

Next-generation sequencing

Lecture 3

Data Analysis Techniques

February 23, 2015

Marylyn D. Ritchie

CHALLENGE= how to handle all of these data bioinformatically?

Bioinformatics Challenges Abound

- How/What to store?
- How to process?
- How to analyze?
- How to assess quality?
 - What is quality?
 - How to QC?
- How to compress?



Data Sizes

Evolution of Instrument Performance

From <1Gb to >1Tb in 4 Years

Internally we have completed multiple runs generating >1Tb of mappable data per run

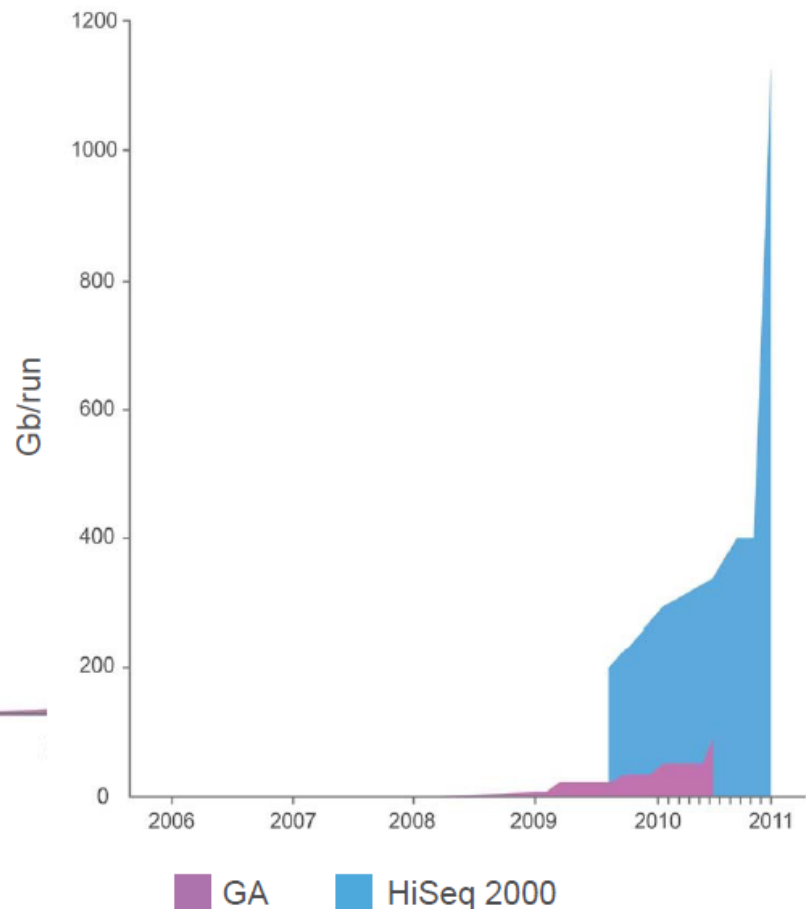
Greater than 80Gb per day!
Greater than 7.5B PE reads!

400

Sequencing Run Parameters

Run format: 2x150 bp

Output full run	1.13Tb
Output per day	81Gb



Reversing the Trend

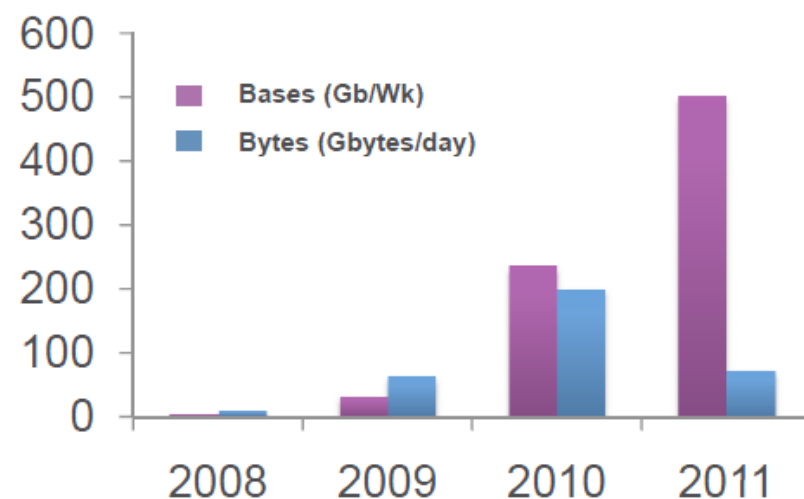
Scaling Output While Reducing the Informatics and Data Management Burden

CONTINUING TO REDUCE THE DATA FOOTPRINT

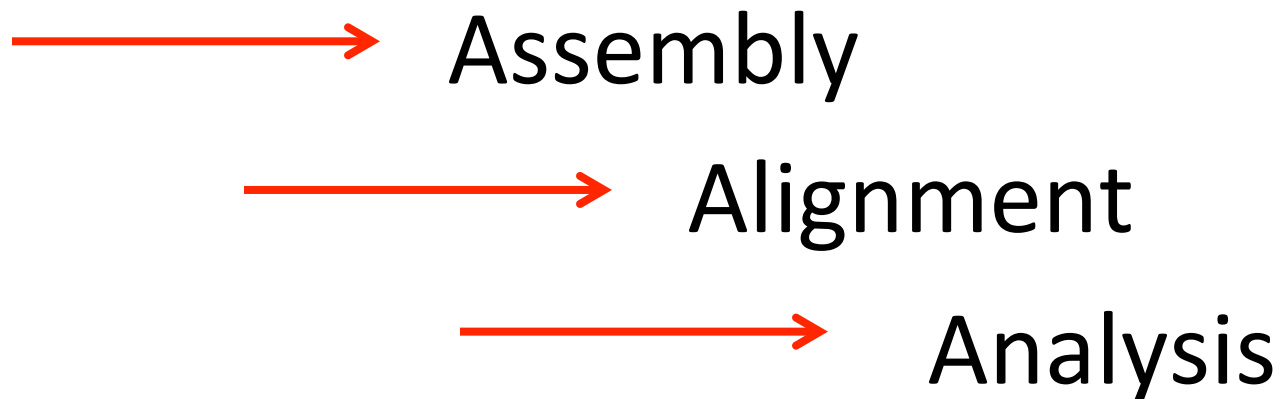
STREAMLINING DATA ANALYSIS

INCREASING PERFORMANCE ON ILLUMINACOMPUTE

Alignment/Mapping	Bytes per Base	Data Size †
Pipeline (2008)	30	3,000 GB
CASAVA (2009)	14	1,400 GB
CASAVA (2010)	6	600 GB
CASAVA (2011)	1	100 GB
Summarized (2011+)	.1	~10 GB



What does the data look like and how is it QC'd?



