Orione Documentation

version 0.1

CRS4 Bioinformatica

July 18, 2013

Contents

Orione

Common tools

Get data

Data can be added to Orione by the following tools:

- *Upload File* from your computer
- *UCSC Main* table browser
- *UCSC Archaea* table browser
- *EBI SRA* ENA SRA
- *Gent Nucl/Prot from Taxa*

NGS: quality control

Tools for **NGS quality control** are:

- *FastQC:Read* QC reports using FastQC
- *Quality format converter* (ASCII-Numeric)
- *Compute quality statistics*
- *Draw quality score boxplot*
- *Draw nucleotides distribution chart*
- *Build base quality distribution*

NGS: manipulation

Tools to **manipulate NGS data in Orione**:

- *FASTQ positional and quality trimming*
- *Paired-end compositional filtering*
- *Remove reads* from FastQ files

FASTX-TOOLKIT for FASTQ data

- *FastQ to FASTA* converter
- *Filter by quality*
- *Remove sequencing artifacts*
- *Barcode splitter*
- *Clip adapter sequences*
- *Collapse sequences*
- *Rename sequences*
- *Reverse-complement*
- *Trim sequences*

ILLUMINA FASTQ

- *FASTQ Groomer* converts between various FASTQ quality formats
- *FASTQ splitter* on joined paired-end reads
- *FASTQ joiner* on paired-end reads

- *FASTQ summary statistics* by column

ROCHE-454 data

- *Select high quality segments*
- *Combine FASTA and QUAL into FASTQ*

Generic FASTQ manipulation

- *Filter FASTQ reads* by quality score and length
- *FASTQ trimmer* by column
- *FASTQ quality trimmer* by sliding window
- *FASTQ masker* by quality score
- *FASTQ interlacer* on paired-end reads
- *FASTQ de-interlacer* on paired-end reads
- *Manipulate FASTQ reads* on various attributes
- *FASTQ to FASTA* converter
- *FASTQ to tabular* converter
- *Tabular to FASTQ* converter

NGS: mapping

BLAT

BLAT produces two major classes of alignments:

- at the DNA level between two sequences that are of 95% or greater identity, but which may include large inserts;
- at the protein or translated DNA level between sequences that are of 80% or greater identity and may also include large inserts.

The output of BLAT is flexible. By default it is a simple tab-delimited file which describes the alignment, but which does not include the sequence of the alignment itself. Optionally it can produce BLAST and WU-BLAST compatible output as well as a number of other formats.

Details

References

BLAT - The BLAST-Like Alignment Tool.
W.J. Kent.
Genome Research 2002, 12(4), 656-664
http://genome.cshlp.org/content/12/4/656

Other **mapping** tools:

- *Lastz* map short reads against reference sequence
- *Map with Bowtie* for Illumina
- *Bowtie2* is a short-read aligner
- *Map with BWA* for Illumina
- *Map with BFAST*
- *Map with Mosaik*

NCBI BLAST+

NCBI BLAST+ available tools:

NCBI BLAST+ blastn

Search nucleotide database with nucleotide query sequence(s).

NCBI BLAST+ blastp

Search protein database with protein query sequence(s).

NCBI BLAST+ blastx

Search protein database with translated nucleotide query sequence(s).

NCBI BLAST+ tblastn

Search translated nucleotide database with protein query sequence(s).

NCBI BLAST+ tblastx

Search translated nucleotide database with translated nucleotide query sequence(s).

BLAST XML to tabular

Convert BLAST XML output to tabular.

NCBI BLAST+ database info

Show BLAST database information from *blastdbcmd*.

NCBI BLAST+ blastdbcmd entry(s)

Extract sequence(s) from BLAST database.

NCBI DustMasker

Mask low complexity regions.

NCBI BLAST+ rpsblast

Search protein domain database (PSSMs) with protein query sequence(s).

NCBI BLAST+ rpsblastn

Search protein domain database (PSSMs) with translated nucleotide query sequence(s).

NCBI BLAST+ makeblastdb

Make BLAST database.

Microbiology

Get microbial data

This tool will allow you to obtain various genomic datasets for any completed **Microbial Genome Project** as listed at NCBI.

Current datasets available include:

- CDS
- tRNA
- rRNA
- FASTA Sequences
- GeneMark Annotations
- GeneMarkHMM Annotations
- Glimmer3 Annotations

NGS: RNA-seq for bacteria

EDGE-pro

EDGE-pro, **Estimated Degree of Gene Expression in PROkaryots** is an efficient software system to **estimate gene expression levels prokaryotic genomes** from RNA-seq data. EDGE-pro uses Bowtie2 for alignment and then estimates expression directly from the alignment results.

