CLAIMS

I/We claim:

1. A method for detecting and quantifying genomic mutations developed *in vivo* in a subject following the subject's exposure to a mutagen, comprising:

providing a sample from the subject, wherein the sample comprises double-stranded DNA molecules, and wherein the subject was exposed to the mutagen 30 days or less prior to extracting the sample comprising the double-stranded DNA molecules from the subject;

generating an error-corrected sequence read for each of a plurality of the double-stranded DNA molecules in the sample, comprising:

generating a set of copies of an original first strand of the double-stranded ~~adapter~~ DNA molecule and a set of copies of an original second strand of the double-stranded~~adapter~~ DNA molecule;

sequencing the set of copies of the original first and second strands to provide a first strand sequence and a second strand sequence; and

comparing the first strand sequence and the second strand sequence to identify one or more correspondences between the first and second strand sequences; and

analyzing the one or more correspondences to determine a mutation spectrum for the double-stranded DNA molecules in the sample.

2. The method of claim 1, further comprising calculating a mutant frequency for the ~~target~~ double-stranded DNA molecules in the sample by calculating the number of unique mutations per duplex base-pair sequenced.

3. The method of claim 1, wherein the ~~target~~ double-stranded DNA molecules were extracted from liver, spleen, blood, lung or bone marrow of the subject.

~~4. The method of claim 1, wherein the subject was exposed to the mutagen 30 days or less prior to the target double-stranded DNA molecules being removed from the subject.~~

4~~5~~. The method of claim 1, wherein the mutation spectrum is generated by unsupervised hierarchical mutation spectrum clustering.

5~~6~~. The method of claim 1, wherein the mutation spectrum is a triplet mutation spectrum.

67. The method of claim 1, wherein generating an error-corrected sequence read for each of thea plurality of the double-stranded DNA molecules includes generating error-corrected sequence reads of one or more targeted genomic regions.

78. The method of claim 67, wherein the one or more targeted genomic regions is a mutation-prone site in the genome.

89. The method of claim 67, wherein the one or more targeted genomic regions is a known cancer driver gene.

910. The method of claim 1, wherein the subject is a transgenic animal, and wherein at least some of the target double-stranded DNA molecules include one or more portions of a transgene.

1011. The method of claim 1, wherein the subject is a non-transgenic animal, and wherein the target double-stranded DNA molecules comprise endogenous genomic regions.

1112. The method of claim 1, wherein the subject is a human, and wherein the target double-stranded DNA molecules are extracted from a blood draw taken from the human.

13. A method for generating a mutagenic signature of a test agent, comprising:

duplex sequencing DNA fragments extracted from a test subject exposed to the test agent; and

generating a mutagenic signature of the test agent, comprising:

calculating a mutant frequency for a plurality of the DNA fragments by calculating the number of unique mutations per duplex base-pair sequenced; and

determining a mutation pattern for the plurality of the DNA fragments, wherein the mutation pattern includes mutation type, mutation trinucleotide context, and genomic distribution of mutations.

14. The method of claim 13, further comprising comparing the mutation signature of the test agent with mutation signatures of one or more known genotoxins.

15. The method of claim 13, wherein the mutation signature of the test agent varies based on one or more of a tissue type, a level of exposure to the test agent, a genomic region, and a subject type.

16. The method of claim 15, wherein the subject type is human cells grown in culture.

17. The method of claim 13, wherein the test animal was exposed to the test compound 30 days or less prior to the animal being sacrificed.

18. The method of claim 13, wherein the mutagenic signature is generated by computational pattern matching.

19. The method of claim 13, wherein the mutation signature is a triplet mutation signature.

20. The method of claim 13, wherein duplex sequencing DNA fragments includes duplex sequencing one or more targeted genomic regions.

21. The method of claim 20, wherein the one or more targeted genomic regions is a mutation-prone site in the genome.

22. The method of claim 20, wherein the one or more targeted genomic regions is a known cancer driver gene.

23. The method of claim 13, wherein the test animal is a transgenic animal, and wherein at least some of the DNA fragments include one or more portions of a transgene.

24. The method of claim 13, wherein the test animal is a non-transgenic animal, and wherein the DNA fragments comprise endogenous genomic regions.

