Optimization of Multigene Analysis of Tumor Samples

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 61/428,568, filed December 30, 2010; U.S. Provisional Application No. 61/428,602, filed December 30, 2010; U.S. Provisional Application No. 61/428,626, filed December 30, 2010; U.S. Provisional Application No. 61/428,638, filed December 30, 2010; U.S. Provisional Application No. 61/467,798, filed March 25, 2011; U.S. Provisional Application No. 61/467,748, filed March 25, 2011; U.S. Provisional Application No. 61/486,006, filed May 13, 2011; U.S. Provisional Application No. 61/486,012, filed May 13, 2011; U.S. Provisional Application No. 61/486,026, filed May 13, 2011; U.S. Provisional Application No. 61/486,033, filed May 13, 2011; and U.S. Provisional Application No. 61/552,884, filed October 28, 2011. The contents of all these prior applications are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

The invention relates to optimized methods for analyzing nucleic acids from tumor samples, *e.g.*, methods having integrated optimized nucleic acid selection, read alignment, and mutation calling.

SUMMARY ~~OF THE INVENTION~~

Methods disclosed herein provide integration of one or more optimized elements including bait-based selection, alignment, and mutation calling, as applied, *e.g.*, to optimized sets of subgenomic intervals, *e.g.*, sets of cancer related segments of the genome described herein. Methods described herein provide for next generation sequencing (NGS)-based analysis of tumors that can be optimized on a cancer-by-cancer, gene-by-gene and/or site-by-site basis.

In one aspect, there is provided a method of analyzing a tumor sample for a somatic mutation, comprising:

(a) acquiring a library comprising a plurality of nucleic acid molecules from a tumor sample;

(b) enriching the library for preselected sequences by contacting the library with a plurality of bait sets to provide selected tumor nucleic acid molecules, thereby producing a library catch;

(c) acquiring a read for a subgenomic interval comprising a somatic mutation from a tumor nucleic acid molecule from said library or library catch by a next generation sequencing method;

(d) aligning said read by an alignment method; and

(e) assigning a nucleotide value from said read for a preselected nucleotide position,

wherein a read from each of X unique subgenomic intervals is aligned with a unique alignment method, wherein unique subgenomic interval means different from the other X-1 subgenomic intervals, and wherein unique alignment method means different from the other X-1 alignment methods, and X is at least 2, 10, 15, 20, 30, 50, 100, 500, or 1,000,

wherein said method of analyzing a tumor sample comprises application of two or more of the following alignment methods:

(i) a first alignment method that is selected responsive to, or is optimized for, a variant which appears at a frequency of 5% or less, in cells in the tumor sample, the nucleic acid molecules of the library, or the nucleic acid molecules of the library catch;

(ii) a second alignment method that is selected responsive to, or is optimized for, a variant which appears at a frequency of greater than 5% and up to 10%, 15% or 20%, in cells in the tumor sample, the nucleic acid molecules of the library, or the nucleic acid molecules of the library catch;

(iii) a third alignment method that is selected responsive to, or is optimized for, a variant which is:

(1) a pharmacogenomic SNP that is associated with the ability of patient to respond to or metabolize a drug;

(2) a genomic SNP that uniquely identifies (fingerprints) a patient; or

(3) a genomic SNP/locus that is used to assess copy number gains/losses of genomic DNA and LOH;

(iv) a fourth alignment method that is selected responsive to, or is optimized for, a structural breakpoint; and

(v) a fifth alignment method that is selected responsive to, or is optimized for, a target for which sparse coverage can improve the ability to detect copy number changes.

In one aspect, there is provided a method of analyzing a tumor sample for a somatic mutation, comprising:

(a) acquiring a library comprising a plurality of nucleic acid molecules from a tumor sample;

(b) enriching the library for preselected sequences by contacting the library with a plurality of bait sets to provide selected tumor nucleic acid molecules, thereby producing a library catch;

(c) acquiring a read for a subgenomic interval comprising a somatic mutation from a tumor nucleic acid molecule from said library or library catch by a next generation sequencing method;

(d) aligning said read by an alignment method; and

(e) assigning a nucleotide value from said read for the preselected nucleotide position, thereby analyzing said tumor sample,

wherein the alignment method comprises aligning the read with a rearrangement alignment sequence which is preselected to align with a preselected rearrangement,

wherein a read from each of at least X unique subgenomic intervals is aligned with a unique alignment method and X is 2, 10, 15, 20, 30, 50, 100, 500, or 1,000,

wherein unique subgenomic interval means different from the other X-1 subgenomic intervals, and wherein unique alignment method means different from the other X-1 alignment methods, and

wherein the reads from the at least X unique subgenomic intervals are analyzed using at least X unique preselected rearrangement alignment sequences, and wherein unique means different from the other X-1.