EDGE-pro includes routines to assign reads aligning to overlapping gene regions accurately. 15% or more of bacterial genes overlap other genes, making this a significant problem for bacterial RNA-seq, one that is generally ignored by programs designed for eukaryotic RNA-seq experiments.

Details

Input files with gene coordinates in PTT and RNT format can be retrieved with the Get EDGE-pro Files tool available in Galaxy, or downloaded from the NCBI ftp repository. EDGE-pro accepts files in Sanger FASTQ format (Galaxy type fastqsanger). Use the FASTQ Groomer to prepare your files.

All 3 types of files (FASTA reference genome, PTT and RNT) must have the same order of chromosomes/plasmids (e.g. if chr1 is before chr2 in genome.fasta file, then chr1 must be before chr2 in ptt and rnt files as well). If there is no PTT or RNT file for one of chromosomes/plasmids, place this chromosome/plasmid at the end of the file.

References

EDGE-pro: Estimated Degree of Gene Expression in Prokaryotic Genomes.
T. Magoc, D. Wood, and S.L. Salzberg.
Evolutionary Bioinformatics (2013) 9: 127-136.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603529/

NGS: de novo assembly

Velvet

Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454.

Velvet currently takes in short read sequences, removes errors then produces high quality unique contigs. It then uses paired-end read and long read information, when available, to retrieve the repeated areas between contigs.

Details

Velvet consists of two parts:

Velveth

takes in a number of sequence files, produces a hashtable, then outputs two files in an output directory (creating it if necessary), Sequences and Roadmaps, which are necessary to velvetg.

Velvetg

can input sequence files in the following formats: `fasta`, `fastq`, `fasta.gz`, `fastq.gz`, `eland`, `gerald`.

The input files are prepared for the velvet assembler using **velveth**.

References

Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
D.R. Zerbino and E. Birney.
Genome Research 18:821-829.
http://genome.cshlp.org/content/18/5/821

Abyss

ABySS, **Assembly By Short Sequences**, is a de novo sequence assembler that is **designed for short reads**.

Details

Two versions are available in Galaxy:

ABySS

assembles short **unpaired reads**.

ABySS paired-end

assembles short **paired reads**.

References

ABySS: A parallel assembler for short read sequence data.
J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, I. Birol.
Genome Research, 2009 June; 19(6): 1117-1123.
http://genome.cshlp.org/content/19/6/1117

Edena

Edena (*V3.130110*) is an **overlaps graph based short reads assembler** and is **suited to Illumina GA reads**. It can assemble both direct-reverse (paired-ends) and reverse-direct (mate-pairs) datasets. This program requires the **reads to be all the same length**, as Illumina GA reads are. This is due to historical reasons and because it greatly simplifies several computational steps. 454 or Sanger reads are therefore not suited to Edena. If you provide multiple files with different read lengths, Edena will trim the 3' end of the reads so that the reads are all the same length as the shortest reads in the file.

The program was developed in a framework of **whole genome bacterial assemblies**. It is therefore more suited for this kind of task though we also successfully used it for other types of projects.

Details

An assembly with Edena is a **two step process**: overlapping and assembling.

Overlapping mode

The reads files are provided to the program which computes the **transitively reduced overlaps graph**. This structure is then stored together with the sequence reads in a binary file suffixed with *.ovl*.

Assembling mode

The *.ovl* file is provided to the program, as well as some assembly parameters. A **set of contigs** in FASTA format is output.

The purpose of having a two step process is that the *.ovl* file is computed **only once** and can then be used to produce assemblies with different parameters.

References

De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.
D. Hernandez, P. François, L. Farinelli, M. Østerås, J. Schrenzel.
Genome Research. 18:802-809, 2008.
http://genome.cshlp.org/content/18/5/802

NGS: post_assembly

Tools

The following utilities are available for **Post Assembly** processes:

Check bacterial contigs

computes some statistics (including N50 and NG50) over a Contigs Collection.

Check bacterial draft

computes some statistics (including N50 and NG50) over a draft consensus bacterial genome.

Extract contigs from draft

extracts Contigs from a Draft Bacterial Genome if longer than indicated as threshold.

SSPACE

SSPACE, **SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension**, is a script able to extend and scaffold pre-assembled contigs using one or more mate pairs or paired-end libraries, or even a combination.

Some parameters are still missing:

- trimming options (-t)
- unpaired single (-u)
- minimal base ratio (-r)

Details

References

Scaffolding pre-assembled contigs using SSPACE.
M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler and W. Pirovano.
Bioinformatics (2011) 27 (4): 578-579.
http://bioinformatics.oxfordjournals.org/content/27/4/578.full

SSAKE

SSAKE is a genomics application for de novo assembly of millions of very short DNA sequences. It is an easy-to-use, robust, reliable and tractable clustering algorithm for very short sequence reads, such as those generated by Illumina Ltd.

