Uncovering the haplotypes of single nucleotide polymorphisms (SNPs) and their population demography is essential for addressing problems such as disease-gene discovery, chromosomal evolution and population formation. The problem of haplotype inference can be formulated as a mixture model, where the set of mixture components corresponds to the pool of ancestor haplotypes, or founders, of the population. The size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history.

The inherent uncertainty of the complexity of the model (e.g., how many components in a mixture model) in the above problem raises a fundamental issue underlying parametric statistical modeling, now practiced throughout the machine learning community in problems such as pattern recognition and data mining. Put it simply, this is analogous to the perennial problem of ”how many clusters?” in the clustering literature, which is particularly salient in large data sets where the number of clusters needs to be relatively large and open-ended—exactly the scenario in population genomic analysis. Current approaches based on fixing the number of clusters and using an information-theoretic score to gauge the appropriate number are clearly not adequate.

In this talk, I shall discuss a new approach to haplotype inference, and to mixture modeling in general, based on a Bayesian formalism known as the Dirichlet process mixture. I show that using a Dirichlet process, one can introduce a nonparametric prior for the mixture model in a way that sets no bounds for the number of mixture components. Following the Bayesian theorem, one can obtain a posterior distribution of the mixture components—naturally corresponding to “ancestors” or “founders” in a molecular evolution setting, and other unobserved variables such as haplotypes, in a well-founded unified statistical framework. I will then report latest developments in joint multi-population haplotype inference using a hierarchical Dirichlet process (HDP) mixture model, which couples multiple heterogeneous populations by facilitating sharing of mixture components across multiple infinite mixtures each over a specific population. I will show connections of HDP to the co-clustering idea that is applicable to similar problems in data mining. Finally, I will discuss the connections of Dirichlet process to species sampling and coalescent process developed in population genetics, which offers a theoretical account for why DP mixture is perhaps the model of choice for haplotype inference and mixture modeling.