Sequence Data Analysis

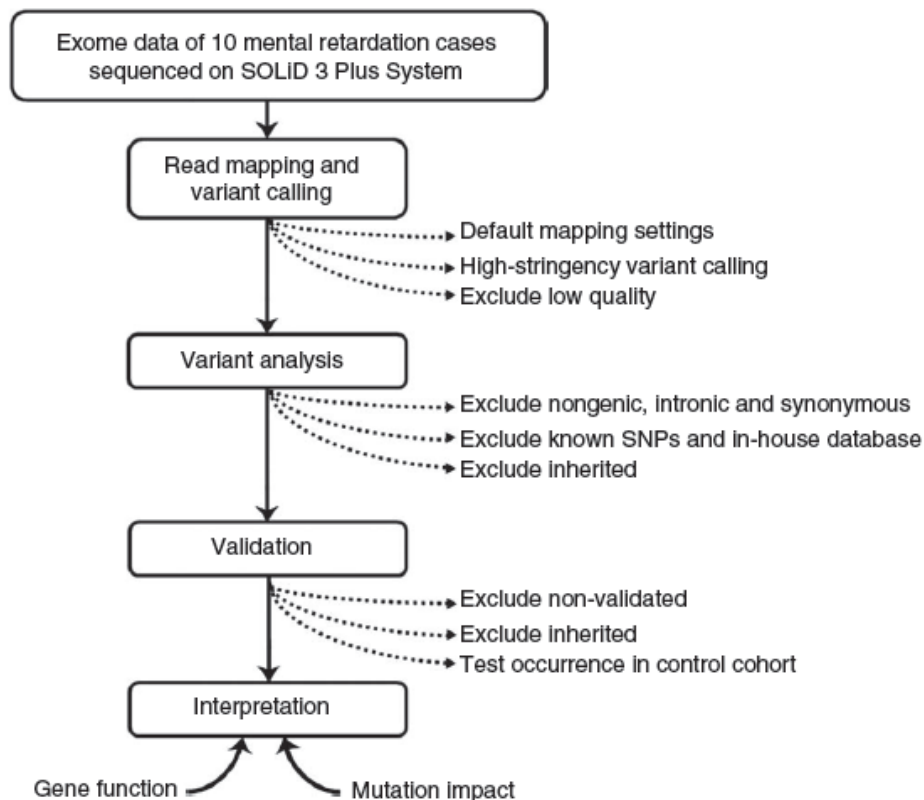
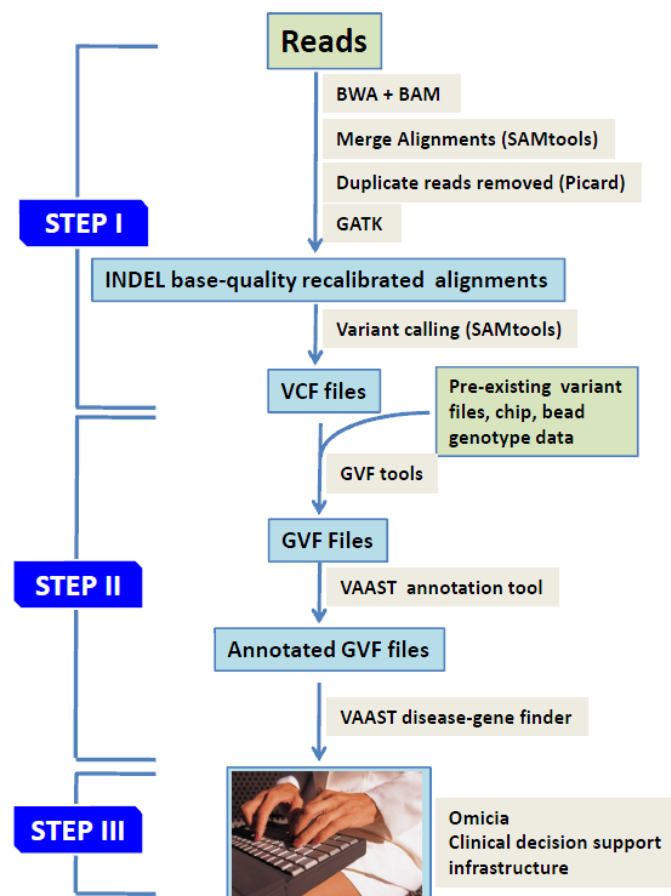
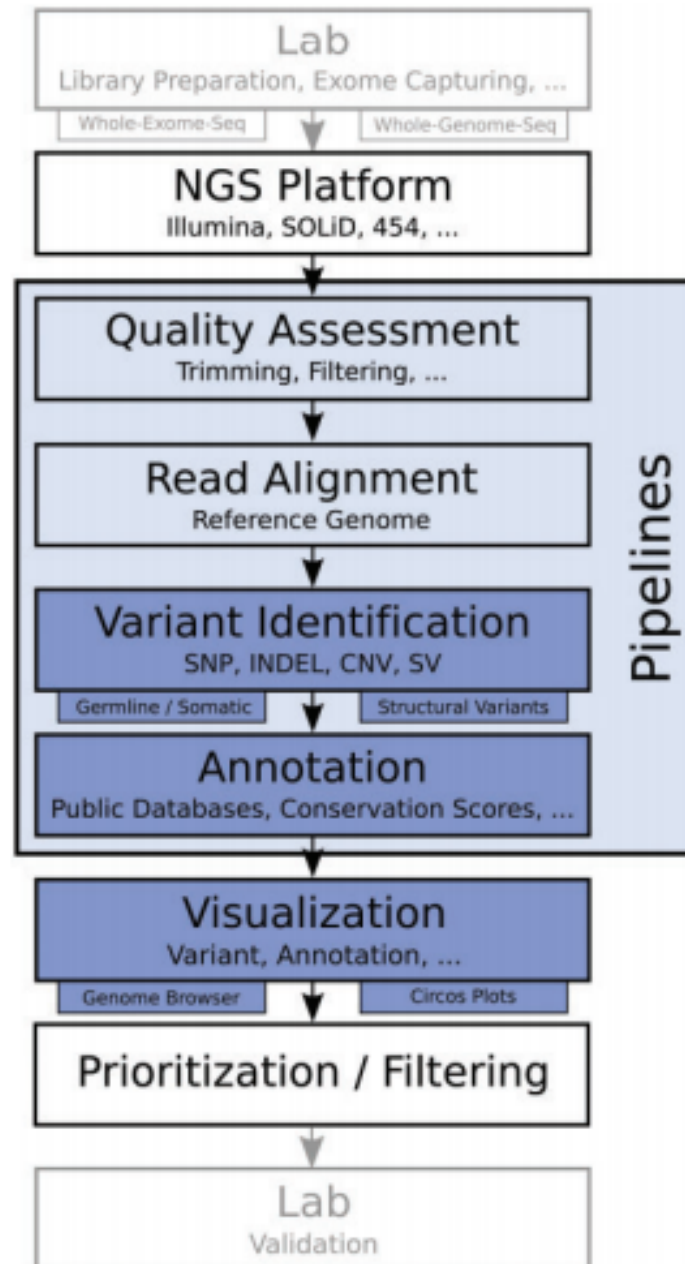


