(cfDNA/cfRNA) obtained via liquid biopsy. In hybridization-based enrichment, tagged (e.g., biotinylated) oligonucleotide probes bearing complementarity to genomic regions of interest are mixed with a DNA sample such that nucleotide base pairing between the probes' sequences and the sequences present in the sample can occur. Thereafter the tagged probes are retrieved and sequenced. It is also possible for the hybridization probes to be physically anchored to a solid surface where they can base-pair with solution phase genomic fragments.

[0005] Numerous hybridization-based enrichment products for use in oncology are widely understood by one of ordinary skill in the art. For example, Agilent's "SureSelect Cancer All-In-One" products facilitate the identification of cancer-relevant genomic variants. Its "SureSelect Cancer All-In-One Lung Assay" encompasses 20 genes (and all of their known somatic mutations) clinically relevant to non-small cell lung cancer while the "SureSelect Cancer All-In-One Solid Tumor Assay" profiles 98 genes relevant to common solid tumor types, including lung, breast, ovarian, colorectal, prostate, sarcoma, and skin. Using such kits one can preferentially enrich these specific genes and known cancer variants and sequence them while the rest of the genome is depleted from downstream analysis.

[0006] It is important to emphasize that the intent of hybridization-based enrichment in the analysis of cancer samples is to specifically enrich regions of the human genome. It has been found that an unexpected—but useful—byproduct of oligonucleotide probe hybridization will be an appreciable level of base-pairing to non-human nucleic acids with sufficient thermodynamic stability to result in those non-human nucleic acids being isolated along with the intended human genomic DNA fragments. It has also been determined that this 'bystander' enrichment can be shown to be reproducible for a given set of hybridization probes and related data derived from targeted sequencing datasets could be employed to discover cancer-associated microbial features. Given the widespread use of hybridization-based enrichment in cancer genomics and the availability of publicly available targeted sequencing datasets, these data could be a readily available source for *in silico* discovery of microbial features with diagnostic utility, as described elsewhere herein.

[0006a] Aspects disclosed herein describe a method of identifying microbial features for determining a disease of the subject, the method comprising: (a) exposing a biological sample of the subject to one or more probes, wherein the one or more probes bind non-specifically to one or more nucleic acid molecules of the biological sample; (b) obtaining a first set of sequencing reads of the one or more nucleic acid molecules bound to the one or more probes; (c) identifying a second set of sequencing reads within the first set of sequencing reads, wherein the second set of sequencing reads comprise non-human sequencing reads obtained through non-specific

hybridizations; and (d) identifying one or more microbial features for determining the disease of the subject from the second set of sequencing reads.

[0006b] Aspects disclosed herein describe a method of validating microbial features, comprising: (a) receiving a first set of one or more microbial features of a first biological sample from a first subject with a disease determined by non-specific interactions of one or more probes with one or more nucleic acid molecules of the first biological sample; (b) training a predictive model with the first set of one or more microbial features of the first biological sample and the disease of the first subject, thereby producing a trained predictive model; (c) receiving a second set of one or more microbial features of a second biological sample of a second subject with a disease; and (d) validating the first set of one or more microbial features by comparing a predicted disease provided by the trained predictive model and the disease of the second subject, wherein the predicted disease provided by the trained predictive model is generated when the second set of one or more microbial features are provided as an input to the trained predictive model.

[0006c] Aspects disclosed herein describe a method, comprising: (a) exposing a biological sample of a first subject with a first disease to one or more probes, wherein the one or more probes bind non-specifically to one or more nucleic acid molecules of the biological sample; (b) sequencing the one or more nucleic acid molecules bound to the one or more probes, thereby generating one or more sequencing reads; (c) mapping the one or more sequencing reads to a genome database, thereby identifying one or more non-human sequencing reads; and (d) generating a predictive model for predicting a second disease of a second subject, wherein the predictive model is trained with one or more microbial features of the one or more non-human sequencing reads and the first disease of the first subject.

[0006d] Aspects disclosed herein describe a method, comprising: (a) exposing a biological sample of a subject with a disease to one or more probes, wherein the one or more probes bind non-specifically to one or more nucleic acid molecules of the biological sample; (b) identifying one or more sequencing reads of the one or more nucleic acid molecule bound to the one or more probes; (c) mapping the one or more sequencing reads to a genome database, thereby identifying one or more non-human sequencing reads of the one or more sequencing reads; and (d) identifying one or more microbial features of the one or more non-human sequencing reads to classify the subject's disease.

[0006e] Aspects disclosed herein describe a system, comprising: (a) one or more processors; and (b) a non-transient computer readable storage medium comprising software, wherein the software comprises executable instructions that, as a result of execution, cause the one or more processors of a computer system to: (i) receive one or more nucleic acid molecule sequencing reads of subject's

biological sample, wherein the subject has a disease, and wherein the one or more nucleic acid molecule sequencing reads are obtained from one or more nucleic acid molecules enriched by one or more probes exposed to the subject's biological sample; (ii) map the one or more nucleic acid molecule sequencing reads to a genome database, thereby identifying one or more non-human sequencing reads of the one or more nucleic acid molecule sequencing reads; and (iii) identify one or more microbial features of the one or more non-human sequencing reads to classify the subject's disease.

[0007] Aspects disclosed herein describe a method of identifying microbial features for diagnosing cancer in a subject based on the analysis of hybridization-based enrichment sequencing data comprising: (a) obtaining hybridization capture enrichment sequencing reads derived from a biological sample; (b) filtering the sequencing reads with a build of a genome database to isolate non-human sequencing reads; (c) generating taxonomic assignments and their associated abundances for the non-human sequencing reads; (d) identifying and removing contaminating microbial features of the taxonomically assigned non-human sequencing reads while retaining other