In one aspect, there is provided a method of analyzing a tumor sample comprising a somatic mutation, comprising:

(a) acquiring a library comprising a plurality of nucleic acid molecules from a tumor sample;

(b) enriching the library for preselected sequences by contacting the library with a plurality of bait sets to provide selected tumor nucleic acid molecules, thereby providing a library catch;

(c) acquiring a read for a subgenomic interval comprising a somatic mutation from a tumor nucleic acid molecule from said library or library catch by a next generation sequencing method;

(d) aligning said read by an alignment method; and

(e) assigning a nucleotide value from said read for the preselected nucleotide position,

wherein a nucleotide value assigned for a nucleotide position in each of X unique subgenomic intervals by a unique calling method,

wherein unique subgenomic interval means different from the other X-1 subgenomic intervals, and wherein unique calling method means different from the other X-1 calling methods, and X is at least 2.

In one aspect, there is provided a method of analyzing a tumor sample, comprising:

(a) acquiring a library comprising a plurality of tumor nucleic acid molecules from a tumor sample;

(b) contacting the library with a bait set to provide selected tumor nucleic acid molecule, thereby providing a library catch;

(c) acquiring a read for a subgenomic interval comprising a somatic mutation from a tumor nucleic acid molecule from said library or library catch by a next generation sequencing method;

(d) aligning said read by an alignment method; and

(e) assigning a nucleotide value from said read for the preselected nucleotide position,

wherein the method comprises contacting the library with at least two of the following bait sets:

(i) a first bait set that selects a high-level target chosen from one or more tumor nucleic acid molecules that comprise a subgenomic interval comprising a somatic mutation that appears at a frequency of about 5% or less of the cells from the sample harboring the alteration in their genome;

(ii) a second bait set that selects a mid-level target chosen from one or more tumor nucleic acid molecules that comprise a subgenomic interval comprising a somatic mutation that appears at a frequency of about 10% or higher of the cells from the sample harboring the alteration in their genome;

(iii) a third bait set that selects a low-level target chosen from one or more pharmacogenomic (PGx) nucleic acid molecules that comprise a subgenomic interval chosen from one or more of:

a) a PGx single nucleotide polymorphism (SNP) that distinguishes the ability of a patient to metabolize different drugs,

b) a genomic SNP that uniquely identifies (fingerprints) a patient, or

c) a genomic SNP/locus that is used to assess copy number gains or losses of genomic DNA and loss-of-heterozygosity (LOH);

(iv) a fourth bait set that selects a nucleic acid molecule that comprises an intron sequence that detects a structural breakpoint; or

(v) a fifth bait set that selects a one-copy deletion of several terminal exons,

wherein each bait set of said plurality has a unique preselected efficiency for selection for its target as compared with the other bait sets in the plurality.

Accordingly, in one embodiment~~aspect~~, the disclosure~~invention~~ features a method of analyzing a tumor sample. The method comprises:

Alignment

Methods disclosed herein can integrate the use of multiple, individually tuned, alignment methods or algorithms to optimize performance in sequencing methods, particularly in methods that rely on massively parallel sequencing of a large number of diverse genetic events in a large number of diverse genes, *e.g.*, methods of analyzing tumor samples. In embodiments, multiple alignment methods that are individually customized or tuned to each of a number of variants in different genes are used to analyze reads. In embodiments, tuning can be a function of (one or more of) the gene (or other subgenomic inerval) being sequenced, the tumor type in the sample, the variant being sequenced, or a characteristic of the sample or the subject. The selection or use of alignment conditions that are individually tuned to a number of subgenomic intervals to be sequenced allows optimization of speed, sensitivity and specificity. The method is particularly effective when the alignments of reads for a relatively large number of diverse subgenomic intervals are optimized.

