Third-Party Analysis Software and Utilities

Several applications are available for analysis and visualization of Illumina sequencing data.

Highlights

- **Several commonly-used feature-rich analysis packages support Illumina workflows**
- **Some third-party tools are a good starting point for next-generation sequencing (NGS) users with limited informatics support**
- **Third-party visualization and process quality control tools are useful in both research and production environments**

Why Use a Third-Party Analysis Tool?

The number of available software packages and third-party utilities have continued to grow along with the types of kits and applications available for Illumina sequencing systems. Some of these packages have become commonly-used tools by next-generation sequencing users, and represents a good starting point for those who might have limited analysis experience or informatics support.

This document describes the following third-party analysis tools that support Illumina sequencing data. Most of the tools listed are open source tools, unless otherwise noted[1].

Alignment	BWA Bowtie MAQ Novoalign[1]
***De novo* Assembly**	ABySS ALLPATHS-LG SSAKE Velvet
SNP/Indel Discovery	GATK SAM Tools SOAPsnp
Genome Annotation Browser	Anno-J Integrative Genomics Viewer UCSC Genome Browser
ChIP-Seq Applications	CisGenome MACS PeakSeq
Transcriptome Analysis	Partek Genomic Suite[1] RNA-Star TopHat
Metagenomics	MEGAN Qiime
Integrated Solutions	Avadis NGS[1] CLCBio Genomics Workbench[1] Galaxy
Utilities and QC Tools	FastQC Picard[1]

1 Commercially-licensed tool

Alignment

Bowtie

bowtie-bio.sourceforge.net/index.shtml

Bowtie is an ultra-fast, memory-efficient short read aligner able to align large sets of short DNA sequences (reads) to large genomes. Bowtie is cabable of aligning 35 bp reads to the human genome at a rate of 25 million reads per hour on a typical workstation. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small. For example, for the human genome, the index is typically about 2.2 GB (for unpaired alignment) or 2.9 GB (for paired-end alignment).

- Supports multiples processors
- Command line for Windows, Mac OS X, Linux, and Solaris
- Developed with C++
- Available under academic license from Johns Hopkins University and listed authors

BWA

bio-bwa.sourceforge.net

BWA maps low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW, and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100 bp, while the remaining two algorithms are better suited for longer sequences ranging from 70 bp to 1 Mbp.

- Supports gapped alignment and paired-end mapping
- Developed with C
- Available under GNU General Public License version 3.0 (GPLv3)

MAQ

maq.sourceforge.net

MAQ builds assemblies by mapping short reads to reference sequences generated by next-generation sequencing instruments. MAQ first aligns reads to reference sequences and then calls the consensus. At the mapping stage, MAQ performs ungapped alignment. For single-end reads, MAQ is able to find all hits with up to two or three mismatches, depending on a command-line option. For paired-end reads, it finds all paired hits with one of the two reads containing up to one mismatch. At the assembling stage, MAQ calls the consensus based on a statistical model.

- Supports single reads and paired-end reads
- Available under GNU General Public License version 3.0 (GPLv3)

Novoalign

www.novocraft.com/main/index.php

Novocraft Technologies is a bioinformatics company specializing in the development of fast and accurate tools for NGS analysis. The Novoalign package is an aligner for single-ended and paired-end reads from Illumina sequencers, and finds global optimum alignments using the full Needleman-Wunsch algorithm, including affine gap penalties.

- Uses base qualities at all steps in the alignment for greater accuracy
- Generates native SAM/BAM alignment output format
- Available under free, evaluation, and commercial licensing terms from Novocraft

De novo Assembly

ABySS

www.bcgsc.ca/platform/bioinfo/software/abyss

Assembly By Short Sequences (ABySS) is a *de novo* sequence assembler that is designed for very short reads. The single-processor version is useful for assembling genomes up to 40–50 Mbases in size. The parallel version is implemented using MPI and capable of assembling larger genomes.