25. A method for assessing a genotoxic potential of a test agent, comprising:

(a) preparing a sequencing library from a sample comprising a plurality of double-stranded DNA fragments from a biological source exposed to the test agent, wherein preparing the sequence library comprises ligating asymmetric adapter molecules to the plurality of double-stranded DNA fragments to generate a plurality of adapter-DNA molecules;

(b) sequencing first and second strands of the adapter-DNA molecules to provide a first strand sequence read and a second strand sequence read for each adapter-DNA molecule;

(c) for each adapter-DNA molecule, comparing the first strand sequence read and the second strand sequence read to identify one or more correspondences between the first and second strand sequences reads; and

(d) determining a mutation signature of the test agent by analyzing the one or more correspondences between the first and second strand sequences reads for each of the adapter-DNA molecules to determine at least one of a mutation pattern, a mutation type, a mutant frequency, a mutation type distribution, and a genomic distribution of mutations in the sample; and

(e) comparing the mutation signature of the test agent to a plurality mutation spectra derived from known genotoxins to determine if the mutation signature is sufficiently similar to a mutation spectrum from a known genotoxin; or

(f) assessing if at least one of the mutant frequency, the mutations type, or the mutation type distribution is above a safe threshold level; or

(g) determining if the mutant frequency exceeds a safe threshold mutant frequency.

26. The method of claim 25, wherein a mutation signature of the test agent comprises a mutant frequency above a safe threshold frequency.

27. The method of claim 25 wherein the mutation signature of the test agent comprises a mutation pattern sufficiently similar to known cancer-associated mutation pattern.

28. The method of claim 25, wherein the biological source is at least one of cells grown in culture, an animal, a human, a human cell line, a transgenic animal, a non-transgenic animal, a human tissue sample, or a human blood sample.

29. The method of claim 25, wherein the biological source was exposed to the test agent 30 days or less prior to extracting the sample comprising a plurality of double-stranded DNA fragments.

30. The method of claim 25, wherein the mutation signature is a triplet mutation signature.

31. The method of claim 25, wherein prior to comparing the first strand sequence read and the second strand sequence read, the method comprises associating the first strand sequence read with the second strand sequence read using one or more of an adapter sequence, sequence read length, and original strand information.

32. The method of claim 25, wherein prior to preparing the sequencing library, the method further comprises exposing the biological source to the test agent.

33. The method of claim 32, wherein prior to exposing the biological source to the test agent, the biological source is or comprises a cancer tissue.

34. The method of claim 32, wherein prior to exposing the biological source to the test agent, the biological source is or comprises a healthy tissue.

35. The method of claim 25, wherein the sample is or comprises a blood sample.

36. The method of claim 25, wherein the sample is or comprises a cancer cell line.

37. The method of claim 25, wherein the biological source comprises cancerous cells, and wherein the substance is tested for selective genotoxicity to at least a portion of the cancerous cells.

38. The method of claim 37, wherein the substance is a therapeutic compound.

39. The method of claim 38, wherein for the portion of the cancerous cells shown to be sensitive to the selective genotoxicity of the therapeutic compound, the method further comprises determining one or more of a mutant frequency and a mutation spectrum for the portion of the cancerous cells prior to exposure to the therapeutic compound.

40. The method of claim 25, wherein the test agent comprises a food, a drug, a vaccine, a cosmetic substance, an industrial additive, an industrial by-product, petroleum distillate, heavy metal, household cleaner, airborne particulate, byproduct of manufacturing, contaminant, plasticizer, detergent, a radiation-emitting product, a tobacco product, a chemical material, or a biological material.

41. A method for determining a subject's exposure to a genotoxic agent, comprising:

comparing a subjects' DNA mutation spectrum with mutation spectra of known mutagenic compounds; and

identifying the mutation spectra of known mutagenic compounds most similar to the subject's DNA mutation spectrum.

42. The method of claim 41, wherein the subject's DNA mutation spectrum is assessed by Duplex Sequencing.

43. The method of claim 41, wherein the subject's DNA mutation spectrum is generated from DNA extracted from the patient's blood.

44. The method of claim 41, wherein the subject's DNA mutation spectrum is a triplet mutation spectrum.

45. The method of claim 41, further comprising sequencing the subject's DNA to generate the subject's DNA mutation spectrum.

46. The method of claim 45, wherein sequencing the subject's DNA includes sequencing one or more known cancer driver genes.

