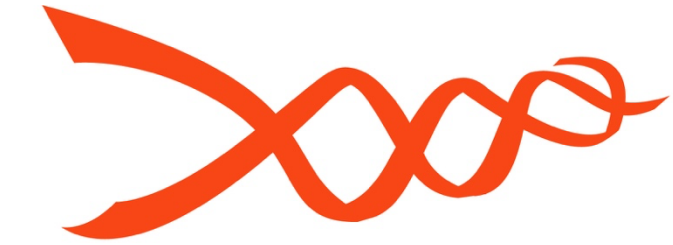

MYORIGINS 3.0

Combining Global and Local Methods for Determining Population Ancestry

White Paper 2021-08-18

Paul Maier • Rui Hu • Göran Runfeldt • Dunia Giniebra • Eric Frichot

Summary

FamilyTreeDNA team is excited to introduce MYORIGINS V3, our new tool for estimating population ancestry. Population ancestry is the proportion of DNA you have inherited from each ancestral population. Depending upon how much admixture occurred between your ancestors, you may have inherited DNA from one or perhaps many populations.

We have updated many aspects of our pipeline, including:

(1) An increase in number of reference populations from 24 to 90,

(2) Improvements in precision and accuracy using our newest methodological advancements,

(3) A chromosome painting:

- You may learn the chromosomal location of each population segment,
- This information may be genealogically valuable.

Contents

Glossary of Genetic and Analytical Terms

Accuracy – Ability to classify something correctly.

Admixture – Occurs when individuals of distinct population ancestries produce offspring, whose DNA is then a mosaic of ancestries (usually within a genealogical timeframe).

Ancestry-informative marker – Marker with large allele frequency differences between populations; thus, they may be informative about a person's population ancestry.

Allele – One of two or more variants of DNA sequence found at a genetic locus (e.g., 'T').

Autosomal – Describing all of the 22 pairs of chromosomes that exclude X, Y, and mitochondrion.

Base pair (bp) – Smallest length of DNA; one complementary pair of DNA bases (nucleotides).

Biallelic – A genetic marker (usually a SNP) possessing only two alleles in the population.

Bifurcating tree – Phylogenetic tree where every divergence contains exactly two daughter branches.

Centimorgan (cM) – A unit of distance along a chromosome. Between two chromosome positions that are spaced 100 cM apart, one recombination event is expected per generation.

Chromosome – One unbroken strand of DNA folded and condensed into the cell nucleus; humans have 23 pairs (one from each parent).

Chromosome painting – A depiction of an individual's ancestry showing the (super-)population of origin for each chromosomal segment.

Conditional Random Field (CRF) – Similar to an HMM but generalized for classification purposes.

Diploid – Refers to the pair of all chromosomes, maternal and paternal; haploid is half of a pair.

Deoxyribonucleic Acid (DNA) – The genetic blueprint of life and basis for inheritance; encoded by four nitrogenous bases (A, C, G, and T).

Ethnicity – Social or cultural group of people; used here to refer to a subgroup of a larger population.

Gene flow – Movement of individuals and their genetic material from one population to another continuously across a period of time; similar to admixture (which may be shorter in duration).

Genetic drift – Population change in allele frequencies over one or more generations that is due to offspring inheriting a random draw of parental alleles; exacerbated by small population size.

Genotype – An individual's genetic makeup from both maternal/paternal sides (e.g., 'T/G'); may refer to one or multiple genetic loci.

Haplotype – Ordered sequence of DNA along only one of the two chromosome copies (maternal or paternal; e.g., 'TAAGACTT').

Hidden Markov Model (HMM) – Statistical model used to predict a sequence of events or states by observing a closely related sequence; the first sequence is "hidden" while the second is "observed."

Hierarchical clustering – Statistical technique for grouping similar features together into a hierarchy.

Homozygous – A genotype with two identical alleles (e.g., 'T/T').

Heterozygous – A genotype with two different alleles (e.g., 'T/G').

Identical-by-descent – A shared segment descending from a common ancestor in genealogical time.

Identical-by-state – A shared segment that is identical but does not share a recent common ancestor.

Leave-one-out cross validation (LOOCV) – Technique for assessing accuracy of a model by removing each reference sample and predicting its result and then comparing to the true value.

Linkage disequilibrium – Correlation between SNP alleles that are physically close together.

Locus – Any defined location in the genome.

Machine learning – A method of artificial intelligence that can efficiently predict unknown values.

Marker – Any locus known to have genetic variation between individuals.

Megabase (Mb) – The physical distance along a chromosome; one million base pair positions.

Mutation – An error in DNA copying that results in a new allele transmitted to offspring; also, may refer to the new allele itself.

Natural selection – Population change in allele frequencies over one or more generations that is due to alleles conferring different probabilities of survival and/or reproductive success.

Panmixia – Completely random interbreeding; any two individuals might have offspring.

Phasing – Sorting out genotype data so that all maternal and paternal alleles are on the correct side. Although the DNA itself requires no phasing, the genotype array data are acquired SNP by SNP such that the original phase is unknown. The results of phasing are maternal and paternal haplotypes.

Phasing (statistical) – Phasing that utilizes a cohort of samples to ascertain which alleles statistically occur together the most; statistical phasing usually produces more switch errors than trio phasing.

Phasing (trio) – Phasing that utilizes samples from the mother and/or father of a subject; trio phasing usually produces very few switch errors except where all three samples are heterozygous.

Phase correction – A step used after phasing to reduce the severity of switch errors.

Pipeline – A workflow of computational steps.

Population – Group of individuals that has intermarried in isolation from other populations to such a degree that they are genetically distinguishable.

Population genetics – Study of genetic variation within and between populations and how it evolves via mutation, genetic drift, gene flow, natural selection, and recombination.

Principal Component Analysis (PCA) – Statistical technique for exploring the variation of a dataset in lower-dimensional space.

Recombination – The process of mixing and matching paternal/maternal haplotypes into new recombinant haplotypes; occurs while producing sperm or egg cells.

Reticulated tree – Phylogenetic tree where branches do not simply diverge, they also merge together.

Single Nucleotide Polymorphism (SNP) – A type of genetic marker composed of only one base pair position.

Specificity – Ability to classify something into a precise group or subgroup.

Statistical noise – Random variation in some data that cannot be explained by known variables.

Switch error – Incorrect phasing from one heterozygous SNP to the next heterozygous SNP.

Triallelic – A genetic marker (usually a SNP) possessing three alleles in the population.

Typological – Categories that are static, unchanging, and unmixed.

Viterbi algorithm – An algorithm used to estimate the most likely hidden sequence for HMMs.

Overview

FamilyTreeDNA is dedicated to providing customers the most useful genealogical information grounded in the best scientific framework available. The science of population genetics and ancient DNA is continually reshaping our understanding of the human story. An explosion of genomic datasets, methodological advances, and increased population sampling have given us an unprecedented toolset for unraveling our history. One of the major epiphanies of the last decade has been that human populations are never typological; each one is itself a mixture of previous ancestral populations [1]. For example, Amerindians are the mixture of Ancient North Eurasians and East Asians some 20–25 thousand years ago [2]. Similarly, modern Europeans are the complex mixture of three prehistoric populations from the Paleolithic, Neolithic, and Bronze Age [3]. Over the past few thousand years, with the development of new technologies and cultures, human population structure has become even more mosaic. We are proud to announce our new MYORIGINS v3 feature, which offers an unparalleled snapshot of our customers' pre-Columbian population ancestry.

Before describing the goals and achievements of MYORIGINS v3, first we must distinguish three types of ancestry analyses: (1) ancient, (2) pre-Columbian, and (3) recent (Fig. 1). ANCIENTORIGINS traces back your ancestry from pre-historic or archaic populations, roughly corresponding to the period before the Common Era (>200,000–2,000 years ago). Although we currently test for ancestry from three European pre-historical populations, we plan to expand this soon. MYORIGINS is designed to estimate ancestry proportions from highly distinct populations that existed prior to major continental travel (roughly 2,000–500 years ago). For example,

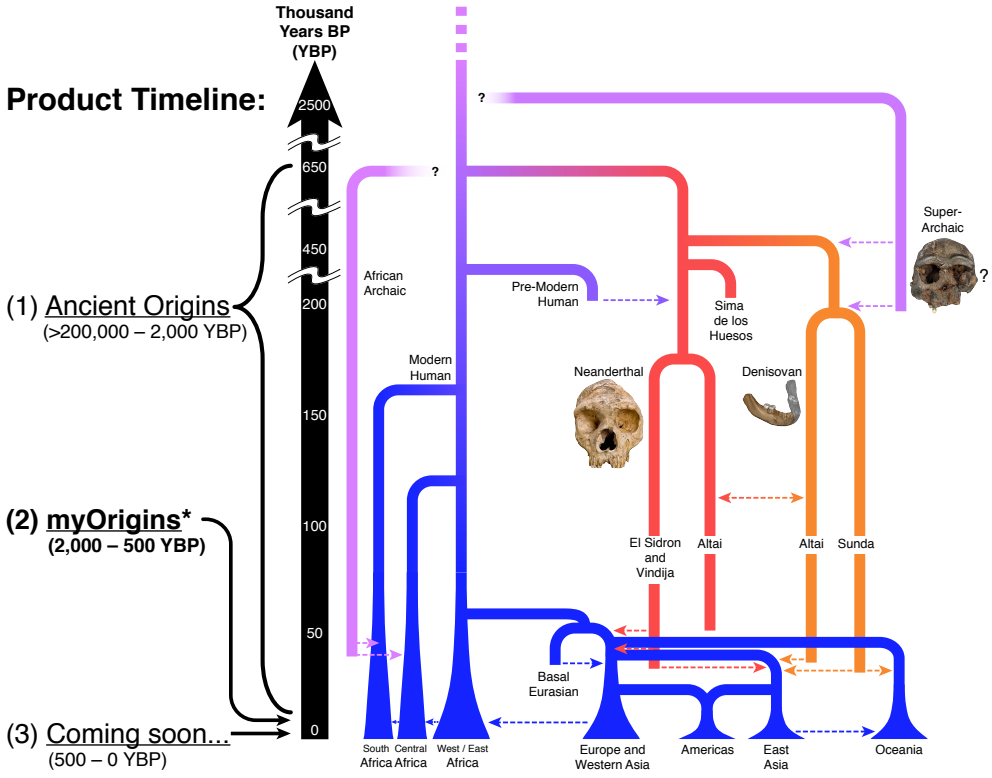

Figure 1. The timeline of three types of ancestry product. Updated and modified with permission based on [4].

Latino origins would include both European and Amerindian components, rather than a single admixed population (e.g., "Ecuadorian"). Finally, another type of product is ideal for populations such as "Ecuadorian" that are too recent, geographically specific, or admixed for MYORIGINS. Unlike MYORIGINS, the methodology for such a product cannot estimate a percentage or proportion, only a match strength (e.g., low, medium, or high). Although we do not currently provide analysis for more recently formed populations, we plan to release such a feature soon.

Ancestry analyses such as MYORIGINS are designed to estimate proportions of DNA that were inherited from ancestral populations. However, such tests require genetic distinguishability between populations to exist. A long time period of isolation is required—whether via geographic or ethnocultural barriers—for ancestry-informative markers to emerge (Fig. 2A). This is because the processes for generating novel genetic variation are a very slowly ticking clock, and they only tick once each generation (25–30 years). These processes include mutation, genetic drift, and recombination. For example, approximately 14% of SNP markers in the human genome are ancestry-informative* on a global scale but fewer than 1% are ancestry-informative in Europe. Thus, continent-level population structure is much easier to detect than sub-continental or ethnic structure.

Additionally, ancestry-informative markers accumulate in small genomic islands (Fig. 2B). In other words, while populations are diverging from one into many, most of their DNA sequences remain statistically indistinguishable except for a few isolated places randomly scattered across the genome [5–11]. Over time these genomic islands grow larger and eventually would include the entire genome if given sufficient time (many 100,000s of years). This means that if you randomly select any segment of autosomal DNA, you are very likely to know its continent of origin but much less likely to know its specific population of origin. Hence, for any set of populations—e.g., Iberian, Russian, and British—a larger number of SNP markers increases the resolution to distinguish them, because more of these genomic islands are sampled (Fig. 2C).

Many methods exist for estimating admixture proportions. So-called "local" methods work by breaking up the genome into small segments and assigning each one to a reference population. Then, the admixture proportions can be calculated by simply aggregating the segments for each group. Many local ancestry methods have been published to date often utilizing Hidden Markov Models (HMMs) or similar graphical models, sometimes in conjunction with machine learning approaches [12–36]. The major benefit of local methods is their ability to identify each segment of each chromosome individually (i.e., make a chromosome painting). Given that random recombination breaks apart maternal and paternal haplotypes each generation, there is a usefulness to knowing how our population proportions are distributed across our DNA. For example, combining a Family Finder match with a known population for that segment of DNA may help narrow down the genealogical common ancestor.

However, local methods are by definition limited by a small number of genetic markers. This means very closely related populations cannot be accurately distinguished for the reasons explained above (Fig. 2). By contrast, "global" methods estimate ancestral populations for the

*We define ancestry-informative here as having a value of Weir and Cockerham's $F_{ST} \geq 0.15$.

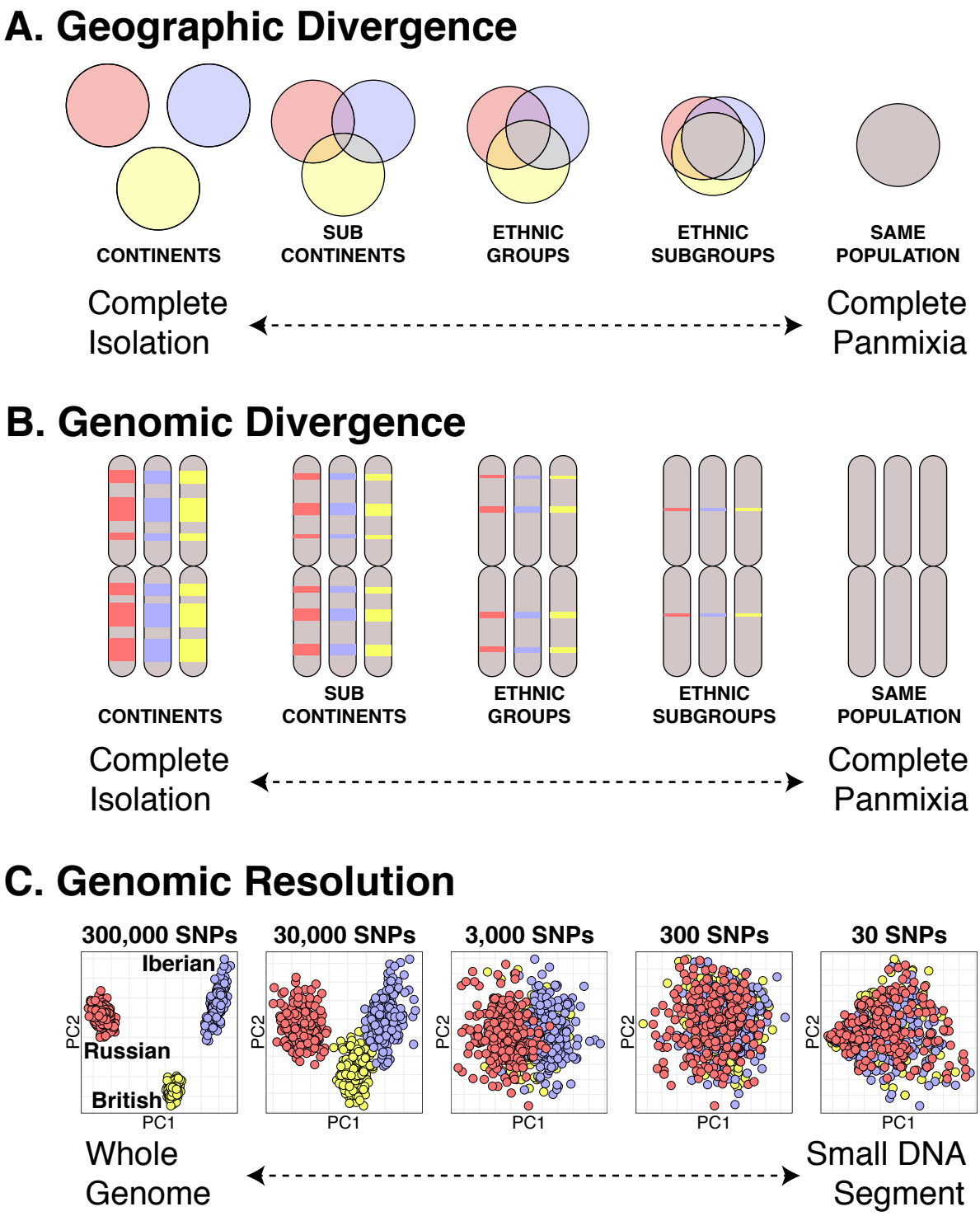

Figure 2. Populations are only distinguishable if: (A) they are isolated for a sufficient period of time, (B) enough genetic markers have diverged during that time, and (C) enough divergent markers are sampled. (A) Populations are often somewhere between the extremes of complete isolation and complete panmixia (free interbreeding). (B) As populations become isolated for longer periods of time, small "islands" of DNA become highly divergent. Over time the DNA islands grow in size. (C) Thus, populations can only be distinguished if these islands of DNA are sampled. Taking three closely related populations—Iberian, Russian, and British—we can see that a moderate number of SNP markers across the whole genome is needed to distinguish them fully. The 300,000 SNPs in this case are randomly chosen markers, free of linkage disequilibrium.

entire genome simultaneously [37–52]. This gives global methods high resolution (many SNPs) to estimate admixture proportions. Instead of classifying each segment of DNA into one of the *N* reference populations, the entire genome is modeled as a vector of *N* proportions. SNP locations are not used, because each SNP is treated as a statistical sample of a genome-wide process of admixture. Typically, global methods work by assuming each population consists of people randomly interbreeding—this assumption vastly simplifies the math. A person's genotype is then simply a random draw of SNP alleles from a frequency distribution in each population. The major benefit of global methods is their higher accuracy in resolving ancestry from closely related populations by using genome-wide SNPs (Fig. 2C). The drawbacks include an inability to identify which DNA segments comprise which proportions (i.e., no chromosome painting). Also, if the assumption of random interbreeding is unrealistic, results may suffer.

