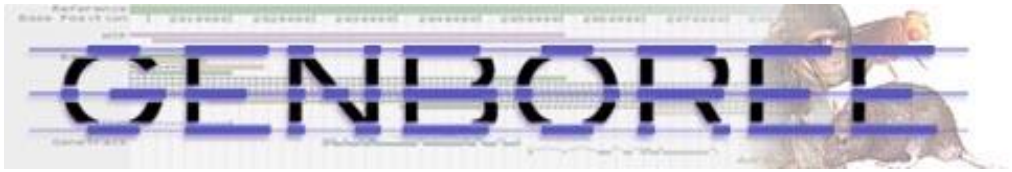


Small and long RNA-Seq pipelines in the Genboree Workbench

NIH Extracellular RNA Communication Consortium
2nd Investigators' Meeting
May 19th, 2014

Sai Lakshmi Subramanian – sailakss@bcm.edu (Primary Contact)
Kevin Riehle – riehle@bcm.edu
Bioinformatics Research Laboratory, Baylor College of Medicine, Houston, TX
Data Coordination Component (DCC) of exRNA Communication Consortium

Robert Kitchen - rob.kitchen@yale.edu
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT
Data Integration and Analysis Component (DCC) of exRNA Communication Consortium

Outline

- Introduction to Genboree
- Genboree Workbench Basics
- Data analysis tools
  - RNA-Seq data – RSEQtools
  - Setting up long RNA-Seq data analysis in the Genboree Workbench
  - Small RNA-Seq data – smallRNA-seq Pipeline
  - Setting up small RNA-Seq data analysis in the Genboree Workbench