- Merges overlapping read pairs
- Uses sequence overlap graph to layout and merge contigs
- Attempts to fill scaffold gap with consensus of all paths between contigs
- Developed with C++

ALLPATHS-LG

www.broadinstitute.org/software/allpaths-lg/blog

ALLPATHS is a short read genome assembler from the Computational Research and Development group at the Broad Institute. ALLPATHS-LG is a whole-genome shotgun assembler that generates high-quality genome assemblies using short reads (~100 bp) such as those produced by the new generation of sequencers. The significant difference between ALLPATHS and traditional assemblers such as Arachne is that ALLPATHS assemblies are not necessarily linear, but instead are presented in the form of a graph. This graph representation retains ambiguities, such as those arising from polymorphism, uncorrected read errors, and unresolved repeats; thereby, providing information that has been absent from previous genome assemblies.

- Supports Illumina-based sequencing output, optimized for, but not necessarily limited to, reads of length 100 bases
- Available under open access license from the Broad Institute

SSAKE

www.bcgsc.ca/platform/bioinfo/software/ssake

The Short Sequence Assembly by K-mer search and 3' read Extension (SSAKE) is a genomics application for aggressively assembling millions of short nucleotide sequences by progressively searching for perfect 3'-most k-mers using a DNA prefix tree. SSAKE is designed to help leverage information from short sequence reads by stringently clustering them into contigs for use in characterizing novel sequencing targets.

- SSAKE is written in PERL and runs on Linux
- Available under the terms of the GNU General Public License

Velvet

www.ebi.ac.uk/~zerbino/velvet

Velvet is a *de novo* genomic assembler specifically designed for short read sequencing technologies. Velvet currently takes short read sequences and removes errors, and then produces high quality unique contigs. Using paired-end reads and long read information, when available, Velvet retrieves the repeated areas between contigs.

- Tested on Linux 64-bit, Mac OS X, and Cygwin; designed for 64-bit Linux-compatible environments
- Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI) near Cambridge in the United Kingdom
- Velvet source code is freely available under the GPL agreement

SNP/Indel Discovery

GATK

www.broadinstitute.org/gatk

The Genome Analysis Toolkit (GATK) analyzes next-generation resequencing data. GATK offers a wide variety of tools with a primary focus on variant discovery and genotyping, as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine, and high-performance computing features make it capable of taking on projects of any size.

- Developed at the Broad Institute
- Designed to run on Linux and other POSIX-compatible platforms (i.e., Mac OS X, Cygwin)
- Supports variant discovery, genotyping, and filtering workflows and protocols
- Full-featured GATK is available for non-commercial use, or licensed through Appistry

SAM Tools

samtools.sourceforge.net

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing, and generating alignments in a per-position format. By default, SNPs are called with a Bayesian model identical to the one used in MAQ, as well as a simplified SOAPsnp model. Indels are called with a simple Bayesian model. The caller performs local realignment to recover indels that occur at the end of a read, but appear to be contiguous mismatches.

- Supports BAM format and FASTA reference format
- Open source available under BSD and MIT licensing; free for non-commercial use

SOAPsnp

soap.genomics.org.cn/soapsnp.html

SOAPsnp assembles consensus sequences for the genome of a newly-sequenced individual based on the alignment of the raw sequencing reads against the known reference. The SNPs are identified on the consensus sequence through comparison with the reference. SOAPsnp uses a method based on Bayes' theorem, the reverse probability model, to call consensus genotype by carefully considering the data quality, alignment, and recurring experimental errors.

- Command line driven program written in C/C++ that generally runs on a 64-bit Linux system
- Available under the terms of the GNU General Public License

Genome Annotation Browser

Anno-J

www.annoj.org

Anno-J is a Web 2.0 application designed for visualizing deep sequencing data and other genome annotation data. It is intended to run in modern W3C-compliant browsers, and allows flexible configuration of plugins and data streams from providers located on the internet. Anno-J is a 100% Javascript application that enables the inclusion of logical components by pointing to them in an Anno-J instance, and allows custom designed plug-ins without consideration for stylistic aspects or data management, and allows remote hosting of plug-ins.

- Built around the REST philosophy and designed to leverage the distributed nature of the internet
- Supports gene models with one level deep sub-components
- Available under a Creative Commons by-nc-sa 3.0 license

Integrative Genomics Viewer

www.broadinstitute.org/igv

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large integrated genomic data sets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations. IGV 2.0 requires that Java 6, or later, is installed. Previous versions of IGV require Java 5, or later.