Figure 1 Experimental work flow for detecting and prioritizing sequence variants. For all ten mental retardation trios, prioritization of variants observed in the probands was based on selection for non-synonymous changes of high quality only and exclusion of all variants previously observed in healthy individuals, together with those variants that were inherited from an unaffected parent. Interpretation of *de novo* variants was based on gene function and the impact of the mutation.



Genotyping vs. Sequencing

- Genotyping is primer-based
 - What comes after “...ATGATCTTATTAA”?
 - Pro: High quality answers
 - Con: Need to know the primer a priori
- Sequencing is DNA replication based
 - I have “GCCCTGGACA” and “GGGATGGACA” and “GCTATAGTCT” ... what does that mean?
 - Pro: Can detect novel variation
 - Con: Highly susceptible to error, many steps
- Sequencing is more powerful, but many things can go wrong, from DNA -> VCF



Quality assessment

- Evaluate the quality of raw reads and to remove, trim or correct reads that do not meet the defined standards
- Need to filter out:
 - Base calling errors, INDELs, poor quality reads and adaptor contamination
- Generally, these steps include:
 - visualization of base quality scores and nucleotide distributions
 - trimming of reads and read filtering based on base quality score and sequence properties such as primer contaminations
 - N content and GC bias.

Quality assessment tools

Name	OS	Input	Output	Supported platforms	Report	Tag (1) removal	Filtering	Trimming
ContEST [1]	Lin, Mac, Win	BAM, VCF, FASTA (ref)	TXT	Illumina, ABI SOLiD, 454	no	no	no	no
FastQC [2]	Lin, Mac, Win	(CS) FASTQ, SAM, BAM	HTML	Illumina, ABI SOLiD	yes	no	no	no
FASTX-Toolkit [3]	Lin, Mac, web interface	FASTA, FASTQ	FASTA, FASTQ	Illumina	yes	yes	yes	yes
Galaxy [4]	Lin, Mac, web interface, Cloud instance	FASTQ	FASTQ	Illumina	yes	yes	yes	yes
htSeqTools [5]	Lin, Mac, Win	FASTQ	Graphs	Illumina	yes	no	no	no
NGSQC [6]	Lin	FASTA (ref), FASTQ, CSFASTA, QUAL FASTA	HTML	Illumina, ABI SOLiD	yes	no	no	no
PIQA [7]	Lin, Mac, Win	FASTQ, bustard, output, SCARF	HTML, TXT	Illumina	yes	no	no	no
PRINSEQ [8]	Lin, Mac, Win, web interface	FASTA, FASTQ, QUAL FASTA	FASTA, FASTQ, QUAL FASTA, HTML	Illumina, 454	yes	no	yes	yes
SolexaQA [9]	Lin, Mac	FASTQ	FASTQ, PNG	Illumina, 454	yes	no	no	yes
TagCleaner [10]	Lin, Mac, web interface	FASTA, FASTQ	FASTA	454	no	yes	no	no

