

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 BRAFV600E mutation in a biological sample of a subject, 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, wherein said plurality of training samples comprises a first set of samples containing said BRAFV600E mutation and a second set of samples not containing said BRAFV600E mutation, and (ii) performing supervised learning on said tissue classifier with said training set to yield said trained algorithm, 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 for expression levels of said biomarkers of (a); and
- (c) applying said trained algorithm of (a) to said expression levels of (b), wherein said applying comprises: (i) comparing said expression levels of (b) to said expression data of (a); and (ii) classifying said biological sample as containing or not containing the BRAFV600E 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

CLAIMS

1. A method to identify a BRAFV600E mutation in a biological sample of a subject, 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, wherein said plurality of training samples comprises a first set of samples containing said BRAFV600E mutation and a second set of samples not containing said BRAFV600E mutation, and (ii) performing supervised learning on said tissue classifier with said training set to yield said trained algorithm, 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 for expression levels of said biomarkers of (a); and

(c) applying said trained algorithm of (a) to said expression levels of (b), wherein said applying comprises: (i) comparing said expression levels of (b) to said expression data of (a); and (ii) classifying said biological sample as containing or not containing the BRAFV600E mutation at a specificity of at least about 95%.

2. The method of claim 1, wherein classifying said biological sample as containing or not containing said BRAFV600E mutation further comprises a prediction of a presence or absence of a disease associated with said BRAFV600E mutation.

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

4. The method of claim 1, wherein said plurality of training samples comprises tissue samples.

5. The method of claim 1, further comprising classifying said 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, further comprising classifying said 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.
8. The method of claim 2, wherein said disease is thyroid cancer, and wherein said method further comprises classifying said 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.
9. The method of claim 8, wherein said classifying comprises an estimate of the proportion of said disease and/or a specific tissue type in said biological sample.
10. The method of claim 1, wherein said 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.
11. The method of claim 1, wherein said biological sample is a fine needle aspiration of thyroid tissue.
12. The method of claim 1, wherein said expression levels are measured by microarray, SAGE, blotting, RT-PCR, sequencing, and/or quantitative PCR.
13. The method of claim 2, wherein said biomarkers comprise RNA.
14. The method of claim 13, wherein said RNA is mRNA, rRNA, tRNA, and/or miRNA.
15. The method of claim 13, wherein said biomarkers correspond to a gene over-expressed in said disease.

16. The method of claim 1, wherein said classifying said biological sample is with at least 95% accuracy.
17. The method of claim 1, wherein said classifying said biological sample is with at least 99% accuracy.
18. The method of claim 1, wherein said classifying said biological sample is with 100% accuracy.
19. The method of claim 1, further comprising classifying said biological sample as containing or not containing non-thyroid cells prior to the classifying of (c).
20. The method of claim 19, wherein, based on said pre-screening, a rate of false positives returned by said trained algorithm is reduced.
21. The method of claim 2, wherein said disease is thyroid cancer, and wherein said trained algorithm is used as a diagnostic for said thyroid cancer.
22. The method of claim 1, further comprising classifying said biological sample as containing or not containing lymphoma using a clinical classifier comprising a plurality of genes selected from Table 1.
23. The method of claim 2, further comprising classifying said biological sample as having an aggressive prognosis based upon said comparing of (c).