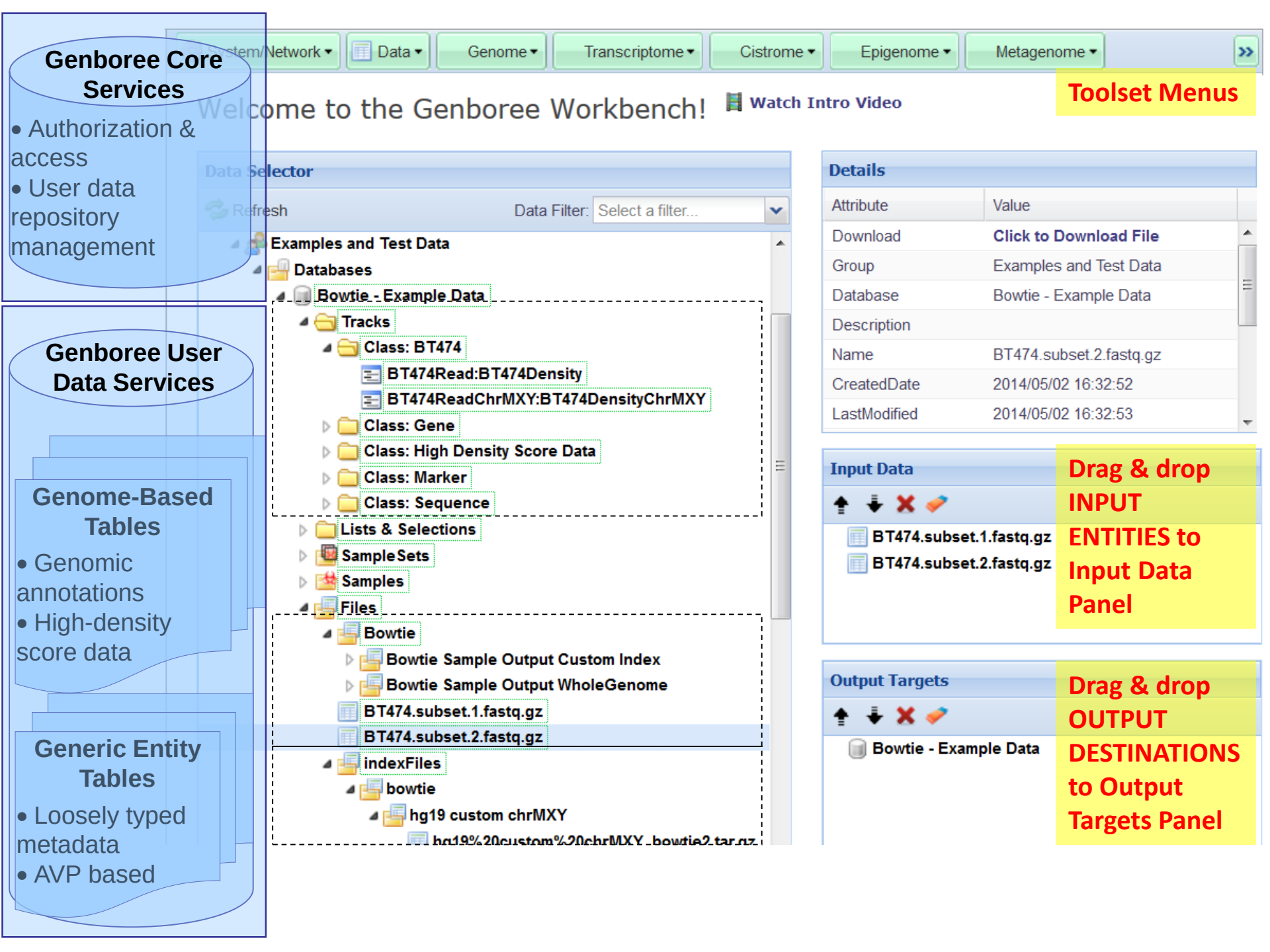
Genboree Workbench Basics

- Genboree URL - http://www.genboree.org/

- Watch Video Tutorial – Introduction to Genboree Workbench
https://docs.google.com/file/d/0Bz3_YiJBA_j3Tk1uOFlIazdMbkk/edit

- How do I create a new account in Genboree?
  http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=493

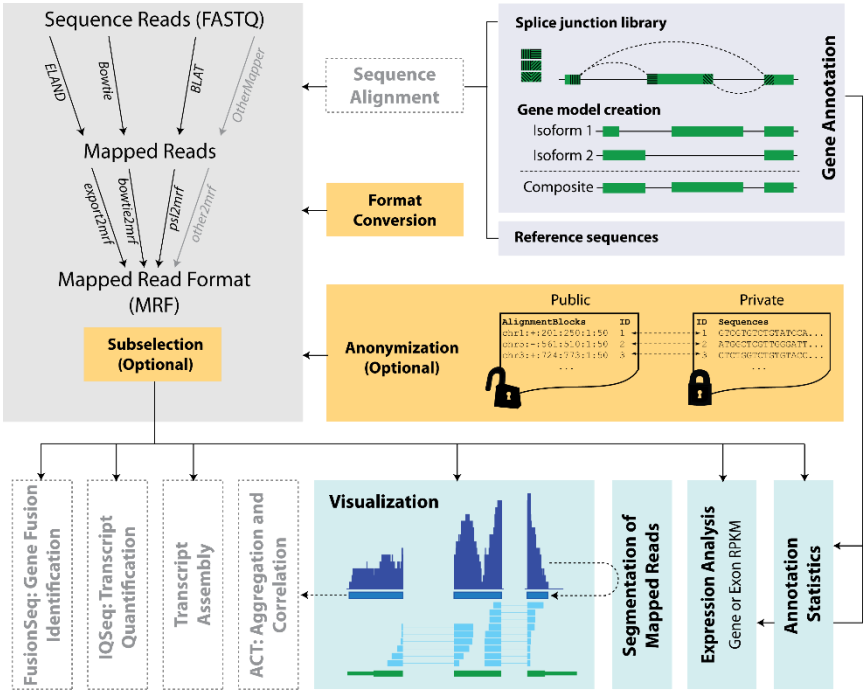
- How to find Help within the Genboree Workbench?
  http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=497
  - Help >> Getting Started
  - Help >> Tool Map
  - Help >> Tool Help Resources