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAATAGTAAATCCATTGTTCACACTCACAGTTT
+
!''*((( (**+))%%%)%%)(%%%).1***-+*''))**55CCF>>>>>CCCCCCCC65
```



Step 1: Output + Alignment

- Alignment is the process of assigning a position in the genome to each read
- Output from sequencers is FASTQ format
 - Each read lists all bases
 - Each base has an associated quality
 - No associated reference
- Need to align each read to the chosen reference genome
 - Reference must be consistent throughout the project
 - We typically use bwa (Burrows-Wheeler Aligner)
 - Other options are Novoalign

Step 1: Alignment Considerations

- Alignment is VERY computationally intensive
 - Claim 3 hrs, 6 GB for a full human genome
 - We have seen 2 hrs, 12 GB on 4 threads for a targeted exome (PGX project)
- Input for alignment is FASTQ
- Output of alignment is a SAM (or BAM) file
- Using a reference with decoy sequences can give better results
 - Decoy sequences attract common forms of contamination (e.g. herpes simplex)

Alignment

- After quality assessment is completed
- Aligned to a reference genome

Alignment

PE

Name	OS	Input	Output	Supported platforms	Indexing method	Gapped alignment
BarraCUDA [12]	Lin	FASTQ	SAM	Illumina	FM index (BWT)	yes
BFAST [13]	Lin	FASTQ	SAM	Illumina, ABI SOLiD, 454	Multiple (hash, tree, ...)	yes
Bowtie [14]	Lin, Mac, Win	FASTQ, FASTA	SAM	Illumina, ABI SOLiD	FM index (BWT)	no
Bowtie2 [15]	Lin, Mac, Win	FASTQ, FASTA, QSEQ	SAM	Illumina, 454	FM index (BWT)	yes
BWA [16]	Lin	(CS)FASTQ, FASTA	SAM	Illumina, ABI SOLiD(1)	FM index (BWT)	yes
BWA-SW [17]	Lin	FASTQ, FASTA	SAM	454	FM index (BWT)	yes
ELAND [18]	Lin	FASTQ, FASTA	SAM	Illumina	-	no
MAQ [19]	Lin	FASTQ, FASTA	Maq	Illumina	Hash based	yes
Mosaik [20]	Lin, Mac, Win	FASTQ, FASTA	SAM, BED, several others	Illumina, ABI SOLiD, 454	-	yes
mrFAST [21]	Lin	FASTQ, FASTA	SAM, DIVET	Illumina	Hash based	yes
mrsFAST [22]	Lin	FASTQ, FASTA	SAM, DIVET	Illumina	Hash based	no
Novoalign [23]	Lin, Mac	FASTQ, (CS)FASTA	SAM, TXT	Illumina, ABI SOLiD	-	yes
SOAP2 [24]	Lin	FASTQ, FASTA	SOAP (2)	Illumina	FM index (BWT)	yes
SOAP3 [25]	Lin	FASTQ, FASTA	SAM	Illumina	FM index (BWT)	no
SSAHA2 [26]	Lin, Mac	FASTA	SAM, GFF	Illumina, ABI SOLiD, 454	Tree index	yes
Stampy [27]	Lin, Mac (3)	FASTQ, FASTA	SAM	Illumina, 454	FM index (BWT)	-
YOABS [28]	Lin	-	-	Illumina	FM & Tree index	yes

Step 2: Variant Calling

- Variant Calling is the process of determining a person's genotype at a position.
- Input is BAM / SAM format, output VCF
- Many options available
 - We will focus on GATK's HaplotypeCaller, vers 3.x
 - Multi-sample calling is preferable
- Overall process:
 - For each sample, generate a GVCF using the option “-ERC GVCF -variant_index_type LINEAR -variant_index_parameter 128000”
 - Also, use vectorized calculations “-pairHMM VECTOR -LOGLESS_CACHING”

Step 2: Variant Merging

- Generating the GVCFs is an embarrassingly parallel problem, merging creates VCFs
 - Generating GVCF takes ~ 30 minutes for PGX targeted exome
 - Ensure genotype-level annotations in GVCF
- Use GATK's GenotypeGVCFs tool
 - Time increases with # of samples (approx 1 minute / sample for PGX)
 - Significant memory requirements (14 GB for 3,000 PGX samples)
 - Add Variant-level annotations here

Variant Calling

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1	Sample2	Sample3
2	4370	rs6057	G	A	29	.	NS=2;DP=13;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:52,51	1 0:48:8:51,51	1/1:43:5:..
2	7330	.	T	A	3	q10	NS=5;DP=12;AF=0.017	GT:GQ:DP:HQ	0 0:46:3:58,50	0 1:3:5:65,3	0/0:41:3
2	110696	rs6055	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
2	130237	.	T	.	47	.	NS=2;DP=16;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:56,51	0/0:61:2
2	134567	microsat1	GTCT	G,GTACT	50	PASS	NS=2;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Variant Calling

Table 1: Variant identification

Name	OS	BAM/SAM input	Other inputs	Output	Identifies	Data set	Result ^a
Germline callers							
CRISP	Lin	Yes	–	VCF	SNP, INDEL	KTS	24 034 SNPs, 259 INDELs
GATK (UnifiedGenotyper)	Lin	Yes	–	VCF	SNP, INDEL	KTS	49 476 SNPs, 1959 INDELs
SAMtools	Lin	Yes	FASTA	VCF	SNP, INDEL	KTS	21 852 SNPs, 332 INDELs
SNVer	Lin, Mac, Win	Yes	–	VCF	SNP, INDEL	KTS	22 105 SNPs, 234 INDELs
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL	KTS	34 984 SNPs, 1896 INDELs
Somatic callers							
GATK (SomaticIndelDetector)	Lin	Yes	–	VCF	INDEL	WES	151 INDELs
SAMtools	Lin	Yes	FASTA	BCF	SNP, INDEL	WES	Canceled ^b
SomaticSniper	Lin	Yes	–	VCF, somatic sniper output	SNP, INDEL	WES	6926 SNPs
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL, CNV	WES	1685 SNPs, 324 INDELs
CNV identification tools							
CNVnator	Lin	Yes	FASTA	CSV	CNV	cnv.sim	39 CNVs
RDXplorer	Lin, Mac	Yes	FASTA	CSV	CNV	cnv.sim	4 CNVs ^c
CONTRA	Lin, Mac	Yes	FASTA	VCF, CSV	CNV	WES	3 CNVs
ExomeCNV	Lin, Mac, Win	Yes	pileup + BED + FASTA	CSV	CNV, LOH	WES	137 CNVs