MYORIGINS v3 combines dual strengths of global and local ancestry methods to improve results. Our new pipeline has three main steps. (1) A global ancestry method with high computational efficiency narrows down the list of possible populations and estimates proportions for very closely related populations (e.g., West Slavic vs. East Slavic). The computational efficiency of this step allows us to include an unprecedented 90 populations. (2) Each pair of chromosomes (maternal and paternal) is phased and broken into small segments. (3) A local ancestry method classifies each DNA segment into "super-populations" (e.g., Western Europe vs. Eastern Europe). Super-populations are used for this last step, because this is the genetic distinctness required to accurately classify small DNA segments. MYORIGINS v3 results include 90 population proportions along with a chromosome painting, which can be used in conjunction with Family Finder match results to identify common ancestors for genealogical work.

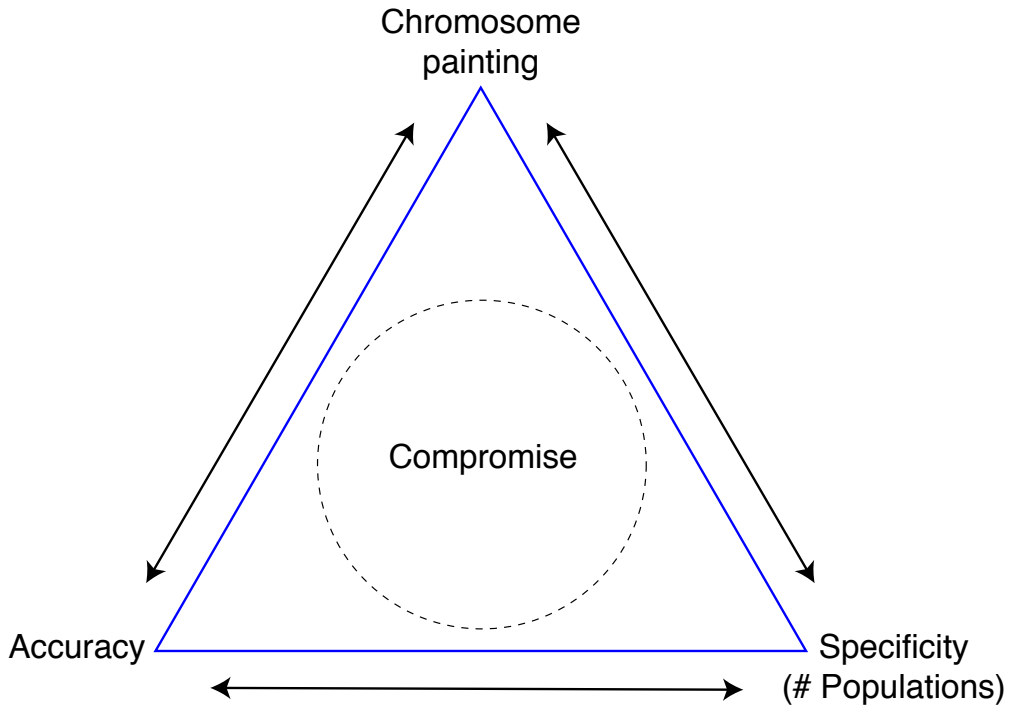

Figure 3. Tradeoffs. When designing a population ancestry analysis, tradeoffs exist between (1) accuracy, (2) specificity via number of populations, and (3) whether or not a chromosome-specific ancestry is estimated.

We had four main goals while improving MYORIGINS:

(1) Increase accuracy,
(2) Increase specificity,
(3) Increase the number of populations,
(4) Estimate a chromosome painting (i.e., use local ancestry methodology).

There are important tradeoffs between these four goals (Fig. 3). Adding more specific populations such as "Ireland" and "Great Britain" instead of simply "British Isles" can reduce the accuracy of both. This is because the average gene flow over 2,000 years has been substantial following founding by closely related populations such as Romano-British, Picts, Gaels, Normans, and Anglo-Saxons. However, we mitigated this by only choosing populations that could be estimated with an acceptable level of accuracy. Similarly, increasing the number of populations from 24 to 90 can reduce accuracy considerably if the new population boundaries are closer together than before. We weighed the potential gain of each new population against the lost accuracy or specificity. Finally, a chromosome painting would potentially reduce accuracy if we estimated local ancestry of DNA segments for populations. However, if we only estimated continent-level segments, this would reduce specificity. Hence, we compromised by painting at the intermediate level of super-populations.

With the exciting introduction of MYORIGINS v3, we provide a new map that is very representative of all human diversity on Earth (Fig. 4) and vastly increases the number of populations offered on each continent (Table 1). We believe it will give our customers an indispensable toolkit for understanding their ancestry, conducting genealogy, and putting their origins into a larger perspective about human origins.

Table 1. Comparison of population number within each continental region.

Continent	MYORIGINS V2 Populations	MYORIGINS V3 Populations
Africa	4	21
Europe & Middle East	12	27
Asia & Oceania	6	33
Americas	2	9
Total	**24**	**90**

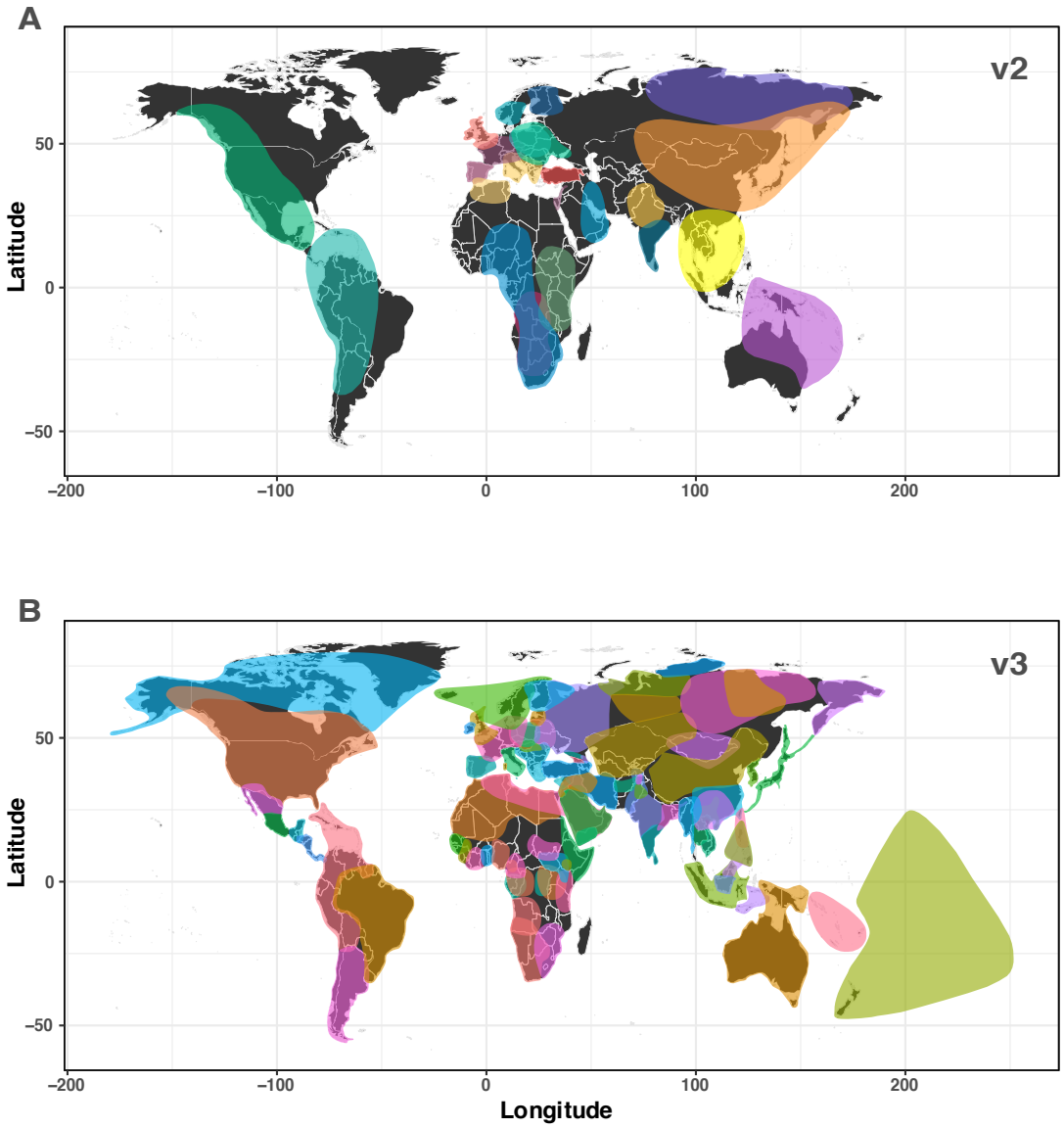

Figure 4. Map showing the geographical extent of each ancestral population ("Origin"). Compared to MYORIGINS v2 (A), MYORIGINS v3 (B) is much more representative of global genetic diversity both in terms of geographical representation and number of populations (V2: 24 populations; V3: 90 populations).

What is a Population?

Before delineating the boundaries of populations, it is a good idea to first define what we mean by "population." The evolutionary-genetic definition [53] has the following criteria:

- A population must be cohesive within,
- A population must be distinct from other populations.

Putting that into more formal population genetic language:

- A population is a group of individuals that is panmictic (randomly interbreeding),
- A population has sufficiently low gene flow with other populations (typically less than one migrant entering the population per generation, averaged over thousands of years [54]).

In reality, groups of individuals have complex histories of isolation, movement, marriage patterns, and demographic changes. This can make the boundaries of closely related populations very fuzzy as discussed above (Fig. 2).

Population boundaries are also fuzzy because everyone descends from many locations on Earth [55]. This seems extremely counterintuitive but is easy to mathematically prove. Imagine a man who is 100% Scandinavian living today in Sweden. His genealogical ancestors double each generation back in time. One thousand years ago (roughly 33 generations ago), he had $2^{33} = 8.5$ billion genealogical ancestors. However, in the year 1000 C.E., Europe only had a population of ~50 million. Some simple math* shows that everyone who was alive in Europe around the year 1000 C.E. is an ancestor of everyone in Europe today, or of no one. Therefore, the Swedish man has the same set of ancestors as a man whose family is 100% Iberian. If we all have the same ancestors, then where do our genetic differences come from? To understand this, you need to appreciate two facts:

(1) Our genealogical ancestors are related to us on multiple different lines,
(2) Our genetic ancestors are a random sample of our genealogical ones (Fig. 5).

Although two Europeans have almost identical sets of genealogical ancestors from 1000 C.E., they are related to those ancestors along different lines [56]. The Swede and Iberian may share millions of genealogical ties. However, the Swede has perhaps 1,000 ties to a Scandinavian ancestor, and the Iberian man has perhaps only 10 ties to that same Scandinavian ancestor. Thanks to the randomness of genetic recombination, the Swede is 100× more likely to inherit Scandinavian DNA than Iberian DNA. The bottom line: your DNA descends from a random subset of your ancestors, but your DNA composition tends to reflect the ancestors who were closest to you in geographic space (Fig. 5). Populations are groups of statistically similar DNA—they are not simple categories.

*The "Identical Ancestors Point" (IAP) was $1.77 \times \log_2(\text{Pop. Size})$ generations ago or 1,350 years ago in Europe. The "Time to Most Recent Common Ancestor (TMRCA) was $\log_2(\text{Pop. Size})$ generations ago or just 775 years ago in Europe. This assumes panmixia.

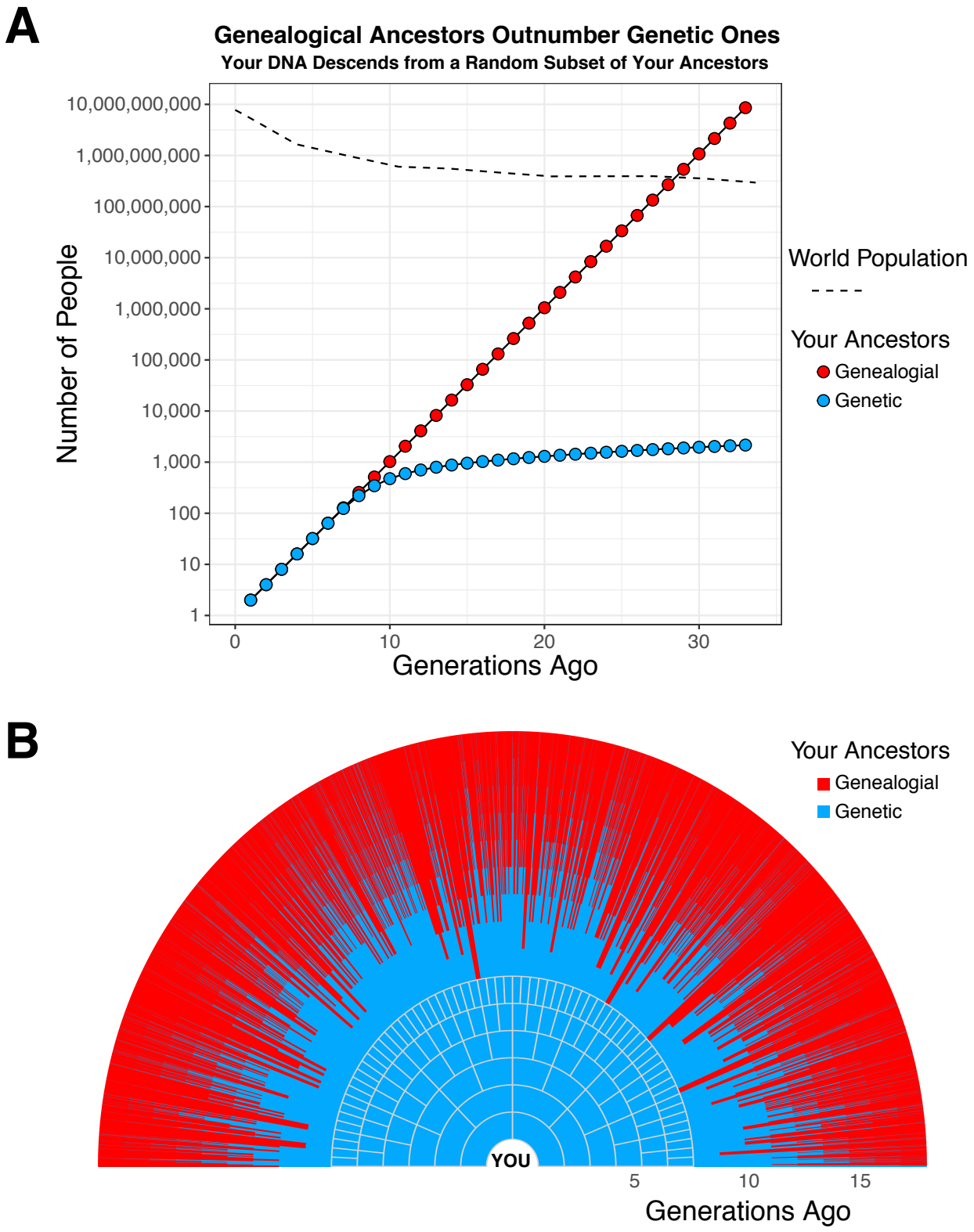

Figure 5. (A) You have an exponentially increasing number of genealogical ancestors, but a much smaller number of genetic ones. Your genealogical ancestors outnumber the world population less than 1,000 years ago. This is because most of your ancestors are duplicated in your family tree. (B) Most of your ancestors from 15 generations ago contributed no DNA to you, due to random genetic recombination, and finite space in the genome.

Reference Panel

In order to infer ancestry from ancestral populations, first a set of reference populations must be constructed as proxies for those ancestral populations. The first step is collecting samples from populations that are potentially suitable (i.e., distinct enough with adequate sample size). The next step is discovering population structure: which populations are actually distinct, which samples are unadmixed, and which reference populations have enough samples after screening. Finally, we need to hierarchically group the new reference population set into super-populations so that global and local ancestry methods can be seamlessly combined (see Overview).

Data Sources

We derived our MYORIGINS V3 reference samples from a combination of sources:

- FamilyTreeDNA internal data and private collections,
- Publicly available databases from international consortia of researchers, including the 1000 Genomes Project [57] and the Human Genome Diversity Project [58,59]
- Other publicly available data.