- Supports custom genome references as well as a variety of applications, such as ChIP-Seq, RNA-Seq, copy number, mutation data, and sequence alignment data
- Permission to use IGV is granted under the GNU LGPL license

UCSC Genome Browser

genome.ucsc.edu

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group at the University of California Santa Cruz (UCSC). The Genome Browser stacks annotation tracks beneath genome coordinate positions, which allows rapid visual correlation of different types of information. Use the UCSC Genome Browser to look at a whole chromosome for gene density, open a specific cytogenetic band to see a positionally-mapped disease gene candidate, or zoom in to a particular gene to view its spliced expressed sequence tags (ESTs) and possible alternative splicing.

- Supports text and sequence-based searches for quick access to any region of specific interest
- Secondary links from individual entries within annotation tracks lead to sequence details and supplementary off-site databases
- Genome Browser and Blat software are free for academic, nonprofit, and personal use; a license is required for commercial use

ChIP-Seq Applications

CisGenome

www.biostat.jhsph.edu/~hji/cisgenome

CisGenome is an integrated tool for tiling array, ChIP-seq, genome and cis-regulatory element analysis. Use CisGenome to study context-dependency of cis-regulatory elements, and find collaborating proteins of the transcription factor of interest by taking advantage of public ChIP data from diverse cell types for a variety of proteins.

- Designed for experiments with IP and control samples
- Available dynamic read rate model that estimates background read sampling rate using a local window
- Download and use permission granted for non-commercial research use only

MACS

liulab.dfci.harvard.edu/MACS

Model-based Analysis of ChIP-Seq (MACS) analyzes short reads. MACS empirically models the length of the sequenced ChIP fragments, which tend to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, which allows for more sensitive and robust prediction. MACS compares favorably to existing ChIP-Seq peak-finding algorithms. It is a publicly available open source used for ChIP-Seq, with or without control samples.

- Uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence
- Distributed under the terms of Artistic License

PeakSeq

info.gersteinlab.org/PeakSeq

PeakSeq is a program for identifying and ranking peak regions in ChIP-Seq experiments. As input, PeakSeq takes mapped reads from a ChIP-Seq experiment and mapped reads from a control experiment, and outputs a file with peak regions ranked with increasing Q-values.

- Supports multiple input read formats, such as SAM, ELAND, default Bowtie format, and tagAlign
- Supports BAM by piping SAM output from SamTools
- Developed with C/Perl

Transcriptome Analysis

Partek Genomic Suite

www.partek.com

Partek Genomic Suite (GS) empowers biologists to analyze RNA-Seq, ChIP-Seq, Methyl-Seq and DNA-Seq easily by following dedicated workflows such as Data import, Quality Control, Statistical Analysis, Clustering, Gene Ontology, and Pathway analysis, including highly interactive and intuitive visualizations, dedicated genome browser, and data analysis reporting features. At the core of Partek GS, annotations are processed efficiently to enable the correlation of various data formats and save time.

- Peak representation of protein binding sites of interest from aligned reads
- Supports differential expression and alternative splicing based on known mRNA annotation
- Enables the differential expression of known microRNAs (miRNA) between different samples
- Commercial grade package with free 14-day trial license available

RNA-Star

code.google.com/p/rna-star

STAR aligns RNA-seq reads to a reference genome using uncompressed suffix arrays. Accurate alignment of high-throughput RNA-seq data is a challenging and yet unsolved problem due to the non-contiguous transcript structure, relatively short read lengths, and constantly increasing throughput of sequencing technologies. Currently available RNA-seq aligners suffer from high mapping error rates, low mapping speed, read length limitation, and mapping biases.

- Discover non-canonical splices and chimeric (fusion) transcripts
- Capable of mapping full-length RNA sequences
- Implemented as a standalone C++ code
- Free open source software distributed under GPLv3 license

TopHat

tophat.cbcb.umd.edu

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort between the Institute of Genetic Medicine at Johns Hopkins University, the Department of Mathematics and Department of Molecular and Cell Biology at the University of California, Berkeley, and the Department of Stem Cell and Regenerative Biology at Harvard University.

