

on the mutation profile. In some embodiments, the steps may be repeated at least 2, 10 or 100 times.

[0004a] In one aspect, there is provided a method to identify a BRAF mutation in a biological sample of a subject having or suspected of having thyroid cancer, the method comprising:

(a) training a tissue classifier to yield a trained algorithm, wherein the training comprises:

(i) assaying a plurality of training samples for expression levels of biomarkers selected from Tables 24-27 to obtain a training set comprising expression data, and (ii) training said tissue classifier with said training set, wherein said plurality of training samples comprise fine needle aspirate (FNA) samples that are independent from said biological sample;

(b) assaying said biological sample of said subject having or suspected of having thyroid cancer for expression levels of said plurality of biomarkers of (a); and

(c) applying said trained algorithm of (a) to said one or more expression levels of (b), wherein said applying comprises: (i) comparing the one or more expression levels of (b) to said expression data of (a); and (ii) classifying the biological sample as containing or not containing the BRAF mutation at a specificity of at least about 95%.

[0005] In some embodiments, one or more references comprise frequencies of variation for single base pairs in a reference sequence, wherein the frequencies of variation in the reference sequence are derived from at least 1000 individuals. In some embodiments one or more references of comprise frequencies of variation for single base pairs in a reference sequence, wherein the frequencies of variation in the reference sequence are derived from a known cancer. In some embodiments one or more references comprise frequencies of variation for single base pairs in a reference sequence, wherein the frequencies of variation in the reference sequence are derived from at least 40 samples.

[0006] In some embodiments a call score is assigned to each mutation identified in the RNA. In some embodiments, a mutation profile of is generated using the COSMIC database of known sites of somatic variations in cancer.

[0007] In some embodiments the identification of the presence or absence of one or more mutations is at least 90%, 95%, or 100% accurate.

[0008] This disclosure also provides for a method for detecting and normalizing 3'-5' amplification bias in microarray sample data generated from a nucleic acid sample from a

subject, the method comprising obtaining a biological sample from a subject, wherein the biological sample comprises a nucleic acid sample; amplifying the nucleic acid sample to generate one or more amplicons, wherein the nucleic acid sample is amplified with the aid of one or more probes; generating a nucleic acid sequence read for an individual amplicon among the one or more amplicons; for each individual amplicon among the one or more amplicons, calculating, with the aid of a computer processor, the extent of a 3' bias for a given probe among the one or more probes upon a comparison of a nucleic acid sequence of the given probe to a nucleic acid sequence of the individual amplicon generated in (c); and applying a normalization procedure to correct for the 3' bias for a given probe.

[0009] In some embodiments, the nucleic acid is an mRNA transcript. In some embodiments calculating the extent of the 3' bias further comprises determining the effective distance from the 3' end of the mRNA transcript and the given probe. In some embodiments calculating the extent of the 3' bias further comprises determining the effective distance from one or more sites or sequences in the mRNA transcript and the given probe. In some embodiments calculating the extent of the 3' bias further comprises calculating a distance or median weighted distance between the given probe and one or

CLAIMS

1. A method to identify a BRAF mutation in a biological sample of a subject having or suspected of having thyroid cancer, the method comprising:

(a) training a tissue classifier to yield a trained algorithm, wherein the training comprises: (i) assaying a plurality of training samples for expression levels of biomarkers selected from Tables 24-27 to obtain a training set comprising expression data, and (ii) training said tissue classifier with said training set, wherein said plurality of training samples comprise fine needle aspirate (FNA) samples that are independent from said biological sample;

(b) assaying said biological sample of said subject having or suspected of having thyroid cancer for expression levels of said plurality of biomarkers of (a); and

(c) applying said trained algorithm of (a) to said one or more expression levels of (b), wherein said applying comprises: (i) comparing the one or more expression levels of (b) to said expression data of (a); and (ii) classifying the biological sample as containing or not containing the BRAF mutation at a specificity of at least about 95%.

2. The method of claim 1, wherein classifying the biological sample as containing or not containing said BRAF mutation further comprises a prediction of the presence or absence of a disease associated with said BRAF mutation.

3. The method of claim 1, wherein the trained algorithm comprises a linear support vector machine classifier.

4. The method of claim 1, wherein the plurality of training samples comprises tissue samples, fine needle aspirations, or a combination thereof.

5. The method of claim 1, further comprising classifying the biological sample as containing or not containing whole blood using a clinical classifier comprising a plurality of genes selected from Table 11 or Table 12.

6. The method of claim 2, wherein said disease is thyroid cancer or lymphoma.

7. The method of claim 1, wherein the BRAF mutation is a BRAFV600E mutation.

8. The method of claim 1, further comprising classifying the biological sample as containing or not containing follicular tissue or cells using a clinical classifier comprising a plurality of genes selected from Table 14 or Table 15.

9. The method of claim 2, wherein said disease is thyroid cancer, and wherein said method further comprises classifying the biological sample as containing or not containing said thyroid cancer using a clinical classifier comprising a plurality of genes selected from Table 2, Table 9 or Table 10.

10. The method of claim 2, wherein said classifying comprises an estimate of the proportion of said disease and/or a specific tissue type in said biological sample.
11. The method of claim 1, wherein the biological sample comprises a needle aspiration sample, a fine needle aspiration sample, a core needle biopsy, a vacuum assisted biopsy, a large core biopsy, an incisional biopsy, an excisional biopsy, a punch biopsy, a shave biopsy, or a skin biopsy.
12. The method of claim 1, wherein the biological sample is a fine needle aspiration of thyroid tissue.
13. The method of claim 1, wherein the one or more expression levels are measured by microarray, SAGE, blotting, RT-PCR, sequencing, and/or quantitative PCR.
14. The method of claim 2, wherein the biomarker comprises RNA.
15. The method of claim 14, wherein the RNA is mRNA, rRNA, tRNA, and/or miRNA.
16. The method of claim 14, wherein the biomarker corresponds to a gene over-expressed in the disease.
17. The method of claim 1, wherein the classifying the biological sample is with at least 95% accuracy.
18. The method of claim 1, wherein the classifying the biological sample is with at least 99% accuracy.
19. The method of claim 1, wherein the classifying the biological sample is with 100% accuracy.
20. The method of claim 1, further comprising classifying the biological sample as containing or not containing non-thyroid cells prior to the classifying of (c).
21. The method of claim 20, wherein, based on the pre-screening, a rate of false positives returned by the trained algorithm is reduced as compared.
22. The method of claim 2, wherein said disease is thyroid cancer, and wherein said trained algorithm is used as a diagnostic for the thyroid cancer.
23. The method of claim 1, further comprising classifying the biological sample as containing or not containing lymphoma using a clinical classifier comprising a plurality of genes selected from Table 1.
24. The method of claim 2, further comprising classifying the biological sample as having an aggressive prognosis based upon the comparing of (c).