Accordingly, in one embodiment~~aspect~~, the disclosure~~invention~~ features, a method of analyzing a sample, *e.g.*, a tumor sample. The method comprises:

(a) acquiring a library comprising a plurality members from a sample, *e.g.*, a plurality of tumor members from a tumor sample;

(b) optionally, enriching the library for preselected sequences, *e.g.*, by contacting the library with a bait set (or plurality of bait sets) to provide selected members (sometimes referred to herein as library catch);

(c) acquiring a read for a subgenomic interval from a member, *e.g.*, a tumor member from said library or library catch, *e.g.*, by a method comprising sequencing, *e.g.*, with a next generation sequencing method;

(d) aligning said read by an alignment method, *e.g.*, an alignment method described herein; and

(e) assigning a nucleotide value (*e.g.*, calling a mutation, e.g., with a Bayesian method) from said read for the preselected nucleotide position,

thereby analyzing said tumor sample,

wherein

a read from each of X unique subgenomic intervals is aligned with a unique alignment method, wherein unique subgenomic interval means different from the other X-1 subgenoimc

likely to result in an alignment with a read for a preselected variant, *e.g.*, a rearrangement, *e.g.*, an insertion, deletion, or translocation.

These and other alignment methods are discussed in more detail elsewhere herein, *e.g.*, in the section entitled "Alignment Module." Elements of that module can be included in methods of analyzing a tumor. In embodiments, alignment methods from the "Alignment Module" are combined with mutation calling methods from the "Mutation Calling Module" and/or a bait set from the "Bait Module." The method can be applied to set of subgenomic intervals from the "Gene Selection Module."

Mutation Calling

Methods disclosed herein can integrate the use of customized or tuned mutation calling parameters to optimize performance in sequencing methods, particularly in methods that rely on massively parallel sequencing of a large number of diverse genetic events in a large number of diverse genes, *e.g.*, from tumor samples. In embodiments of the method mutation calling for each of a number of preselected subgenomic intervals is, individually, customized or fine tuned. The customization or tuning can be based on one or more of the factors described herein, *e.g.*, the type of cancer in a sample, the gene in which subgenomic interval to be sequenced is located, or the variant to be sequenced. This selection or use of alignment conditions finely tuned to a number of subgenomic intervals to be sequenced allows optimization of speed, sensitivity and specificity. The method is particularly effective when the alignment of reads for a relatively large number of diverse subgenomic intervals is optimized.

Accordingly, in one embodiment~~aspect~~, the disclosure~~invention~~ features, a method of analyzing a sample, *e.g.*, a tumor sample. The method comprises:

(a) acquiring a library comprising a plurality members from a sample, *e.g.*, a plurality of tumor members from the sample, *e.g.*, the tumor sample;

(b) optionally, enriching the library for preselected sequences, *e.g.*, by contacting the library with a bait set (or plurality of bait sets) to provide selected members, *e.g.*, a library catch;

(c) acquiring a read for a subgenomic interval from a member, *e.g.*, a tumor member from said library or library catch, *e.g.*, by a method comprising sequencing, *e.g.*, with a next generation sequencing method;

Accordingly, in one embodiment~~aspect~~, the disclosure~~invention~~ features, a method of analyzing a sample, *e.g.*, a tumor sample. The method comprises:

(a) acquiring a library comprising a plurality of members (*e.g.*, target members) from a sample, *e.g.*, a plurality of tumor members from a tumor sample;

(b) contacting the library with a bait set to provide selected members (*e.g.*, a library catch);

(c) acquiring a read for a subgenomic interval from a member, *e.g.*, a tumor member from said library or library catch, *e.g.*, by a method comprising sequencing, *e.g.*, with a next generation sequencing method;

(d) aligning said read by an alignment method, *e.g.*, an alignment method described herein; and

(e) assigning a nucleotide value (*e.g.*, calling a mutation, *e.g.*, with a Bayesian method or a method described herein) from said read for the preselected nucleotide position,

thereby analyzing said tumor sample,

wherein the method comprises contacting the library with a plurality, *e.g.*, at least two, three, four, or five, of baits or bait sets, wherein each bait or bait set of said plurality has a unique (as opposed to the other bait sets in the plurality), preselected efficiency for selection. *E.g.*, each unique bait or bait set provides for a unique depth of sequencing. The term "bait set", as used herein, collectively refers to one bait or a plurality of bait molecules.

In an embodiment, the efficiency of selection of a first bait set in the plurality differs from the efficiency of a second bait set in the plurality by at least 2 fold. In an embodiment, the first and second bait sets provide for a depth of sequencing that differs by at least 2 fold.

