

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, 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 one or more expression levels of a biomarker 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 for one or more 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 statistical confidence level 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

[0009] 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.

[0010] 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

WHAT IS CLAIMED IS:

1. A method to diagnose and/or treat a subject suspected of having a disease, the method comprising:
 - (a) obtaining a biological sample from said subject;
 - (b) assaying an expression level of one or more gene expression products in the biological sample;
 - (c) using one or more clinical classifiers to compare said expression level of (b) to a reference expression level of a plurality of genes selected from Tables 24-27 to generate a comparison of expression levels, wherein the comparison is performed by a computer processor;
 - (d) classifying the biological sample as containing or not containing said disease and/or a specific tissue type based upon said comparison of said one or more clinical classifiers, to yield a classification of said biological sample; and
 - (e) diagnosing and/or treating said subject based upon the classification of (d).
2. The method of claim 1, wherein classifying the biological sample as containing or not containing said disease further comprises a prediction of the presence or absence of a mutation associated with said disease.
3. The method of claim 1, wherein the comparison is performed using a trained algorithm or an algorithm that comprises a linear support vector machine classifier.
4. The method of claim 3, wherein the trained algorithm is trained using 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 1, wherein said disease is thyroid cancer or lymphoma.
7. The method of claim 6, wherein said disease is thyroid cancer or lymphoma, and the mutation associated with thyroid cancer or lymphoma 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 1, wherein said disease is thyroid cancer, and wherein said method further comprises classifying the biological sample as containing or not

containing 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 1, wherein classifying the biological sample as containing or not containing said disease and/or a specific tissue type based upon said comparisons of said one or more clinical classifiers further provides an estimate of the proportion of said disease and/or specific tissue type in said sample.

11. The method of claim 1, wherein the biological sample is obtained by needle aspiration, fine needle aspiration, core needle biopsy, vacuum assisted biopsy, large core biopsy, incisional biopsy, excisional biopsy, punch biopsy, shave biopsy, or 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 expression level is assayed by microarray, SAGE, blotting, RT-PCR, sequencing, and/or quantitative PCR.

14. The method of claim 1, wherein the one or more gene expression products are RNA.

15. The method of claim 14, wherein the RNA is mRNA, rRNA, tRNA, and/or iRNA.

16. The method of claim 14, wherein at least one of the gene expression product corresponds to a gene over-expressed in the disease.

17. The method of claim 1, wherein the classifying differentiates disease containing samples from non disease containing samples with at least 95% accuracy.

18. The method of claim 1, wherein the classifying differentiates disease containing samples from non disease containing samples with at least 99% accuracy.

19. The method of claim 1, wherein the classifying differentiates disease containing samples from non disease containing samples with 100% accuracy.

20. The method of claim 1, further comprising pre-screening the biological sample prior to classifying with the one or more clinical classifiers.

21. The method of claim 20, wherein, based on the pre-screening, the rate of false positives returned by the one or more clinical classifiers is reduced as compared.

22. The method of claim 1, wherein said disease is thyroid cancer, and wherein said one or more clinical classifiers are used as a diagnostic for thyroid cancer.

23. The method of claim 1, wherein the treating of (c) further comprises conducting surgery.

24. 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.

25. The method of claim 1, wherein said one or more clinical classifiers comprise at least two or more clinical classifiers, which at least two or more clinical classifiers are used for classifying the biological sample as containing or not containing said disease and/or a specific tissue type.

26. The method of claim 1 or 2, further comprising classifying the sample as having an aggressive prognosis based upon said comparison.

27. A method for diagnosing and/or treating a subject suspected of having a disease, the method comprising:

(a) isolating an ribonucleic acid (RNA) sample from a biological sample obtained from said subject;

(b) identifying one or more mutations within a first region of interest in said RNA sample;

(c) comparing, using a computer processor, a frequency of variation for each base pair position in said first region of interest of said RNA sample to one or more references to identify one or more mutations that are correlated with said disease;

(d) comparing said one or more mutations identified in (b) to said one or more mutations identified in (c) to identify the presence or absence of at least one mutation in said RNA sample;

(e) repeating (a)-(d) for a second region of interest of said RNA sample to generate a mutation profile for said RNA, wherein said second region of interest is different from said first region of interest; and

(f) diagnosing and/or treating said subject based on said mutation profile.

28. The method of claim 27, wherein said one or more references of (c) comprise frequencies of variation for single base pairs in a reference sequence, wherein the frequencies of variation in the reference sequence are associated with at least 1000 individuals.