In order to include a sample, the data needed to be compatible with our Gene by Gene BeadChip array. This means we only considered data produced by other Illumina SNP arrays with a large percentage of intersecting markers (i.e., >95%) or whole-genome sequencing data with sufficient sequencing depth (i.e., mean 20×).

We also derived our 100K phasing panel (see 'Phasing') from FamilyTreeDNA private collections. We selected a balanced sample of approximately 100,000 individuals with population ancestry spanning all of human diversity. This ensures that the phasing panel contains haplotypes that match any potential customer sample.

Only biallelic SNP markers were used (i.e., those with only two alleles) for simplicity. Triallelic SNPs can cause problems for data produced by different technologies—sequencing may recover the true genotype, while SNP arrays may only consider two out of three alleles. We used a minor allele frequency (MAF) cutoff of ≥0.001, ensuring that the genetic diversity in our panel is found widely enough to be considered real and not an artifact of any technology. Across all MYORIGINS v3 reference samples, the MAF was 0.24 ± 0.14, and the genotyping rate across all samples and SNPs was 0.99, for a total of 637,645 SNPs.

Finding Population Structure

We used Principal Component Analysis (PCA) to screen potential samples for existing population structure. PCA is a type of linear model and thus assumes each SNP is uncorrelated with the others. However, many SNPs are densely packed together in haplotypes and therefore are correlated with one another. This is known as linkage disequilibrium. Before conducting PCA, we used the software PLINK [60] to prune any SNPs with squared correlation (R^2) ≥0.7 if they occurred within one megabase of each other. This left a total of 379,880 uncorrelated SNPs for PCA. Samples with close kinship (i.e., first cousin relationships or closer) are another type of genetic correlation that we removed from the dataset, using the software KING [61].

We used various metadata to inform which samples should be included in the PCA. Wherever possible, we used family trees going back 2–6 generations to corroborate the ancestral location of each potential reference. In other cases, ancestry survey responses were used to determine the four grandparents' ethnicities and birth locations. When this information was unavailable, we relied upon expert opinion or previous MYORIGINS results.

Fig. 6 exemplifies the before and after of reference selection using PCA. Several European populations are distinct enough (e.g., Finnish, Sardinian) that admixed samples become obvious and pruning them into good references is easy. Distinct populations tend to form their own isolated cluster along the axes of the PCA biplot, because the distance between points is related to the time to common ancestor [62]. However, many European countries show a pattern of isolation-by-distance, whereby samples are not grouped by population but rather spread across a two-dimensional gradient. For example, northern and southern Germany are as distinct as southern Germany is from central France. This makes reference selection more challenging, because there are multiple ways the boundaries can be drawn between populations.

We used a combination of PCA and the global ancestry software ADMIXTURE [39] to select potential references, draw putative boundaries around populations, and iteratively test the efficacy of those samples and boundaries (Fig. 6). We used five-fold cross validation on supervised ADMIXTURE as our preliminary test for accuracy. In some cases, clusters looked distinct in PCA space (e.g., France and Germany), but ADMIXTURE showed accuracy to be poor unless they were combined (e.g., Central Europe).

After we selected 8,053 references from our 90 MYORIGINS v3 populations (Table 2), we needed to hierarchically group them into more inclusive super-populations for chromosome painting (see Overview). We used several methods to generate a putative population tree of human life: TreeMix [63], Speedymix (e.g., Appendix A), hierarchical clustering on pairwise F_{ST}, and scientific literature [63–68]. The super-population groupings are shown in Fig. 7. It is important to note: numerous studies [69–73] have shown that human population history is reticulated—not bifurcating—however, we use a bifurcating tree for simplicity. For example, our population tree depicts Polynesians as a bifurcation from other East Asians; however, in reality, Polynesians share dual ancestry [74] from Island Southeast Asians (70%) and Melanesia/New Guinea (30%).

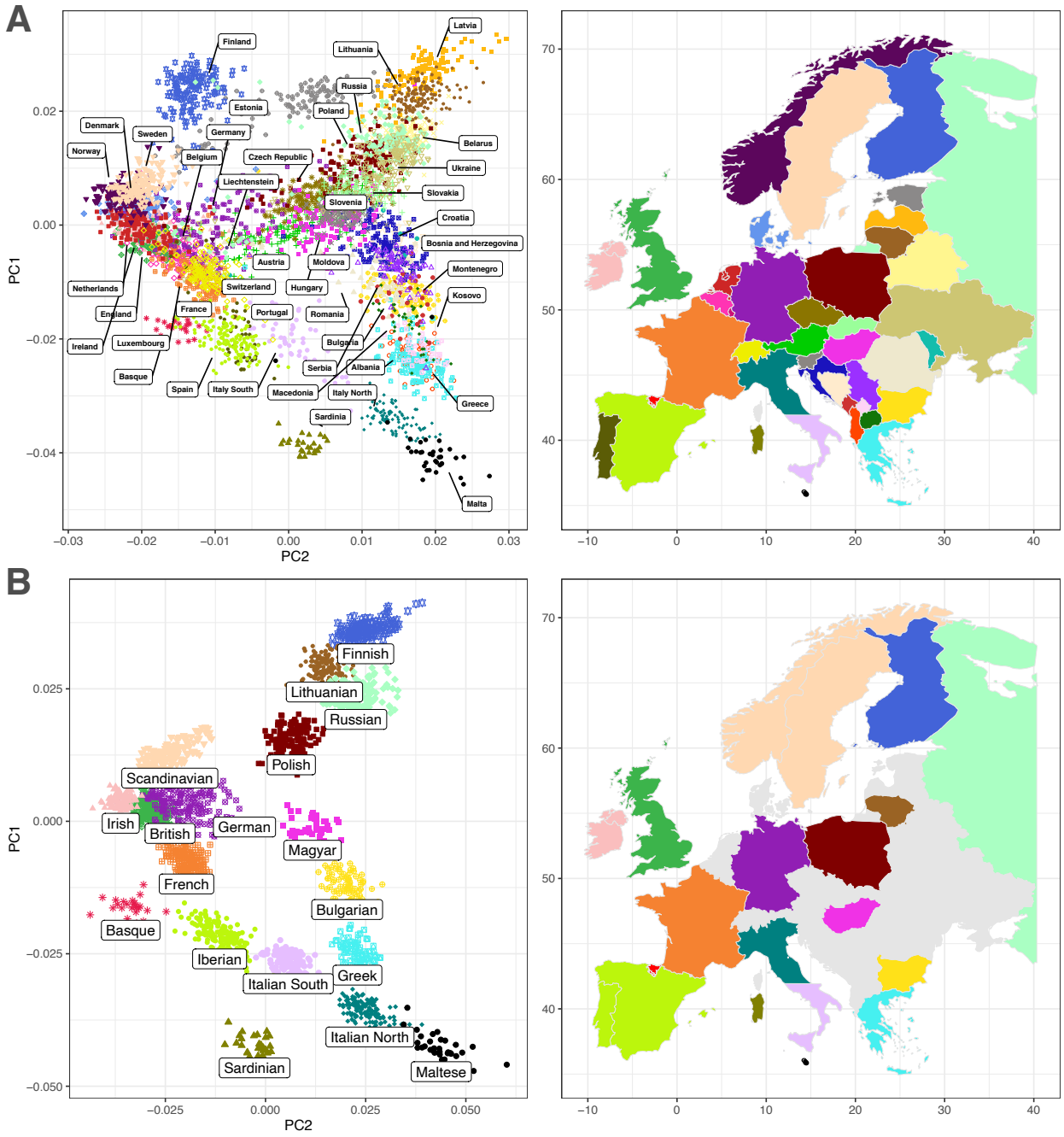

Figure 6. Reference selection using Principal Component Analysis (PCA); Europe is shown here as an example. (A) Samples with both parents originating from each country in Europe are initially chosen as potential references. The extreme level of genetic overlap between neighboring countries, sometimes called "isolation-by-distance," is apparent. (B) After several analyses are done to decide which populations are sufficiently distinct (see text), reference samples are selected as proxies for those ancestral populations. Example for illustrative purposes only.

Table 2. Population names and sample sizes for MYORIGINS v3. For super-populations, see Fig. 7.

Population	Sample Count
San Forager	42
African Rainforest Forager (East)	32
African Rainforest Forager (West)	76
African Rainforest Forager (North)	14
Senegal, Gambia & Guinea-Bissau	113
Guinea & Sierra Leone	85
Liberia & Ivory Coast	34
Ghana, Togo & Benin	32
Nigeria	123
Northern Congo Basin	121
Atlantic Equatorial Africa	200
Southern Congo Basin	41
Southern Africa	98
Western Lake Victoria Basin	56
Eastern Lake Victoria Basin	108
East African Savannah	43
Nile River Basin	16
Eritrea, Northern Ethiopia & Somalia	120
Southern Ethiopia	35
Maghreb & Egypt	85
Bedouin	18
Southern Levant	114
Druze	27
Arabian Peninsula	70
Yemenite Jewish	116
Northern Levant	85
Mesopotamia, Armenia & Anatolia	279
Sephardic Jewish	53
Mizrahi Jewish	41
Ashkenazi Jewish	246
Southern Caucasus	50
Northern Caucasus	44
Eastern Caucasus	103
Basque	24
Malta	33
Sardinia	27
Italian Peninsula	329
Greece & Balkans	160
Iberian Peninsula	256
Magyar	43
West Slavic	146
East Slavic	219
Central Europe	690
England, Wales & Scotland	364
Ireland	104
Scandinavia	156
Baltic	161
Finland	235
Indus Valley	62
Afghanistan & Northern Pakistan	25
Western India	46
Northern India	30
Southern India	108
Eastern India	80
Mongolia	164
Southern Siberia	36
Kalash	24
Northwestern Siberia	26
Western Siberian Plains	75
Central & Eastern Siberia	30
Taimyr Peninsula	13
Yakut	19
Northeastern Siberia	26
Inuit	27
Amerindian – North America	30
Amerindian – North Mexico	14
Amerindian – Yucatan Peninsula	12
Amerindian – Central & South Mexico	11
Amerindian – Central America	48
Amerindian – Andes & Caribbean	57
Amerindian – Argentina & Chile	19
Amerindian – Amazon	19
Japan	178
Korean Peninsula	248
Northern Han	63
Southern Han	83
Thailand and Southern China	91
Laos, Vietnam & Cambodia	111
Yao	11
Myanmar	20
Malaysia & Western Indonesia	72
Northern Borneo	20
Southern Borneo	125
South Wallacea Islands	182
Philippine Lowlands	81
Philippine Austronesian	19
Philippine Melanesian	11
Polynesia	31
Melanesia	9
Sahul	30
Total	**8053**

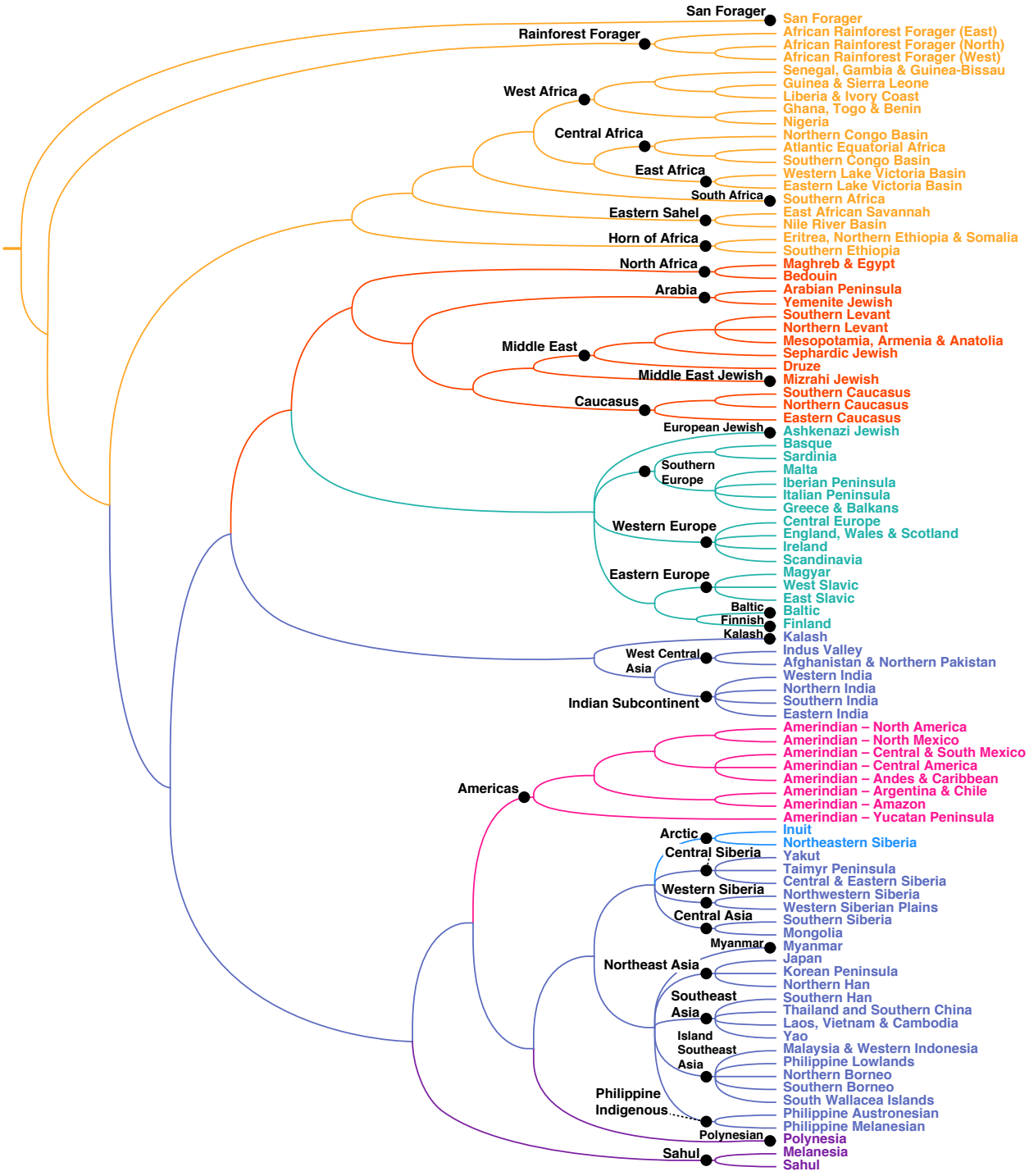

Figure 7. Population tree for 90 MYORIGINS v3 populations from a consensus of analyses such as TreeMix, Speedymix, hierarchical clustering on pairwise F_{ST}, and academic literature. The 34 super-population groupings in MYORIGINS v3 are shown. Branch lengths have no meaning in this cladogram. Note: the true population tree of humanity is reticulated due to admixture; however, a bifurcating tree is used here for simplicity. **Therefore, this tree is simply a clustering tool for organizing populations into super-populations and cannot accurately reflect the complex multi-population admixtures that occurred in deeper human ancestry.**

Overview of MYORIGINS v3 Pipeline

The MYORIGINS v3 pipeline integrates the population specificity of global ancestry inference with the genomic specificity of local ancestry inference as discussed above (see Overview). This means we can accurately estimate the proportion of DNA a customer inherited from very specific populations but only if genome-wide SNPs are deployed. Therefore, we do not attempt to "paint" chromosome segments at this level of population specificity. Broader and older groups—super-populations—contain a higher density of SNP and haplotype frequency differences, allowing us to estimate a chromosome painting at this level in the population hierarchy.

Using dual global and local estimators has the additional benefit of combining multiple checks. Global methods are limited by model assumptions such as the independence of SNP markers, and therefore, undiscovered correlation between markers can slightly bias results. In contrast, local methods make no such assumption, and in fact perform best with densely correlated SNPs (i.e., haplotypes). We therefore expect a slight increase in accuracy by normalizing a customer's global ancestry proportions based on their local ancestry proportions.

Another advantage of our dual estimators is our ability to screen a customer's populations. Global ancestry results include a list of irrelevant populations, i.e., those with zero proportion. This is advantageous because local methods make noisier predictions than global methods. There is a very limited number of SNPs residing in local DNA segments, and this can cause misclassifications. Hence, we reduce this greatly by only selecting relevant reference panels in our local ancestry analysis.

Global and local ancestry methods in the MYORIGINS v3 pipeline are thus mutually reinforcing, and our workflow leverages this principle (Fig. 8). Briefly, the steps are:

- Global ancestry inference
  (1) A customer's sample is combined with our reference panel at 379,880 SNPs.
  (2) Speedymix calculates ancestral proportions from 90 populations worldwide.
  (3) A list of relevant reference populations is selected for the next step.

- Local ancestry inference
  (1) We phase the customer's unphased genotype of 637,645 densely packed SNPs.
  (2) Breaking up each chromosome into small windows, segments are classified into relevant super-populations (out of 34 total).
  (3) A conditional random field smooths over misclassifications.
  (4) We correct phasing errors using a unique hidden Markov model.

- Global-local ancestry integration
  (1) Globally estimated population proportions are normalized into locally estimated super-population proportions.
  (2) Chromosomes are sorted so that a chromosome painting may be displayed.

In the following sections, we expand on these steps in more detail.

