Lecture Title and Date

Summary of HumanGermline Variation & Somatic Variation – 2/12/2025

Objectives of the Lecture

1. Interpret key variant statistics from the 1000 Genomes Project, including common versus rare variants and their population distributions;
2. Explain how reference bias and population history lead to differences in variant counts, and why African populations show higher genetic diversity;
3. Contrast germline variation (1000 Genomes) with somatic variation in tumors (PCAWG), including the mutation loads across cancer types and the concepts of driver vs. passenger mutations;
4. Discuss the intraspecies common versus rare variants in healthy versus disease link genomes and interspecies gene conservation;
5. Understand different types of germline variants prevalence and how structural variations are detected;

Key Concepts and Definitions

1. **1000 Genomes Project:** A large-scale study cataloging human genetic variation across diverse populations, highlighting approximately 4 million SNPs per individual and the predominance of common variants;
2. **gnomAD:** An expansion of the ExAC project (versions v2–v4) that aggregates large-scale sequencing data. It provides high-resolution catalogs of both single nucleotide variants and structural variants (SVs), reporting, for instance, 1,199,117 SVs from 807,162 individuals and mean counts of rare and unique coding variants across populations.
3. **PCAWG (Pan-Cancer Analysis of Whole Genomes):** An initiative to profile somatic variants across various cancer types, emphasizing the vast range in mutation counts (from a few thousand to over 100,000 per tumor);
4. **Germline vs. Somatic Variants:** Germline variants are inherited and shared among all cells (as cataloged by 1000 Genomes), while somatic variants occur in specific cells, such as in tumors (analyzed by PCAWG);
5. **Reference Bias:** The influence of the reference genome on variant detection. Because the human reference genome is more representative of European ancestry, it can lead to differences in variant counts among populations.

Main Content/Topics

1. **Detailed Analysis of the 1000 Genomes Project:**

– **Populations:** Sequenced 26 populations, encompassing roughly 2,504 individuals.

– **Variant Counts per Individual:** On average, each genome harbors about 4 million variants and among all the germline variants. **[However, due to reference bias, European genomes average around 3.5 million variants while African genomes average about 4.3 million, reflecting higher genetic diversity in African populations]**

– **Overall Catalog:** The project cataloged roughly 85 million unique variants, with approximately 97% of an individual's variants being common and the remainder classified as rare.

– **Coding Variants:** Each individual has around 22,000 coding variants, with European genomes showing roughly 11,000 protein-coding differences compared to one another.

– **Structural Variants (SVs):** Using short-read sequencing, about 2,000 SVs were detected per genome. More recent long-read approaches suggest that a true count could reach up to 30,000 SVs per individual.

2. **Integration of gnomAD Data:**

– gnomAD further refines our understanding of genomic variation by expanding the variant catalog beyond the 1000 Genomes Project. For example, gnomAD v3 reports a comprehensive set of SVs and rare coding variants across a vast number of individuals, offering enhanced resolution in variant frequency estimation;

– Recent gnomAD analyses reveal detailed statistics on rare and unique coding variants, underscoring the importance of high-depth sequencing and the use of long-read technologies. This update is exemplified by studies using PacBio HiFi, which detected on average ~24,653 SVs, ~794,406 indels, and ~3,895,274 SNVs per diploid genome (Ebert et al., Science 2021).

3. **Detailed Analysis of PCAWG (Pan-Cancer Analysis of Whole Genomes):**

– **Sample Size:** PCAWG analyzed tumor–normal pairs from approximately 2,800 cancer patients.

– **Somatic Variant Load:** Somatic mutation counts vary widely by cancer type—some cancers average around 5,000 somatic variants per genome, while others may exceed 100,000 variants, reflecting the high mutational burden in certain tumors.

– **Coding Somatic Variants:** Typically, only about 50 coding somatic mutations are observed per tumor genome, though this number can vary considerably based on the cancer subtype.

– **Structural Variants in Cancer:** Cancer genomes often display complex structural rearrangements. While exact counts can vary, many tumors show a substantial number of SVs, complicating comparisons with germline data from the 1000 Genomes Project.

4. Evolutionary Conservation and Disease Associations:

– **Interspecies Selection:** Metrics such as dN/dS ratios (for coding regions) and GERP scores (for noncoding regions) indicate similarity in gene regions across species, reflecting strong purifying selection as such gene regions having high functional importance.

– **Common vs. Rare Variants interspecies selection:** Common variants are detected by GWAS and found at high allele frequencies often contribute modestly to disease risk due to their ubiquity in the population, while rare variants generally require burden tests to assess cumulative effects because of their low frequency, usually associated with higher functional impacts, potentially indicating stronger intraspecies selection.

Discussion/Comments

1. There are several numbers important to remember for the quiz. For example, remember that an average genome from the 1000 Genomes Project has roughly 4 million variants, with African genomes averaging around 4.3 million variants and European genomes around 3.5 million, while only about 20,000 of these are coding variants;
2. The 1000 Genome project did a really great job diversifying the populations around the world, but for genomAD dataset has around 77.07% European, which then includes the reference bias;
3. In the concept of rare and common variants, Dr. Gerstein has pointed out an interesting concept of strong selection that if a region is under strong selection, the conservation will drive the common variants to be depleted but at the same time since rare variants are somehow stochastic, more rare variants are observed proportionally;
4. Data from gnomAD and updated sequencing technologies (e.g., PacBio HiFi) notes that current methods can detect up to ~24,653 SVs, ~794,406 indels, and nearly 3.9 million SNVs per diploid genome—numbers that provide context for understanding the technological advances in genome sequencing;
5. In contrast, the PCAWG dataset for cancer genomes shows dramatic heterogeneity, with roughly 2,800 tumor/normal pairs yielding a total of ~30 million somatic SNVs. While many cancers average around 5,000 somatic mutations, certain types can exceed 100,000 mutations;
6. Overall, the integrated discussion draws a clear contrast: the 1000 Genomes Project sets the benchmark for normal human variation across 26 populations (2,504 individuals) with around 85 million unique variants cataloged, whereas PCAWG reveals the extreme mutational variability in cancer genomes. This different perspective is critical for drawing a complete picture using both germline and somatic mutations. Consideration of how integrating these datasets can enhance personalized medicine strategies by differentiating inherited versus tumor-specific mutations;

Suggested References

• 1000 Genomes Consortium, Nature (2010, 2012); Mills et al., Nature (2011)
• 1000 Genomes Project Consortium. Nature, 526(7571):68–74, 2015 – A global reference for human genetic variation (DOI:10.1038/nature15393)
• 1000GP Phase3 SV paper, submitted to Nature, 2015
• 1000GP Consortium Summary, submitted to Nature, 2015
• gnomAD Project, available at https://gnomad.broadinstitute.org/
• gnomAD v3 paper: Konrad et al., Nature (2020)
• Khurana E. et al., Nature Reviews Genetics, 17:93–108, 2016
• Gudmundsson et al., Human Mutation, 2021
• gnomAD-SV: Ryan et al., Nature (2020)
• Ebert et al., Science (2021) – Updating SV numbers with current PacBio HiFi technology
• PCAWG Consortium. Nature, 578(7793):82–93, 2020 – Pan-cancer analysis of whole genomes (DOI:10.1038/s41586-020-1969-6)

• Campbell et al., originally on bioRxiv (2017), now published as part of the PCAWG Nature paper (578:82–93, 2020)