Details

References

Assembling millions of short DNA sequences using SSAKE.
R.L. Warren, G.G. Sutton, S.J.M. Jones, and R.A. Holt.
Bioinformatics (2007) 23(4): 500-501.
http://bioinformatics.oxfordjournals.org/content/23/4/500

SOPRA

SOPRA is an **assembly tool for mate pair/paired-end data** generated by high throughput sequencing technologies, e.g. Illumina and SOLiD platforms.

Details

References

SOPRA: Scaffolding algorithm for paired reads via statistical optimization.
A. Dayarian, T.P. Michael, and A.M. Sengupta.
BMC Bioinformatics 2010, 11:345
http://www.biomedcentral.com/1471-2105/11/345

SEQuel

SEQuel is a tool for correcting errors (i.e., insertions, deletions, and substitutions) in contigs output from assembly. While assemblies of next generation sequencing (NGS) data are accurate, they still contain a substantial number of errors that need to be corrected after the assembly process. The algorithm behind SEQuel makes use of a graph structure called the positional de Bruijn graph, which models k-mers within reads while incorporating their approximate positions into the model.

SEQuel substantially reduces the number of small insertions, deletions and substitusion erros in assemblies of both standard (multi-cell) and single-cell sequencing data. SEQuel was tested mainly on Illumina sequence data, in combination with multiple NGS assemblers, such as Euler-SR, Velvet, SoapDeNovo, ALLPATHS and SPAdes.

Details

References

SEQuel: improving the accuracy of genome assemblies.
R. Ronen, C. Boucher, H. Chitsaz, and P. Pevzner.
Bioinformatics (2012) 28 (12): i188-i196
http://bioinformatics.oxfordjournals.org/content/28/12/i188

MuMMer

MuMMer aligns a drfat sequence to a reference one. It also performes SNP detection.

Details

It runs with default parameters, in the next release a fine tuning will be allowed.

References

Versatile and open software for comparing large genomes.
S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S.L. Salzberg.
Genome Biology 2004, 5:R12
http://genomebiology.com/2004/5/2/R12

Mugsy

Mugsy is a **multiple whole genome alignment** tool.

Details

Mugsy generates a MAF (multiple alignment format) file containing the multiple alignments from FASTA inputs.

This implementation runs with two files only: reference vs contigs/draft(s). For multiple alignment a single Multi-FASTA file containing all contigs should be provided.

References

Mugsy: Fast multiple alignment of closely related whole genomes.
S.V. Angiuoli and S.L. Salzberg.
Bioinformatics (2011), 27 (3): 334-342.
http://bioinformatics.oxfordjournals.org/content/27/3/334

Gene Annotation

Glimmer3

Glimmer is a system for **finding genes in microbial DNA**, especially the genomes of bacteria, archaea, and viruses. Glimmer, **Gene Locator and Interpolated Markov ModelER**, uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA. The IMM approach uses a combination of Markov models from 1st through 8th-order, weighting each model according to its predictive power. Glimmer uses 3-periodic nonhomogenous Markov models in its IMMs.

Details

Glimmer3 is presented as a suite of programs that can be used independently:

Long-ORFs

identifies long, non-overlapping open reading frames (ORFs) in a DNA sequence file.

Extract

reads FASTA-format *sequence-file* and extract from it the subsequences specified by *coords*.

Build-ICM

produces a probability model of coding sequences, called an interpolated context model or ICM.

Glimmer

reads DNA sequences in *sequence-file* and predict genes in them using the Interpolated Context Model. Outputs are a detail file and a prediction file.

Anomaly

reads DNA sequence in *sequence-file* and for each region specified by the coordinates in *coords*, check whether the region represents a normal gene, i.e., it begins with a start codon, ends with a stop codon, and has no frame shifts.

References

Identifying bacterial genes and endosymbiont DNA with Glimmer.
A.L. Delcher, K.A. Bratke, E.C. Powers and S.L. Salzberg.
Bioinformatics 23 (6): 673-679.
http://bioinformatics.oxfordjournals.org/content/23/6/673.full

Prokka

Prokka is a software tool to annotate bacterial, archaeal and viral genomes very rapidly, and produce output files that require only minor tweaking to submit to Genbank/ENA/DDBJ.

Details

Prokka creates several output files:

gff

This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.

gbk

This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.

fna

Nucleotide FASTA file of the input contig sequences.

faa

Protein FASTA file of the translated CDS sequences.

ffn

Nucleotide FASTA file of all the annotated sequences, not just CDS.

sqn

An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.

fsa

Nucleotide FASTA file of the input contig sequences, used by **tbl2asn** to create the `.sqn` file. It is mostly the same as the `.fna` file, but with extra Sequin tags in the sequence description lines.

tbl

Feature Table file, used by **tbl2asn** to create the `.sqn` file.

err

Unacceptable annotations - the NCBI discrepancy report.

log

Contains all the output that Prokka produced during its run. This is a record of what settings you used.