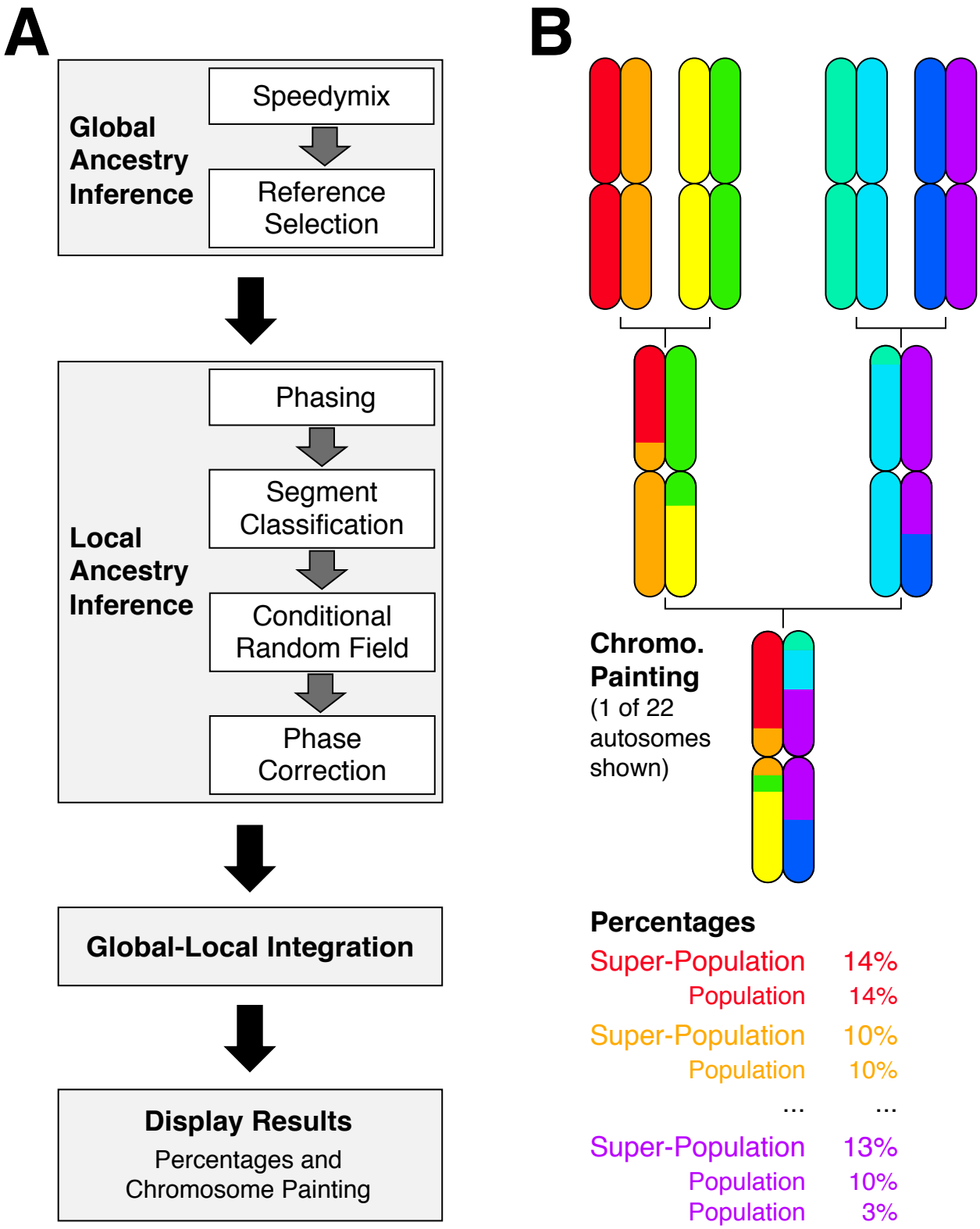

Figure 8. (A) Workflow for MYORIGINS v3. See text for details. (B) Hypothetical results for a customer whose eight great-grandparents are from different super-populations. In this example, the last great-grandparent (purple color) comes from two different populations. Only 1 of 22 autosomes are shown for simplicity.

Global Ancestry – Speedymix

We estimate global (genome-wide) ancestry for each customer using the references and SNP set described above (see Finding Population Structure). Our methodology is similar to sNMF [41] but with some important modifications—we call our software package Speedymix. The basic idea of Speedymix (Fig. 9) is that each SNP genotype in a customer's data (**X**) exists with a probability equal to the fraction of his/her genome that came from each population (**Q**), multiplied by the frequency of that genotype in each ancestral population (**G**).

$$\Pr(\mathbf{X}) = \mathbf{QG} \tag{1}$$

Or more formally: the probability that individual i possesses j derived alleles at locus l is

$$p_{il}(j) = \sum_{k=1}^{K} q_{ik} \times g_{kl}(j), \qquad j \in \{0, 1, 2\} \tag{2}$$

where q_{ik} is the fraction of individual i's genome from population k, and g_{kl} is the frequency of that genotype (either 0, 1, or 2 derived alleles; i.e., 0/0, 0/1, or 1/1) at locus l in population k.

This brings us to the goal: estimate a customer's ancestry proportions (**Q**). Based on Equation 1, we seek values of **QG** that minimize the difference between the most probable data Pr(**X**) and the actual data (**X**). This is accomplished using least-squares methods that minimize the value

$$\mathrm{LS}(\mathbf{Q}, \mathbf{G}) = \left|\left|\mathbf{X} - \mathbf{QG}\right|\right|_{\mathrm{F}}^{2} \tag{3}$$

where $\left|\left|\mathbf{M}\right|\right|_F$ is the Frobenius norm of a matrix **M**. We use the Alternating Least Squares (ALS) algorithm of nonnegative matrix factorization [75] to minimize the least-squares criterion in Equation 3. We select initial values for the **Q** matrix, then iteratively estimate least-squares values of **G**, followed by **Q**, followed by **G**, etc. After each cycle values of **G** and **Q** are normalized so all proportions and frequencies sum to 1.0. We also force all reference samples to ancestry proportions of 1.0 in their respective populations, making the analysis semi-supervised. The ALS algorithm alternates these cycles until it converges. This is determined by values remaining stationary: change in ancestry proportions (<0.01) and least-squares criterion (<10^{-6}).

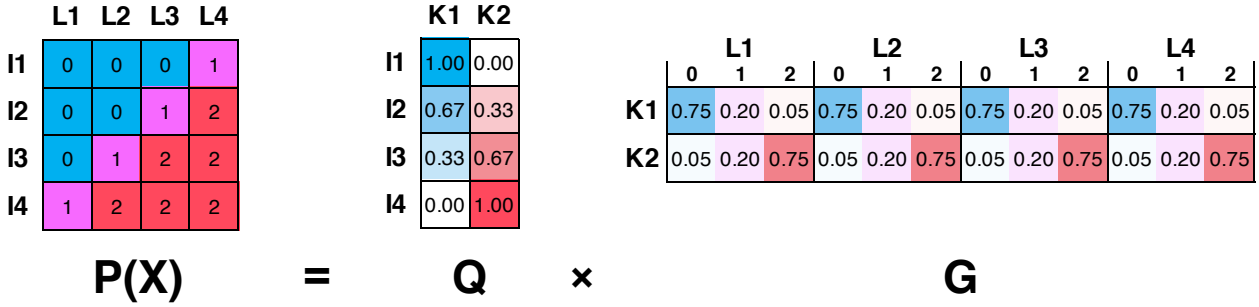

Figure 9. An example Speedymix calculation. Global ancestry proportions **Q** multiplied by genotype frequencies **G** of each population yields a probability of a customer's data $\boldsymbol{Pr}(\mathbf{X})$. Minimizing the difference between **Pr**(**X**) and **X** gives the best **Q** and **G**. This example shows four testing individuals (I), four loci (L), and two populations (K).

Local Ancestry

Phasing

Humans are diploid (2N), meaning they inherit two genomes: one maternal and one paternal. Each haploid (1N) chromosome differs from its pair but only at a small number of positions. These heterozygous positions are genotyped individually, e.g., A/T, C/A, G/C, without any knowledge of the overall sequence that each parent contributed, e.g., AAG, TCC. The order of alleles along each haploid copy of chromosomes is known as the "phase." Phasing—or determining the correct sequence of [ACGT] along haploid chromosomes—is essential for conducting local ancestry analysis. Population origins of a DNA sequence cannot be determined if the sequence is an incoherent blend of two chromosomes.

The first step in local ancestry analysis is phasing each autosomal chromosome pair. We utilize a previously described software package called Eagle [76], which has been shown to outperform comparable programs such as Beagle and SHAPEIT, particularly when using large reference panels [77]. A large panel of unrelated individuals is essential for accurate phasing, and we use our 100K phasing panel. Eagle combines two different techniques for phasing: (1) searching a reference panel for short haplotypes that are exact matches (i.e., distant cousins that share DNA identical-by-descent (IBD) from a common ancestor); (2) modeling haplotype frequency in the panel to calculate the probability of each possible haplotype (Fig. 10). Eagle combines these two techniques for increased computational speed and accuracy.

The Eagle algorithm contains three steps to phase a chromosome:

(1) It scans the reference panel for >4 cM matching segments, i.e., those matching at least one allele at every SNP. Potential matches are scored according to their likelihood of being close or distant relatives (and not matching due to chance). All matches above a threshold likelihood score are then pruned to remove any matches that are inconsistent with other matches. Finally, phase is assigned to a customer's genotype using the IBD matches. For every SNP that is heterozygous in the customer, a homozygous match is used to determine phase. If the SNP is heterozygous in all matches, then allele frequency is used to phase the SNP probabilistically.

(2) It splits the chromosome into overlapping windows of approximately 1 cM, and once again, scans the reference panel for matches. This time it finds the best pair of complementary matches for both maternal and paternal haplotypes in the customer's sample. The idea is to vastly increase the number of potential matching segments by allowing extremely distant relationships (e.g., 1 cM may be shared by 20th cousins). Several SNP mismatches are tolerated in this step to accommodate phasing errors from Step 1. The pair of complementary matches with fewest errors is used to locally refine the customer's phase in each window.

(3) Finally, it models haplotype frequency and recombination in the reference panel to statistically phase sites that were not adequately phased in the first two steps. Using up to 80 reference matches that represent both maternal and paternal haplotypes, it

phases 0.3 cM windows with a Hidden Markov Model (HMM). The model is based on a previously described model of recombination [78], which exploits the fact that unknown haplotypes are likely to be similar to frequently occurring haplotypes. Therefore, the HMM chooses appropriate references to phase the customer's sample by penalizing recombination and mutation between the two matches. Two iterations of the Viterbi algorithm are used to find the most likely haplotype phase of the HMM.

The end result of Eagle phasing is a set of SNP genotypes on each chromosome with maternal and paternal alleles separated into different haplotypes. Although Eagle phasing is ~99.7% accurate, it cannot be perfect. There are still switch errors (maternal/paternal transitions) approximately every few cM [76]. Since local ancestry methods classify tiny DNA segments that are much smaller than this, switch errors have little effect on the final ancestry proportions. However, at the end of the local ancestry pipeline, these switch errors must be corrected by using the inferred sequence of super-population ancestry (see Phase Correction).

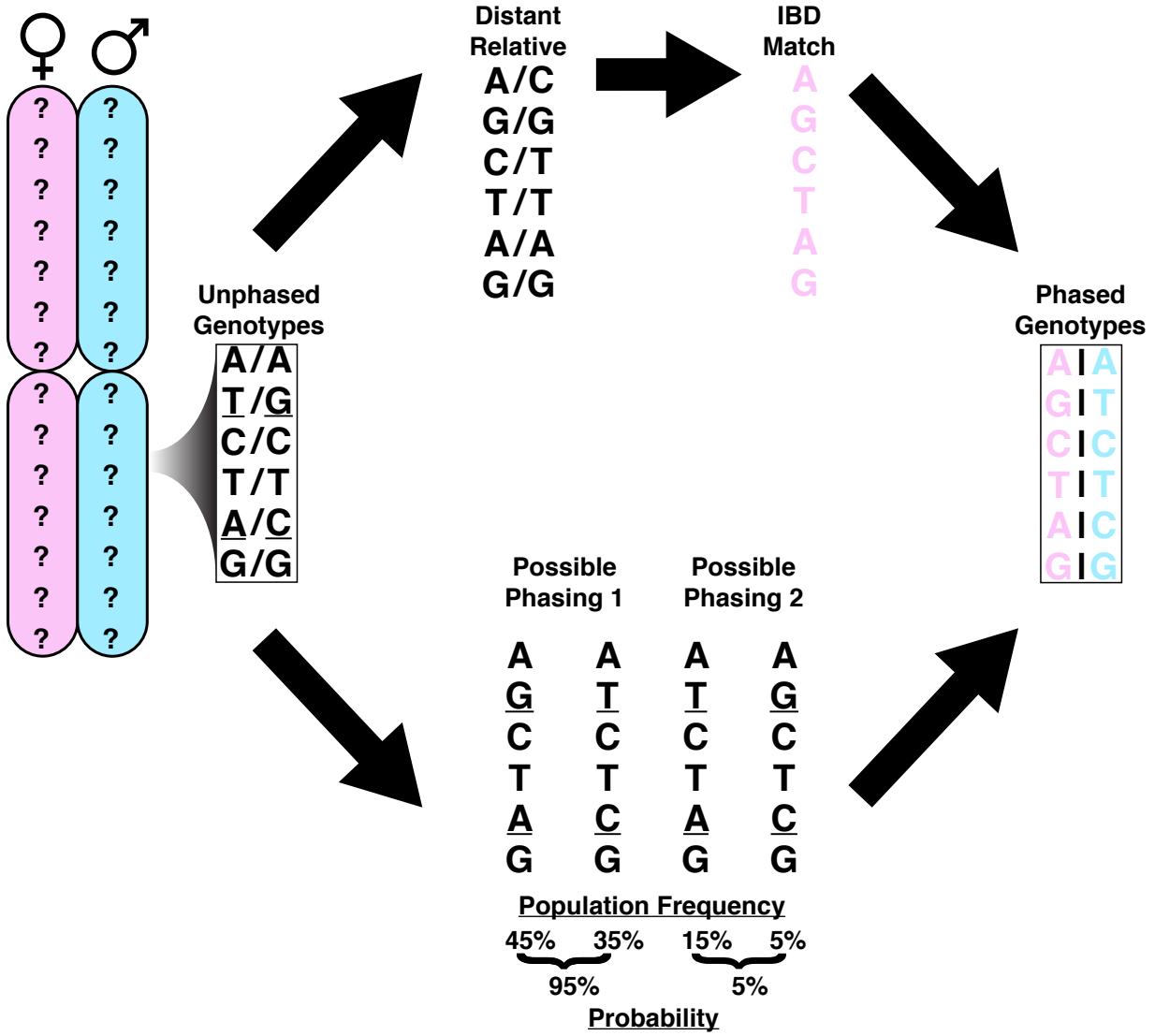

Figure 10. Eagle combines two different methods of phasing: exact IBD matches (top) and haplotype frequencies in a reference panel to generate phasing probabilities (bottom).

Segment Classification

Local ancestry inference (LAI) is a method for determining population ancestry along small segments of each chromosome. In contrast to a global ancestry method such as Speedymix, which can only determine overall ancestry proportions, LAI can also determine the genetic coordinates where each ancestral population segment was recombined. Chromosome painting is somewhat synonymous with LAI but generally refers to the result instead of the method.

LAI has existed for nearly two decades and was originally designed as an extension of the most popular global ancestry method [19]. The idea behind LAI is to leverage the correlation between SNP markers that are in close proximity to one another. Recombination cannot fully break up the association between adjacent SNPs over shorter time intervals, a phenomenon known as linkage disequilibrium. Hence, the order of SNPs along a DNA segment can be highly informative about its population origin. Short segments of DNA that are shared by many individuals in a particular population due to ancient shared ancestry are known as identical-by-state (IBS).

Some of the earliest applications of LAI were to identify locations of genes associated with human diseases, a process known as admixture mapping [15,20,21]. However, the method has also continually been refined in order to study human admixture proportions or measure the amount of time elapsed since admixture. Over two dozen methods of LAI now exist [12–36]. Very often, a Hidden Markov Model (HMM) or one of its derivatives is used to model ancestry as a hidden state, based on either the observed order of SNPs or haplotypes [15,19–21]. Sometimes, other classification methods are used in conjunction with or instead of an HMM: e.g., Markov chain Monte Carlo [19,21,30], iterated conditional modes [16], PCA [24,30], random forests [33], dynamic programming [14], and deep learning [28].

In MYORIGINS v3 we classify phased haplotype segments with our own proprietary machine learning technique (Fig. 11). We have found it to outperform one of the most popular LAI methods [33]. We break a customer's phased chromosomal data into segments of 500 SNPs in overlapping windows spaced apart by 200 SNPs. Then, we classify each segment into one of several super-populations using a multi-class clustering method. Our technique is ideal for discriminating between groups with data that are complex and high-dimensional (such as SNP haplotypes). We use the same reference panel as in Speedymix except all references are phased at the full set of 637,645 SNPs.

Our pipeline is novel and unique in its paired use of global and local estimates. Due to the inherently noisy nature of LAI, which must classify small segments with limited information, super-populations are used instead of populations. Using groups that are older and more distinct allows us to more accurately classify segments. Unlike other existing LAI methods, ours reduces noise in results by eliminating irrelevant super-populations during the Speedymix step. Additionally, we apply weights (C) for each relevant super-population to correct for sample size imbalance: $C_k = N/(KN_k)$, where N_k is the sample size of class k, and K is the total number of classes. The super-population confidence scores are mapped into probabilities using a sigmoid function to be used in the next step of our pipeline.