SV identification tools							
BreakDancer	Lin, Mac	Yes	config file	CSV, BED	INDEL, INV, TRANS, CNV	WGS (tumor + normal)	6219 DELs, 0 INs, 7 INVs, 17 303 ITX, 5037 CTX ^d
Breakpointer	Lin	Yes	–	GFF	INDEL	WGS (tumor)	^d
CLEVER	Lin	Yes	FASTA	CLEVER format	INDEL	WGS (tumor)	^d
GASVPro (GASVPro-HQ)	Lin, Mac	Yes	–	clusters file	INDEL, INV, TRANS	WGS (tumor)	2529 DELs, 207 INVs
SVMerge	Lin	Yes	FASTA	BED	INDEL, INV, CNV	–	Aborted ^e

Step 3: Filtration / Recalibration

- Raw VCFs typically include many errors, so filtration is essential
- For whole genome/exome, use GATK's VariantRecalibrator for automatic filtering
- For targeted exome, must use hard filters.
Good generic candidates are:
 - "QD" (Qual by Depth) for variant-level filters
 - "QUAL" for variant-level filters
 - "GQ" (Genomic Quality) for genotype-level filters
- **IMPORTANT:** If using hard filters, make sure to filter individual calls!

Summary + Resources

- General pipeline is FASTQ -> BAM -> VCF -> Filtered VCF
- PGX Pipeline located on RCC at ~/group/projects/eMERGE-PGX/scripts
- Other Tools / Resources
 - [GATK Best Practices](#)
 - [GATK Forums](#)
 - [Picard tools](#) (SAM/BAM processing)
 - [BWA help](#)
 - [SeqAnswers Forum](#)

A survey of tools for variant analysis of next-generation genome sequencing data

Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperl, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski

Submitted: 20th August 2012; Received (in revised form): 4th December 2012

Abstract

Recent advances in genome sequencing technologies provide unprecedented opportunities to characterize individual genomic landscapes and identify mutations relevant for diagnosis and therapy. Specifically, whole-exome sequencing using next-generation sequencing (NGS) technologies is gaining popularity in the human genetics community due to the moderate costs, manageable data amounts and straightforward interpretation of analysis results. While whole-exome and, in the near future, whole-genome sequencing are becoming commodities, data analysis still poses significant challenges and led to the development of a plethora of tools supporting specific parts of the analysis workflow or providing a complete solution. Here, we surveyed 205 tools for whole-genome/whole-exome sequencing data analysis supporting five distinct analytical steps: quality assessment, alignment, variant identification, variant annotation and visualization. We report an overview of the functionality, features and specific requirements of the individual tools. We then selected 32 programs for variant identification, variant annotation and visualization, which were subjected to hands-on evaluation using four data sets: one set of exome data from two patients with a rare disease for testing identification of germline mutations, two cancer data sets for testing variant callers for somatic mutations, copy number variations and structural variations, and one semi-synthetic data set for testing identification of copy number variations. Our comprehensive survey and evaluation of NGS tools provides a valuable guideline for human geneticists working on Mendelian disorders, complex diseases and cancers.

Keywords: Mendelian disorders; cancer; variants; bioinformatics tools; next-generation sequencing

A survey of tools for variant analysis of next-generation genome sequencing data

Table 2
Variant annotation

Name	OS	Input	Output	SNP	INDEL	CNV	GUI	CLI	Web	Function/Location Parameters	DB IDs	Number of scores
ANNOVAR	Lin, Mac, Win, web interface	VCF, pileup, CompleteGenomics, GFF3-SOLID, SOAPsnp, MAQ, CASAVA	TXT	Yes	Yes	Yes	No	Yes	No	9 (func) + 11(exonic-func)	Yes	GERP++ conservation, LRT, MutationTaster, PhyloP conservation, PolyPhen, SIFT
AnnTools	Lin, Mac	VCF, pileup, TXT	VCF	Yes	Yes	Yes	No	Yes	No	5 (position) + 4 (functional class)	Yes	-
NGS-SNP	Lin, Mac	VCF, pileup, MAQ, diBayes, TXT	TXT	Yes	No	No	No	Yes	No	17	Yes	Condel, PolyPhen, SIFT
SeattleSeq	web interface	VCF, MAQ, CASAVA, GATK BED, custom	VCF, SeattleSeq	Yes	Yes	No	No	No	Yes	11(dbSNP) + 5 (GVS)	Yes	GERP, Grantham, phastCons, PolyPhen
snpEff	Lin, Mac, Win	VCF, pileup/TXT (deprecated)	VCF, TXT, HTML overview	Yes	Yes	No	No	Yes	No	34	Yes	-
SVA	Lin	VCF, SV.events file, BCO	CSV	Yes	Yes	Yes	Yes	Yes	No	17 (SNP), 17 (INDEL), 10 (CNV)	Yes	-
VARIANT	web interface	VCF, GFF2, BED	web report, TXT	Yes	Yes	No	No	Yes	Yes	26	Yes	-
VEP	Lin, web interface	VCF, pileup, HGVS, TXT, variant identifiers	TXT	Yes	Yes	No	No	Yes	Limited	28	Yes	Condel, PolyPhen, SIFT

Other quality control considerations

- Impact from large amounts of data
 - data management
 - QC analysis

Software for SNP QC

plink...

Last original PLINK release is v1.07 (10-Oct-2009); [PLINK 1.](#)

Whole genome association analysis toolset