In an embodiment, the method comprises contacting one, or a plurality of the following bait sets with the library:

a) a bait set that selects sufficient members comprising a subgenomic interval to provide for about 500X or higher sequencing depth, *e.g.*, to sequence a mutation present in no more than 5% of the cells from the sample;

b) a bait set that selects sufficient members comprising a subgenomic interval to provide for about 200X or higher, *e.g.*, about 200X-about 500X, sequencing depth, *e.g.*, to sequence a mutation present in no more than 10% of the cells from the sample;

equimolar mix of baits, and then introducing a differential excess of internally-leveled group 1 to the overall bait mix relative to internally-leveled group 2.

In an embodiment, the method comprises the use of a plurality of bait sets that includes a bait set that selects a tumor member, *e.g.*, a nucleic acid molecule comprising a subgenomic interval from a tumor cell (also referred to herein as "a tumor bait set"). The tumor member can be any nucleotide sequence present in a tumor cell, *e.g.*, a mutated, a wild-type, a PGx, a reference or an intron nucleotide sequence, as described herein, that is present in a tumor or cancer cell. In one embodiment, the tumor member includes an alteration (*e.g.*, one or more mutations) that appears at a low frequency, *e.g.*, about 5% or less of the cells from the tumor sample harbor the alteration in their genome. In other embodiments, the tumor member includes an alteration (*e.g.*, one or more mutations) that appears at a frequency of about 10% of the cells from the tumor sample. In other embodiments, the tumor member includes a subgenomic interval from a PGx gene or gene product, an intron sequence, *e.g.*, an intron sequence as described herein, a reference sequence that is present in a tumor cell.

In another embodimentaspect, the disclosureinvention features, a bait set described herein, combinations of individual bait sets described herein, *e.g.*, combinations described herein. The bait set(s) can be part of a kit which can optionally comprise instructions, standards, buffers or enzymes or other reagenats.

Gene Selection

Preselected subgenomic intervals for analysis, *e.g.*, a group or set of subgenomic intervals for sets or groups of genes and other regions, are described herein.

Thus, in embodiments a method comprises sequencing, *e.g.*, by a next generation sequencing method, a subgenomic interval from at least five, six, seven, eight, nine, ten, fifteen, twenty, twenty-five, thirty or more genes or gene products from the acquired nucleic acid sample, wherein the genes or gene products are chosen from: ABL1, AKT1, AKT2, AKT3, ALK, APC, AR, BRAF, CCND1, CDK4, CDKN2A, CEBPA, CTNNB1, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FLT3, HRAS, JAK2, KIT, KRAS, MAP2K1, MAP2K2, MET, MLL, MYC, NF1, NOTCH1, NPM1, NRAS, NTRK3, PDGFRA, PIK3CA, PIK3CG, PIK3R1, PTCH1, PTCH2, PTEN, RB1, RET, SMO, STK11, SUFU, or TP53, thereby analyzing the tumor sample.

Accordingly, in one embodiment~~aspect~~, the disclosure~~invention~~ features, a method of analyzing a sample, *e.g.*, a tumor sample. The method comprises:

(a) acquiring a library comprising a plurality members from a sample, *e.g.*, a plurality of tumor members from a tumor sample;

(b) optionally, enriching the library for preselected sequences, *e.g.*, by contacting the library with a bait set (or plurality of bait sets) to provide selected members (*e.g.*, a library catch);

(c) acquiring a read for a subgenomic interval from a member, *e.g.*, a tumor member from said library or library catch, *e.g.*, by a method comprising sequencing, *e.g.*, with a next generation sequencing method;

(d) aligning said read by an alignment method, *e.g.*, an alignment method described herein; and

(e) assigning a nucleotide value (*e.g.*, calling a mutation, *e.g.*, with a Bayesian method or a method described herein) from said read for the preselected nucleotide position,

thereby analyzing said tumor sample,

wherein the method comprises sequencing, *e.g.*, by a next generation sequencing method, a subgenomic interval from at least five, six, seven, eight, nine, ten, fifteen, twenty, twenty-five, thirty or more genes or gene products from the sample, wherein the genes or gene products are chosen from: ABL1, AKT1, AKT2, AKT3, ALK, APC, AR, BRAF, CCND1, CDK4, CDKN2A, CEBPA, CTNNB1, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FLT3, HRAS, JAK2, KIT, KRAS, MAP2K1, MAP2K2, MET, MLL, MYC, NF1, NOTCH1, NPM1, NRAS, NTRK3, PDGFRA, PIK3CA, PIK3CG, PIK3R1, PTCH1, PTCH2, PTEN, RB1, RET, SMO, STK11, SUFU, or TP53.

In an embodiment, step (b) is present. In an embodiment, step (b) is absent.