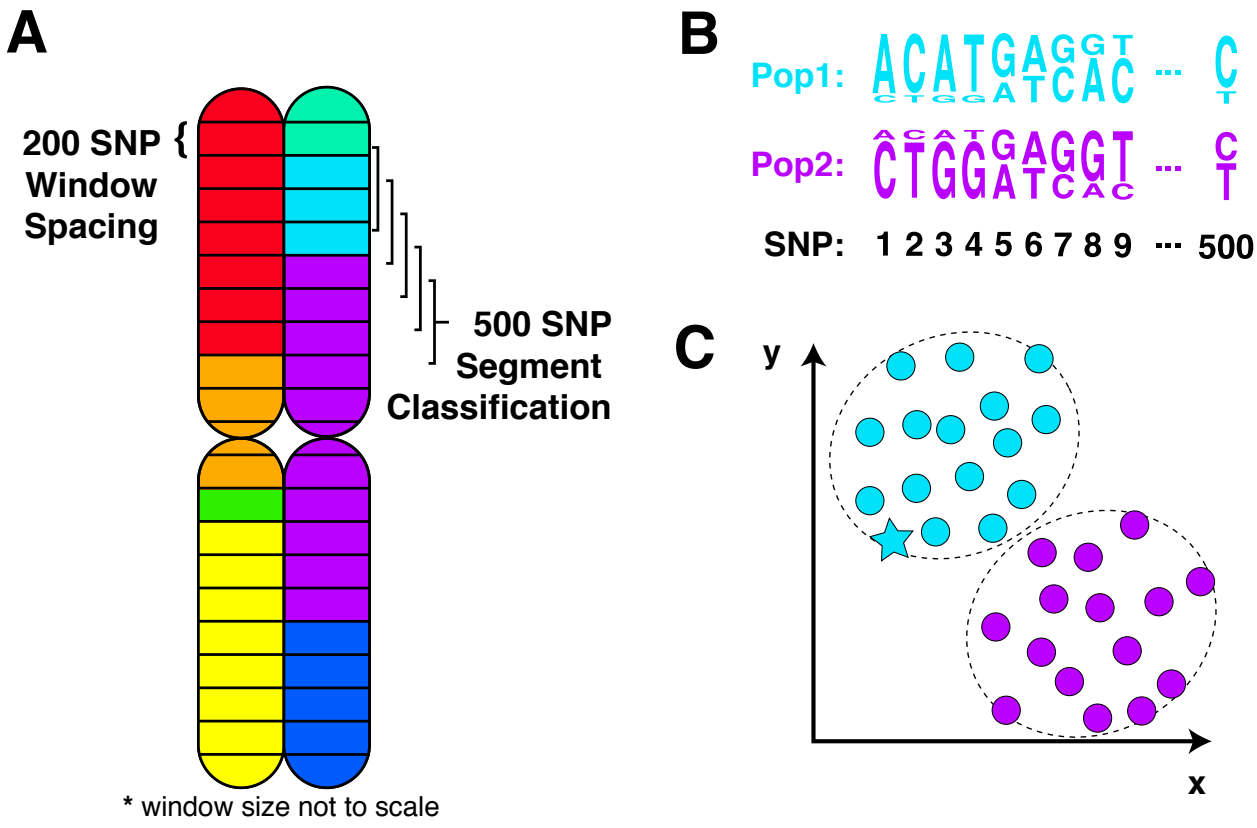

Figure 11. Segment classification steps. (A) Each pair of phased chromosomes is divided into windows spaced 200 SNPs apart, and overlapping segments of 500 SNPs are taken from each window for classification. (B) ACGT letter size represents frequency of that nucleotide in hypothetical super-populations 1 and 2 along a 500 SNP haplotype. Haplotype differences between super-populations are quantified and modeled. (C) Our proprietary classification technique is trained on all super-populations to maximize the distance between them in multi-dimensional space. The star indicates a customer sample that was predicted into hypothetical super-population 1.

Conditional Random Field

Once each segment of each chromosome is assigned probabilities of each super-population, we use those probabilities to parameterize a linear chain conditional random field (CRF). This process uses pattern recognition of an entire chromosome to better predict the sequence of ancestries. The CRF "smooths over" each segment classification by incorporating information from neighboring segments on the chromosome. Although the small number of SNPs available to segment classifiers can result in more noise (Fig. 2C), the CRF step adjusts misclassifications by maximizing the probability of the entire sequence (Fig. 12). Our linear-chain CRF follows [33] except that our segment probabilities are estimated by our method instead of random forests.

For each chromosome, we model the ancestries of individual haplotypes $\{1\ldots N\}$ across 200 SNP windows $\{1\ldots W\}$ as an $N \times W$ matrix $\mathbf{A}$, where $\mathbf{A}_{i,w}$ is the most probable super-population $k \in \{1\ldots K\}$ for individual i at window w. Similarly, we model all haplotypes as an $N \times W$ matrix $\mathbf{H}$, where $\mathbf{H}_{i,w}$ is the haplotype h for individual i at window w.

The log-linear probability of ancestries across individual i's entire haploid chromosome is

$$\Pr\left(\mathbf{A}_{i,*}\middle|\mathbf{H}_{i,*}:\Theta\right) = \tag{4}$$

$$\frac{1}{Z\left(\mathbf{H}_{i,*}\right)}\exp\left\{\sum_{w=1}^{W}\sum_{k=1}^{K}\sum_{h\in\mathcal{H}_w}\theta^{A}_{w,k,h}1_{\{A_{i,w}=k\}}1_{\{H_{i,w}=h\}} + \sum_{w=1}^{W-1}\sum_{k=1}^{K}\sum_{k'=1}^{K}\theta^{T}_{w,k,k'}1_{\{A_{i,w}=k\}}1_{\{A_{i,w+1}=k'\}}\right\}$$

where $\mathcal{H}_w$ includes all possible haplotypes in window w, $1_{\{x=y\}}$ equals 1 if $x = y$ and 0 otherwise, and $Z\left(\mathbf{H}_{i,*}\right)$ is a partition function for normalizing the probability:

$$Z\left(\mathbf{H}_{i,*}\right) = \tag{5}$$

$$\sum_{A_{i,*}}\exp\left\{\sum_{w=1}^{W}\sum_{k=1}^{K}\sum_{h\in\mathcal{H}_w}\theta^{A}_{w,k,h}1_{\{\mathrm{A}_{i,w}=k\}}1_{\{H_{i,w}=h\}} + \sum_{w=1}^{W-1}\sum_{k=1}^{K}\sum_{k'=1}^{K}\theta^{T}_{w,k,k'}1_{\{\mathrm{A}_{i,w}=k\}}1_{\{A_{i,w+1}=k'\}}\right\}$$

The parameter θ^{A} is a probability for the ancestry of the haplotype in each window:

$$\theta^{A}_{w,k,h} = ln(Pr(\mathbf{A}_{i,w} = k \mid \mathbf{H}_{i,w} = h)) \tag{6}$$

The parameter θ^{T} is the joint probability of ancestry in two adjacent windows:

$$\theta^{T}_{w,k,k'} = ln(Pr(\mathbf{A}_{i,w} = k, \quad \mathbf{H}_{i,w+1} = k')) \tag{7}$$

Parameter values of θ^{A} are estimated by the multi-class segment probabilities (previous step), whereas the θ^{T} values are estimated by a previously described [19] linkage model of admixture:

$$\Pr\left(A_{i,w} = k, H_{i,w+1} = k'\right) = \begin{cases} q_k(exp(-d_w G) + (1 - exp(-d_w G))q_{k'}) & \text{if } k = k' \\ q_k((1 - exp(-d_w G))q_{k'}) & \text{otherwise} \end{cases} \quad (8)$$

where d_w is the distance between the midpoints of windows w and $w + 1$, G is the number of generations since admixture, and q_k and $q_{k'}$ are the chromosome-wide admixture proportions for the super-population in the current window (k) and the next window (k').

The logic for this linkage model is as follows. Recombination is responsible for breaking apart and fusing together DNA haplotypes from different super-populations. One breakpoint does not influence the location of a future breakpoint; thus, recombination can be modeled as a random Poisson process. The term dG can be thought of as the expected number of recombination events within the window since admixture occurred. The top portion of equation (9) accounts for the possibility that no recombination has occurred in the window, or there has been at least one breakpoint but with the adjacent windows coming from the same super-population. The bottom portion of equation (9) accounts for the possibility of at least one breakpoint, resulting in a switch from super-population k to k'. We use a uniform distribution for values of q_k to simplify the model, although chromosome-wide admixture proportions could be included in the future.

For each customer's haploid chromosomes, we use the Viterbi algorithm along with the CRF probability to infer the most likely chain of ancestries.

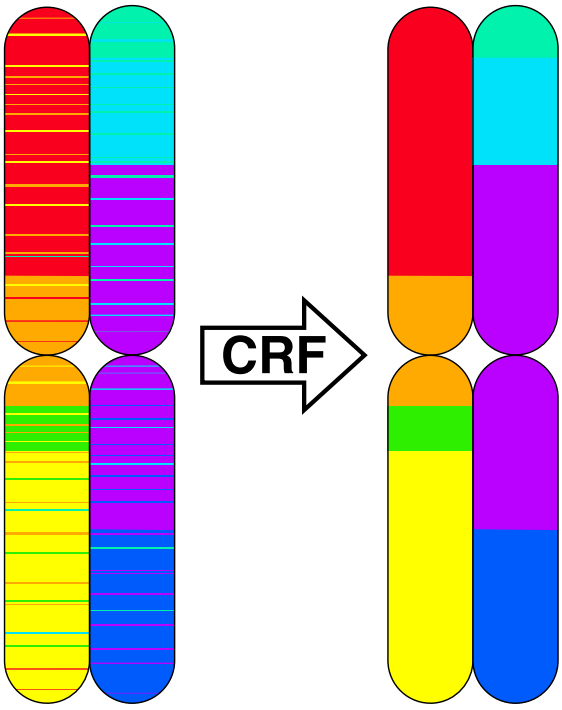

Figure 12. Effect of the Conditional Random Field: noisy classifications are "smoothed out" by a linkage model.

Phase Correction

Statistical phasing is >99% accurate, but this small minority of switch errors add up across 100,000s of SNP positions. Fortunately, the end result can be corrected by essentially "phasing" the final super-population labels. We use a Hidden Markov Model (HMM) parameterized by some basic expectations of which pair of maternal/paternal labels are most probable in each window, given the labels in the previous window (Fig. 13). Phase correction, the removal of switch errors using an HMM, is nearly as old as local ancestry inference itself [13,14,28,33].

First, we initialize the space for hidden states and observations. For each window w along a chromosome, the observation is a diploid pair of predicted super-population labels, and the hidden state is the true pair of super-population labels. If a customer's result includes proportions from K super-populations, then the observation space i/i' and hidden state space j/j' are both pairs of super-population labels where $i, i', j, j' \in \{1 \ldots K\}$.

We assume that all observations are potentially unphased but contain the correct labels. Therefore, our HMM model only allows the chromosome strands to be flipped. The emission probabilities of observed states given hidden states is:

$$\Pr\left(\mathbf{O}_{ii'} | \mathbf{S}_{jj'}\right) = \begin{cases} 0.5 & \text{if } (i = j \text{ and } i' = j') \text{ or } (i = j' \text{ and } i' = j) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

To illustrate this: when label pair 1/2 is observed, the hidden state for it can only be 1/2 or 2/1 with equal probabilities.

The main correction comes from our transition probabilities. Recombination happens rarely compared with the number of windows, so we assign higher probability for hidden states to belong to the same chromosome strand:

$$\Pr\left(\mathbf{S}_{ii',w} | \mathbf{S}_{jj',w-1}\right) = \quad (10)$$
$$\begin{cases} p_1 & \text{if } (i = j \text{ and } i' = j') \\ 0 & \text{if } (i = j' \text{ and } i' = j) \\ (1 - p_1)p_2 & \text{if } (i = j \text{ and } j \neq j') \text{ or } (j = j' \text{ and } i \neq i') \\ (1 - p_1)(1 - p_2)p_3 & \text{if } (i = j' \text{ and } j \neq i') \text{ or } (j = i' \text{ and } i \neq j') \\ (1 - p_1)(1 - p_2)(1 - p_3) & \text{if } (i \neq j \text{ and } j \neq j') \text{ or } (j \neq i' \text{ and } i \neq j') \end{cases}$$

where $p_1 = 0.85$, $p_2 = 0.85$, and $p_3 = 0.75$ (see Table 3).

Table 3. Transition probabilities for each type of change between diploid pairs of super-population labels.

Type	Example	Probability
No change	1/2 to 1/2	0.85
Strand flip	1/2 to 2/1	0.00
Partial overlap	1/2 to 1/3	0.1275
Partial overlap after strand flip	1/2 to 3/1	0.016875
Other	1/2 to 3/4	0.005625
Total		**1.00**

We employ one additional type of penalty on top of the aforementioned transition probability matrix and also a gap filling procedure prior to estimating the HMM via the Viterbi algorithm.

(1) Penalty from hierarchical clustering. Across all windows, we count how frequently pairs of super-population labels appear together. We then use that frequency table to conduct hierarchical clustering analysis. The resulting tree structure is informative about whether or not maternal and paternal labels can easily be separated or "bucketed." We apply a penalty in our transition probability matrix $\Pr(\mathbf{S}_{ii',w}|\mathbf{S}_{jj',w-1})$ to labels being phased in a way that violates this bucketing. The magnitude of the penalty is dependent upon how well separated the buckets are.

(2) Gap filling procedure. The HMM assumes all classifications have been fixed by the CRF; however, there may still be misclassified "gaps" of one or two windows. These sometimes occur if a true recombination breakpoint does not perfectly align with the windows we use for classification. For example, consider this result:

1/2, 1/2, 1/2, 1/2, 1/2, **2/2**, 2/1, 2/1, 2/1, 2/1, 2/1

This is likely caused by a phasing error that requires a correction to:

1/2, 1/2, 1/2, 1/2, 1/2, **1/2**, 1/2, 1/2, 1/2, 1/2, 1/2

However, the window in bold features a misclassification of 2/2. We only allow our HMM to correct phasing, not minor classification errors—we find that these two components are best corrected separately. Instead, we fill these gaps of 1–2 windows prior to estimating the HMM.

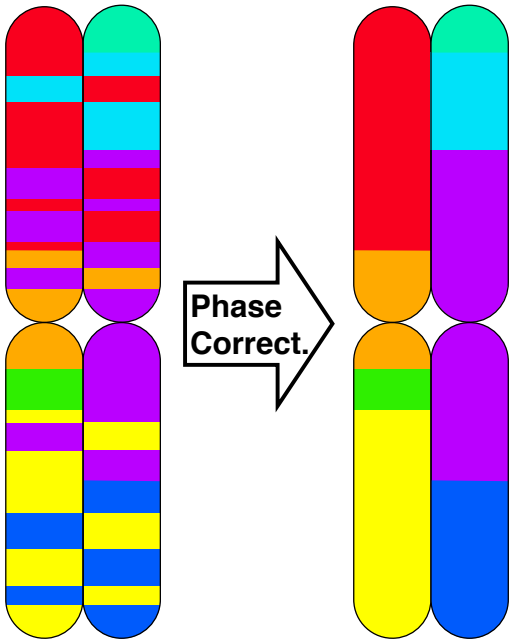

Figure 13. Phase Correction: The HMM transition probabilities remove switch errors caused by imperfect phasing.

Global-Local Ancestry Integration

The final step merges global and local ancestry results. Population percentages estimated by Speedymix are normalized into the super-population percentages estimated by local methods (Table 4). This takes advantage of the higher accuracy of local classification methods when used on suitably distinct (super-)populations and the higher population specificity of global methods when given genome-wide information (see Overview and Overview of MYORIGINS v3 Pipeline).

There will generally be very minor discrepancies between the two results, and any difference is typically an improvement (see Validation). This is because the local classifier is fed the exact genomic coordinates of each segment and can therefore apportion them better. For example, the hypothetical customer in Table 4 has three Ashkenazi grandparents and one grandparent from Western Europe. That last grandparent is half Scandinavian and half Irish. Let us assume the true percentages are Ashkenazi: 75%, Scandinavian: 12.5%, and Irish: 12.5%. In this case, Speedymix slightly overestimates their Ashkenazi percentage from a slight overrepresentation of SNPs in those genomic regions. The local classifier corrects this overestimate and the 1:1 ratio of Scandinavian/Irish populations are finally integrated, thus combining strengths of each method.

When calculating final percentages, there are two possible units of total DNA: centimorgans or megabases. The former measures the length of DNA by how frequently it recombines, whereas the latter measures the physical length. Although both units are reasonable ways to quantify a customer's population percentages, we think that physical length is more intuitive. This makes the statement ("X% of my DNA is from Y population") more accurate. Therefore, we sum and normalize results in units of megabases.

The final result includes a chromosome painting that shows the ancestry of each DNA segment (Fig. 14). Each chromosome pair is sorted by the major genome-wide percentage. We exclude our calculation from two genomic regions that are SNP-poor (within chromosomes 1 and 9). These regions are in close proximity to the centromere and more conserved, i.e., less likely to mutate. We also exclude the short arms of Chromosomes 13–15 and 21–22 from our SNP array. These short arms contain an abundance of repeated sequences, low recombination rate, and low SNP density; hence, they are excluded from our SNP genotyping array.