29. The method of claim 27, wherein said one or more references of (c) comprise frequencies of variation for single base pairs in a reference sequence, wherein the frequencies of variation in the reference sequence are associated with a known disease.

30. The method of claim 27, wherein said biological sample is obtained from a tissue suspected of having the disease, wherein said one or more references of (c) comprise frequencies of variation for single base pairs in a reference sequence, and wherein the frequencies of variation in the reference sequence are associated with at least 40 samples from a tissue of a type that is different from a type of said tissue suspected of having said disease.
31. The method of claim 27, further comprising assigning a call score to each mutation identified in said RNA sample.
32. The method of claim 27, wherein the mutation profile of (e) is generated using the COSMIC database of known sites of somatic variations in cancer.
33. The method of claim 27, wherein the identification of the presence or absence of one or more mutations is at least 90% accurate.
34. The method of claim 27, wherein the identification of the presence or absence of one or more mutations is at least 95% accurate.
35. The method of claim 27, wherein the identification of the presence or absence of one or more mutations is at least 99% accurate.
36. The method of claim 27, further comprising repeating (e) at least 2 times.
37. The method of claim 27, further comprising repeating (e) at least 10 times.
38. The method of claim 27, further comprising repeating (e) at least 100 times.
39. 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:
- (a) obtaining a biological sample from a subject, wherein said biological sample comprises a nucleic acid sample;
 - (b) using one or more probes, amplifying said nucleic acid sample to generate one or more amplicons;
 - (c) generating a nucleic acid sequence read for an individual amplicon among said one or more amplicons;
 - (d) for each individual amplicon among said one or more amplicons, calculating, with the aid of a computer processor, the extent of a 3' bias for a given probe among said one or more probes upon a comparison of a nucleic acid sequence of said given probe to a nucleic acid sequence of said individual amplicon generated in (c); and
 - (e) applying a normalization procedure to correct for said 3' bias for a given probe.

40. The method of claim 39, wherein the nucleic acid sample is an mRNA transcript.
41. The method of claim 40, wherein calculating the extent of the 3' bias further comprises determining an effective distance from the 3' end of the mRNA transcript and said given probe.
42. The method of claim 40, wherein calculating the extent of the 3' bias further comprises determining an effective distance from one or more sites or sequences in said mRNA transcript and said given probe.
43. The method of claim 40, wherein calculating the extent of the 3' bias further comprises calculating a distance or median weighted distance between said given probe and one or more downstream polyA sites or sequences within the mRNA transcript, wherein the weighted distance is determined by read counts associated with each polyA site in the mRNA transcript.
44. The method of claim 40, wherein calculating the extent of the 3' bias further comprises comparing variability of paired intensity profiles of two or more identical probes, wherein the intensity profiles are obtained from two or more independent sets of microarray data, wherein each microarray data set is generated from an identical biological sample.
45. The method of claim 44, wherein comparing variability of paired intensity profiles of two or more identical probes further comprises performing a per-transcript alignment of probes within the mRNA transcript to calculate an effective distance.
46. The method of claim 39, wherein the normalization procedure further comprises generating a normalization target distribution.
47. The method of claim 39, wherein the normalization procedure further comprises quantile normalization, wherein probes are grouped into bins, and wherein a quantile normalization is applied to each probe within each bin to normalize the median intensity of probes across a bin.
48. The method of claim 39, wherein the normalization procedure removes application bias from sample data.
49. The method of claim 39, further comprising applying summarization methods to normalized probe intensities.
50. The method of claim 49, wherein applying said summarization methods to normalized probe intensities improves the detection of differential gene expression in the microarray sample data.

51. A method for the detection of heterogeneity in microarray data, the method comprising:

- (a) generating hypothetical microarray data from a mixture of one or more samples in silico;
- (b) generating one or models from said hypothetical microarray data;
- (c) obtaining microarray data from a mixture of one or more samples performed in vitro;
- (d) comparing, with the aid of a computer processor, said one or more models of (b) to the data obtained in (c); and
- (e) based upon the comparison of (d), assessing the strength of said one or more models.

52. The method of claim 51, wherein assessing the strength of said one or more models comprises comparing a mean squared error between the model generated in (b) and the data obtained in (c).

53. The method of claim 51, further comprising selecting a model generated in (b), wherein said selecting is determined by a predictive ability of the model.

54. The method of claim 53, wherein the predictive ability of the model is determined by comparing the model to experimental data.

