nature genetics

Taking pan-cancer analysis global

Although federated cooperation is politically desirable, uniform data quality and standards are essential and should not be reinvented from scratch. The International Cancer Genome Consortium (ICGC) will do well to start with the data standards of The Cancer Genome Atlas (TCGA) and the Pediatric Cancer Genome Consortium if it is to succeed in genomic analysis across cancer types.

Whereas TCGA is driven by US government contracts, ICGC is a coalition of the willing, drawing upon multiple funding sources across the world. It is therefore to be expected that, for the latter, multiple solutions will be volunteered and that its agreements will be developed by consensus.

From the journal's point of view, robust results resting on explicit and reliable standards for data production and analysis are essential, but—provided pipelines are objectively and transparently evaluated—we do not care whether they are developed by consensus or adopted from best practice from another institute or project. Within ICGC, there are many outstanding institutions (including but not limited to the UK's Sanger Institute) that have been producing high-quality genomic data for many years. Other centers are relative newcomers, keen to demonstrate their throughput and accuracy. All have a place in developing ICGC's equivalent of the Broad Institute's Firehose pipeline that fed TCGA's pan-cancer analysis effort on the first dozen cancer types (http://www.nature.com/ng/focus/tcga/index.html). Five data centers worldwide have stepped up to host ICGC genomic data, and benchmarking exercises for components of the data pipelines have begun.

However, not all data producers will be able to participate using their existing standards and practices, nor is a compromise involving a poll or average of existing sequence analysis pipelines the solution. Some methods and combinations are more accurate, and there will have to be agreement on standards. One obstacle to standardization is that better tools are the enemy of the good. We already have almost too many creative ways to assemble whole-genome sequence reads, too many ways to call variants and too many algorithms for the identification of structural variants and recurrent mutations, and all of these programs are evolving and being fine-tuned (*Nat. Rev. Genet.* **14**, 321–332, 2013). Consequently, only a few centers have had the staff, resources and time to evaluate simple questions such as: "If I run the same sequence reads from a single cancer genome through this pipeline of assembly and variant calling twice, can I expect 70–80% concordance between the results?" But it is on the answers to basic questions such as this one that comparisons between different individual tumors as well as between and among cancer types depend. Among the great diversity of study designs for next-generation sequence analysis, many steps have the potential to deliver non-determinate results (different sets of variants can be called by the same pipeline under identical parameters), and the opportunity for non-determinacy arises, especially when parallel or multithreaded computing steps are invoked to save processing time.

One way to evaluate pipelines and standards for sequence variant discovery is via head-to-head competition on a range of tasks. Cloud computing resources and open community collaborations have recently made these collegial contests a reality, for example, in the Rheumatoid Arthritis Responder Challenge (*Nat. Genet.* **45**, 468–469, 2013). The journal is keen to help in incentivizing the ICGC sequence variant calling pipeline, and we hereby announce our interest in participating, if a publishing partner for the challenge is what it takes to move its development forward. Running such challenges in a standard environment will do a lot to objectively determine the robustness of software and pipelines for consortium-wide use (and not just in the environments in which such tools were originally used). Of course, new players will emerge, but we should not be afraid to admit that some solutions that are good enough are already available.

We support the view that it is a bad idea for any genomics endeavor to ignore the experience in standards and practices developed in the HapMap Project and 1000 Genomes Project (*Nat. Rev. Genet.* **14**, 321–332, 2013). In addition, the standards developed by foregoing cancer genome projects should be explicitly tested as benchmarks in any exercise seeking to set consortium-wide standards for data production and analysis. Incorporating existing solutions that work will enable the ICGC pan-cancer analysis initiative to start at a higher level. The tower of Babel will be easier to build with bricks than with sand. ■