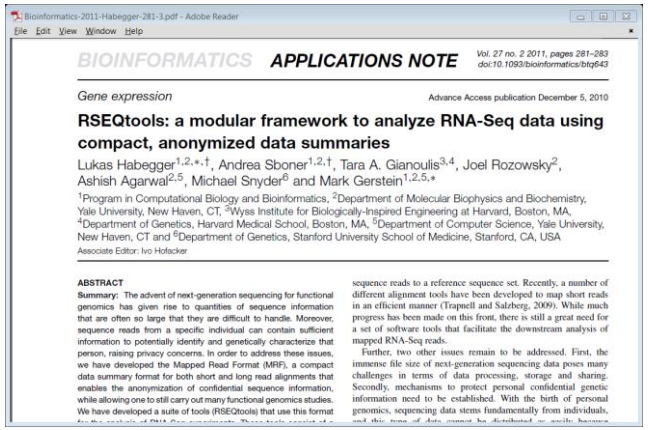

Genboree Core Services
• Authorization & access
• User data repository management
Genboree User Data Services
Genome-Based Tables
• Genomic annotations
• High-density score data
Generic Entity Tables
• Loosely typed metadata
• AVP based
Data
Genome
Transcriptome
Cistrome
Epigenome
Metagenome
Welcome to the Genboree Workbench!
Watch Intro Video
Toolset Menus
Data Selector
Refresh
Data Filter: Select a filter...
Examples and Test Data
Databases
Bowtie - Example Data
Tracks
Class: BT474
BT474Read:BT474Density
BT474ReadChrMXY:BT474DensityChrMXY
Class: Gene
Class: High Density Score Data
Class: Marker
Class: Sequence
Lists & Selections
SampleSets
Samples
Files
Bowtie
Bowtie Sample Output Custom Index
Bowtie Sample Output WholeGenome
BT474.subset.1.fastq.gz
BT474.subset.2.fastq.gz
indexFiles
bowtie
hg19 custom chrMXY
Details
Attribute
Value
Download
Click to Download File
Group
Examples and Test Data
Database
Bowtie - Example Data
Description
Name
BT474.subset.2.fastq.gz
CreatedDate
2014/05/02 16:32:52
LastModified
2014/05/02 16:32:53
Input Data
BT474.subset.1.fastq.gz
BT474.subset.2.fastq.gz
Drag & drop INPUT ENTITIES to Input Data Panel
Output Targets
Bowtie - Example Data
Drag & drop OUTPUT DESTINATIONS to Output Targets Panel

Genboree Workbench Basics

- Create Group
  http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=489
- Create Database
  http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=491
- Create Project
  http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=492
- Upload Data Files
  http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=495

exRNA Data analysis tools





BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 2 2011, pages 281–283
doi:10.1093/bioinformatics/btq643

Gene expression

Advance Access publication December 5, 2010

RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries

Lukas Habegger[1,2,*,†], Andrea Sboner[1,2,†], Tara A. Gianoulis[3,4], Joel Rozowsky[2], Ashish Agarwal[2,5], Michael Snyder[6] and Mark Gerstein[1,2,5,*]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, [3]Wyss Institute for Biologically-Inspired Engineering at Harvard, Boston, MA, [4]Department of Genetics, Harvard Medical School, Boston, MA, [5]Department of Computer Science, Yale University, New Haven, CT and [6]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: The advent of next-generation sequencing for functional genomics has given rise to quantities of sequence information that are often so large that they are difficult to handle. Moreover, sequence reads from a specific individual can contain sufficient information to potentially identify and genetically characterize that person, raising privacy concerns. In order to address these issues, we have developed the Mapped Read Format (MRF), a compact data summary format for both short and long read alignments that enables the anonymization of confidential sequence information, while allowing one to still carry out many functional genomics studies. We have developed a suite of tools (RSEQtools) that use this format

sequence reads to a reference sequence set. Recently, a number of different alignment tools have been developed to map short reads in an efficient manner (Trapnell and Salzberg, 2009). While much progress has been made on this front, there is still a great need for a set of software tools that facilitate the downstream analysis of mapped RNA-Seq reads.

Further, two other issues remain to be addressed. First, the immense file size of next-generation sequencing data poses many challenges in terms of data processing, storage and sharing. Secondly, mechanisms to protect personal confidential genetic information need to be established. With the birth of personal genomics, sequencing data stems fundamentally from individuals,