47. A kit able to be used in error corrected duplex sequencing of double stranded polynucleotides to identify genotoxins, the kit comprising:

at least one set of polymerase chain reaction (PCR) primers and at least one set of adaptor molecules, wherein the primers and adaptor molecules are able to be used in error corrected duplex sequencing experiments; and

instructions on methods of use of the kit in conducting error corrected duplex sequencing of DNA extracted from a subject's sample to identify if the subject has been exposed to at least one genotoxin.

48. The kit of claim 47, wherein the reagent comprises a DNA repair enzyme.

49. The kit of claim 47, wherein each of the adapter molecules in the set of adaptor molecules comprises at least one single molecule identifier (SMI) sequence and at least one strand defining element.

50. The kit of claim 47, further comprises a computer program product embodied in a non-transitory computer readable medium that, when executed on a computer, performs steps of determining an error-corrected duplex sequencing read for one or more double-stranded DNA molecules in a sample, and determining the mutant frequency, mutation spectrum, and/or triplet spectrum of at least one genotoxin using the error-corrected duplex sequencing read.

51. The kit of claim 50, wherein the computer program product further determines the mechanism of action of the genotoxin in mutating a subject's DNA; and therapeutic or prophylactic treatments suitable for administering to the subject based upon the genotoxin mechanism of action.

52. A method for diagnosing and treating a subject exposed to a genotoxin, comprising:

a) determining whether a subject was exposed to a genotoxin by:

i) obtaining a biological sample from the subject;

ii) providing duplex error-corrected sequencing reads for a plurality of double-stranded DNA sequences extracted from the sample;

iii) determining the mutant frequency, mutation spectrum, and/or triplet mutation spectrum of the DNA sequences;

iv) determining if the mutant frequency, mutation spectrum and/or triplet mutation spectrum is indicative of the subject having been exposed to a genotoxin;

b) if the subject has been exposed to the genotoxin, then providing a prophylactic and/or a therapeutic treatment to prevent or inhibit the onset of a disease or disorder associated with the genotoxin.

53. A method for identifying a threshold level of safe exposure to a genotoxin, and providing treatment, comprising:

a) determining a genotoxin's threshold level of safe exposure;

b) determining whether a subject was exposed to the genotoxin at a level greater than the threshold level of safe exposure by:

i) obtaining a biological sample from the subject;

ii) providing duplex error-corrected sequencing reads for a plurality of double-stranded DNA sequences extracted from the biological sample;

~~iii) determining the mutant frequency, mutation spectrum, and/or triplet mutation spectrum of the DNA sequences;~~

~~iv) determining if the mutant frequency, mutation spectrum and/or triplet mutation spectrum are indicative of the subject having been exposed to a specific genotoxin;~~

~~v) computing the level of exposure of the subject to the genotoxin based on the mutant frequency, mutation spectrum and/or triplet mutation spectrum; and~~

~~c) if the subject has been exposed to more than the genotoxin's threshold level of safe exposure, then providing a prophylactic and/or a therapeutic treatment to prevent or inhibit the onset of a disease or disorder associated with the genotoxin.~~

~~54. A system for detecting and identifying mutagenic events and/or nucleic acid damage events resulting from genotoxic exposure of a sample, comprising:~~

~~a computer network for transmitting information relating to sequencing data and genotoxicity data, wherein the information includes one or more of raw sequencing data, duplex sequencing data, sample information, and genotoxin information;~~

~~a client computer associated with one or more user computing devices and in communication with the computer network;~~

~~a database connected to the computer network for storing a plurality of genotoxin profiles and user results records;~~

~~a duplex sequencing module in communication with the computer network and configured to receive raw sequencing data and requests from the client computer for generating duplex sequencing data, group sequence reads from families representing an original double-stranded nucleic acid molecule and compare representative sequences from individual strands to each other to generate duplex sequencing data; and~~

~~a genotoxin module in communication with the computer network and configured to compare duplex sequencing data to reference sequence information to identify mutations and generate genotoxin data comprising at least one of a mutant frequency, a mutation spectrum, and a triplet mutation spectrum.~~

~~55. The system of claim 54, wherein the genotoxin profiles comprise genotoxin mutation spectrum from a plurality of known genotoxins.~~

12~~56~~. A non-transitory computer-readable storage medium comprising instructions that, when executed by one or more processors, performs the~~a~~ method of any one of claims 1-11~~53~~ for determining if the~~a~~ subject is exposed to at least one genotoxin and/or determining an identity of at least one genotoxin.

