CLAIMS

1. A method of generating a hash for an omics data set, comprising:

   identifying in an omics data set a plurality of single nucleotide polymorphisms (SNPs) in respective selected locations;

   determining allele frequencies for the plurality of SNPs, and assigning respective values to the plurality of SNPs based on the allele frequencies; and

   generating an output file with a signature-hash that comprises the values for the plurality of SNPs and that further comprises metadata related to the selected locations.

2. The method of claim 1, wherein the omics data set comprises raw sequence reads.

3. The method of any one of the preceding claims, wherein the omics data set has a format selected from the group of a SAM format, BAM format, and GAR format, and/or

   wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

   wherein the values are based on a non-linear scale, and/or

   wherein the values are expressed as hexadecimal values, and/or

   wherein the values for the plurality of SNPs are in a single string, and/or

   wherein the metadata are located in a separate header, and/or

   wherein the metadata comprise scale information for the values.

4. The method of any one of the preceding claims, further comprising a step of associating the signature-hash with the omics data set.

5. The method of claim 1, wherein the omics data set has a format selected from the group of a SAM format, BAM format, and GAR format, and/or

   wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

   wherein the values are based on a non-linear scale, and/or

wherein the values are expressed as hexadecimal values, and/or

wherein the values for the plurality of SNPs are in a single string, and/or

wherein the metadata are located in a separate header, and/or

wherein the metadata comprise scale information for the values.

6. The method of claim 1, further comprising a step of associating the signature-hash with the omics data set.

7. A method of comparing a plurality of omics data sets, comprising:

obtaining or generating a first signature-hash for a first omics data set, and obtaining or generating a second signature-hash for a second omics data set;

wherein each of the first and second signature-hashes comprise a plurality of values corresponding to allele frequencies for a plurality of SNPs in selected locations of the second omics data sets and further comprise metadata related to the selected locations; and

comparing the plurality of values for the first and second signature-hashes to determine a degree of relatedness.

8. The method of claim 7, wherein the first and second omics data sets have a format selected from the group of a SAM format, BAM format, and GAR format, and/or

wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

wherein the values are based on a non-linear scale, and/or

wherein the values are expressed as hexadecimal values, and/or

wherein the first omics data set comprises the first signature-hash, and wherein the second omics data comprises the second signature-hash, and/or

wherein the degree of relatedness is based on SNP frequency, gender, ethnicity, and mutation type, and/or

wherein a predetermined degree of relatedness is indicative of common provenance.

9. The method of claim 7, wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

wherein the values are based on a non-linear scale, and/or

wherein the values are expressed as hexadecimal values, and/or

wherein the first omics data set comprises the first signature-hash, and wherein the second omics data comprises the second signature-hash, and/or

wherein the degree of relatedness is based on SNP frequency, gender, ethnicity, and mutation type, and/or

wherein a predetermined degree of relatedness is indicative of common provenance.

10. A method of identifying a single omics data set in a plurality of omics data sets having respective hashes, comprising:

obtaining or generating a single hash having a predetermined degree of relatedness to the single omics data set;

wherein each of the hashes comprises a plurality of values corresponding to allele frequencies for a plurality of SNPs in selected locations of an omics data set and further comprises metadata related to the selected locations;

comparing the plurality of values for the single hash with values of the hashes of each of the plurality of omics data sets; and

identifying the single omics data set in the plurality of omics data sets on the basis of a degree of relatedness between the values of the single hash and values of the hashes of each of the plurality of omics data sets.

11. The method of claim 10, wherein the single hash is obtained or generated from an additional omics data set, and/or

wherein the predetermined degree is identity of at least 90% of the plurality of values, and/or

wherein the predetermined degree is similarity of at least 90% of the plurality of values, and/or

wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

further comprising a step of retrieving the single omics data set, and/or

wherein the step of comparing uses the metadata.

12. The method of claim 10, wherein the predetermined degree is identity of at least 90% of the plurality of values, and/or

wherein the predetermined degree is similarity of at least 90% of the plurality of values, and/or

wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

further comprising a step of retrieving the single omics data set, and/or

wherein the step of comparing uses the metadata.

13. A method of identifying source contamination in an omics file, comprising:

providing a plurality of omics data sets having respective signature-hashes;

wherein each of the signature-hashes comprises a plurality of values corresponding to allele frequencies for a plurality of SNPs in selected locations of an omics data set and further comprises metadata related to the selected locations;

identifying at least some of the plurality of values of one of the omics data set in another omics data set.

14. The method of claim 13, wherein at least two of the plurality of omics data sets are from the same patient and are representative of at least two distinct points in time, and/or

wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

wherein the step of identifying comprises a step of subtraction of corresponding values between at least two omics data sets, and/or

further comprising a step of identifying metadata in the one of the omics data set.

15. The method of claim 13, wherein the selected locations are selected for at least one of SNP frequency, gender, ethnicity, and mutation type, and/or

wherein the step of identifying comprises a step of subtraction of corresponding values between at least two omics data sets, and/or

further comprising a step of identifying metadata in the one of the omics data set.