# SOMATIC COPY NUMBER VARIATION DETECTION

## CROSS REFERENCE TO RELATED APPLICATIONS

**[0001]** The present application claims priority to U.S. Provisional Application No. 62/398,354, entitled "SOMATIC COPY NUMBER VARIATION DETECTION" and filed September 22, 2016, and to U.S. Provisional Application No. 62/447,065, entitled "SOMATIC COPY NUMBER VARIATION DETECTION" and filed January 17, 2017, the disclosures of which are incorporated herein by reference for all purposes.

## BACKGROUND

**[0002]** The present disclosure relates generally to the field of data related to biological samples, such as sequence data. More particularly, the disclosure relates to techniques for determining copy number variation based on sequencing data.

**[0003]** Genetic sequencing has become an increasingly important area of genetic research, promising future uses in diagnostic and other applications. In general, genetic sequencing involves determining the order of nucleotides for a nucleic acid such as a fragment of RNA or DNA. Some techniques involve whole genome sequencing, which involves a comprehensive method of analyzing a genome. Other techniques involve targeted sequencing of a subset of genes or regions of the genome. Targeted sequencing focuses on regions of interest, generating a smaller and more compact data set. Further, targeted sequencing reduces sequencing costs and data analysis burdens while also allowing deep sequencing at high coverage levels for detection of variants in the regions of interest. Examples of such variants may include somatic mutations, single nucleotide polymorphisms, and copy number variations. Detection of variants may provide clinicians with information about disease likelihood or susceptibility. Accordingly, there is a need for improved detection of variants in sequencing data.

SUMMARY

**[0003a]** According to a first aspect of the present disclosure, there is provided a method of normalizing copy number, including: receiving a sequencing request from a user to sequence one or more regions of interest in a biological sample; acquiring baseline sequencing data from the one or more regions of interest from a plurality of baseline biological samples that are not matched to the biological sample; determining copy number normalization information using the baseline sequencing data, wherein the copy number normalization information comprises at least one copy number baseline for a region of interest of the one or more regions of interest; and providing the copy number normalization information to the user.

**[0003b]** According to a second aspect of the present disclosure, there is provided a method of detecting copy number variation, including: acquiring sequencing data from a biological sample, wherein the sequencing data comprises a plurality of raw sequencing read counts for a respective plurality of regions of interest; normalizing the sequencing data to remove region-dependent coverage bias ,wherein the normalizing comprises: for each region of interest, comparing a raw sequencing read count of one or bins in a region of interest of the biological sample to a baseline median sequencing read count to generate a baseline-corrected sequencing read count for the one or more bins in the region of interest, wherein the baseline median sequencing read count for one or more bins in the region of interest is derived from a plurality of baseline samples that are not matched to the biological sample and is determined from only the most representative portions of the baseline sequencing data for each region of interest; and removing GC bias from the baseline-corrected sequencing read count to generate a normalized sequencing read count for each region of interest; and determining copy number variation in each region of interest based on the normalized sequencing read count of the one or more bins in each region of interest.

31. The method of claim 24, wherein the baseline median sequencing read count for each region of interest is determined by clustering the baseline sequencing data.

32. The method of claim 24, wherein a first baseline median sequence coverage count for a first region of interest is derived from a first subset of the plurality of baseline samples and wherein a second baseline median sequence coverage count for a second region of interest is derived from a second subset of the plurality of baseline samples that is different from the first subset.

33. The method of claim 24, comprising removing or masking outlier bins in the sequencing data before normalizing the sequencing data.

34. The method of claim 24, wherein normalizing the sequencing data comprising applying loess regression to the sequencing data fit the sequencing data to a curve after removing or masking the outlier bins.

35. The method of claim 24, wherein the region-dependent bias comprises one or more of GC bias, PCR bias, or DNA quality bias.