discussion and Future Work

We have presented a novel haplotype reconstruction framework HapTree-X that is tailored towards transcriptome sequencing datasets, leveraging both sequence contiguity information within reads that overlap multiple heterozygous-SNPs as well as differential allele-specific expression (DASE) levels. We presented a maximum likelihood model for DASE-based haplotype reconstruction and a computational phasing algorithm that integrates haplotype reconstruction through differential expression and sequence contiguity.

Future extensions to HapTree-X framework includes incorporation of complex alternative and chimeric splicing events for higher resolution estimation of DASE within an augmented mathematical model of isoform-specific differential allelic expression within genes. As more RNA-seq data

sets become available, HapTree-X will be useful for personalized medicine through identification of disease association of gene haplotypes as well as genotype imputation studies.

References

1. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, *et al.*, "The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data," *Genome research*, vol. 20, no. 9, pp. 1297–1303, 2010.

2. R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork, "The importance of phase information for human genomics," *Nature Reviews Genetics*, vol. 12, no. 3, pp. 215–223, 2011.

3. D. C. Crawford and D. A. Nickerson, "Definition and clinical importance of haplotypes," *Annu. Rev. Med.*, vol. 56, pp. 303–320, 2005.

4. E. W. Petersdorf, M. Malkki, T. A. Gooley, P. J. Martin, and Z. Guo, "Mhc haplotype matching for unrelated hematopoietic cell transplantation," *PLoS medicine*, vol. 4, no. 1, p. e8, 2007.

5. B. Berger, J. Peng, and M. Singh, "Computational solutions for omics data," *Nature Reviews Genetics*, vol. 14, no. 5, pp. 333–346, 2013.

6. D. Serre, S. Gurd, B. Ge, R. Sladek, D. Sinnett, E. Harmsen, M. Bibikova, E. Chudin, D. L. Barker, T. Dickinson, *et al.*, "Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression," *PLoS genetics*, vol. 4, no. 2, p. e1000006, 2008.

7. A. Efros and E. Halperin, "Haplotype reconstruction using perfect phylogeny and sequence data," *BMC bioinformatics*, vol. 13, no. Suppl 6, p. S3, 2012.

8. A. L. Williams, D. E. Housman, M. C. Rinard, and D. K. Gifford, "Rapid haplotype inference for nuclear families," *Genome biology*, vol. 11, no. 10, p. R108, 2010.

9. O. Delaneau, C. Coulonges, and J.-F. Zagury, "Shape-it: new rapid and accurate algorithm for haplotype inference," *BMC bioinformatics*, vol. 9, no. 1, p. 540, 2008.

10. B. L. Browning and S. R. Browning, "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals," *The American Journal of Human Genetics*, vol. 84, no. 2, pp. 210–223, 2009.

11. L. Eronen, F. Geerts, and H. Toivonen, "Haplorec: efficient and accurate large-scale reconstruction of haplotypes," *BMC bioinformatics*, vol. 7, no. 1, p. 542, 2006.

12. J. E. Allen and S. Whelan, "Assessing the state of substitution models describing noncoding rna evolution," *Genome biology and evolution*, vol. 6, no. 1, pp. 65–75, 2014.

13. S. R. Browning and B. L. Browning, "High-resolution detection of identity by descent in unrelated individuals," *The American Journal of Human Genetics*, vol. 86, no. 4, pp. 526–539, 2010.

14. D. Aguiar and S. Istrail, "Haplotype assembly in polyploid genomes and identical by descent shared tracts," *Bioinformatics*, vol. 29, no. 13, pp. i352–i360, 2013.

15. V. Bansal and V. Bafna, "Hapcut: an efficient and accurate algorithm for the haplotype assembly problem," *Bioinformatics*, vol. 24, no. 16, pp. i153–i159, 2008.

16. V. Bansal, A. L. Halpern, N. Axelrod, and V. Bafna, "An mcmc algorithm for haplotype assembly from whole-genome sequence data," *Genome research*, vol. 18, no. 8, pp. 1336–1346, 2008.

17. D. Aguiar and S. Istrail, "Hapcompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data," *Journal of Computational Biology*, vol. 19, no. 6, pp. 577–590, 2012.

18. D. He, A. Choi, K. Pipatsrisawat, A. Darwiche, and E. Eskin, "Optimal algorithms for haplotype assembly from whole-genome sequence data," *Bioinformatics*, vol. 26, no. 12, pp. i183–i190, 2010.

19. E. Berger, D. Yorukoglu, J. Peng, and B. Berger, "Haptree: A novel bayesian framework for single individual polyplotyping using ngs data," *PLoS computational biology*, vol. 10, no. 3, p. e1003502, 2014.

20. A. Quinn, P. Juneja, and F. M. Jiggins, "Estimates of allele-specific expression in drosophila with a single genome sequence and rna-seq data," *Bioinformatics*, p. btu342, 2014.

21. S. Selvaraj, J. R. Dixon, V. Bansal, and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing," *Nature biotechnology*, 2013.

22. S. Selvaraj, J. R. Dixon, V. Bansal, and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing," *Nature biotechnology*, 2013.

23. . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

24. K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, *et al.*, "Encode data in the ucsc genome browser: year 5 update," *Nucleic acids research*, vol. 41, no. D1, pp. D56–D63, 2013.

25. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

26. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, *et al.*, "Gencode: the reference human genome annotation for the encode project," *Genome research*, vol. 22, no. 9, pp. 1760–1774, 2012.

27. H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.