1357. The non-transitory computer-readable storage medium of claim 1256, further comprising computing the mutation spectrum, mutant frequency, and/or triplet mutation spectrum of a detected agent, from which the identity of the at least one genotoxin is determined.

1458. A computer system for performing thea method of any one of claims 1-1153 for determining if thea subject is exposed to and/or an identity of at least one genotoxin, the system comprising: at least one computer with a processor, memory, database, and a non-transitory computer readable storage medium comprising instructions for the processor(s), wherein said processor(s) are configured to execute said instructions to perform operations comprising the methods of any one of claims 1-1153.

1559. The system of claim 1458, further comprising a networked computer system comprising:

a. a wired or wireless network;

b. a plurality of user electronic computing devices able to receive data derived from use of a kit comprising reagents to extract, amplify, and produce a polynucleotide sequence of thea subject's sample, and to transmit the polynucleotide sequence via a network to a remote server; and

c. a remote server comprising the processor, memory, database, and the non-transitory computer readable storage medium comprising instructions for the processor(s), wherein said processor(s) are configured to execute said instructions to perform operations comprising the methods of any one of claims 1-1153; and

d. wherein said remote server is able to detect and identify mutagenic events and/or nucleic acid damage events resulting from genotoxic exposure of a sample.

1660. The system of claim 1559, wherein the database and/or a third-party database accessible via the network, further comprises a plurality of records comprising one or more of a genotoxin profile of known genotoxins, a genotoxin profile of at least one subject's sample, and wherein the genotoxin profile comprises a mutation or a site of DNA damage.

~~61. A non-transitory computer readable medium whose contents cause at least one computer to perform a method for providing duplex sequencing data for double-stranded nucleic acid molecules in a sample from a genotoxicity screening assay, the method comprising:~~

~~receiving raw sequence data from a user computing device; and~~

~~creating a sample-specific data set comprising a plurality of raw sequence reads derived from a plurality of nucleic acid molecules in the sample;~~

~~grouping sequence reads from families representing an original double-stranded nucleic acid molecule, wherein the grouping is based on a shared single molecule identifier sequence;~~

~~comparing a first strand sequence read and a second strand sequence read from an original double-stranded nucleic acid molecule to identify one or more correspondences between the first and second strand sequences reads; and~~

~~providing duplex sequencing data for the double-stranded nucleic acid molecules in the sample.~~

62. The computer-readable medium of claim 58, further comprising identifying nucleotide positions of non-complementarity between the compared first and second sequence reads, wherein the method further comprises:

in positions of non-complementarity, identifying and eliminating or discounting process errors; and

in positions of non-complementarity that are not identified as process errors, identifying remaining positions of non-complementarity as sites of possible in-vivo DNA damage resulting from exposure to a genotoxin.

63. A non-transitory computer readable medium whose contents cause at least one computer to perform a method for detecting and identifying mutagenic events resulting from genotoxic exposure of a sample, the method comprising:

comparing duplex sequence data to reference sequence information;

identify mutations in the duplex sequence data, wherein a mutation is identified as a region of non-agreement with the reference information;

determining a mutant frequency in the duplex sequence data;

generating a mutation spectrum from the duplex sequence data;

generating a triplet mutation spectrum from the duplex sequence data; and

compare the mutation spectrum and/or the triplet mutation spectrum to a plurality of known genotoxin data sets.

64. A non-transitory computer readable medium whose contents cause at least one computer to perform a method for detecting and identifying a carcinogen or carcinogen exposure in a subject, the method comprising:

identifying sequence variants in a target genomic region using duplex sequencing data generated from a sample from the subject;

calculating a variant allele frequency (VAF) of a test sample and a control sample;

determining if a VAF is higher in a test group than in a control group;

in samples having a higher VAF, determining if a sequence variant is a non-singlet;

in samples having a higher VAF, determining if the sequence variant is a driver mutation; and

characterizing samples having a non-singlet and/or a driver mutation as being suspicious for being a carcinogen.

65. A non-transitory computer-readable medium of claim 68, further comprising assessing a safety threshold for the carcinogen and/or determining a risk associated with developing a genotoxin-associated disease or disorder following the exposure in the subject.