Table 4. Example normalization of results. This hypothetical customer has three Ashkenazi grandparents and one British/Irish grandparent. The true percentages are achieved by normalizing global and local ancestry results.

Global Ancestry		Local Ancestry		Final Integrated Result	
Ashkenazi Jewish	80%	European Jewish	75%	European Jewish Ashkenazi Jewish	75%
Scandinavia	10%	Western Europe	25%	Western Europe Scandinavia	12.5%
Ireland	10%			Western Europe Ireland	12.5%

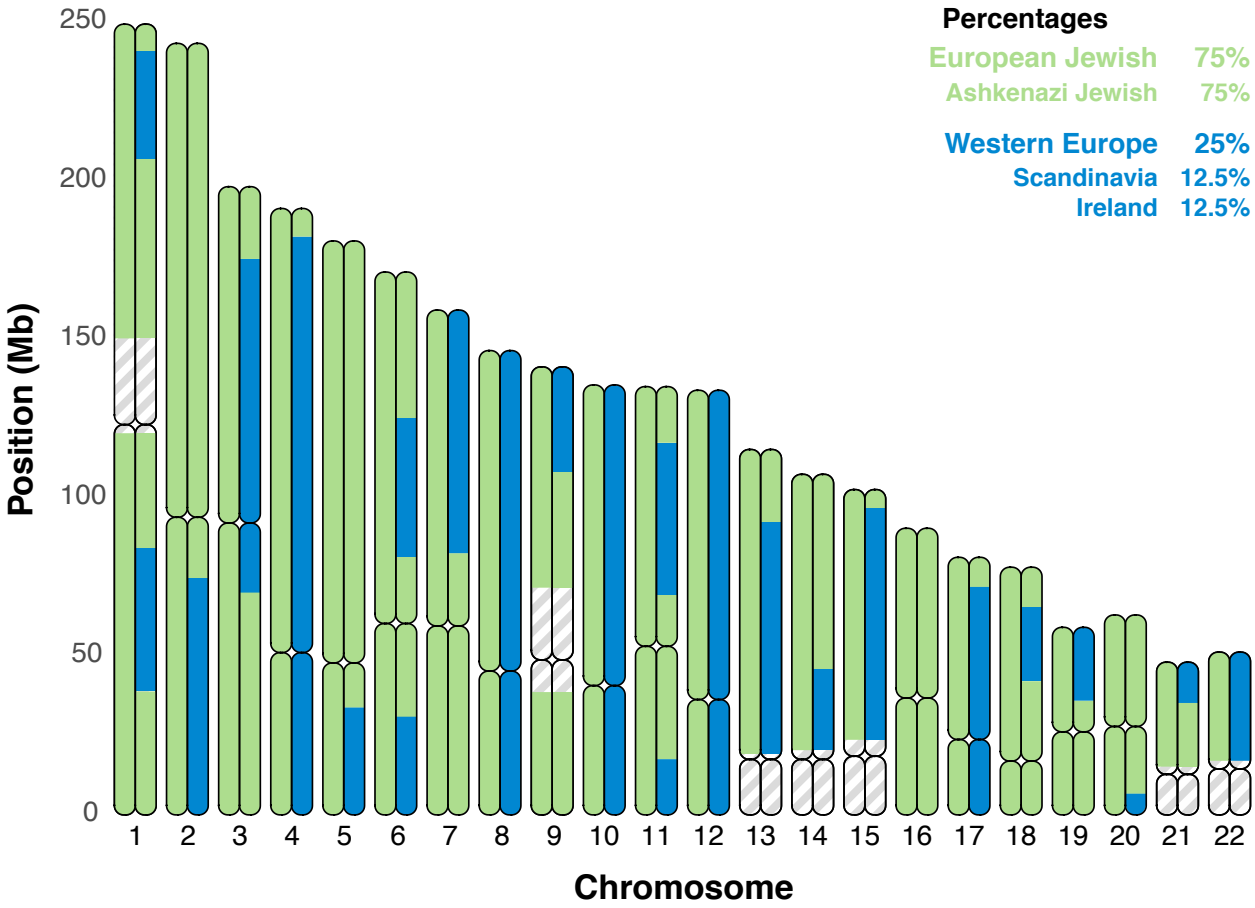

Figure 14. Example chromosome painting for a customer with three Ashkenazi grandparents and one ½ Scandinavian ½ Irish grandparent. The missing segments on Chr. 1 and 9 are SNP-poor areas that cannot be classified confidently. Segments missing from Chr. 13–15, 21–22 are not included on our SNP genotyping array.

Validation

We used leave-one-out cross validation (LOOCV) to assess the performance of MYORIGINS v3. For every one of our 8,053 references, we removed it from the panel and predicted its ancestry into all 90 populations. A proportion of 1.0 in its own population would be perfect (Fig. 15).

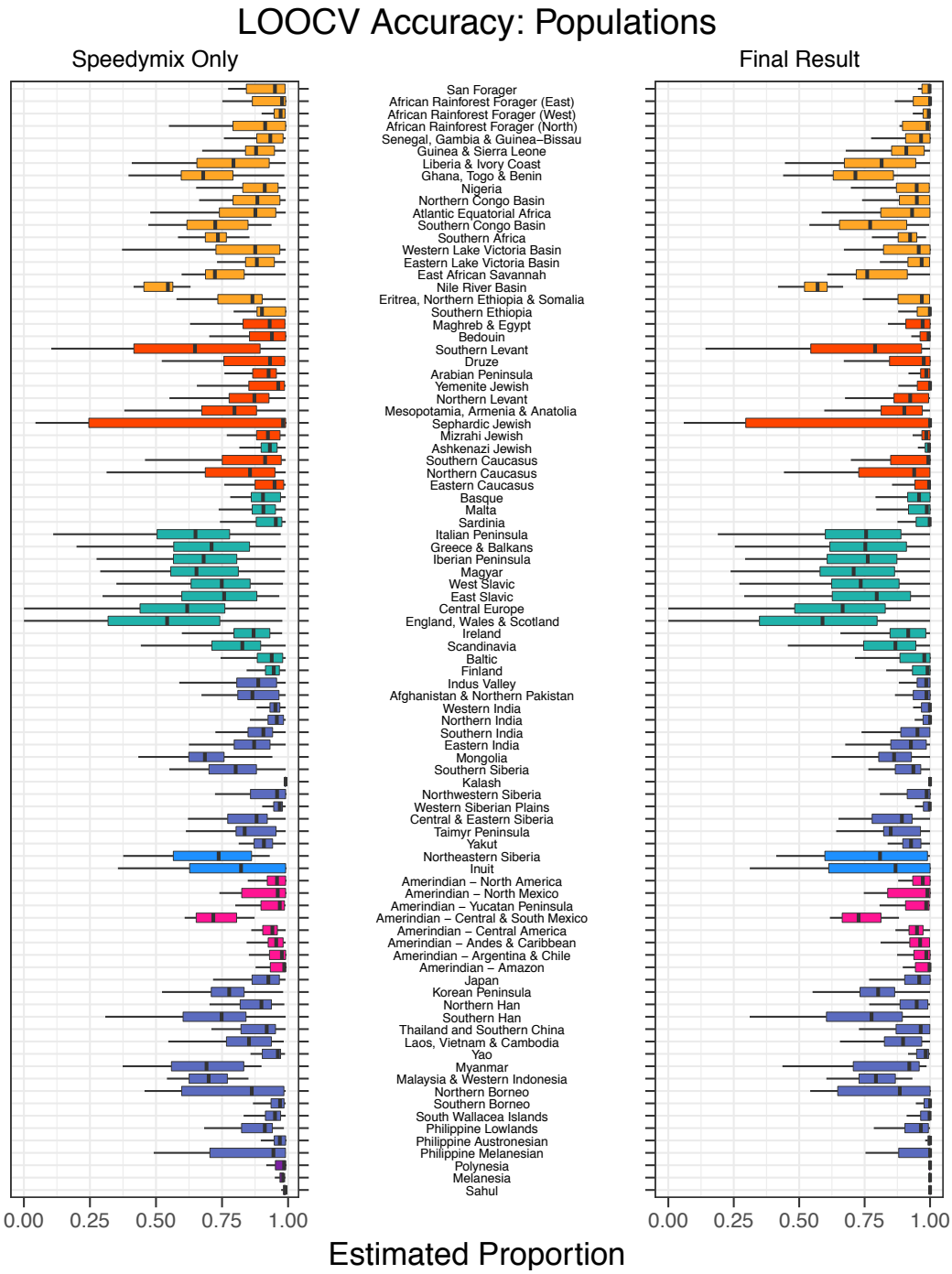

Figure 15. Estimated population proportions of references by Speedymix (left) and the entire pipeline (right). Boxplots show median and interquartile range (middle 50% of values) for each reference in its own population.

Our pipeline has a mean accuracy of 0.89 ± 0.03. In other words, the mean percentage estimated for a reference sample in its correct population is 89%. This is an improvement from using our global method Speedymix alone (0.84 ± 0.02). The largest variation in accuracy is found in continents with numerous freely migrating populations, such as Europe. Hundreds to thousands of years of shared gene flow and isolation-by-distance can make genetic variation between groups very indistinct (see Finding Population Structure). Inaccurate classifications tend to be between geographically neighboring populations that share many recent common ancestors, such as Great Britain and Central Europe or Nile River Basin and East African Savanah (Fig. 16).

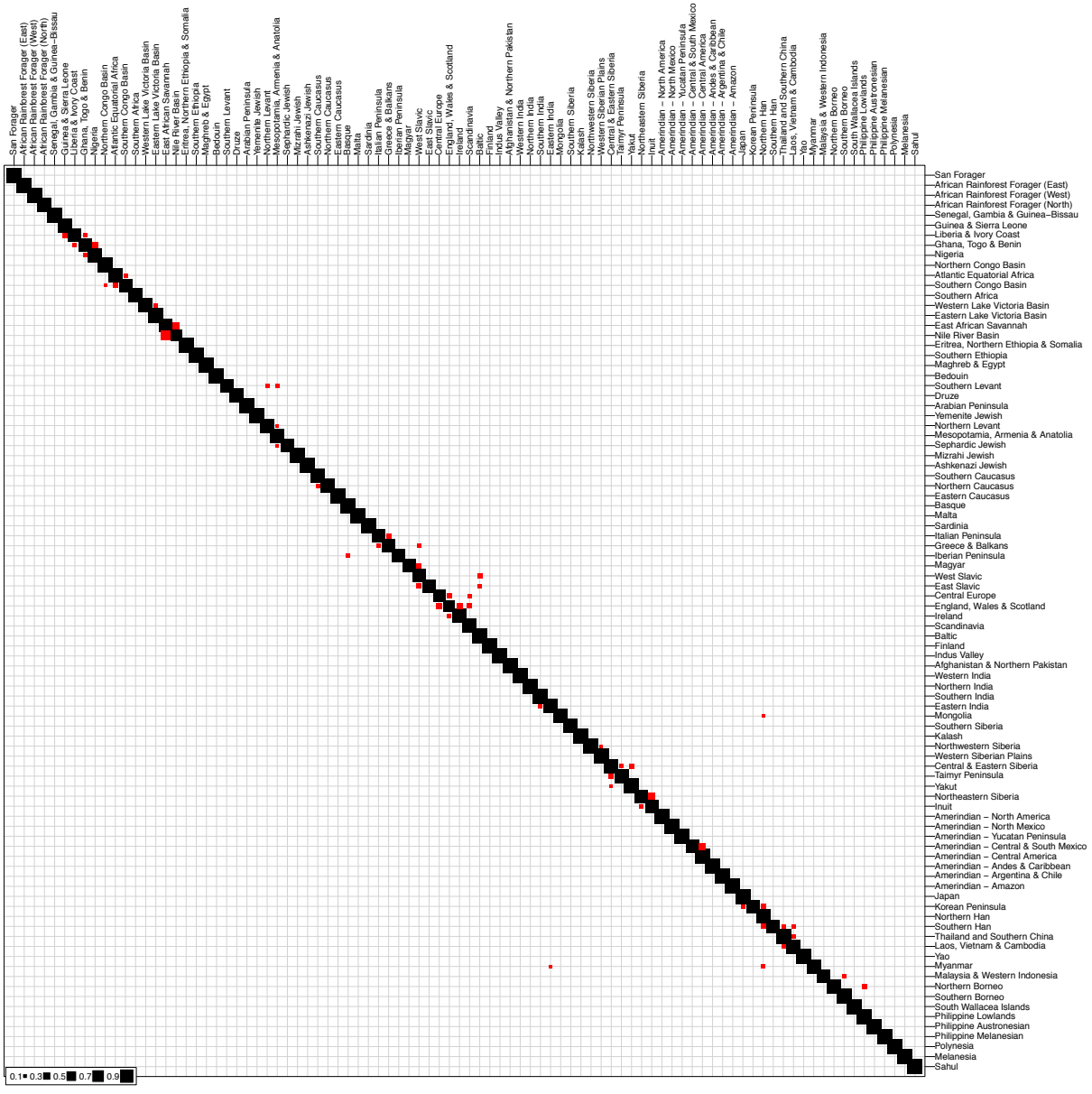

Figure 16. Confusion matrix for Fig. 15. Each row shows the true populations; columns show the estimated populations. Each square indicates the mean estimated proportion of the correct population (black), or incorrect population (red). Only incorrect estimates >0.05 are shown.

At the level of super-populations (Fig. 17), our pipeline has a mean accuracy of 0.96 ± 0.02, which is also an improvement from using our global method Speedymix alone (0.90 ± 0.03).

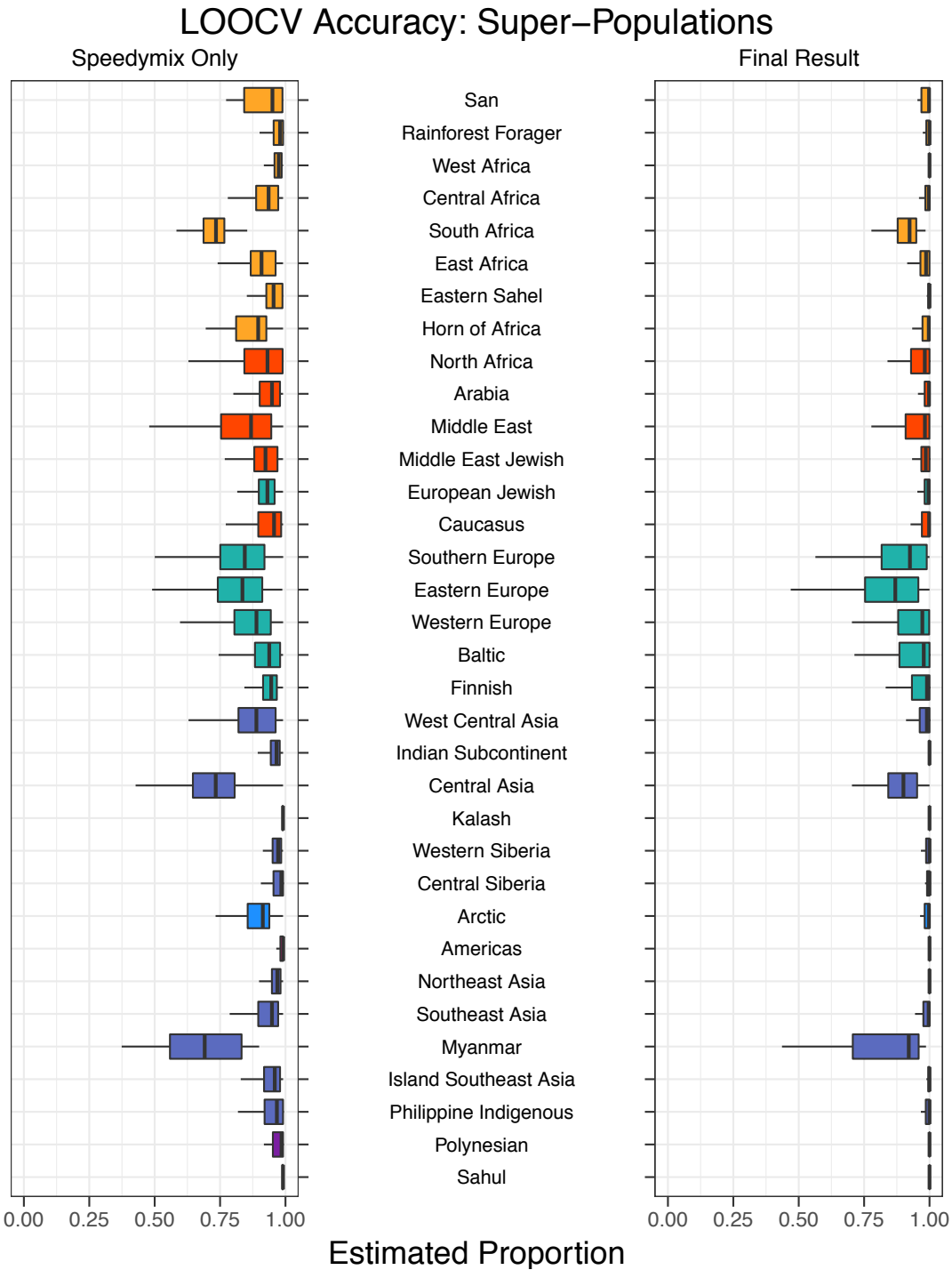

Figure 17. Estimated super-population proportions of references by Speedymix (left) and the entire pipeline (right). Boxplots show median and interquartile range (middle 50% of values) for each reference in its own super-population.

