Lecture Title and Date

Variant Identification (2/12)

Objectives of the Lecture

- By the end of this lecture, students should be able to:
    - Comprehend the complete process of genome resequencing and variant identification
    - Understand the distinction between Single Nucleotide Polymorphisms (SNPs) and Structural Variants (SVs)
    - Apply Bayesian methods for variant detection and probability calculation
    - Analyze various computational approaches for detecting structural variants, including paired-end mapping, split-read analysis, and read depth evaluation

Key Concepts and Definitions

- Single Nucleotide Polymorphism (SNP): a type of variants in genome where a single nucleotide is different
- Structural Variant (SV): variants in genome longer than 50 bp in length
- Bayes' Theorem: a mathematical formula to calculate conditional probabilities based on given knowledge/data.

Main Content/Topics

- Genome resequencing involves aligning sequenced reads to a reference genome to identify discrepancies
    - SNPs are identified by aligning reads to the reference genome
    - Structural variants are then identified by computation
    - Once the variants are identified, the sequences are assembled
    - The assembled sequences are phased and assigned to the corresponding strand (due to the diploid nature of humans)
- SNP detection is achieved through alignment and Bayesian probability calculations, assuming binomial distribution for heterozygous sites
    - According to Bayes' Theorem, $P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$
    - P(G|D): Given the observed reads D, what is the probability of observing the genotype of interest G
        - Calculated from P(D|G), P(D), and P(G)
    - P(D|G): Given the genotype, what is the likelihood of getting the data
        - Calculated assuming binomial distribution
- Structural variants are identified using multiple methods:
    1. Paired-end mapping

        - Comparing distances between paired-end reads to detect insertions, deletions, and inversion.
        - If the distance between the paired ends (span) is shorter than expected, then there must have been an insertion between the paired ends
        - If the span is longer than expected, then there must have been a deletion between the paired ends
    2. Split-read analysis
        - Identifying the breakpoints in the reads where the alignment to the reference genome is broken
            - If the read is mapped without a gap at the breakpoint on the reference genome, it is an insertion.
            - If a gap appears when the read is mapped to the reference genome, then there is a deletion.
    3. Read depth analysis
        - Employing Hidden Markov Models (HMM) to predict genomic states based on coverage levels
            - Predict whether two, one, or no strand exists at a given point in the target genome

Discussion/Comments

- Individuals have very different variants in the genome. It is difficult to map reads back to the reference genome.
- Bayesian Analysis is highly applicable to many fields. Good to know the general idea.
- Real data is noisy. HMM is one of the many ways to analyze noisy data.
- Structural variants are more complex to identify than SNPs, with deletions being the easiest to detect

Suggested References

Alkan, C., Coe, B. & Eichler, E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363–376 (2011). https://doi.org/10.1038/nrg2958

Figure 1 illustrates the main classes of structural variants in the genome, and Figure 2 visually summarizes the signatures of structural variants identifiable with various methods of computation. Details of each computational method covered in class are available in the section under “Sequencing-based Computational Approaches”.

van de Schoot, R., Depaoli, S., King, R. *et al.* Bayesian statistics and modelling. *Nat Rev Methods Primers* **1**, 1 (2021). https://doi.org/10.1038/s43586-020-00001-2

This Nature Method Primers paper provides the basic concept and applications of Bayesian Statistics with great examples.