RSEQtools

A modular and flexible framework to perform common RNA-Seq analysis tasks

Current implemented pipeline

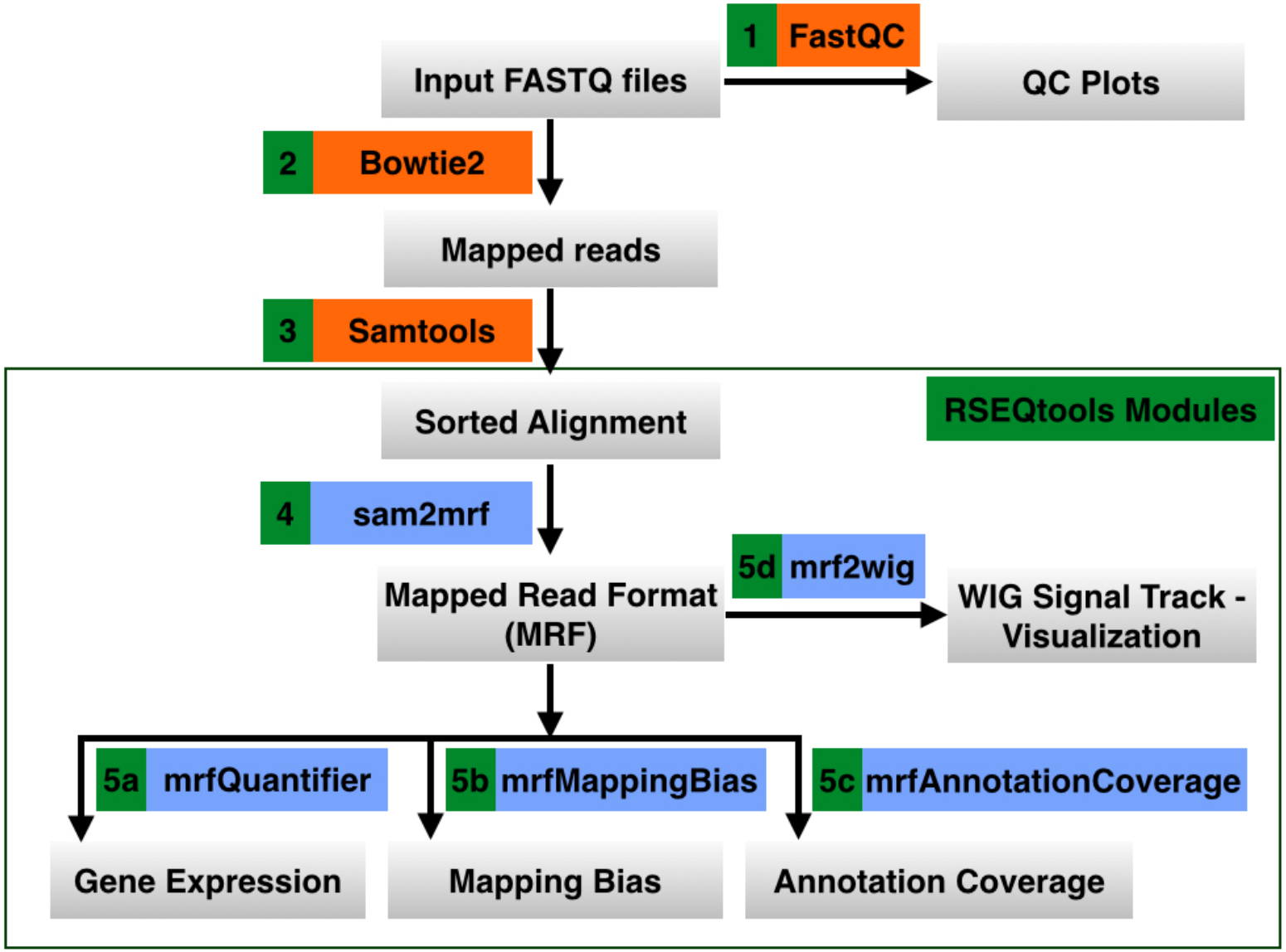
Generates genome-wide signal maps (hg18/19) as well as gene level expression quantifications (RPKM) for UCSC known genes.