[Introduction](#) | [Basics](#) | [Download](#) | [Reference](#) | [Formats](#) | [Data management](#) | [Summary stats](#) | [Filters](#) | [Stratification](#) | [IBS/IBD](#) | [Association](#) | [Family-based](#) | [Permutation](#) | [LD calculations](#) | [Haplotypes](#) | [Conditional tests](#) | [Proxy association](#) | [Imputation](#) | [Dosage data](#) | [Clumping](#) | [Gene Report](#) | [Epistasis](#) | [Rare CNVs](#) | [Common CNPs](#) | [R-plugins](#) | [SNP annotation](#) | [Simulation](#) | [Profiles](#) | [ID helper](#) | [Resources](#) | [Flow chart](#) | [Misc.](#) | [FAQ](#) | [gPLINK](#)

1. Introduction

2. Basic information

- [Citing PLINK](#)
- [Reporting problems](#)
- [What's new?](#)
- [PDF documentation](#)

3. Download and general notes

- [Stable download](#)
- [Development code](#)
- [General notes](#)
- [MS-DOS notes](#)
- [Unix/Linux notes](#)
- [Compilation](#)
- [Using the command line](#)
- [Viewing output files](#)
- [Version history](#)

4. Command reference table

- [List of options](#)
- [List of output files](#)
- [Under development](#)

5. Basic usage/data formats

- [Running PLINK](#)
- [PED files](#)
- [MAP files](#)
- [Transposed files](#)
- [Long-format files](#)
- [Binary PED files](#)
- [Alternate phenotypes](#)
- [Covariate files](#)
- [Cluster files](#)
- [Set files](#)

6. Data management

- [Recode](#)

New (15-May-2014): PLINK 1.9 is now available for beta-testing!

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

The focus of PLINK is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with gPLINK and Haploview, there is some support for the subsequent visualization, annotation and storage of results.

PLINK (one syllable) is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the [support of others](#).

New in 1.07: [meta-analysis](#), [result annotation](#) and analysis of [dosage data](#).

Data management

- [Read data in a variety of formats](#)
- [Recode and reorder files](#)
- [Merge two or more files](#)
- [Extracts subsets \(SNPs or individuals\)](#)
- [Flip strand of SNPs](#)
- [Compress data in a binary file format](#)

Summary statistics for quality control

Quick links

[PLINK tutorial](#)

[gPLINK](#)

[Join e-mail list](#)

[Resources](#)

[FAQs](#) | [PDF](#)

[Citing PLINK](#)

[Bugs, questions?](#)



Software for SNP QC

PENNSTATE



The Ritchie Lab

A laboratory of the Center for System Genomics

[Home](#) [Research](#) [People](#) [Software](#) [Publications](#) [Outreach and Coordination](#) [Contact us](#)

PLATO Downloads



What is PLATO ?

The PLatform for the Analysis, Translation, and Organization of large-scale data (PLATO) is a standalone program written in C++ that is designed to be a flexible and extensible analysis tool for a wide variety of genetic data. PLATO includes a configurable set of QC and analysis steps that can be used for the filtering and analysis of data in a single command step. Further, through the abstraction of genetic data, PLATO allows for the easy addition of customized analysis or filtering steps requiring only a basic level of computing expertise.

Why use PLATO ?

With the wide array of genotypic and phenotypic data available, there is no single analytical method that is appropriate for all data. In fact, no single method can be optimal for all datasets, especially when the genetic architecture for diseases can vary substantially. PLATO serves as an integrative platform that can accommodate multiple analytical methods for analysis as we learn more about genetic architecture. By allowing for user customization through the use of command line options, PLATO can adapt to many different kinds of data and analyses. Additionally, PLATO has the ability to be run in parallel for some steps, reducing the computing time of the analyses on the multi-core machines that have become standard.

[Notes about PLATO 2.0](#)

<https://ritchielab.psu.edu/plato>

Software for Sequence QC

PLINK/SEQ

A library for the analysis of genetic variation data

[Home](#) | [Overview](#) | [Download](#) | [Installation](#)

A toolset for working with human genetic variation data

PLINK/SEQ is an open-source C/C++ library for working with human genetic variation data. The specific focus is to provide a platform for analytic tool development for variation data from large-scale resequencing and genotyping projects, particularly whole-exome and whole-genome studies. It is independent of (but designed to be complementary to) the existing [PLINK](#) package.

Downloads

The latest version of PLINK/SEQ (v0.10, released 14-July-2014) is available on the [download](#) page. This page contains source (C/C++) code as well as pre-compiled binary executables for Linux (x86_64) and MacOS (built on Mavericks).

Getting Started

- This [overview](#) provides a high-level description of the aims, scope and design of the library.
- After [downloading](#) and [installing](#) the library, see this [gentle introduction](#)
