

CLAIMS:

1. A system for identifying candidate splice variants from a patient comprising:
a memory;
at least one processor; and
at least one non-transitory computer readable medium containing instructions that, when executed by the at least one processor, cause the at least one processor to perform operations comprising:
determining one or more sample splice junctions from a plurality of RNA sequence reads from a single formalin-fixed paraffin-embedded (FFPE) tumor sample obtained from the patient;
retrieving a set of baseline splice junctions, the set of baseline splice junctions determined from a plurality of healthy RNA samples not obtained from the same patient as the single FFPE tumor sample;
comparing the one or more sample splice junctions to the set of baseline splice junctions;
identifying one or more filtered sample splice junctions, wherein the filtered sample splice junctions are sample splice junctions that do not overlap with the set of baseline splice junctions; and
determining that one or more of the identified filtered sample splice junctions are candidate oncogenic events.
2. The system of claim 1, further comprising outputting the list of candidate oncogenic events.
3. The system of claim 1 or claim 2, wherein the plurality of healthy RNA samples comprises healthy RNA samples taken from a cross section of one or more of: geographical regions, ages, genders, ethnic groups, tissue types, and sample preservation qualities.
4. The system of any one of claims 1 to 3, wherein the plurality of healthy RNA samples comprises samples from one or more of: lung tissue, adrenal gland tissue, bladder tissue, breast tissue, ovary tissue, liver tissue, prostate tissue, skin tissue, and spleen tissue.
5. The system of any one of claims 1 to 4, wherein the plurality of healthy RNA samples comprises samples from donors across a range of ages.

6. The system of any one of claims 1 to 5, wherein the baseline splice junctions from the plurality of healthy RNA samples are determined prior to the determining the one or more sample splice junctions from the single FFPE tumor sample.

7. The system of any one of claims 1 to 6, wherein the baseline splice junctions are from a same genomic region as the one or more sample splice junctions.

8. The system of claim 1, wherein the plurality of healthy RNA samples are from non-tumor tissue.

9. The system of any one of claims 1 to 8, wherein the sample splice junctions and the baseline splice junctions are both determined using a common test.

10. The system of any one of claims 1 to 9, wherein determining the one or more sample splice junctions comprises:

determining the plurality of RNA sequence reads from the single FFPE tumor sample;
retrieving, a DNA reference sequence aligned with the RNA sequence reads from the single FFPE tumor sample; and

determining one or more sample splice junctions as missing contiguous locations in the RNA sequence reads compared with the DNA reference sequence.

11. The system of any one of claims 1 to 10, wherein the filtered sample splice junctions do not overlap with non-cancerous splice variants determined from a splice graph that captures multiple alternate combinations of exons for a given gene.

12. The system of any one of claims 1 to 10, wherein the set of baseline splice junctions are determined without determining a splice graph that captures multiple alternate combinations of exons for a given gene.

13. A computer implemented method, comprising:

determining, using at least one processor, one or more sample splice junctions from a plurality of RNA sequence reads from a single formalin-fixed paraffin-embedded (FFPE) tumor sample obtained from a patient;

retrieving, by the at least one processor from a memory, a set of baseline splice junctions determined from a plurality of healthy RNA samples not obtained from the same patient as the single FFPE tumor sample;

comparing the one or more sample splice junctions to the set of baseline splice junctions;

identifying, by the at least one processor, one or more filtered sample splice junctions, wherein the filtered sample splice junctions are sample splice junctions that do not overlap with the baseline splice junctions; and

determining that one or more of the identified sample splice junctions are candidate oncogenic events.

14. The method of claim 13, further comprising outputting the list of the candidate oncogenic events.

15. The method of claim 13 or claim 14, further comprising:

determining, by the at least one processor, RNA reads from the single FFPE tumor sample;

retrieving, by the at least one processor from the memory, a DNA reference sequence aligned with the RNA sequence reads from the single FFPE tumor sample; and

determining, by the at least one processor, the one or more sample splice junctions as missing contiguous locations in the RNA sequence reads compared with the DNA reference sequence.

16. The method of any one of claims 13 to 15, wherein the plurality of healthy RNA samples comprises healthy RNA samples taken from a cross section of one or more of: geographical regions, ages, genders, ethnic groups, tissue types, and sample preservation qualities.

17. The method of any one of claims 13 to 16, wherein the filtered sample splice junctions do not overlap with non-cancerous splice variants determined from a splice graph that captures multiple alternate combinations of exons for a given gene.