Developed by Gerstein Lab – DIAC, exRNA consortium

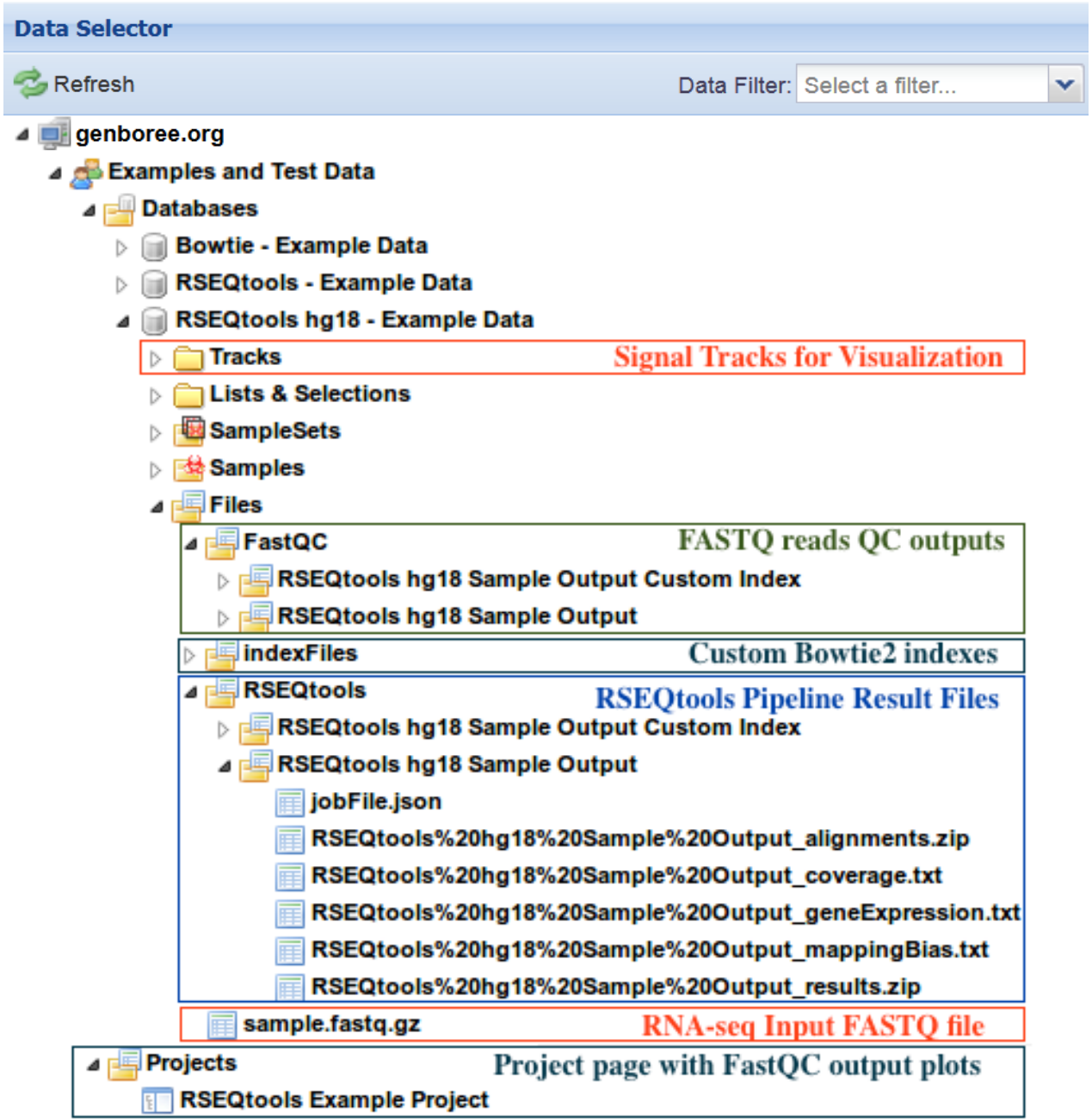
RNA-Seq Analysis using RSEQtools

- **INPUT**: Takes 1 or 2 FASTQ sequences
- **Gene Annotations**: UCSC KnownGene for hg18, hg19
- **OUTPUT**:
  - Alignments in SAM and BAM formats - compressed in *AnalysisName*_alignments.tar.gz
  - All result files compressed in *AnalysisName*_results.tar.gz
  - Key result files for quick access (also found in *AnalysisName*_results.tar.gz):
    - File with gene expression values
    - Mapping bias file
    - Annotation Coverage File
  - Signal Tracks of mapped reads (for visualization in genome browsers) - Uploaded as signal track in user db (if specified in the UI)
  - FastQC Output Plots
  - Custom Bowtie2 index, if generated