Certain super-populations, such as Myanmar, represent a mix between historical groups (Northeast Asia and Indian Subcontinent [79]) leading to slightly reduced accuracy (Fig. 18). Others may contain a small amount of recent admixture with other groups (e.g., Arctic and Western Europe), slightly lowering accuracy. However, an overall accuracy of 0.96 indicates that super-populations are highly predictable and a much more genetically distinct grouping in the population hierarchy (see Fig. 2), which validates using these for chromosome painting.

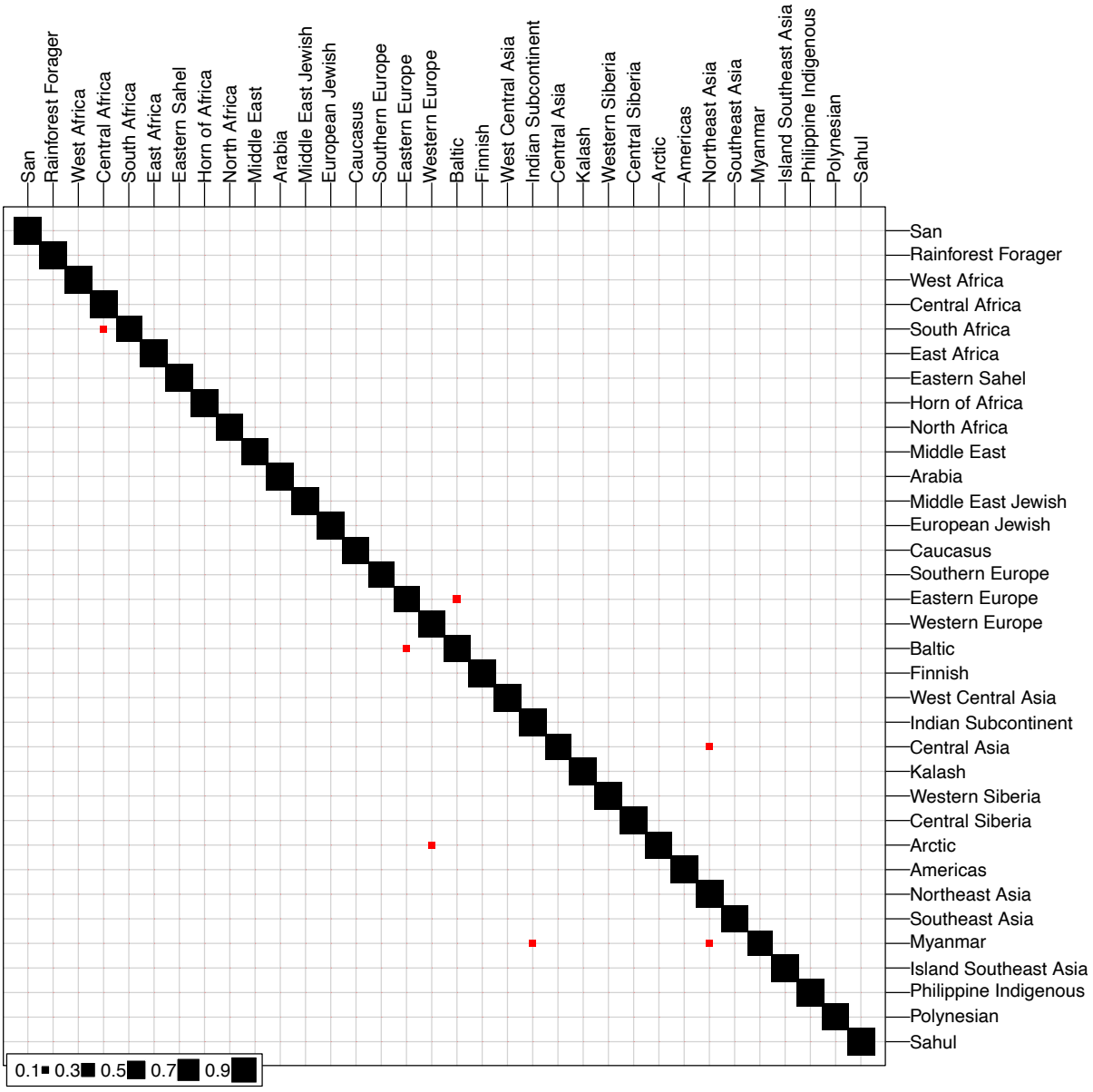

Figure 18. Confusion matrix for Fig. 17. Each row shows the true super-populations; columns show the estimated super-populations. Each square indicates the mean estimated proportion of the correct super-population (black) or incorrect super-population (red). Only incorrect estimates >0.05 are shown.

We also estimated that our previous version of MYORIGINS had a mean accuracy of 0.84 ± 0.06. Given the near quadrupling of our population number and the known tradeoff between accuracy and specificity (see Overview), this amounts to a large improvement overall (Fig. 19).

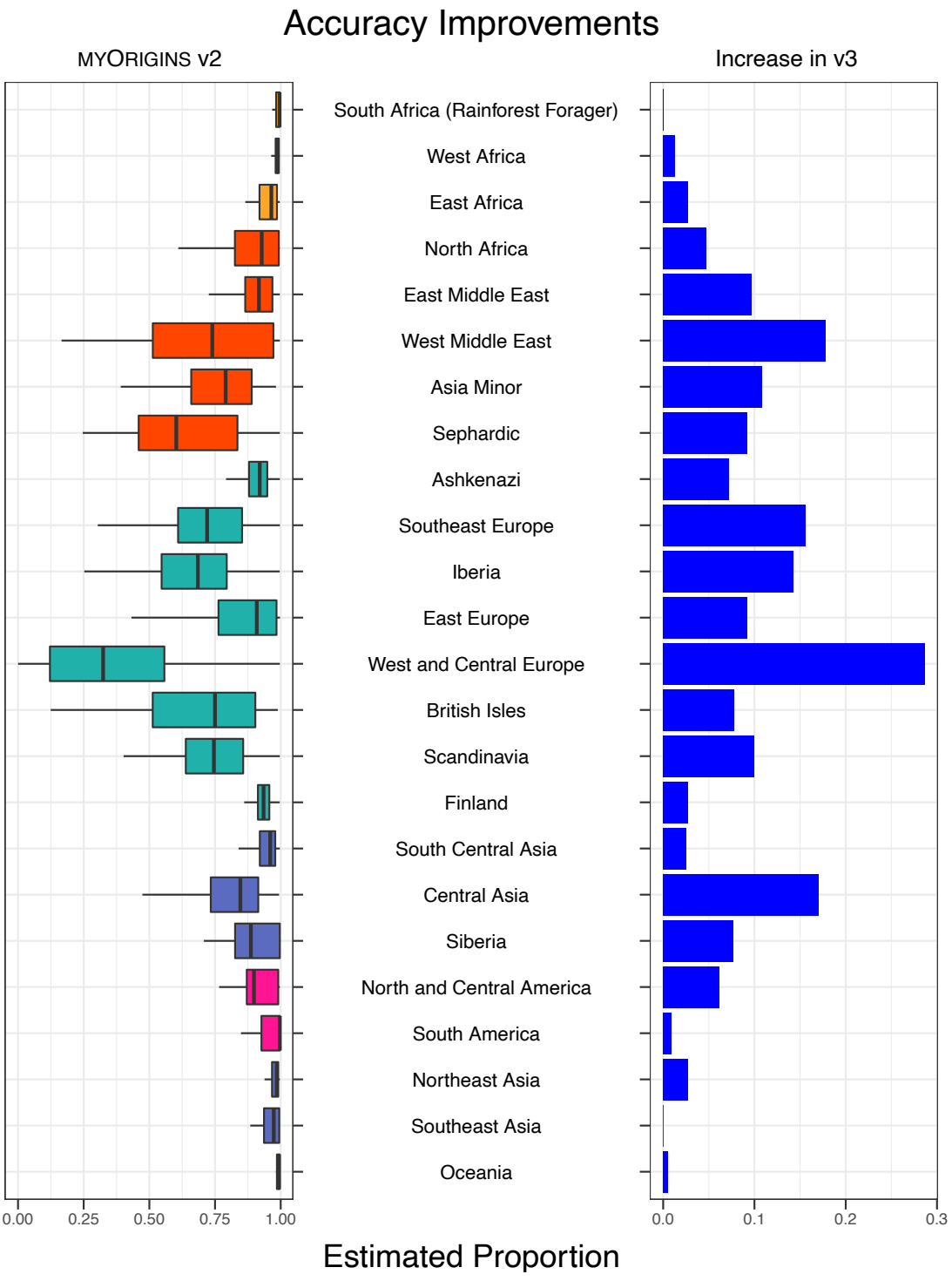

Figure 19. Accuracy increase between the previous MYORIGINS v2 and the new MYORIGINS v3. Population names shown here are the previous (v2) names and are compared to an analogous group in v3.

Next, we validated the performance of our pipeline on admixed customers. Although the above LOOCV procedure on references is informative, it does not test the ability to accurately predict percentages other than 100%. Therefore, we created a validation set of admixed samples with known population percentages of 25%, 50%, and 75%. We created pedigrees of four references as grandparents from two different populations. Then we simulated genetic recombination to create admixed grandchildren. Finally, we removed these four references from the panel and predicted each simulated customer.

We used three main types of admixture, which we refer to here as F1, F1B, and F2 (Table 5). These indicate the population identity of the four grandparents. F1s are 50% of each population, and both parents are 100%. F2s are 50% from each population, but each parent is also 50%. Finally, F1Bs have one grandparent from a unique population with percentages of 75% / 25%. Although these are the genealogical percentages, genetic recombination is random; hence, the genetic percentages are more variable. For example, the mean and S.D. of genetic percentage for F1Bs was actually 75.1 ± 4.6% and 24.9 ± 4.6%, and for F2s was 50.0 ± 6.4%.

Table 5. Simulated admixed samples created from different combinations of two populations.

Type	Sample Size	Father's Father's Population	Father's Mother's Population	Mother's Father's Population	Mother's Mother's Population
F1	10	1	1	2	2
		2	2	1	1
F1B	20	1	1	1	2
		1	1	2	1
		1	2	1	1
		2	1	1	1
F2	20	1	2	1	2
		2	1	2	1
		2	1	1	2
		1	2	2	1
Total	50				

With $N = 50$ validation samples to test for each pair of populations and $90 \times 89/2 = 4005$ possible population combinations, we opted not to test all $> 200{,}000$ samples. Instead, we selected 12 representative pairs of populations that spanned a range of genetic distances from close to far (Table 6). This gave us a reasonable demonstration of the pipeline's performance across a range of geographic backgrounds, combinations, and levels of genetic similarity. We assessed performance at both the population and super-population level. For example, we assessed an Irish sample's estimated percentage of Ireland and also of Western Europe.

Table 6. Populations used for simulated admixture. We estimated results from different population distances.

Distance	Population 1	Population 2
Far	Ireland	Nigeria
	Scandinavia	Amerindian – North America
	Iberian Peninsula	Japan
	Eastern Lake Victoria Basin	Polynesia
Medium	Northern Han	Philippine Lowlands
	Ashkenazi Jewish	East Slavic
	Eastern Caucasus	Mizrahi Jewish
	Southern India	Central Europe
Close	Greece & Balkans	East Slavic
	Mesopotamia, Armenia & Anatolia	Arabian Peninsula
	Laos, Vietnam & Cambodia	Japan
	Western Lake Victoria Basin	Eritrea, Northern Ethiopia & Somalia

For the mixed samples we tested, our pipeline has a mean population accuracy of 0.91 ± 0.07 and a mean super-population accuracy of 0.94 ± 0.05. See Figs. 20–22. We define accuracy here to be the weighted accuracy of each component, i.e., $1 - (p_1\delta_1 + p_2\delta_2)$, where p_k is the true proportion of population k, and δ_k is the absolute difference between the true end estimated proportion. Although there is some minor difference between types of admixture (F1, F1B, F2), most of the variation in accuracy depends upon population distance (far, medium, close) and the mean accuracy of specific populations (Figs. 15–18). For example, the mean accuracy of "far" population mixes (0.92) is higher than for both "medium" and "close" populations (0.90). Similarly, the mean accuracy of mixes with European populations (0.89) and without them (0.93) demonstrates that population differences in Figs. 15–18 are relevant here too.

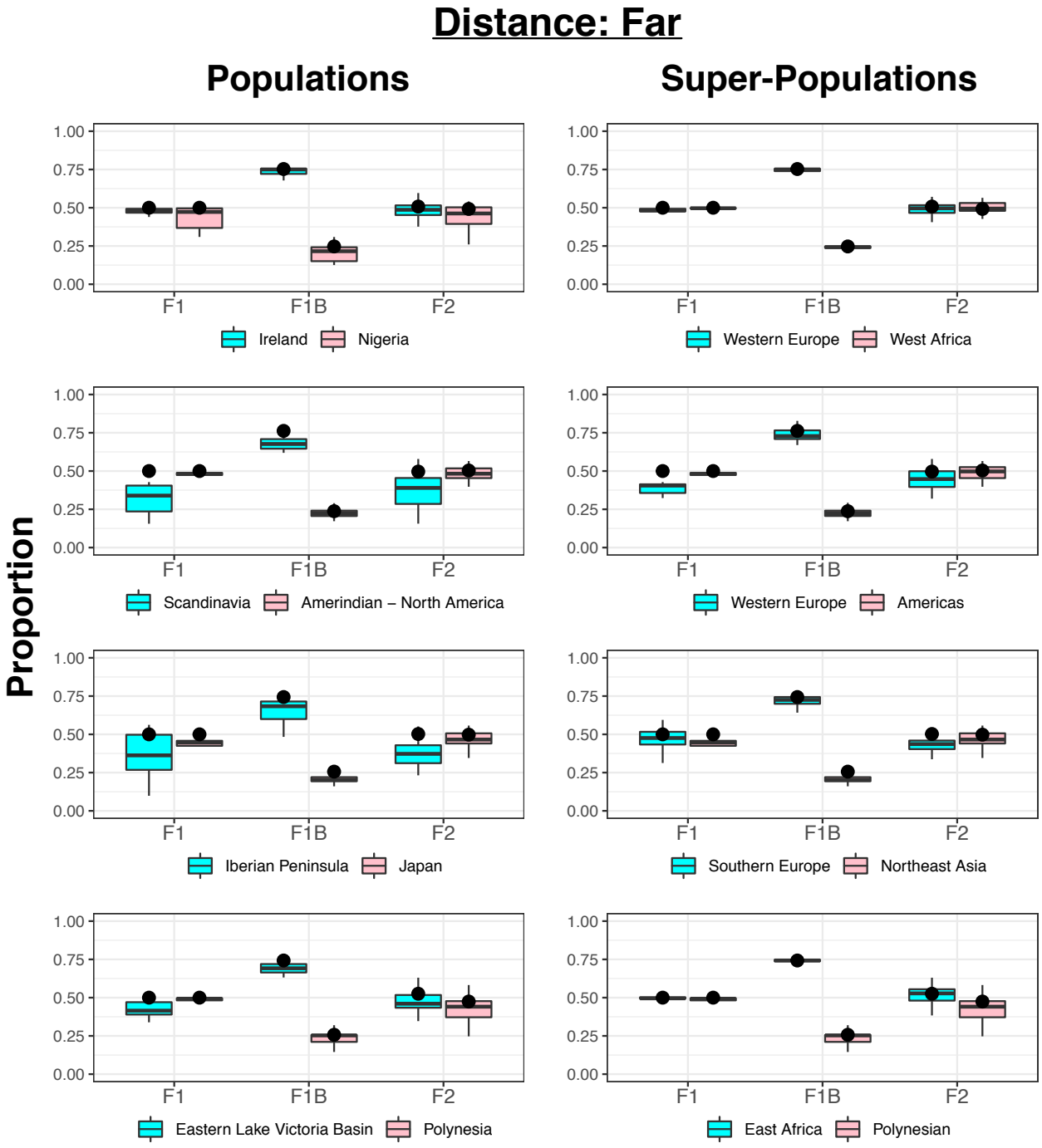

Figure 20. Estimated accuracy for samples with simulated admixture (with far distance between populations). Results for populations are shown on the left, and results for the corresponding super-populations on the right. Definitions of admixture types (F1, F1B, and F2) are in-text and Table 5. Black circles indicate the true proportion; hence, any scatter around the circle indicates estimation error.