References

Prokka: Prokaryotic Genome Annotation System.
T. Seemann.
In preparation.

tRNAscan

tRNAscan-SE was designed to make rapid, sensitive **searches of genomic sequence** feasible using the selectivity of the Cove analysis package. We have optimized search sensitivity with eukaryote cytoplasmic and eubacterial sequences, but it may be applied more broadly with a slight reduction in sensitivity .

Details

This tool requires FASTA format.

References

tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence.

T.M. Lowe and S.R. Eddy.
Nucleic Acids Research (1997) 25 (5): 0955-964.
http://nar.oxfordjournals.org/content/25/5/0955.full

The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.
P. Schattner, A.N. Brooks and T.M. Lowe.
Nucleic Acids Research (2005) 33 (suppl 2): W686-W689.
http://nar.oxfordjournals.org/content/33/suppl_2/W686.full

Variant Calling

NGS: SNP effects

This tool calculate the effect of variants (SNPs/MNPs/Insertions) and deletions.

Details

References

http://snpEff.sourceforge.net

NGS: Indel analysis

Filter Indels

Allows extracting indels from SAM produced by BWA. Currently it can handle SAM with alignments that have only one insertion or one deletion, and will skip that alignment if it encounters one with more than one indel. It matches CIGAR strings (column 6 in the SAM file) like 5M3I5M or 4M2D10M, so there must be a match or mismatch of sufficient length on either side of the indel.

Details

References

Extract Indels

Given a SAM file containing indels, converts these to an interval file with a column indicating whether it is an insertion or a deletion, and then also can create a BED file for each type (one for insertions, one for deletions). The interval file can be combined with other like files to create a table useful for analysis with the Indel Analysis Table tool. The BED files can be useful for visualizing the reads.

Details

References

Indel analysis table

Creates a table allowing for analysis and comparison of indel data. Combines any number of interval files that have been produced by the tool that converts indel SAM data to interval format. Includes overall total counts for all or some files. The tool has the option to not include a given file's counts in the total column. This could be useful for combined data if the counts for certain indels might be included more than once.

Details

References

Indel analysis

Given an input sam file, this tool provides analysis of the indels. It filters out matches that do not meet the frequency threshold. The way this frequency of occurence is calculated is different for deletions and insertions. The CIGAR string's "M" can indicate an exact match or a mismatch.

Details

References

NGS: SAM tools

Tools to manipulate SAM files:

Filter SAM

allows parsing of SAM datasets using bitwise flag (the second column).

Convert SAM

converts positional information from a SAM dataset into interval format with 0-based start and 1-based end. CIGAR string of SAM format is used to compute the end coordinate.

SAM-to-BAM

produces an indexed BAM file based on a sorted input SAM file.

BAM-to-SAM

produces a SAM file from a BAM file.

Merge BAM files

uses the Picard merge command to merge any number of BAM files together into one BAM file while preserving the BAM metadata such as read groups.

MPileup SNP and indel caller

generates BCF or pileup for one or multiple BAM files. Alignment records are grouped by sample identifiers in @RG header lines. If sample identifiers are absent, each input file is regarded as one sample.

Generate pileup

uses SAMTools' pileup command to produce a pileup dataset from a provided BAM dataset. It generates two types of pileup datasets depending on the specified options.

Filter pileup on coverage and SNPs

allows one to find sequence variants and/or sites covered by a specified number of reads with bases above a set quality threshold. The tool works on six and ten column pileup formats produced with SAMTools pileup command.

Pileup-to-Interval

reduces the size of a results set by taking a pileup file and producing a condensed version showing consecutive sequences of bases meeting coverage criteria. The tool works on six and ten column pileup formats produced with SAMTools pileup command.

flagstat

uses the SAMTools toolkit to produce simple stats on a BAM file.

rmdup

removes potential PCR duplicates: if multiple read pairs have identical external coordinates, only retain the pair with highest mapping quality. In the paired-end mode, this command ONLY works with FR orientation and requires ISIZE is correctly set. It does not work for unpaired reads (e.g. two ends mapped to different chromosomes or orphan reads).

Slice BAM

accepts an input BAM file and an input BED file and creates an output BAM file containing only those alignments that overlap the provided BED intervals.

References

The Sequence Alignment/Map format and SAMtools.
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup.
Bioinformatics (2009) 25 (16): 2078-2079.
http://bioinformatics.oxfordjournals.org/content/25/16/2078

NGS: GATK tools

ALIGNMENT UTILITIES

Depth of Coverage

processes a set of bam files to determine coverage at different levels of partitioning and aggregation. Coverage can be analyzed per locus, per interval, per gene, or in total; can be partitioned by sample, by read group, by technology, by center, or by library; and can be summarized by mean, median, quartiles, and/or percentage of bases covered to or beyond a threshold. Additionally, reads and bases can be filtered by mapping or base quality score.

