caArray 095 - Lists of Must-Have and Nice-to-Have Features to Support Next-Generation Sequencing Data

Question: What Are Some Proposed Next-Generation Sequencing Features for Future caArray Releases?

Topic: caArray Usage

Release: future caArray release

Date entered: 05/03/2012

Answer

Several research groups have inquired about adding support to next-generation sequencing data in caArray. We have hosted two requirements gathering meetings with potential NGS data users at Columbia University. The users we interviewed belong to two different categories and represent different types of needs. One represents the core facilities and the other represents investigator labs that are heavy users of NGS data. The following two tables summarize features that they deem important. Please add your comments and features you would like to have to the corresponding forum thread.

Must-Have Features

Features	*Explanations*
Support FASTQ, BAM, and VCF Formats	NGS data are stored in those three formats. A FASTQ file contains both base calls and quality scores for sequence reads. It is usually compressed as gz file. A BAM file contains the alignment of multiple reads against a reference genome, as well as base calls and quality scores for each sequence read. It is usually compressed as well. The VCF file contains variant calls.
Support large files (100GB)	A FASTQ file is typically 100GB in size for whole genome sequencing and 5GB in size for exon sequencing (at medium depth coverage). A BAM file is normally around 100 GB.
Need to track capture protocol and sequencing platform	For many applications (e.g., exome sequencing, ChIP-seq) specific target genomic DNA regions (e.g., exons) have to be captured and sequenced. Knowledge of both the capture protocol and the sequencing platform is important.
Need to save multiple protocols for RNA-seq	There are many protocols that can be used to capture and filter RNA, such as ribosome RNA reduction, polyA pull-down. All applicable protocols should be saved along with an experiment/project.
Need to capture the number of reads and the depth of coverage	Metadata about what's in a FASTQ and/or BAM file is informative for researchers.
Support both single-end sequencing and paired-end sequencing	For single-end sequencing, one FASTQ and one BAM file are produced. For paired-end sequencing, two FASTQ and one BAM file are produced.
Support many-to-one mapping between FASTQ files and samples.	For whole genome sequencing, usually several FASTQ files are produced for a single sample.
Record the version of the reference genome sequence used in read alignment	The version of the reference genome used for mapping and base calling is extremely important and should be provided for each BAM file.
Support for cufflinks output file	It is desirable to be able to retrieve FPKM data for genes or transcripts.

Nice-To-Have Features

Features	*Explanations*
Connect to Galaxy or other data analysis tools	Once data are stored in caArray, this connectivity can make it easy for researchers to conduct subsequent data analysis.
Ability to search with an aggregate of attributes	Right now caArray only allows searching with one attribute such as experiment title, sample name, etc. It will be nice to enable search with multiple attributes.
Ability to download through Aspera or something more efficient than ftp for large archives and perform MD5 checksum	Data uploading and downloading will be time consuming given the file size. Faster is always better.

Have a comment?

Please leave your comment in the caArray End User Forum.