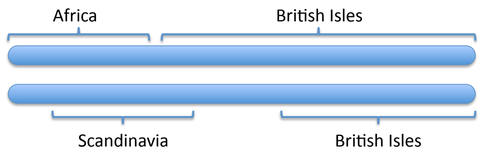
A scalable pipeline for local ancestry inference using thousands of reference individuals

C. B. Do[1,*], E. Y. Durand[1,*], J. M. Macpherson[1], B. Naughton[1], J. L. Mountain[1].

[1]23andMe, Inc, Mountain View, CA. *Contributed equally.

Summary

The ancestry deconvolution problem



Input: unphased genotype data for admixed individual
Output: ancestral origin of each chromosomal region

- Straightforward when populations are sufficiently distinct.
- Hard for closely-related populations (e.g., within Europe).
- Want efficient algorithm that scales to reference panels containing thousands of individuals.
- Want well-calibrated confidence estimates to avoid making predictions where algorithm is too uncertain.

Overview

We describe a **modular four-stage pipeline** for efficiently and accurately identifying the ancestral origin of chromosomal segments in admixed individuals:

1. Phase genotypes with population- or trio-based methods.
2. Assign ancestry labels to short, phased windows.
3. Correct phasing/labeling errors with probabilistic model.
4. Recalibrate confidence scores and assign final labels.

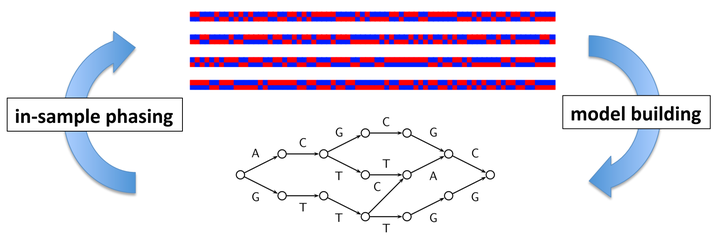
Methods

Stage 1: Phasing

Population-based phasing
Run BEAGLE++ on individual.

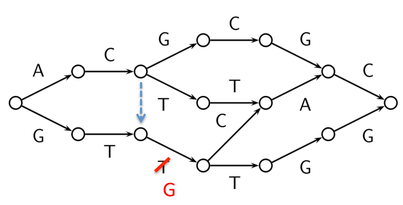
Trio-based phasing
1. Run BEAGLE++ on mom, dad, and individual.
2. Run TRIO++ to perform error-correction.

Stage 1a: Population-based phasing



Disadvantages of existing phasing algorithms
1. Phasing a few thousand individuals together takes several hours.
2. What if you want to phase a few new individuals? Reproducibility?

Out-of-sample phasing (BEAGLE++)



Key idea
Build a reference haplotype graph ONCE, and then reuse every time.

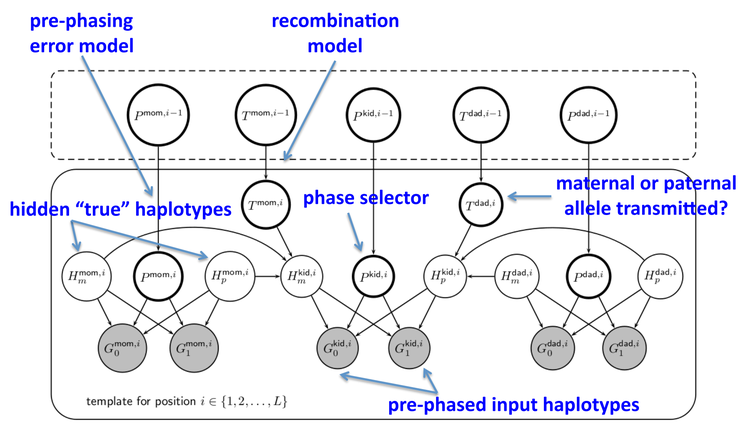
Our modifications
1. Allow recombinations
2. Allow genotyping error

Implementation details
1. Need really large haplotype graphs to ensure good generalization performance for out-of-sample phasing. So, we built **per-chromosome** haplotype graphs over ~500k markers across a dataset of **>100k individuals** derived from the customer database of 23andMe, Inc.
2. Standard dynamic programming ($O(LW^2)$) is too slow when haplotype graph is thick ($W > 10^3$). Use sparse dynamic programming heuristic in which partial paths with low probability are pruned early.
3. During graph construction, represent graph using compressed segments (i.e., sequences of consecutive edges with no incoming or outgoing edges to internal nodes), and store bit-packed genotype data separately.