Print Reads from BAM files

can dynamically merge the contents of multiple input BAM files, resulting in merged output sorted in coordinate order.

REALIGNMENT

Realigner Target Creator

for use in local realignment, emits intervals for the Local Indel Realigner to target for cleaning. Ignores 454 reads, MQ0 reads, and reads with consecutive indel operators in the CIGAR string.

Indel Realigner

performs local realignment of reads based on misalignments due to the presence of indels. Unlike most mappers, this walker uses the full alignment context to determine whether an appropriate alternate reference (i.e. indel) exists and updates SAMRecords accordingly.

BASE RECALIBRATION

Count Covariates on BAM files

is designed to work as the first pass in a two-pass processing step. It does a by-locus traversal operating only at sites that are not in dbSNP. We assume that all reference mismatches we see are therefore errors and indicative of poor base quality. This walker generates tables based on various user-specified covariates (such as read group, reported quality score, cycle, and dinucleotide) Since there is a large amount of data one can then calculate an empirical probability of error given the particular covariates seen at this site, where p(error) = num mismatches / num observations The output file is a CSV list of (the several covariate values, num observations, num mismatches, empirical quality score) The first non-comment line of the output file gives the name of the covariates that were used for this calculation. Note: ReadGroupCovariate and QualityScoreCovariate are required covariates and will be added for the user regardless of whether or not they were specified Note: This walker is designed to be used in conjunction with TableRecalibrationWalker.

Table Recalibration on BAM files

is designed to work as the second pass in a two-pass processing step, doing a by-read traversal. For each base in each read this walker calculates various user-specified covariates (such as read group, reported quality score, cycle, and dinuc) Using these values as a key in a large hashmap the walker calculates an empirical base quality score and overwrites the quality score currently in the read. This walker then outputs a new bam file with these updated (recalibrated) reads. Note: This walker expects as input the recalibration table file generated previously by CovariateCounterWalker. Note: This walker is designed to be used in conjunction with CovariateCounterWalker.

Analyze Covariates

creates collapsed versions of the recal csv file and call R scripts to plot residual error versus the various covariates.

GENOTYPING

Unified Genotyper SNP and indel caller

is a variant caller which unifies the approaches of several disparate callers. Works for single-sample and multi-sample data. The user can choose from several different incorporated calculation models.

ANNOTATION

Variant Annotator

annotates variant calls with context information. Users can specify which of the available annotations to use.

FILTRATION

Variant Filtration on VCF files

filters variant calls using a number of user-selectable, parameterizable criteria.

Select Variants from VCF files

Often, a VCF containing many samples and/or variants will need to be subset in order to facilitate certain analyses (e.g. comparing and contrasting cases vs. controls; extracting variant or non-variant loci that meet certain requirements, displaying just a few samples in a browser like IGV, etc.). SelectVariants can be used for this purpose.

VARIANT QUALITY SCORE RECALIBRATION

Variant Recalibrator

takes variant calls as .vcf files, learns a Gaussian mixture model over the variant annotations and evaluates the variant -- assigning an informative lod score.

Apply Variant Recalibration

applies cuts to the input vcf file (by adding filter lines) to achieve the desired novel FDR levels which were specified during VariantRecalibration.

VARIANT UTILITIES

Validate Variants

validates a variants file.

Eval Variants

is a general-purpose tool for variant evaluation (% in dbSNP, genotype concordance, Ti/Tv ratios, and a lot more).

Combine Variants

combines VCF records from different sources; supports both full merges and set unions. Merge: combines multiple records into a single one; if sample names overlap then they are uniquified. Union: assumes each rod represents the same set of samples (although this is not enforced); using the priority list (if provided), emits a single record instance at every position represented in the rods.

References

A framework for variation discovery and genotyping using next-generation DNA sequencing data.
M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler and M.J. Daly.
Nature Genetics (2011) 43, 491–498.
http://www.nature.com/ng/journal/v43/n5/full/ng.806.html

VCF tools

Intersect two VCF files

uses vcfPytools' intersect command to generate the intersection of two VCF files. Two input files are required as input and the intersection of these two files is generated and sent to the output. These files must be sorted by genomic coordinate to function correctly, although the reference sequence order is no important. The intersection can be calculated on two VCF files or a VCF and a BED file. If the priority file argument is set (this must be equal to one of the input VCF files), then the record written to the output will come from this file. If this argument is not set, the record with the highest quality is written out.