RNA-Seq Analysis pipeline in the Genboree Workbench



Example Dataset for RNA-Seq analysis using RSEQtools



exRNA Data analysis tools

smallRNA-seq Pipeline

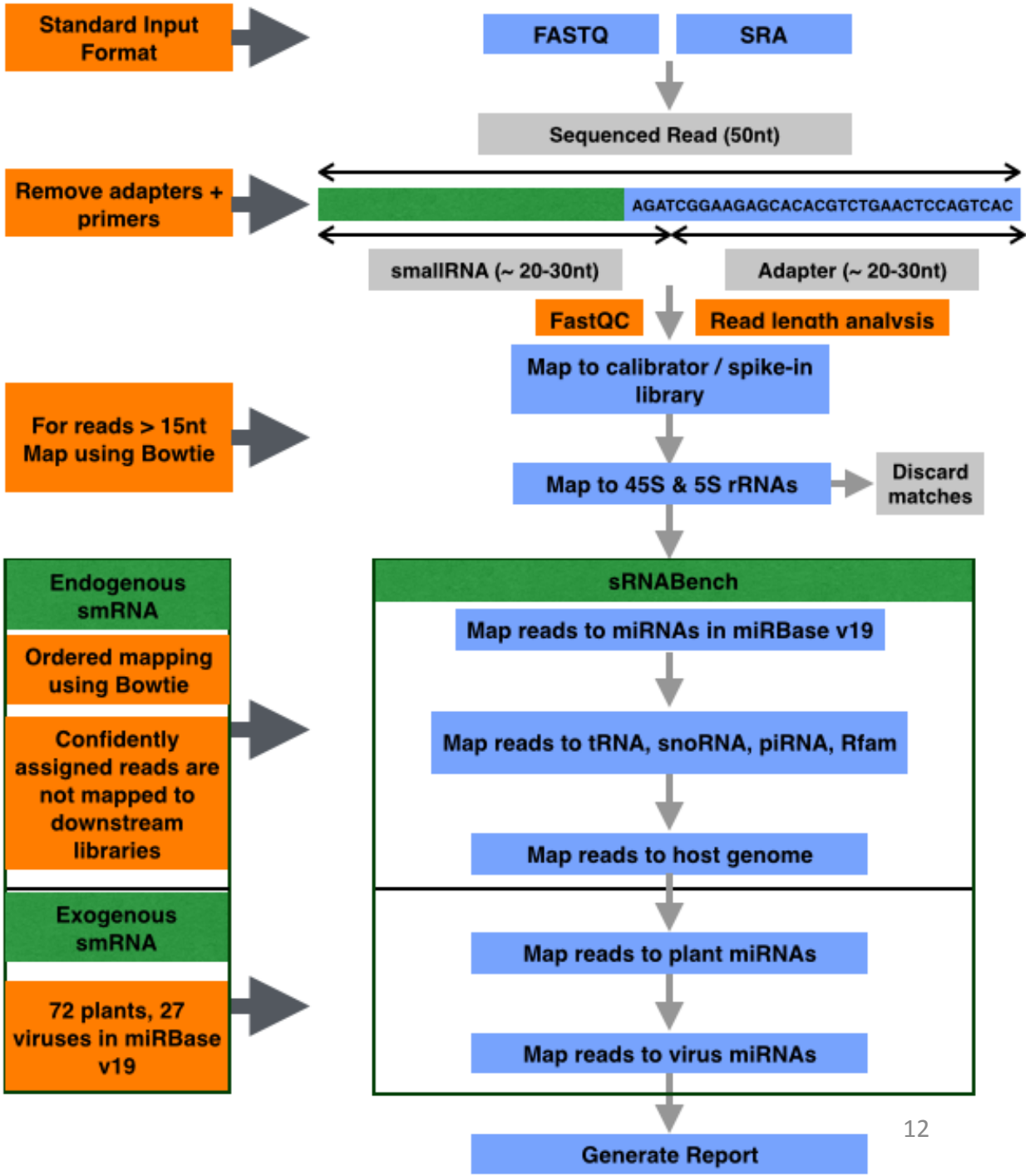
- pre-processing: remove adapter & primer sequences
- contaminants: no poly-A purification leaves rRNAs, etc
- mapping: small reads tend to multi-map to large genomes
- diversity: many species of small RNA (not just miRNAs!)
- quantification: different normalization

