

E.6 Genome Analysis

The goal of ICGC ARGO is to advance discovery; therefore, nucleic acid analysis must go significantly beyond the assessment of limited gene sets using panels such as those currently performed in most clinical diagnostic laboratories. Cognizant of the tractability of biospecimen quantity, quality and fixation methods, minimum requirements are one of the approaches detailed below, complemented where appropriate with methodologies listed in Boxes 4 and 5 below. The exact composition of analyses for each individual project may be defined through discussion with the ICGC ARGO Management Committee. Given that technologies are rapidly evolving and that novel platforms are in development, it will be necessary to review the recommendations on a regular basis. It is also preferable to avoid being unduly prescriptive about study designs and platform choices, as it is conceivable that several different approaches could be appropriate to deliver the same ultimate goals.

It is critical to the overall success of ICGC that datasets obtained from one class of cancer (generated using a specific technology platform) are comparable to the datasets obtained from other programs and projects (even if generated using a different technologies). It is particularly important, therefore, that members contribute sequencing data of the highest quality, including sufficient sequencing depth to detect a high proportion of somatic mutations in each of the samples interrogated.

ICGC members are expected to deliver one of the following types of genomic sequencing analyses:

- Whole genome sequencing (see Box 1)
- Whole exome sequencing (see Box 2)
- Clinical genome targeted sequencing (see Box 3)

It is envisaged that mutational catalogues for each tumor type or subtype will include the full range of somatic variant types, including single base substitutions, insertions, deletions, copy number changes, and structural variants including translocations and gene disrupting events.

Whilst transcriptome sequencing represents an important additional analysis (see Box 4), genomic DNA sequencing of tumor samples represent the core elements of the project and are therefore considered as mandatory. Other relevant studies, which are encouraged to complement the data from genomic DNA sequencing, are listed in Box 5.

Box 1. Guidelines on whole genome sequencing analyses

The aim of whole genome sequencing is to capture high-quality information on all variant classes across all genomic features, including small variants, gene and chromosome level copy number alterations, structural variants and mutational signatures. Sequence coverage calculations should be based on a target of $\geq 90\%$ of somatic alterations identified in each sample. The required coverage should take into account tissue cellularity and, where applicable to the specific research project, detection of subclonal somatic variants. Experience indicates

that at least 60-fold median coverage of the tumor sample will be required (with minimum 30-fold median coverage of the germline sample), with higher tumor coverage required for samples where the tumor cellularity is low. Analysis pipelines should be appropriately validated for all variant classes.

Box 2. Guidelines on whole exome sequencing

Sequencing should include all coding exons and splice junctions along with other genomic regions of biological interest (e.g. regulatory regions). Sequence coverage calculations should be based on a target of $\geq 90\%$ of somatic alterations identified in each sample.

Box 3: Guidelines on clinical genome targeted sequencing

Where whole genome sequencing is either unaffordable or not feasible (for example the analysis of formalin-exposed DNA, where tissue availability is limited or for low cellularity tumors), then a well-designed targeted capture sequencing assay is the best option. This approach is able to report on all variant classes (including small coding and non-coding variants, copy number alterations and structural variants) and is compatible with small amounts of FFPE-derived DNA. The objective should be to capture $\geq 90\%$ of the relevant genomic information for each tumor type sequenced. The ICGC Technical Working Group can provide assistance with assay specification, including provision of pre-designed and tested assays that meet the requirements ICGC ARGO.

Box 4: Guidelines on transcriptome sequencing

It is recommended that transcriptome analysis should include expression of all protein coding genes, with consideration given to the coverage of important non-coding features such as microRNAs. Analysis of the transcriptome may be more critical in some cancer types than in others, for example in breast and colorectal cancer where clinically useful classification systems have been described. Various technology platforms are available, including commercial options, some of which are compatible with RNA obtained from formalin-exposed tissue.

Box 5: Additional complementary analyses

- Epigenetic analyses (including histone modification and DNA methylation)
- Proteomic analyses
- Metabolomic analyses (including analysis of tumor and serum)

E.6.1 Germline analysis

For whole genome and whole exome sequencing, analysis of a tumor-normal pair remains the state-of-the-art for filtering germline polymorphisms. For targeted capture sequencing, either paired tumor-normal sequencing or tumor-only sequencing is acceptable, provided that in the latter situation, analysis pipelines contain sufficient steps to filter the majority of germline events. The ICGC Technical Working Group can provide advice on analytics and germline databases.

E.6.2 Quality control

The ICGC consortium is dependent on a high-quality genomic dataset. Members are encouraged to put in place processes and mechanisms to ensure the quality of the data, covering both the laboratory and sequencing steps and the variant calling pathway. Variant calling software, in particular, requires thorough validation to ensure the veracity of the data for all classes of genomic variant.