Annotate a VCF file (dbSNP, hapmap)

uses vcfPytools' annotate command annotate a VCF file. Currently, either a hapmap or a dbsnp file should be provided, not both. dbSNP option will annotate the VCF file with dbSNP rsid values. The input dbSNP file must also be in VCF v4.0 format. Only dbSNP entries with VC=SNP are included. hapmap option will annotate the VCF file info string to include HM3 if the record is included hapmap. If the ref/alt values do not match the hapmap file, the info string will be populated with HM3A.

Filter a VCF file

uses vcfPytools' filter command. Quality option will check the variant quality for each record and if it is below the defined value, the filter field will be populated with the filter entry Q[value]. Any value in the info string can be used for filtering by using the 'Filter by info' option. This option takes three values: the info string tag, the cutoff value and whether to filter out those records with less than (lt) or greater than (gt) this value.

Extract reads from a specified region

uses vcfPytools' extract command to extract reads from a specified region of a VCF file.

VCF to MAF Custom Track for display at UCSC

converts a Variant Call Format (VCF) file into a Multiple Alignment Format (MAF) custom track file suitable for display at genome browsers.

Extract consensus Sequence from VCF

extracts Consensus sequence from VCF file.

BCFtools view

converts BCF format to VCF format.

References

Metagenomics

Metaphlan

MetaPhlAn (Metagenomic Phylogenetic Analysis) is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data. MetaPhlAn relies on unique clade-specific marker genes identified from reference genomes, allowing orders of magnitude speedups and unambiguous taxonomic assignments.

Although MetaPhlAn can use both BlastN and BowTie2 in the read-to-marker mapping step, this Galaxy module uses only BowTie2 for computational reasons.

Other available tools are:

MetaPhlAn to PhyloXML

converts the results of metagonome profiling performed with MetaPhlAn from tabular to PhyloXML format. The tool accepts as input only the result of rel_ab analysis (profiling in terms of relative abundaces), not reads_map or clade_profiles analysis.

Get Proteomes from MetaphlanOut

retrieves proteomes from NCBI of the genus and species indicated by a Metaphlan output file.

Details

The input file must be a multi-fasta file containing metagenomic reads loaded with the "Get Data" module in the left panel. Reads can be as short as ~40 nt although lengths higher than 70 nt are recommended.

A synthetic metagenome you can use as sample input is available at http://huttenhower.sph.harvard.edu/sites/default/files/LC1.fna

The output is a two column tab-separated plain file reporting the predicted microbial clades present in the metagenomic samples and the corresponding relative abundances.

All taxonomic levels from domain to species will be reported and higher taxonomic levelis contain the sum of the abundances of its taxonomic leaf nodes (usually species) and, possibly, some lower level "unclassified" clades.

References

Metagenomic microbial community profiling using unique clade-specific marker genes.
N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, C. Huttenhower.
Nature Methods 9(8), 811-814 (2012)
http://www.nature.com/nmeth/journal/v9/n8/full/nmeth.2066.html

Other tools

Other available **metagenomics** tools are:

Fetch taxonomic representation

fetches taxonomic information for a list of GI numbers (sequences identifiers used by the National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov).

Filter taxonomy data

filters Taxonomy Datafile according to the relative abundance.

Summarize taxonomy

computes a summary of all taxonomic ranks for the given taxonomic representation.

Draw phylogeny

produces a graphical representations of phylogenetic tree in PDF format.

Find diagonistic hits

takes data generated by Taxonomy manipulation->Fetch Taxonomic Ranks as input and outputs either a list of sequence reads unique to a particular taxonomic rank, or a list of taxonomic ranks and the count of unique reads corresponding to each rank.

Find lowest diagnostic rank

identifies the lowest taxonomic rank for which a mategenomic sequencing read is diagnostic. It takes datasets produced by Fetch Taxonomic Ranks tool (aka Taxonomy format) as the input.

Poisson two-sample test

checks if the number of reads that fall in a particular taxon in location 1 is different from those that fall in the same taxon in location 2.

Create Krona chart

converts the standard result file of MetaPhlAn or Summarize taxonomy tools in a zoomable pie chart using Krona.

NGS: RNA analysis

RNA-seq

TopHat for Illumina

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.

Details

Tophat accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prepare your files.

Tophat produces two output files:

junctions

a UCSC BED track of junctions reported by TopHat. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.

accepted_hits

a list of read alignments in BAM format.

Two other possible outputs, depending on the options you choose, are insertions and deletions, both of which are in BED format.

References

TopHat: discovering splice junctions with RNA-Seq.
C. Trapnell, L. Pachter, S.L. Salzberg.
Bioinformatics (2009) 25 (9): 1105-1111
http://bioinformatics.oxfordjournals.org/content/25/9/1105

TopHat2

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.

