CLAIMS

What is claimed is:

1. A method, implemented at a computer system that includes one or more processors and system memory, of quantifying a nucleic acid sample comprising nucleic acid of two or more contributors, the method comprising:

(a) receiving, by the computer system, nucleic acid sequence reads obtained from the nucleic acid sample and mapped to one or more alleles at one or more polymorphism loci;

(b) determining, using the nucleic acid sequence reads and by the one or more processors, allele counts for each of the one or more alleles at the one or more polymorphism loci;

(c) applying a probabilistic mixture model to the allele counts–that uses probability distributions to model the allele counts at the one or more polymorphism loci, the probability distributions comprising an error distribution accounting for errors in the nucleic acid sequence reads at each loci of the one or more polymorphism loci;

(d) quantifying, using the probabilistic mixture model and by the one or more processors, one or more fractions of nucleic acid of the two or more contributors in the nucleic acid sample;

(e) determining a posterior probability that a specific contributor among the two or more contributors has a specific genotype at the one or more polymorphism loci.

2. The method of claim 1, further comprising determining a total number of contributors in the two or more contributors.

3. The method of claim 1, wherein one or more genotypes of the two or more contributors were unknown.

4. The method of claim 3, further comprising determining one or more allele configurations at each of the one or more polymorphism loci, each allele configuration comprising allele status of two or more alleles for each of the two or more contributors.

5. The method of claim 4, further comprising determining estimated probabilities for the one or more allele configurations.

6. The method of claim 1, wherein determining the posterior probability that the specific contributor among the two or more contributors has a specific genotype comprises:

(i) multiplying prior probabilities of genotype configurations by likelihoods of the genotype configurations, wherein the prior probabilities are based on population data of allele frequencies of the genotype configurations;

(ii) normalizing a product of (i) by a sum over genotype space, wherein the genotype space is based on possible genotypes as a function of a relationship between the two or more contributors; and

(iii) summing over genotype configurations containing the specific genotype to obtain the posterior probability.

7. The method of claim 1, wherein the specific genotype comprises a multiple-locus genotype, the method further comprising:

summing, over all contributors, a posterior probability that a contributor has the specific genotype at all loci; and

determining, based on the summed probability, the specified multiple-locus genotype appears in any contributor.

8. The method of claim 7, wherein the nucleic acid sample is a forensic sample and the data of the multiple-locus genotype is obtained from a person of interest, the method further comprising determining that the person of interest is a contributor of the nucleic acid sample.

9. The method of claim 1, wherein the nucleic acid sample comprises DNA molecules and/or RNA molecules.

10. The method of claim 9, wherein the nucleic acid sequence reads were obtained by sequencing the DNA molecules and/or RNA molecules using unique molecular indices.

11. The method of claim 1, wherein the probability distributions comprise a first binomial distribution.

12. The method of claim 11, wherein the first binomial distribution is expressed as follows:

$$n_{ij} \sim BN(n_i, p_{ij})$$

wherein

n_{ij} is an allele count for allele j at locus i;

n_i is a total allele count at locus i; and

p_{ij} is a probability parameter indicating the probability of allele j at locus i.

13. The method of claim 12, wherein the probability parameter p_{ij} is a function of:

(i) a fraction of nucleic acid of one of the two or more contributors in the nucleic acid sample, or β;

(ii) genotypes of the two or more contributors, or G; and/or

(iii) errors in the nucleic acid sequence reads, or θ.

14. The method of claim 13, wherein the probabilistic mixture model uses a beta distribution to model the errors in the nucleic acid sequence reads.

15. The method of claim 14, wherein the beta distribution is defined by a mean parameter, μ, and a concentration parameter, k.

16. The method of claim 15, wherein the concentration parameter has a prior representing different noise conditions, and the concentration parameter varies across loci.

17. The method of claim 14, wherein the probabilistic mixture model uses a combined first binomial distribution and the beta distribution to obtain a marginal distribution of n_{ij} that follows a beta-binomial distribution to generate a combined error model that models the errors in the nucleic acid sequence reads and errors in DNA extraction of the nucleic acid sample.

18. The method of claim 17, wherein the beta-binomial distribution has the form:

$$BB(n_{ij} \mid n_i, \mu, k).$$

19. The method of claim 1, wherein (c) comprises quantifying the one or more fractions of nucleic acid of the two or more contributors in the nucleic acid sample by maximizing a likelihood function of the nucleic acid sequence reads.

20. The method of claim 19, wherein (c) comprises:

calculating a plurality of likelihood values using a plurality of potential fraction values and a likelihood function of the allele counts determined in (b);

identifying a potential fraction vector associated with a maximum likelihood value; and

quantifying the one or more fractions of nucleic acid of the two or more contributors in the nucleic acid sample using the identified potential fraction vector.

21. The method of claim 19, wherein the likelihood function depends on $P(G|\pi)$, which is a prior probability of the genotype of the two or more contributors given a population allele frequency (π).

22. The method of claim 21, wherein the prior probability $P(G|\pi)$ is calculated using marginal distributions that satisfy the Hardy-Weinberg equilibrium.

