IOWA INITIATIVE IN HUMAN GENETICS BIOINFORMATICS SHORT COURSE 2012 "CHEAT SHEET"

DEFINITIONS

Bait files: depict sequence capture regions that are typically downloaded from the sequence capture kit manufacturer.
Binary Alignment/Map (BAM) file: a file format that is a binary version of SAM, making the file size more compact.
Browser Extensible Data (BED) file: a standardized format for a tab-delimited file containing, at a minimum, chromosomal coordinates for depicting regions of the genome.
CDCV (complex disease common variant hypothesis): the rationale behind GWAS (genome-wide association study) designs.
CDRV (complex disease rare variant hypothesis) assuming that rare variants make a greater contribution to complex disease than do common variants; requires deep sequencing.
Depth of coverage: number of sequencing reads aligned to a specific genomic location. Higher depth of coverage generally increases confidence in variant calls.
Edge prioritization: instead of prioritizing genes in isolation generate hypotheses about potential interactions among the top candidates and 'seed' genes.
Cluster: a group of linked computers, working together thus in many respects forming a single computer.
Exome sequencing: sequencing every exon of every gene in the genome.
FASTQ: a file format for storing short read massively parallel sequencing data.
Galaxy: A web-based platform that makes command-line tools available to biologists, is flexible, sharable, and can be run from (almost) any computer.
Massively parallel sequencing: a next-generation DNA sequencing technology that allows millions or billions of base-pairs to be sequenced simultaneously.
Multiplexing: the application of nucleotide barcodes followed by subsequent pooling of multiple DNA samples. Allows pooling of samples to increase throughput and take advantage of sequencer output.
Sequence Alignment/Map (SAM) file: a standardized and widely accepted format for storing large nucleotide sequence alignments that is designed to be: flexible (accommodates alignment information from various sequencers and alignment programs, compact in size, and easily convertable.
Sequence Capture: also known as targeted sequence capture or targeted genomic enrichment. The massively parallel replacement for PCR. The process of simultaneously isolating thousands or millions of regions of the genome prior to massively parallel sequencing.
Target Interval File: a file containing any areas of the genome that you think are biologically relevant (can be a bait file, can be a subset of a bait file, or can be user-generated).
Variant Call Format (VCF) file: a standardized and widely accepted file type for storing genotype data generated from variant calling algorithms.

TOOLS/WEBSITES

General

Course website/Galaxy Wiki: http://tinyurl.com/iihg-course
University of Iowa Galaxy: (https://galaxy.hpc.uiowa.edu/) Only accessible from campus, HAWKID required. Runs on Helium High Performance Computing Cluster.
Penn State Galaxy: (http://usegalaxy.org) Freely accessible, relatively small amount of storage space and
UCSC Genome Browser: http://genome.ucsc.edu/
Integrated Genomics Viewer (IGV): (http://www.broadinstitute.org/igv/) Freely available read alignment/ variant call visualization tool from the Broad institute.
University of Iowa Helium Cluster: https://www.icts.uiowa.edu/confluence/display/ICTSit/Helium+Cluster+Overview+and+Quick+Start+Guide

Databases

1000 genomes: (http://www.1000genomes.org/) Contains data from the 1,000 genomes project, which contains significantly less than 1,000 genomes at the current time (primarily exomes). Useful for population-level filtering based on minor allele frequency.
NHLBI Exome Variant Server (EVS): (http://evs.gs.washington.edu/EVS/) Contains data from 6,500 exomes completed at the University of Washington. Useful in population-level filtering based on minor allele frequency. Database is searchable via a web tool and downloadable.
KEGG database: (http://www.genome.jp/kegg/kegg1.html) Kyoto Encyclopedia of Genes and Genomes -- computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information)
dbNSFP (https://sites.google.com/site/jpopgen/dbNSFP) - database for Nonsynonymous SNPs' functional predictions; predictions from SIFT, PolyPhen2, LRT, MutationTaster and conservation score (PhyloP) for every potential SNV in the genome

Annotation tools

ANNOVAR: a web-based (wANNOVAR, http://wannovar.usc.edu) or command-line based (ANNOVAR, http://www.openbioinformatics.org/annovar/) tool for annotation of varaints. Accepts files in VCF format. Annotates with several databases including 1000 genomes, EVS and dbNSFP.
SeattleSeq: (http://snp.gs.washington.edu/SeattleSeqAnnotation131/) A web-based annotation platform from the University of Washington. Accepts files in VCF format. Annotates with several databases and output looks similar to EVS.

Gene prioritization

Gene Prioritization Portal (http://homes.esat.kuleuven.be/~bioiuser/gpp//): describes 33 publically available prioritization tools by the inputs they require (i.e. genes or key words), the outputs they produce (i.e. a prioritized list or a gene selection through filtering) and the data they use (i.e. text-mining, expression data)
GeneWanderer (http://compbio.charite.de/genewanderer/GeneWanderer): measures the relative location of each candidate gene in a genomic interval (e.g. found by linkage analysis) to genes known to be involved in the phenotype under investigation in a huge protein-protein interaction network