NGS software tools — quo vadis?

Next-generation sequencing (NGS) is changing biomedical research and its potential for clinical applications is unfolding at a breathtaking pace. My impression is that making use of NGS data is more difficult than generating it in the first place. After surveying a number of recent exome sequencing papers, I found that often a complex user certification-procedure was required to obtain genomic data of clinical relevance. Often data are only available upon request to the paper’s authors. If web links or database IDs are provided they are quite often broken or obsolete, and in some other cases, no statement about data availability is made at all.

You had better put on your “I-have-patience” cap to get a hold of the data.

Data access is only half the story. Consider the recent publication of the ENCODE data. The full data set encompasses a whopping 41 terabytes, thereby de facto limiting the download, storage and processing of the data to a few major, well-equipped institutions.

After you have the data, the next step is choosing the right tools from a huge and rapidly growing collection of NGS-specific computational methods, data manipulation utilities and visualization tools.

A typical exome analysis pipeline involves, at the very minimum:

- Quality control of reads
- Alignment-based read mapping to the reference genome
- Duplicate removal
- Variant calling
- Complex multi-step process of functional annotation of variants

A brief analysis of recent reviews on bioinformatics approaches in exome sequencing shows there are virtually dozens of alternative algorithms implementing each step.
A mathematical genius is not required to figure out the number of combinatorial possibilities for combining all these tools into a pipeline is extremely large.

Which tools should you choose?

Literature reviews and benchmark studies are certainly very helpful, but they rarely provide a definitive answer (which tool is best) and, in fact, different tools may be optimal in different settings.

To make matters worse, you are dealing with a moving target as new bioinformatics tools are published virtually on a weekly basis.

Under very real conditions, it is rarely possible to avoid trying several tools out and making one’s own judgment.

Fortunately, pre-built, end-to-end pipelines, which attempt to automate the most typical analysis steps, have started to emerge, but even the number of alternative integrated pipelines is already quite high, including locally installable packages such as the Broad Institute’s widely used GATK as well as cloud-based solutions, both academic and commercial. In particular, the latter category is a very interesting alternative for small research groups and clinical organizations. Research institutions heavily involved in NGS usually build custom pipelines, while those using pre-configured pipelines discover the need to fine-tune parameters, patch the code or develop new, specialized processing steps.

Workflow management tools can accelerate the development of custom analytical pipelines, a leading example being the Konstanz Information Miner (KNIME™), which allows for easy integration of tools into highly complex pipelines, enabling efficient re-use of software. The popular Galaxy platform also offers pipeline building functionality. Workflow management tools and software libraries, such as SeqAn, are an ideal combination, but their use requires professional software development skills. In an interesting parallel development, provider-centric software solutions and even NGS app stores are emerging, exemplified by the cloud-based Illumina BaseSpace™ analysis service.

I believe that in the course of time, there will be considerable consolidation of NGS tools, similar to the developments in microarray analysis during the last decade.

The focus will be shifting to downstream interpretation of the data, beyond mere sequence analysis and statistics. Dealing with this challenge requires more than experimental data repositories and network visualization.

A knowledge management system allowing scientists to design, develop and query the semantic data models describing the systems they are working on and to refine these models continuously as new data becomes available, is what’s needed.

The diverse, specialized databases underlying this process must be seamlessly linked and queried within the system.

The Biomax Pedant-Pro™ Sequence Analysis Suite and the BioXM™ Knowledge Management Environment perform precisely this task.

The Pedant-Pro software can process and analyze DNA sequencing data and enables comparative analysis on a genome-wide scale. Using semantic integration and natural language queries, the BioXM platform integrates the data with results obtained from the Pedant-Pro software and, thanks to its unlimited scaleability and unmatched data integration capability, allows users to link and query any number of data sets relevant to their area of research.