23. The method of claim 21, wherein the prior probability is calculated considering a dummy allele with a fixed prior probability representing mechanistic drop-out.

24. The method of claim 11, wherein the probabilistic mixture model uses a second binomial distribution to model stutter errors in the allele data.

25. The method of claim 24, wherein the second binomial distribution is expressed as follows:

$$s_{ik} \sim BN(n_{i(k+1)}, r_i)$$

wherein

s_{ik} is a stutter allele count at locus i of a stutter allele that appears to be allele k but actually results from a stutter error of allele $k+1$;

$n_{i(k+1)}$ is an original allele count of allele $k+1$ at locus i; and

r_i is a stutter rate for locus i.

26. The method of claim 25, wherein the stutter rate *r* varies across loci and has a prior representing different noise conditions, the prior being shared across loci.

27. The method of claim 25, wherein to generate a combined error model that models the errors in the nucleic acid sequence reads and errors in DNA extraction of the nucleic acid sample fractions of nucleic acid of the two or more contributors in the nucleic acid sample using a likelihood function comprising a product of likelihoods of non-stutter allele counts and likelihoods of stutter allele counts.

28. The method of claim 25, wherein the probabilistic mixture model adds a fixed number of molecules to an allele count assigned to allele *k+1* when determining a number of molecules from which stutter can potentially originate.

29. The method of claim 1, wherein the probabilistic mixture model uses a dummy out-of-sample allele to model natural drop-out.

30. The method of claim 29, wherein the prior of the dummy out-of-sample allele is proportional to a number of unobserved alleles.

31. The method of claim 30, wherein the number of unobserved alleles is estimated by:

interpolating all integers between the shortest and longest observed integer-valued alleles, adding any observed non-integer-valued alleles, and

returning the maximum of the resulting value and a threshold value.

32. The method of claim 1, wherein the probabilistic mixture model prunes genotype configurations from data used to quantify the fractions of nucleic acid of the two or more contributors in the nucleic acid sample.

33. The method of claim 32, wherein pruning genotype configurations comprises: limiting genotype configurations that are plausible by constructing a list of required alleles and excluding loci with not enough contributors to explain all required alleles.

34. The method of claim 33, wherein the list of required alleles consists essentially of alleles having allele counts above a threshold and too high to be plausible due to stutter drop-in.

35. The method of claim 34, wherein the threshold is a sum of (i) a maximum non-stutter allele count, and (ii) a value multiplied by a count of potential stutter donor alleles.

36. The method of claim 32, wherein pruning genotype configurations comprises: removing genotype configurations that have poor matches between the allele data and expected allele counts.

37. The method of claim 36, wherein the genotype configurations that have poor matches have root mean squared error (RMSE) values larger than one or more thresholds.

38. The method of claim 1, wherein the alleles at the one or more polymorphism loci comprise single nucleotide polymorphism (SNP) alleles and/or short tandem repeat (STR) alleles.

39. A computer system comprising system memory and one or more processors configured to:

(a) receive nucleic acid sequence reads obtained from the nucleic acid sample and mapped to one or more alleles at one or more polymorphism loci;

(b) determine, using the nucleic acid sequence reads, allele counts for each of the one or more alleles at the one or more polymorphism loci;

(c) apply a probabilistic mixture model to the allele counts, and that uses probability distributions to model the allele counts at the one or more polymorphism loci, the probability distributions comprising a first probability distribution accounting for errors in the nucleic acid sequence reads, a second probability distribution accounting for nucleic acid extraction errors, and a third probability distribution accounting for nucleic acid amplification errors;

(d) quantify, using the probabilistic mixture model, one or more fractions of nucleic acid of the two or more contributors in the nucleic acid sample; and

(e) determining a posterior probability that a specific contributor among the two or more contributors has a specific genotype at the one or more polymorphism loci.

40. The system of claim 39, further comprising a tool for extracting nucleic acid from the nucleic acid sample.

41. The system of claim 39, the one or more processors are further configured to determine a total number of contributors in the two or more contributors.

42. The system of claim 39, the one or more processors are further configured to determine an allele configuration at each of the one or more polymorphism loci, the allele configuration comprising allele status of two or more alleles for each of the two or more contributors.

43. A non-transitory computer-readable medium storing program code that, when executed by one or more processors of a computer system, causes the computer system to implement a method of quantifying a nucleic acid sample comprising nucleic acid of two or more contributors, said program code comprising:

(a) code for receiving nucleic acid sequence reads obtained from the nucleic acid sample and mapped to one or more alleles at one or more polymorphism loci;

(b) code for determining, using the nucleic acid sequence reads, allele counts for each of the one or more alleles at the one or more polymorphism loci;

(c) code for applying a probabilistic mixture model to the allele counts, and that uses probability distributions to model the allele counts at the one or more polymorphism loci, the probability distributions at the one or more polymorphism loci accounting for errors in the nucleic acid sequence reads at the one or more polymorphism loci;

(d) code for quantifying, using the probabilistic mixture model, one or more fractions of nucleic acid of the two or more contributors in the nucleic acid sample; and

(e) code for determining a probability that a specific contributor among the two or more contributors has a specific genotype at the one or more polymorphism loci.