- For a more in-depth introduction, see the [tutorial using 1000 Genomes data](#).

Getting started

- [PLINK/SEQ 101](#)
- [Extended tutorial](#)

Key concepts

- [Project structure](#)
- [Variants and samples](#)
- [Meta-information](#)
- [Masks](#)

PSEQ documentation

- [Basic syntax](#)
- [Project management](#)
- [Main data input](#)
- [Auxiliary data input](#)
- [Viewing data](#)

Software for Sequence QC

VCFTools

[Home](#)[Sourceforge page](#)[Examples & Documentation](#)[Downloads](#)

Welcome to VCFTools

VCFTools is a program package designed for working with VCF files, such as those generated by the 1000 Genomes Project. The aim of VCFTools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files.

This toolset can be used to perform the following operations on VCF files:

- Filter out specific variants
- Compare files
- Summarize variants
- Convert to different file types
- Validate and merge files
- Create intersections and subsets of variants

VCFTools consists of two parts, a **perl module** and a **binary executable**. The perl module is a general Perl API for manipulating VCF files, whereas the binary executable provides general analysis routines.

Documentation

A list of **usage examples** can be found [here](#).

Sourceforge

The VCFTools project is hosted on [Sourceforge](#).

Variant call format specification

VCFTools is compatible with VCF versions 4.0, 4.1 and 4.2.

For more information regarding the VCF format, please visit the [VCF specification page](#).

Contact

For help regarding VCFTools or the VCF format, please see the [mailing lists](#).

Citations and Licensing

Information about licensing and publications can be found [here](#).

Links

Other useful links can be found on [this page](#).

<http://vcftools.sourceforge.net/>

Galaxy

- <http://main.g2.bx.psu.edu/>



Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the [free public server](#) or [your own instance](#), you can perform, reproduce, and share complete analyses.

Use Galaxy



[Use the free public server](#)

Get Galaxy



Install [locally](#) or [in the cloud](#)

Learn Galaxy



[Screencasts](#), [Galaxy 101](#),
[...](#)

Get Involved



[Mailing lists](#), [Tool Shed](#),
[wiki](#)

OMICtools

[Home](#)
[Reviews](#)
[News](#)
[FAQ](#)
[Media](#)
[About](#)
[Submit tools](#)

A workflow for omic data analysis (NGS, microarray, PCR, MS, NMR)



OMICtools can help a) experimental researchers/clinicians find appropriate tools for their needs b) developers to stay up to date and to avoid redundancy c) funding agencies to ensure that the submitted projects are high value-added. Do you want help us to improve OMICtools? [Call for curators](#)

Browse by omic applications


[Sequencing \(2102\)](#)

[Microarray \(497\)](#)

[Mass spectrometry \(339\)](#)

[NMR spectroscopy \(97\)](#)

[PCR \(111\)](#)

[nCounter System \(4\)](#)

[Cytometry \(70\)](#)

[Common tools \(261\)](#)

[Drug discovery \(388\)](#)

[Genome editing \(30\)](#)

[Biomolecular structure \(340\)](#)

[Health & Diseases \(139\)](#)

[Functional analysis \(1842\)](#)

[Educational resources \(201\)](#)

Useful links



Current Analysis Strategies

- Single variant association will not work
 - Computational burden
 - Underpowered to capture association for variants with low (rare) allele frequencies
 - Unable to explain complex heritability (GxG, GxE)
 - Type 1 error inflation (more variants to test without more power to test them)
 - Inherently based on LD which is very low in rare variants

Current Analysis Strategies

- Single

- (

- (

- |

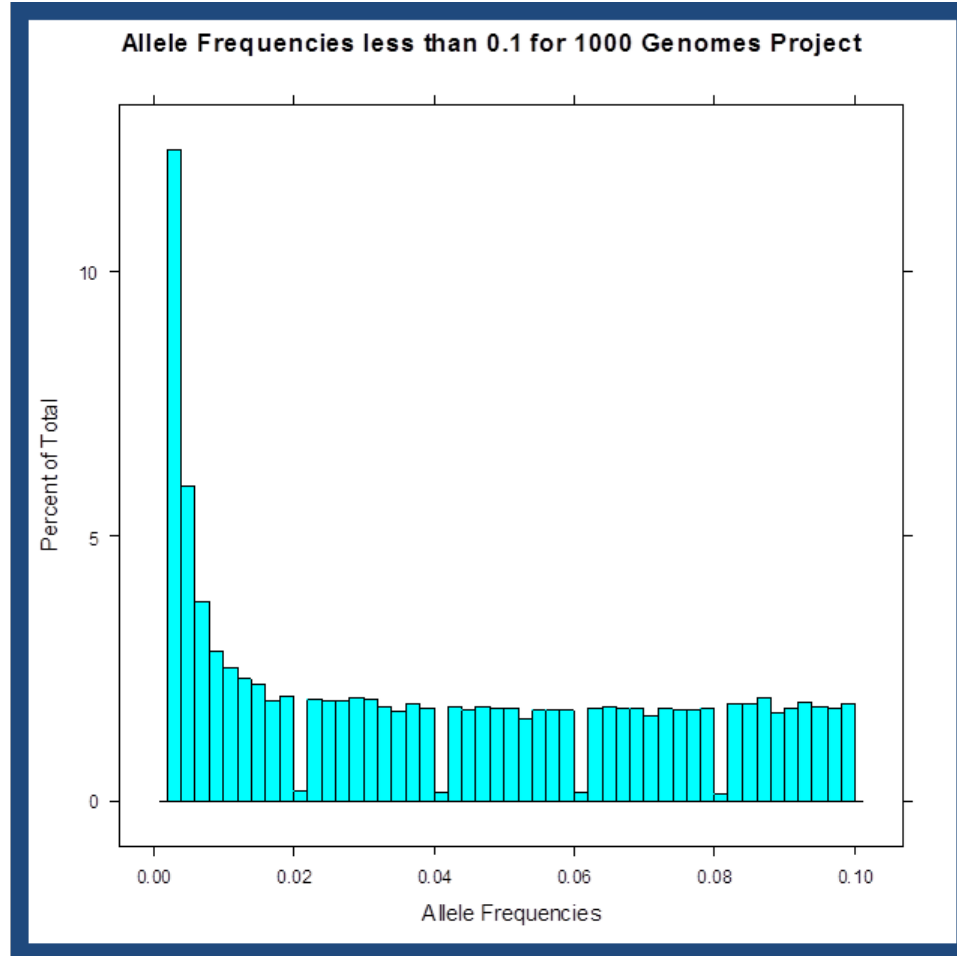
- (

- (

- |

- (

- |



with

it


**1000 Genomes Pilot data CEU frequency distribution
(MAF < 0.10)**

Current Analysis Strategies

- The most prevalent methods involve filtering:
 - Analyzing pedigrees
 - Candidate gene collapsing strategies

Current Analysis Strategies

- Analyzing family data



nature
genetics


[nature.com](#) ▶ [journal home](#) ▶ [archive](#) ▶ [issue](#) ▶ [brief communication](#) ▶ [abstract](#)

ARTICLE PREVIEW

[view full access options](#) ▶

NATURE GENETICS | BRIEF COMMUNICATION

Identity-by-descent filtering of exome sequence data identifies *PIGV* mutations in hyperphosphatasia mental retardation syndrome


Peter M Krawitz, Michal R Schweiger, Christian Rödelisperger, Carlo Marcelis, Uwe Kölsch, Christian Meisel, Friederike Stephani, Taroh Kinoshita, Yoshiko Murakami, Sebastian Bauer, Melanie Isau, Axel Fischer, Andreas Dahl, Martin Kerick, Jochen Hecht, Sebastian Köhler, Marten Jäger, Johannes Grünhagen, Birgit Jonske de Condor, Sandra Doelken, Han G Brunner, Peter Meinecke, Eberhard Passarge, Miles D Thompson, David E Cole  *et al.*

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

Nature Genetics **42**, 827–829 (2010) | doi:10.1038/ng.653