Details

Tophat accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prepare your files.

Tophat produces two output files:

junctions

a UCSC BED track of junctions reported by TopHat. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.

accepted_hits

a list of read alignments in BAM format.

Two other possible outputs, depending on the options you choose, are insertions and deletions, both of which are in BED format.

References

TopHat: discovering splice junctions with RNA-Seq.
C. Trapnell, L. Pachter, S.L. Salzberg.
Bioinformatics (2009) 25 (9): 1105-1111
http://bioinformatics.oxfordjournals.org/content/25/9/1105

Cufflinks

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one.

Details

References

Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms.
C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter.
Nature Biotechnology (2010) 28, 511–515
http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html

Cuffcompare

Cuffcompare helps you:

- compare your assembled transcripts to a reference annotation,
- track Cufflinks transcripts across multiple experiments (e.g. across a time course).

Details

References

Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms.
C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter.
Nature Biotechnology (2010) 28, 511–515
http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html

Cuffmerge

Cuffmerge is part of Cufflinks.

Details

Cuffmerge takes Cufflinks' GTF output as input, and optionally can take a "reference" annotation (such as from Ensembl).

References

Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms.
C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter.
Nature Biotechnology (2010) 28, 511-515
http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html

Cuffdiff

Cuffdiff finds significant changes in transcript expression, splicing, and promoter use.

Details

Cuffdiff takes Cufflinks or Cuffcompare GTF files as input along with two SAM files containing the fragment alignments for two or more samples.

References

Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms.
C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter.
Nature Biotechnology (2010) 28, 511–515
http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html

Filter combined transcripts

The tool uses a tracking file (produced by cuffcompare) to filter a GTF file of transcripts (usually the transcripts produced by cufflinks). Filtering is done by extracting transcript IDs from tracking file and then filtering the GTF so that the output GTF contains only transcript found in the tracking file. Because a tracking file has multiple samples, a sample number is used to filter transcripts for a particular sample.

Details

References

miRNA prediction

miRDeep2 is a software package which can be used for identification of novel and known miRNAs in deep sequencing data and for miRNA expression profiling across samples.

Here miRDeep2 is coupled with the SeqTrimMap algorithm that maps reads using a sequential trimming strategy which allows an efficient mapping of short reads.

Details

This tool uses miRDeep2 and SeqTrimMap, which are licensed separately.

References

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.
M.R. Friedländer, S.D. Mackowiak, N. Li, W. Chen, N. Rajewsky.
Nucleic Acids Research (2012) 40 (1): 37-52.
http://nar.oxfordjournals.org/content/40/1/37

Detection of microRNAs in color space.
A. Marco and S. Griffiths-Jones.
Bioinformatics (2012) 28 (3): 318-323.
http://bioinformatics.oxfordjournals.org/content/28/3/318

Chip-seq

NGS: peak calling

MACS

This tool allows ChIP-seq peak calling using MACS.

Depending upon selected options, 2 to 6 history items will be created; the first output will be a standard BED file and the last will be an HTML report containing links to download additional files generated by MACS. Up to two each of wig and interval files can be optionally created; the interval files are parsed from the xls output.

Details

References

Model-based Analysis of ChIP-Seq (MACS).
Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li and X.S. Liu.
Genome Biology 2008, 9:R137.

http://www.ncbi.nlm.nih.gov/pubmed/18798982

D. Blankenberg et al., in preparation.

MACS14

This tool allows ChIP-seq peak calling using MACS.

Depending upon selected options, 2 to 6 history items will be created; the first output will be a standard BED file and the last will be an HTML report containing links to download additional files generated by MACS. Up to two each of wig and interval files can be optionally created; the interval files are parsed from the xls output.

Details

References

Model-based Analysis of ChIP-Seq (MACS).
Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li and X.S. Liu.
Genome Biology 2008, 9:R137.
http://www.ncbi.nlm.nih.gov/pubmed/18798982

D. Blankenberg et al., in preparation.

SICER

SICER first and foremost is a filtering tool. Its main functions are:

- Delineation of the significantly ChIP-enriched regions, which can be used to associate with other genomic landmarks.
- Identification of reads on the ChIP-enriched regions, which can be used for profiling and other quantitative analysis.

Details

By default, SICER creates files that do not conform to standards (e.g. BED files are closed, not half-open). This could have implications for downstream analysis. To force the output of SICER to be formatted properly to standard file formats, check the "Fix off-by-one errors in output files" option.

References

A clustering approach for identification of enriched domains from histone modification ChIP-Seq data
C. Zang, D.E. Schones, C. Zeng, K. Cui, K. Zhao and W. Peng.
Bioinformatics (2009) 25 (15): 1952-1958.
http://bioinformatics.oxfordjournals.org/content/25/15/1952

CCAT

This tool allows ChIP-seq peak/region calling using CCAT.