55. The method of claim 51, further comprising, using the one or more models to improve selectivity and/or sensitivity in the detection of heterogeneity in a sample.

56. A method to diagnose and/or treat a subject suspected of having a disease, the method comprising:

- (a) obtaining a biological sample from said subject;
- (b) assaying an expression level of one or more gene expression products in said biological sample;
- (c) using one or more clinical statistics to compare said expression level of (b) to a reference expression level of a plurality of genes from Table 11 and/or Table 12 to generate a comparison of expression levels, wherein the comparison is performed by a computer processor;
- (d) classifying said biological sample as containing or not containing a blood component based upon said comparison to yield a classification of said biological sample; and
- (e) diagnosing and/or treating said subject based upon the classification of (d).

57. The method of claim 56, wherein the comparison is performed using a trained algorithm or an algorithm that comprises a linear support vector machine classifier.
58. The method of claim 57, wherein the trained algorithm is trained using tissue samples, fine needle aspirations, or a combination thereof.
59. The method of claim 56, wherein the biological sample is obtained by needle aspiration, fine needle aspiration, core needle biopsy, vacuum assisted biopsy, large core biopsy, incisional biopsy, excisional biopsy, punch biopsy, shave biopsy, or skin biopsy.
60. The method of claim 56, wherein the biological sample is a fine needle aspiration of thyroid tissue.
61. The method of claim 56, wherein the expression level is assayed by microarray, SAGE, blotting, RT-PCR, sequencing, and/or quantitative PCR.
62. The method of claim 56, wherein the one or more gene expression products comprise RNA.
63. The method of claim 62, wherein the RNA is mRNA, rRNA, tRNA, and/or miRNA.
64. The method of claim 56, wherein the biological sample is classified as containing or not containing a blood component at an accuracy of least 95% accuracy.
65. The method of claim 56, wherein the biological sample is classified as containing or not containing a blood component at an accuracy of least 99% accuracy.
66. The method of claim 56, wherein the comparison of (c) yields a reduced rate of false positive identification of non-blood tissue.
67. A method to diagnose and/or treat a subject suspected of having a disease, the method comprising:
- (a) obtaining a biological sample from said subject;
 - (b) assaying an expression level of one or more gene expression products in the biological sample;
 - (c) using one or more clinical statistics to compare said expression level of (b) to a reference expression level of a plurality of genes of Table 14 and/or Table 15 to generate a comparison of expression levels, wherein the comparison is performed by a computer processor;
 - (d) classifying the biological sample as containing or not containing follicular tissue based upon said comparison to yield a classification of said biological sample; and

- (c) diagnosing and/or treating said subject based upon the classification of (d).
68. The method of claim 67, wherein the comparison is performed using a trained algorithm or an algorithm that comprises a linear support vector machine classifier.
69. The method of claim 68, wherein the trained algorithm is trained using tissue samples, fine needle aspirations, or a combination thereof.
70. The method of claim 67, wherein the biological sample is obtained by needle aspiration, fine needle aspiration, core needle biopsy, vacuum assisted biopsy, large core biopsy, incisional biopsy, excisional biopsy, punch biopsy, shave biopsy, or skin biopsy.
71. The method of claim 67, wherein the biological sample is a fine needle aspiration of thyroid tissue.
72. The method of claim 67, wherein the expression level is assayed by microarray, SAGE, blotting, RT-PCR, sequencing, and/or quantitative PCR.
73. The method of claim 67, wherein the one or more gene expression products comprise RNA.
74. The method of claim 73, wherein the RNA is mRNA, rRNA, tRNA, and/or miRNA.
75. The method of claim 67, wherein the biological sample is classified as containing or not containing follicular tissue at an accuracy of at least 95%.
76. The method of claim 67, wherein the biological sample is classified as containing or not containing follicular tissue at an accuracy of at least 99%.
77. The method of claim 67, wherein the comparison of (c) yields a reduced rate of false positive identification of non-follicular tissue.
78. The method of claim 1, 27, 56 or 67, wherein the disease is cancer.

CLAIMS

WHAT IS CLAIMED IS:

1. A method to identify a BRAF 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 one or more expression levels of a biomarker 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 for one or more 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 statistical confidence level 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 pre-screening the biological sample 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 1, wherein said trained algorithm comprises at least two or more trained algorithms, which at least two or more trained algorithms are used for classifying the biological sample as containing or not containing said BRAF mutation.

25. The method of claim 2, further comprising classifying the biological sample as having an aggressive prognosis based upon the comparing of (c).