Stage 1b: Trio phasing

Phasing is straightforward for close relatives (i.e., mother, father, child) except for doubly heterozygous sites. Want model that doesn't re-do work of population-based phaser.

Dynamic Bayesian network model (TRIO++)



Stage 2: Local classification

Divide haplotypes into 100-SNP windows, and classify window using SVM with a specialized string kernel.

Substring-based feature encoding

1	0	0	1	1	1	1	0
		0	0	0	✗		
		0	0	1	✗		
		0	1	0	✗		
		0	1	1	✓		
		1	0	0	✗		
		1	0	1	✗		
		1	1	0	✗		
		1	1	1	✗		

- Exponentially many possible patterns.
- Feature selection algorithms ineffective in practice – the more features the better (consistent with lots of weakly informative haplotypes; highly informative population-specific haplotypes are rare).
- Use **dynamic programming**-based algorithm to evaluate kernel efficiently.

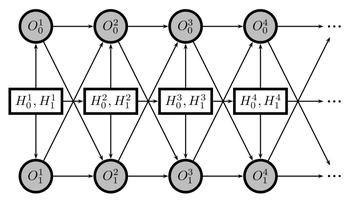
Window size should be large enough to give some amount of predictive power, but small enough to be of homogeneous ancestry while containing few phasing errors. In practice, phasing accuracy is often the limiting factor:

Double recombination or phasing error?

France	France	France	France	Italy	Italy	France	France	France	France
Italy	Italy	Italy	Italy	UK	UK	Italy	Italy	Italy	Italy

Are these predictions really independent?

Stage 3: Error correction



An **autoregressive hidden Markov model** that:
1. Identifies and corrects phasing switch errors.
2. Models first-order dependencies between consecutive SVM predictions within blocks of uniform ancestry.
3. Uses $O(LK^2)$ dynamic programming algorithm, where K is the number of populations (~20).
4. Produces per-window ancestry posterior probabilities.

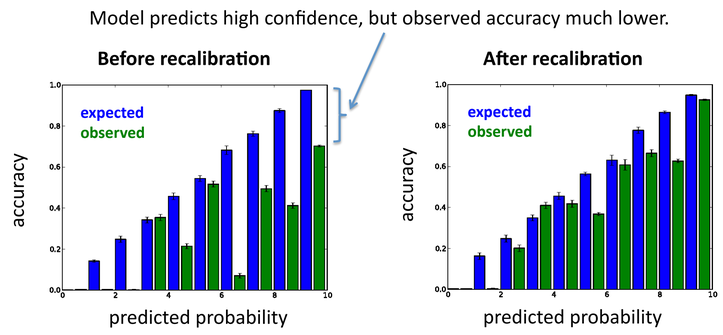
Parameter learning

- Emission parameters (i.e., P(O I H)): confusion matrix of SVM errors (estimated on a holdout set).
- Multiple transition parameter models (i.e., P(H)): unsupervised EM on 9 subsets of possibly admixed 23andMe customers (grouped using PCA/AIMs):