In another embodiment, subgenomic intervals of one of the following sets or groups are analyzed. *E.g.*, subgenomic intervals associated with a tumor or cancer gene or gene product, a reference (*e.g.*, a wild type) gene or gene product, and a PGx gene or gene product, can provide a group or set of subgenomic intervals from the tumor sample.

In an embodiment, the method acquires a read, *e.g.*, sequences, a set of subgenomic intervals from the tumor sample, wherein the subgenomic intervals are chosen from at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 or all of the following:

In an embodiment, the method comprises annotating a called mutation, *e.g.*, annotating a called mutation with an indication of mutation structure, *e.g.*, a mis-sense mutation, or function, *e.g.*, a disease phenotype.

In an embodiment, the method comprises acquiring nucleotide sequence reads for tumor and control nucleic acid.

In an embodiment, the method comprises calling a nucleotide value, *e.g.*, a variant, *e.g.*, a mutation, for each of X subgenomic intervals, *e.g.*, with a Bayesian calling method or a non-Bayesian calling method.

In an embodiment, multiple samples, *e.g.*, from different subjects, are processed simultaneously.

The methods disclosed herein can be used to detect alterations present in the genome or transcriptome of a subject, and can be applied to DNA and RNA sequencing, *e.g.*, targeted RNA and/or DNA sequencing. Thus, another embodiment~~aspect~~ featured in the disclosure~~invention~~ includes methods for targeted RNA sequencing, *e.g.*, sequencing of a cDNA derived from an RNA acquired from a sample, *e.g.*, an FFPE-sample, to detect an alteration described herein. The alteration can be rearrangement, *e.g.*, a rearrangement encoding a gene fusion. In other embodiments, the method includes detection of a change (*e.g.*, an increase or decrease) in the level of a gene or gene product, *e.g.*, a change in expression of a gene or gene product described herein. The methods can, optionally, include a step of enriching a sample for a target RNA. In other embodiments, the methods include the step of depleting the sample of certain high abundance RNAs, *e.g.*, ribosomal or globin RNAs. The RNA sequencing methods can be used, alone or in combination with the DNA sequencing methods described herein. In one embodiment, the method includes performing a DNA sequencing step and an RNA sequencing step. The methods can be performed in any order. For example, the method can include confirming by RNA sequencing the expression of an alteration described herein, *e.g.*, confirming expression of mutation or a fusion detected by the DNA sequencing methods of the disclosure~~invention~~. In other embodiments, the method includes performing an RNA sequencing step, followed by a DNA sequencing step.

In another embodiment~~aspect~~, the disclosure~~invention~~ features a method comprising building a database of sequencing/alignment artifacts for the targeted subgenomic regions. In embodiment the database can be used to filter out spurious mutation calls and improve specificity. In an embodiment the database is built by sequencing unrelated non-tumor (*e.g.*, FFPE) samples or cell-lines and

Thus, methods featured in the disclosure~~invention~~ allow a quick turn around time, *e.g.*, within 7, 14 or 21 days of receipt of sample.

Methods described herein can also be used to evaluate a histologically normal sample, *e.g.*, samples from surgical margins. If one or more alterations as described herein is detected, the tissue can be re-classified, *e.g.*, as malignant or pre-maligant, and/or the course of treatment can be modified.

In certain embodiment~~aspects~~, the sequencing methods described herein are useful in non-cancer applications, *e.g.*, in forensic applications (*e.g.*, identification as alternative to, or in addition to, use of dental records), paternity testing, and disease diagnosis and prognosis, *e.g.*, for cystic fibrosis, Huntington's Disease, Alzheimer's Disease, among others. For example, identification of genetic alterations by the methods described herein can indicate the presence or risk of an individual for developing a particular disorder.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this disclosure~~invention~~ belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present disclosure~~invention~~, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the detailed description, drawings, and from the claims.

DESCRIPTION OF THE FIGURES

The drawings are first described.

FIG. 1A-F is a flowchart depiction of an embodiment of a method for multigene analysis of a tumor sample.

FIG. 2 depicts the impact of prior expectation and read depth on mutation detection.

FIG. 3 depicts the mutation frequencies in more than 100 clinical cancer samples.

FIG. 4 is a linear representation of a coverage histogram. The number of targets (y-axis) are depicted as a function of coverage (x-axis). Line #1 represents the coverage using a bait set that includes biotinylated, array-derived RNA oligonucleotide baits spiked with biotinylated,