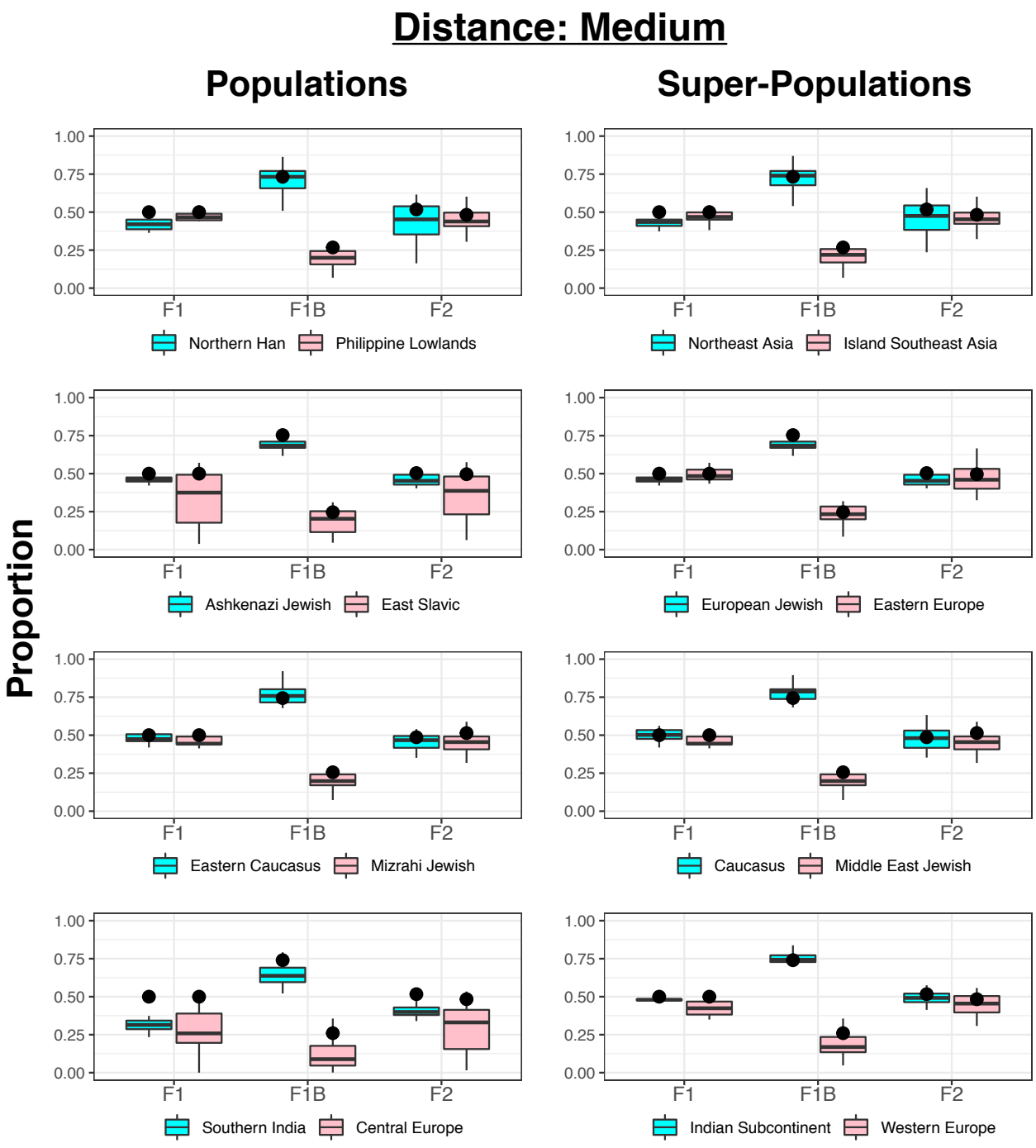

Figure 21. Estimated accuracy for samples with simulated admixture (with medium distance between populations). Results for populations are shown on the left, and results for the corresponding super-populations on the right. Definitions of admixture types (F1, F1B, and F2) are in-text and Table 5. Black circles indicate the true proportion; hence, any scatter around the circle indicates estimation error.

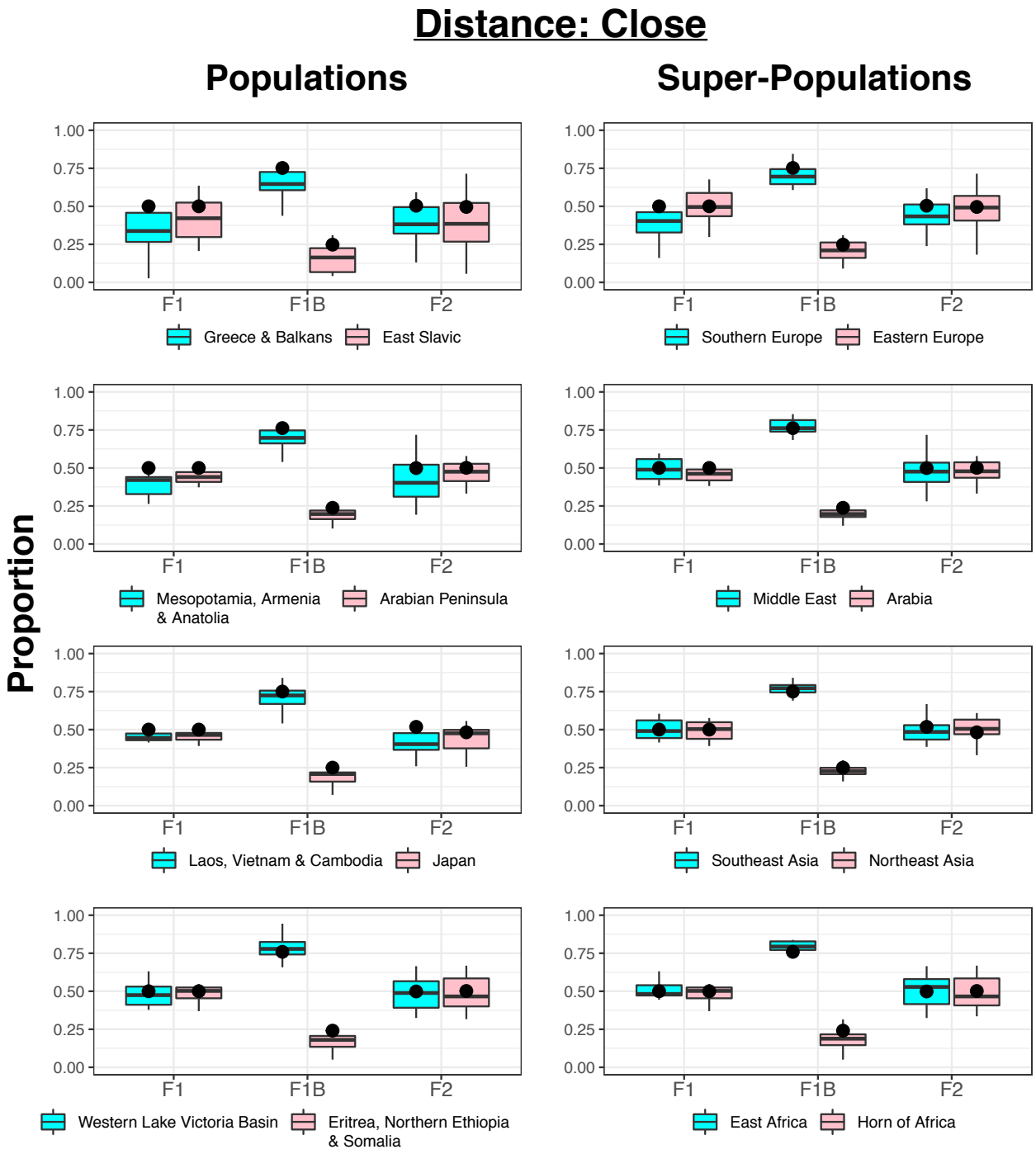

Figure 22. Estimated accuracy for samples with simulated admixture (with close distance between populations). Results for populations are shown on the left, and results for the corresponding super-populations on the right. Definitions of admixture types (F1, F1B, and F2) are in-text and Table 5. Black circles indicate the true proportion; hence, any scatter around the circle indicates estimation error.

Future Improvements

Although this release is already very exciting, we anticipate several updates for the future:

- Increased reference panel – Our reference panel will continue to grow as people seek out their new MYORIGINS v3 results. Much of our improved power to predict specific and unique populations is powered by our customers. In the future, we will continue to offer more populations and improved references as they become available. Group Project Administrators can assist this ever-expanding effort by recommending to us multiple (ideally 30+) FamilyTreeDNA customer kit numbers from not closely related people that have a strong genealogical connection (all four grandparents) from an under-represented, specific, or unique population.

- Trio phasing – The method of phasing that we employ for this release of MYORIGINS v3 is known as statistical "population" phasing. The accuracy of a customer's phasing depends on our having a sizable panel of distant relatives to the customer, because each match can only phase part of their DNA. Trio phasing is another approach that directly uses one or both parents of a customer to phase their DNA. In the future, we may allow customers with linked parent-child relationships in the family tree to improve their results with trio phasing (both parents available) or duo phasing (one parent available).

- Genealogy tools – We plan to include MYORIGINS v3 results in the Chromosome Browser. This will allow customers to compare their matched segments with population segments to help narrow down the common ancestor's identity and location.

- Y-DNA and mtDNA integration – Our diverse autosomal database with data for 90 populations can be integrated with the largest databases of Big Y and mtFull haplogroups! We plan to offer frequency tables and maps that relate MYORIGINS v3 populations to haplogroups.

References

1 Reich, D. (2018) *Who we are and how we got here: ancient DNA and the new science of the human past*, Pantheon.

2 Moreno-Mayar, J.V. *et al.* (2018) Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* 553, 203–207

3 Haak, W. *et al.* (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211

4 Nielsen, R. *et al.* (2017) Tracing the peopling of the world through genomics. *Nature* 541, 302–310

5 Wu, C.I. (2001) The genic view of the process of speciation. *J. Evol. Biol.* 14, 851–865

6 Wu, C.I. and Ting, C.T. (2004) Genes and speciation. *Nat. Rev. Genet.* 5, 114–122

7 Nosil, P. *et al.* (2009) Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* 18, 375–402

8 Turner, T.L. *et al.* (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3, 1572–1578

9 Ellegren, H. *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491, 756–760

10 Malinsky, M. *et al.* (2015) Genomic islands of speciation separate Cichlid ecomorphs in an East African crater lake. *Science* 350, 1493–1498

11 Maier, P.A. (2018) , Evolutionary past, present, and future of the Yosemite toad (*Anaxyrus canorus*): a total evidence approach to delineating conservation units. Ph.D. dissertation, University of California Riverside

12 Tang, H. *et al.* (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* 79, 1–12

13 Sundquist, A. *et al.* (2008) Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res.* 18, 676–682

14 Dias-Alves, T. *et al.* (2018) Loter: a software package to infer local ancestry for a wide range of species. *Mol. Biol. Evol.* 35, 2318–2326

15 Zhu, X. *et al.* (2006) A classical likelihood based approach for admixture mapping using EM algorithm. *Hum. Genet.* 120, 431–445

16 Sankararaman, S. *et al.* (2008) Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* 82, 290–303

17 Pasaniuc, B. *et al.* (2013) Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* 29, 1407–1415

18 Price, A.L. *et al.* (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5, e1000519

19 Falush, D. *et al.* (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587

20 Montana, G. and Pritchard, J.K. (2004) Statistical tests for admixture mapping with case-control and cases-only data. *Am. J. Hum. Genet.* 75, 771–789

21 Patterson, N. *et al.* (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* 74, 979–1000

22 Paşaniuc, B. *et al.* (2009) Inference of locus-specific ancestry in closely related populations. *Bioinformatics* 25, 213–221

23 Baran, Y. *et al.* (2012) Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367

24 Brisbin, A. *et al.* (2012) PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* 84, 343–364

25 Guan, Y. (2014) Detecting structure of haplotypes and local ancestry. *Genetics* 196, 625–642

26 Corbett-Detig, R. and Nielsen, R. (2017) A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS Genet.* 13, 1–40

27 Salter-Townshend, M. and Myers, S. (2019) Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics* 212, 869–889

28 Montserrat, D.M. *et al.* (2020) LAI-Net: local-ancestry inference with neural networks, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1314–1318

29 Schumer, M. *et al.* (2020) Versatile simulations of admixture and accurate local ancestry inference with mixnmatch and ancestryinfer. *Mol. Ecol. Resour.* 20, 1141–1151

30 Lawson, D.J. *et al.* (2012) Inference of population structure using dense haplotype data. *PLoS Genet.* 8, 11–17

31 Rodriguez, J.M. *et al.* (2013) Ancestry inference in complex admixtures via variable-length markov chain linkage models. *J. Comput. Biol.* 20, 199–211

32 Churchhouse, C. and Marchini, J. (2013) Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet. Epidemiol.* 37, 1–12

33 Maples, B.K. *et al.* (2013) RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288

34 Brown, R. and Pasaniuc, B. (2014) Enhanced methods for local ancestry assignment in sequenced admixed individuals. *PLoS Comput. Biol.* 10, e1003555

35 Ma, Y. *et al.* (2014) Accurate inference of local phased ancestry of modern admixed populations. *Sci. Rep.* 4, 1–5

36 Chacón-Duque, J.C. *et al.* (2018) Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* 9, 1–13

37 McKeigue, P.M. *et al.* (2000) Estimation of admixture and detection of linkage in admired populations by a Bayesian approach: Application to African-American populations. *Ann. Hum. Genet.* 64, 171–186

38 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–59

39 Alexander, D.H. *et al.* (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664

40 Raj, A. *et al.* (2014) FastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589

41 Frichot, E. *et al.* (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196, 973–983

42 Gopalan, P. *et al.* (2016) Scaling probabilistic models of genetic variation to millions of humans. *Nat. Genet.* 48, 1587–1590

43 Jay, F. *et al.* (2015) POPS: a software for prediction of population genetic structure using latent regression models. *J. Stat. Softw.* 68, 1–19

44 Cheng, J.Y. *et al.* (2016) Ohana, a tool set for population genetic analyses of admixture components. *bioRxiv* DOI: 10.1101/071233

45 Dawson, K.J. and Belkhir, K. (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* 78, 59–77

46 Pella, J. and Masuda, M. (2001) Bayesian methods for analysis of stock mixtures from genetic characters. *Fish. Bull.* 99, 151–167

47 Tang, H. *et al.* (2005) Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* 28, 289–301

48 Guillot, G. *et al.* (2005) A spatial statistical model for landscape genetics. *Genetics* 170, 1261–1280

49 Corander, J. and Marttinen, P. (2006) Bayesian identification of admixture events using multilocus molecular markers. *Mol. Ecol.* 15, 2833–2843

50 Françǫis, O. *et al.* (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* 174, 805–816

51 Gao, H. *et al.* (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 176, 1635–1651

52 Huelsenbeck, J.P. and Andolfatto, P. (2007) Inference of population structure under a dirichlet process model. *Genetics* 175, 1787–1802

53 Waples, R.S. and Gaggiotti, O. (2006) What is a population? an empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol. Ecol.* 15, 1419–1439

54 Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16, 97–159

55 Rohde, D.L.T. *et al.* (2004) Modelling the recent common ancestry of all living humans. *Nature* 431, 562–566

56 Ralph, P. and Coop, G. (2013) The geography of recent genetic ancestry across Europe. *PLoS Biol.* 11, e1001555

57 Siva, N. (2008) 1000 Genomes project. *Nat. Biotechnol.* 26, 256

58 Cavalli-Sforza, L.L. (2005) The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* 6, 333–340

59 Li, J.Z. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100–1104

60 Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575

61 Manichaikul, A. *et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873

62 McVean, G. (2009) A genealogical interpretation of principal components analysis. *PLoS Genet.* 5, e1000686

63 Pickrell, J.K. and Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8, e1002967

64 Duda, P. and Zrzavý, J. (2016) Human population history revealed by a supertree approach. *Sci. Rep.* 6, 1–10

65 Wong, E.H.M. *et al.* (2017) Reconstructing genetic history of Siberian and Northeastern European populations. *Genome Res.* 27, 1–14

66 Malaspinas, A.S. *et al.* (2016) A genomic history of Aboriginal Australia. *Nature* 538, 207–214

67 Fan, S. *et al.* (2019) African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.* 20, 1–14
68 Lipson, M. *et al.* (2020) Ancient West African foragers in the context of African population history. *Nature* 577, 665–670
69 Lipson, M. *et al.* (2014) Reconstructing Austronesian population history in Island Southeast Asia. *Nat. Commun.* 5, 4689
70 Lazaridis, I. *et al.* (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513, 409–413
71 Pickrell, J.K. *et al.* (2012) The genetic prehistory of southern Africa. *Nat. Commun.* 3, 1–6
72 Reich, D. *et al.* (2009) Reconstructing Indian population history. *Nature* 461, 489–494
73 Pierron, D. *et al.* (2014) Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proc. Natl. Acad. Sci. U. S. A.* 111, 936–941
74 Hudjashov, G. *et al.* (2018) Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. *Sci. Rep.* 8, 1–12
75 Berry, M.W. *et al.* (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.* 52, 155–173
76 Loh, P.R. *et al.* (2016) Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* 48, 811–816
77 Choi, Y. *et al.* (2018) Comparison of phasing strategies for whole human genomes. *PLoS Genet.* 14, 1–26
78 Li, N. and Stephens, M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213–2233
79 Summerer, M. *et al.* (2014) Large-scale mitochondrial DNA analysis in Southeast Asia reveals evolutionary effects of cultural isolation in the multi-ethnic population of Myanmar. *BMC Evol. Biol.* 14, 1–12

Appendix

Appendix A. Unsupervised Speedymix analysis of 90 reference populations. This analysis shows hierarchical population structure across the world by forcing global genetic diversity to be partitioned into *K* clusters. Continental-level structure is well-demonstrated at *K*=10, and more sub-continental and ethnic structure is shown between *K* of 20–60. Colors are randomly selected, and balanced sample sizes of N ≤ 30 samples are randomly chosen for each population.