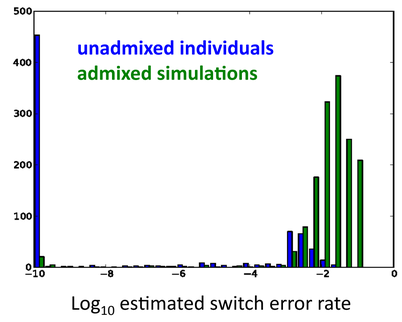
Combine via Bayesian model averaging

Stage 4: Recalibration and clustering

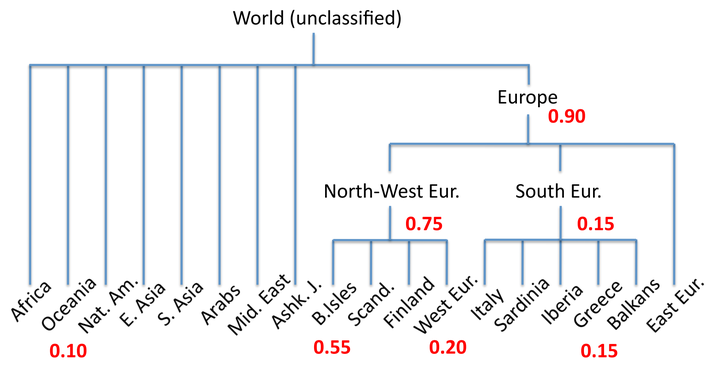
Apply a regularized variant of isotonic regression to perform recalibration of posterior probabilities to ensure that model predictions are not "overconfident."



Phasing errors greatly affect accuracy/calibration; split individuals into two groups based on estimated "effective switch error", and recalibrate each group separately.

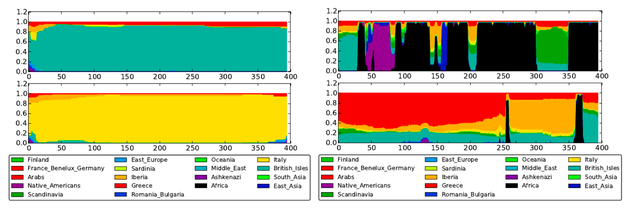


Then assign labels based on hierarchy in order to reach a desired level of confidence (e.g., 80% post-calibration).



Results

Examples



1st generation admixture: UK + Italy — African American

Preliminary benchmarks

	before error correction				after error correction				after recalibration	
	unadmixed		simulations		unadmixed		simulations		unadm	sims
	recall	prec	recall	prec	recall	prec	recall	prec	coverage	coverage
African	0.82	0.83	0.77	0.70	1.00	0.99	0.94	0.93	0.98	0.87
Native American	0.71	0.19	0.64	0.45	0.99	1.00	0.90	0.92	1.00	0.81
Ashkenazi	0.40	0.57	0.41	0.32	0.91	0.98	0.88	0.92	0.82	0.62
East Asian	0.57	0.68	0.53	0.36	1.00	0.99	0.88	0.86	0.99	0.79
Balkans	0.08	0.04	0.08	0.10	0.71	0.75	0.29	0.45	0.45	-
Eastern Europe	0.12	0.16	0.11	0.13	0.80	0.75	0.55	0.47	0.50	-
Middle East	0.10	0.08	0.10	0.10	0.89	0.78	0.55	0.47	0.79	-
British Isles	0.10	0.25	0.10	0.11	0.96	0.53	0.68	0.22	0.46	-
Scandinavia	0.12	0.12	0.12	0.13	0.60	0.62	0.35	0.51	0.28	-
Finland	0.31	0.15	0.30	0.24	0.75	0.90	0.66	0.82	0.74	0.31
Oceania	0.47	0.06	0.42	0.43	1.00	1.00	0.87	0.88	1.00	0.80
South Asian	0.23	0.34	0.22	0.24	0.97	1.00	0.74	0.78	0.98	0.04
Sardinia	0.21	0.01	0.20	0.12	0.39	0.45	0.16	0.64	0.26	0.03
Italy	0.06	0.14	0.06	0.09	0.93	0.73	0.57	0.21	0.68	-
Iberian	0.09	0.10	0.09	0.13	0.60	0.91	0.36	0.53	0.51	-
Greece	0.09	0.02	0.09	0.10	0.15	0.32	0.12	0.32	-	-
Arab	0.19	0.08	0.18	0.22	0.93	0.86	0.63	0.75	0.81	0.31
Western Europe	0.06	0.17	0.06	0.11	0.11	0.73	0.06	0.34	0.06	-

recall = proportion of windows from population X predicted as X
precision = proportion of windows predicted as X from population X
coverage = recall, requiring recalibrated confidence score > 0.80