Details

References

A signal-noise model for significance analysis of ChIP-seq with negative control.
H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C.-L. Wei, F. Lin and W.-K. Sung.
Bioinformatics (2010) 26 (9): 1199-1204.
http://bioinformatics.oxfordjournals.org/content/26/9/1199

GeneTrack indexer

This tool will create a visualization of the bed file that is selected.

Details

References

GeneTrack - a genomic data processing and visualization framework.
I. Albert, S. Wachi, C. Jiang, B.F. Pugh.
Bioinformatics (2008) 24 (10): 1305-1306.
http://bioinformatics.oxfordjournals.org/content/24/10/1305

Peak predictor

This tool will generate genome wide peak prediction from an index file.

Details

References

GeneTrack - a genomic data processing and visualization framework.
I. Albert, S. Wachi, C. Jiang, B.F. Pugh.
Bioinformatics (2008) 24 (10): 1305-1306.
http://bioinformatics.oxfordjournals.org/content/24/10/1305

Motif tools

MEME

Details

References

Fitting a mixture model by expectation maximization to discover motifs in biopolymers.
T.L. Bailey, C. Elkan.

Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
http://www.cs.utoronto.ca/~brudno/csc2417_10/10.1.1.121.7056.pdf

FIMO

Details

References

FIMO: scanning for occurrences of a given motif.
Grant CE, Bailey TL, Noble WS.
C.E. Grant, T.L. Bailey, W.S. Noble.
Bioinformatics (2011) 27(7): 1017-8.
http://bioinformatics.oxfordjournals.org/content/27/7/1017

Sequence logo

The tool uses Weblogo3 in Galaxy to generate a sequence logo. The input file must be a fasta file in your current history.

Details

References

Other tools

Text manipulation

Data can be manipulated with the following toolsxxx:

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns

- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset
- Secure Hash / Message Digest on a dataset
- Filter on ambiguities in polymorphism datasets
- Arithmetic Operations on tables

FASTQ manipulation

FASTQ files can be manipulated with the following tool:

- Extract multiconsensus sequences from multiFastQ

FASTA manipulation

FASTA files can be manipulated with the following tools:

- Compute sequence length
- Filter sequences by length
- Concatenate FASTA alignment by species
- FASTA-to-Tabular converter
- Tabular-to-FASTA converts tabular file to FASTA format
- FASTA Width formatter
- RNA/DNA converter
- Collapse sequences
- Extract Genomic DNA using coordinates from assembled/unassembled genomes
- CD-Hit

NGS: Picard

Conversion

- FASTQ to BAM creates an unaligned BAM file
- SAM to FASTQ creates a FASTQ file

QC/Metrics for SAM/BAM

- BAM Index Statistics
- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data

BAM/SAM Cleaning

- Add or Replace Groups

- Reorder SAM/BAM
- Replace SAM/BAM Header
- Paired Read Mate Fixer for paired data
- Mark Duplicate reads

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF

- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions
- Filter GTF data by attribute values_list

Join, Subtract and Group

- Join two Datasets side by side on a specified field
- Compare two Datasets to find common or distinct rows
- Subtract Whole Dataset from another dataset
- Group data by a column and perform aggregate operation on other columns
- Column Join

Convert formats

- AXT to concatenated FASTA Converts an AXT formatted file to a concatenated FASTA alignment
- AXT to FASTA Converts an AXT formatted file to FASTA format
- BED to GFF converter
- FASTA to Tabular converter
- GFF to BED converter
- MAF to BED Converts a MAF formatted file to the BED format
- MAF to Interval Converts a MAF formatted file to the Interval format
- MAF to FASTA Converts a MAF formatted file to FASTA format
- Tabular to FASTA converts tabular file to FASTA format
- FASTQ to FASTA converter
- Wiggle to Interval converter
- SFF converter

Operate on genomic intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset

- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval
- Profile Annotations for a set of genomic intervals

NGS: Simulation

This tool simulates an Illumina run and provides plots of false positives and false negatives. It allows for a range of simulation parameters to be set. Note that this simulation sets only one (randomly chosen) position in the genome as polymorphic, according to the value specified. Superimposed on this are *sequencing errors*, which are uniformly (and randomly) distributed. Polymorphisms are assigned using the detection threshold, so if the detection threshold is set to the same as the minor allele frequency, the expected false negative rate is 50%.

Evolution

- ClustalW multiple sequence alignment program for DNA or proteins
- MUSCLE multiple aligner
- Branch Lengths Estimation
- Neighbor Joining Tree Builder
- dN/dS Ratio Estimation
- Mutate Codons with SNPs
- aaChanges amino-acid changes caused by a set of SNPs
- phyloP interspecies conservation scores