Developed by Gerstein Lab – DIAC, exRNA consortium

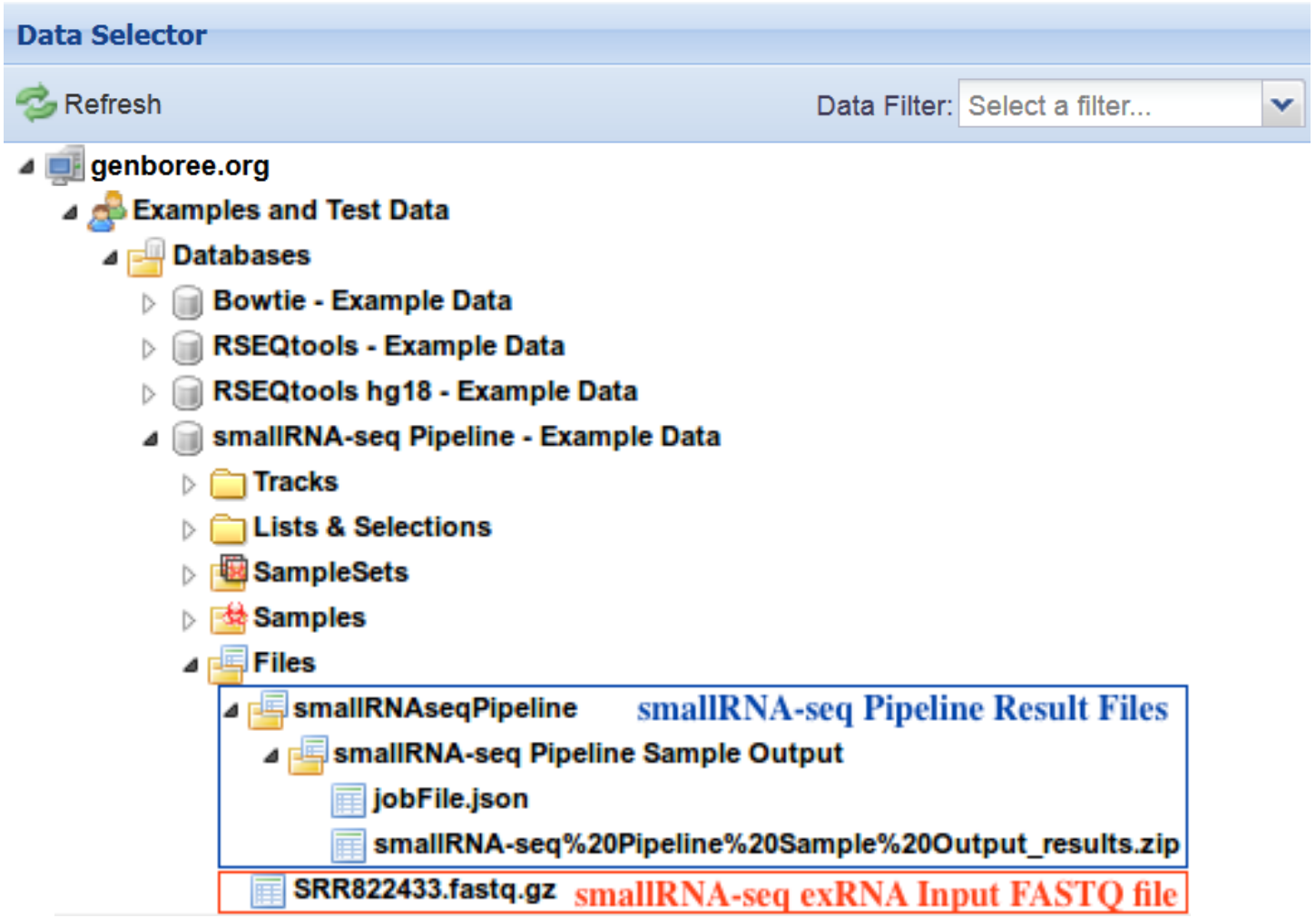
smallRNA-seq Pipeline

- **INPUT**: Takes a single-end FASTQ sequence file
- **Supported Genome build**: hg19
- **OUTPUT**:
  - All result files compressed in *AnalysisName*_results.tar.gz
    - Read length distribution
    - Endogenous Mapping results – rRNA, miRNA, tRNA, snoRNA, piRNA, Rfam
    - Exogenous Mapping results – Plant and virus miRNAs in miRBase
    - 4 files containing sequence reads following
      - Adapter removal
      - rRNA removal
      - Endogenous mapping
      - Exogenous mappins

smallRNA-seq pipeline in the Genboree Workbench



Example Dataset for smallRNA-seq Pipeline



Live Demo of long and small RNA-seq pipelines in the Genboree Workbench

Tutorial Screencasts to setup your analysis in the Genboree Workbench

- Long RNA-Seq data analysis - https://docs.google.com/file/d/0Bz3_YiJBA_j3WHBpQmp2S0ljb0U/edit

- small RNA-Seq data analysis - https://docs.google.com/file/d/0Bz3_YiJBA_j3b1EzR3ZEWGFEam8/edit