Notes from Network meeting, May 2, Seattle, WA -- TCGA Topics

I. Mutation Validation – Investigation of Orthogonal Platforms

- David Wheeler suggested that TCGA release a description of best approaches for mutation validation. Baylor's current practice includes repeating probe-capture and Illumina sequencing. This method has excellent scalability, can cope with false negative errors, and is cost effective. The effect of systematic errors is unknown. It has been suggested that an alternative enrichment and sequencing chemistry be evaluated to understand random and systematic errors.
- BCM is generating Illumina replicates for the KIRC validation sets that were done by pyrosequencing (Ion Torrent).
- **ACTION:** During a future teleconference, the Steering Committee will discuss drafting a recommendation for the best approach to mutation validation.

II. Benchmarking Exercise Report

- Benchmarking is critical to evaluate mutation calling tools and monitor progression.
- A mutation caller is a classifer that is characterized by ROC (Receiver Operators Characteristic) curves depending on the allele fraction, coverage of tumor and normal, and sequencing/alignment noise.
- There are two types of false positives: no event or germline event. Both types of false positives should be measured when the normal is wildtype or variant.
- We must require at least 3 alternate reads for a false positive rate less than 0.05×10^{-6}.
- The virtual tumor approach allows for specificity and sensitivity to be measured. However, some false positive rates and sensitivity calculations will require benchmarking exercises with real tumors.
- MuTect is well suited for studying impure and heterogeneous tumors because it can detect low allele faction mutations. Gaddy commented that in the future, MuTect will be able to define the probability to detect clonal and subclonal mutations at every base in the genome.
- Benchmark 4 (BMK4) is able to simulate tumor subclones and normal contamination. It was noted that normal contamination leads to decreased somatic concordance; increasing normal contamination increases false positives.
- The mutation types included in BMK4 include: SNV, INDEL, SV, and CNV.
- At least 7 more submissions to BMK4 are expected.
- **ACTION:** Groups who have not yet already, should submit data for Benchmark 4.
- David Haussler commented that TCGA does not yet have the ground truth. Once more WGS data is available, proper validation may be completed.
- David Haussler suggested that benchmarking should be incorporated into the standard pipeline for each Center. At least three Centers need to contribute to mutation calling for each data freeze. The benchmarking data will be used in the final analysis by the AWG.
- **ACTION:** NHGRI will work with the GSCs to ensure that at least 3 groups contribute to mutation calling for every tumor data freeze.

- **ACTION:** Brad and Heidi will follow up with Baylor and WashU for pre-commitment to contribute to network mutation calling for HNSC and other upcoming project data freezes.

III. Analysis of WGS Data

- Li Ding noted that WashU has completed 298 total high-pass (>30x) tumor whole genome data and 929 total low-pass (~5x) tumor whole genome data. WashU is working to define what can be learned from existing Whole Genome Sequencing (WGS) data.
- WGS is necessary for a comprehensive understanding of the entire cancer genome.
- WGS includes noncoding regions which may be relevant to the pathogenesis of cancer, e.g., germline variants in the promoter region may increase the risk of cancer. WGS also enhances RNAseq gene fusion discovery.
- Li commented that besides being a causative agent in cancer, the presence of a virus may also affect treatment outcomes.
- On average, WGS is 10x more expensive than WEx.

IV. Publication Strategy Threads

- Threads at *Nature* will provide TCGA with an opportunity to explore the wealth of information collectively described by the TCGA-related papers published across all *Nature* journals. The Threads complement the papers by highlighting topics that are otherwise covered only in subsections of individual papers. Each thread consists of relevant paragraphs, figures and tables from across the papers, united around a specific theme.
- This effort will have a similar structure to the ENCODE Threads and will be separate from Pan-Can.
- ENCODE did not involve specific authors for their Threads, however, *Nature* is open to discussions with TCGA to install an authorship model that would recognize junior investigators who contribute to editing.
- The TCGA Project Team is enthusiastic that this endeavor will be rewarding and increase publicity surrounding TCGA.
- Raju Kucherlapati stressed that no new content (e.g., data, text, interpretation) will be generated for the Threads, but scientific editing is required. Authorship will acknowledge editing efforts.
- The Steering Committee discussed allowing external collaborators to participate as editors.
- *Nature* and the Project Team expect the Threads to evolve over time. New content will be added to the Threads as relevant papers are published. The primary editorial author will be responsible for updating the content.
- Josh Stuart suggested that at the next in-person meeting, the agenda include a session for junior investigators to submit thread ideas to be reviewed and selected by the Steering Committee.
- Richard Gibbs acknowledged that although the principal concept of Threads would be beneficial to TCGA, the long term commitment and limited bandwidth will make it difficult for TCGA investigators to participate. Stacey Gabriel suggested hiring a contract science writer to do this work rather than junior investigators.

- There was general consensus among the Steering Committee that Threads is a worthwhile pursuit, but that the Steering Committee needs to find an appropriate avenue for participation and topic selection.
- **ACTION:** Volunteer editors to assemble Threads at *Nature* will follow up with Kenna Shaw, Raju Kucherlapati, and John Weinstein.

V. ICGC & TCGA: Building a Stronger Partnership for Pan-Cancer Analysis/Challenges of Data Management

- ICGC DCC has received data for ~7300 cancer genomes.
- ICGC has experienced difficulties with callers providing discordant results.
- The next steps for the ICGC Benchmarking Exercise will include high confidence SNV and indel calls for experimental verification. A sampling of medium confidence calls will also be verified. DNA from the German Medulloblastoma Project will be distributed to ICGC members for a round of dry and wet lab benchmarking.
- ICGC hopes to coordinate Benchmarking efforts with TCGA to reduce duplication of effort; this may be achieved by merging or overlapping working groups.
- ICGC has proposed a system whereby dbGaP autoforwards applications to access TCGA data to the ICGC's DACO. DACO then independently reviews the application and grants approval for ICGC data. In the future, the ICGC hopes to see a similar courtesy submission for data applications submitted to ICGC.
- ICGC controlled-access data are stored at the European Genome-phenome Archive (EGA).
- Lincoln Stein suggested that TCGA and ICGC co-locate raw data in cloud with viable private and public environments.

VI. Analysis Reporting/QC/Follow-up on Batch Effects Recommendations

- The Steering Committee previously approved a policy to make analysis scripts publicly available for levels 2-4 data. The version information for both the scripts and data should also be available. At a minimum, the algorithms used to generate figures should be documented in sufficient detail so that they can be utilized by external investigators.
- **ACTION:** The NCI Program Office will work with Broad/Firehose to determine the feasibility of ensuring scripts generated by the Centers function properly for external users.
- It was acknowledged that there are details in DWG operation that cannot be encapsulated in an automated system. Manual curation may be necessary at some steps.
- The Project Team proposed that the Melanoma AWG be the first upcoming publication to make all scripts available for content and figures.

VII. Importance of Germline Data Analysis in Cancer Genomics

- Dr. Chanock was unable to attend the May Meeting due to illness. This topic will be readdressed during a future Steering Committee teleconference.