- Finds splice junctions without a reference annotation
- Identifies potential exons, since many RNA-Seq reads will contiguously align to the genome
- Distributed under the terms of Artistic License

Metagenomics

MEGAN

ab.inf.uni-tuebingen.de/software/megan/welcome.html

MEGAN is a tool for studying the taxonomical content of a set of DNA reads, typically collected in a metagenomics project. In a pre-processing step, a sequence comparison of all reads with a suitable database of reference DNA or protein sequences must be performed to produce an input file for the program. MEGAN is suitable for DNA reads (metagenome data), RNA reads (metatranscriptome data), peptide sequences (metaproteomics data), and 16S rRNA data (amplicon sequencing) using a suitable synonyms file that maps SILVA IDs to taxon IDs.

- Interactively explore the data set
- Interactively inspect the assignment of reads to a specific node
- Free academic licensing is available under certain conditions

Qiime

qiime.org

QIIME is an open source software package for comparison and analysis of microbial communities, primarily based on high-throughput amplicon sequencing data (such as SSU rRNA) generated on a variety of platforms, but also supporting analysis of other types of data, such as shotgun metagenomic data. QIIME starts with raw sequencing output from initial analyses, such as OTU picking, taxonomic assignment, and construction of phylogenetic trees from representative sequences of OTUs, and generates downstream statistical analysis, visualization, and production of publication-quality graphics.

- Supports analysis of data from one or more sequencing platforms
- Parallelization capabilities for computationally-intensive steps
- Open source software

Integrated Solutions

Avadis NGS

www.avadis-ngs.com

Avadis is the data mining and visualization platform at the core of all bioinformatics products developed by Strand Scientific Intelligence. The Avadis platform has been enhanced with state-of-the-art algorithms for analysis, visualization, and management of next-generation sequencing data on a wide range of computing infrastructures.

- Supports extensive workflows for alignment, RNA-Seq, small RNA-Seq, DNA-Seq, and ChIP-Seq experiments
- Includes an embedded feature-rich Genome Browser with custom visualizations that provide an intuitive feel for analysis results
- Commercial-grade package with free 20-day trial license is available

CLCBio Genomics Workbench

www.clcbio.com

CLCBio Genomics Workbench is a comprehensive and user-friendly analysis package for analyzing, comparing, and visualizing next-generation sequencing data. It includes a number of features within the fields of genomics, transcriptomics, and epigenomics, and includes all features of CLC Main Workbench.

- Support all major next-generation sequencing platforms
- Supports read mapping as well as *de novo* assembly of hybrid data
- Available for Windows, Mac OS X, and Linux platforms
- Commercial-grade package with 6-month free trial license available

Galaxy

galaxyproject.org

Galaxy is an open, web-based platform for data intensive biomedical research. Programming experience is not required to specify parameters, run tools, and perform workflows.

The Galaxy team is a part of BX at Penn State, and the Department of Biology and Department of Mathematics and Computer Science at Emory University.

- Extensive user documentation that is applicable to any public or local Galaxy instance is available
- Free, public, internet accessible resource
- Data transfer and data storage are not encrypted

Utilities and QC Tools

FastQC

www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC provides a simple way to perform quality control checks on raw sequence data from high-throughput sequencing pipelines. It provides a modular set of analyses to assess data for problems before performing further analysis.

- Imports data from BAM, SAM, or FastQ files (any variant)
- Produces summary graphs and tables to quickly assess data
- Free open source software distributed under GPLv3 license

Picard

picard.sourceforge.net

Picard comprises Java-based command-line utilities that manipulate SAM files, and a Java API (SAM-JDK) for creating new programs that read and write SAM files. Both SAM text format and SAM binary (BAM) format are supported.

- Command-line tools are packaged as executable jar files
- Most of the commands are designed to run in 2 GB of JVM, so the JVM argument -Xmx2g is recommended
- Source code is available under Apache V2.0 and MIT Licensing

Illumina, Inc. • Headquartered in San Diego, California, U.S.A. • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2013-010 Current as of 30 April 2013

illumina®