Received 29 March 2010 | Accepted 03 August 2010 | Published online 29 August 2010

Current Analysis Strategies

- Analyzing family data



nature
genetics
nature.com

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy

James R. Lupski, M.D., Ph.D., Jeffrey G. Reid, Ph.D., Claudia Gonzaga-Jauregui, B.S., David Rio Deiros, B.S., David C.Y. Chen, M.Sc., Lynne Nazareth, Ph.D., Matthew Bainbridge, M.Sc., Huyen Dinh, B.S., Chyn Jing, M.Sc., David A. Wheeler, Ph.D., Amy L. McGuire, J.D., Ph.D., Feng Zhang, Ph.D., Pawel Stankiewicz, M.D., Ph.D., John J. Halperin, M.D., Chengyong Yang, Ph.D., Curtis Gehman, Ph.D., Danwei Guo, M.Sc., Rola K. Irikat, B.S., Warren Tom, B.S., Nick J. Fantin, B.S., Donna M. Muzny, M.Sc., and Richard A. Gibbs, Ph.D.

Nature Genetics 42, 827–829 (2010) | doi:10.1038/ng.653
Received 29 March 2010 | Accepted 03 August 2010 | Published online 29 August 2010

Current Analysis Strategies

- Analyzing family data



Genetics *IN* Medicine®

Official Journal of the American College of Medical Genetics

Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease

Worthey, Elizabeth A. PhD^{1,2}; Mayer, Alan N. MD, PhD^{2,3}; Syverson, Grant D. MD²; Helbling, Daniel BSc¹; Bonacci, Benedetta B. MSc²; Decker, Brennan BSc¹; Serpe, Jaime M. BSc²; Dasu, Trivikram PhD²; Tschannen, Michael R. BSc¹; Veith, Regan L. MSc²; Basehore, Monica J. PhD⁴; Broeckel, Ulrich MD, PhD^{1,2,3}; Tomita-Mitchell, Aoy PhD^{1,2,3}; Arca, Marjorie J. MD^{3,5}; Casper, James T. MD^{2,3}; Margolis, David A. MD^{2,3}; Bick, David P. MD^{1,2,3}; Hessner, Martin J. PhD^{1,2}; Routes, John M. MD^{2,3}; Verbsky, James W. MD, PhD^{2,3}; Jacob, Howard J. PhD^{1,2,3,6}; Dimmock, David P. MD^{1,2,3}

I.S.,

.D.,
3.S.,

Current Analysis Strategies

- Candidate gene collapsing strategies

OPEN  ACCESS Freely available online

PLoS GENETICS

A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic

Bo Eskerod Madsen¹, Sharon R. Browning^{2*}

¹ Bioinformatics Research Center (BiRC), University of Aarhus, Aarhus C, Denmark, ² Department of Statistics, The University of Auckland, Auckland, New Zealand

Current Analysis Strategies

- Candidate gene collapsing strategies

OPEN  ACCESS Freely available online

PLoS GENETICS

ARTICLE

Pooled Association Tests for Rare Variants in Exon-Resequencing Studies

Alkes L. Price,^{1,2,3,6} Gregory V. Kryukov,^{3,4,6} Paul I.W. de Bakker,^{3,4} Shaun M. Purcell,^{3,5} Jeff Staples,^{3,4} Lee-Jen Wei,² and Shamil R. Sunyaev^{3,4,*}

PENNSTATE



Current Analysis Strategies

- Candidate gene collapsing strategies

OPEN  ACCESS Freely available online

PLoS GENETICS

ARTICLE

Resource

A probabilistic disease-gene finder for personal genomes

Mark Yandell,^{1,3,4} Chad Huff,^{1,3} Hao Hu,^{1,3} Marc Singleton,¹ Barry Moore,¹ Jinchuan Xing,¹ Lynn B. Jorde,¹ and Martin G. Reese²

¹Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah and School of Medicine, Salt Lake City, Utah 84112, USA; ²Omicia, Inc., Emeryville, California 94608, USA

PENNSYLVANIA



Collapsing methods

Authors	Pub. Year	Test	Notes
Morgenthaler and Thilly	2007	Cohort Allelic Sums Test (CAST)	<ul style="list-style-type: none">•First published collapsing method•No covariates•Assumes variants have same magnitude and direction of effect•No quantitative phenotypes•Can't return direction of association•Calculates variant sums in cases and controls-applies chi-square statistics
Li and Leal	2008	Combined Multivariate and Collapsing Method (CMC)	<ul style="list-style-type: none">•Can combine rare and common variants•Can incorporate covariates•Can incorporate direction of effect•Uses a multivariate statistical test
Madsen and Browning	2009	Weighted Sum Statistics (WSS)	<ul style="list-style-type: none">•Weight each variant using its allele frequency•Perform ranks sum test between cases and controls
Price et al.	2010	Variable Threshold Approach (VT)	<ul style="list-style-type: none">•Variable threshold approach based on allele frequency•Incorporated PolyPhen2

Collapsing methods

Authors	Pub. Year	Test	Notes
Liu and Leal	2010	Kernel Based Adaptive Cluster (KBAC)	<ul style="list-style-type: none">•Can analyze whole-genome data•Test gives multi-marker genotypes with higher sample risks higher weights to separate causal from non-causal multi-marker genotypes•Can be used with permutation (or regression)•Advantageous in the presence of variant misclassification and gene interactions•Exclusive to rare variants
Yandell et al.	2011	Variant Annotation, Analysis & Search Tool (VAAST)	<ul style="list-style-type: none">•Incorporates allele frequency and amino acid significance data to functional prediction of variant•Can manage large datasets <p>Calculates significance between case and control rare variant vectors using kernels (statistical functions to determine the weight of a given rare variant genotype)</p>
Quintana et al.	2011	Bayesian Risk Index (BRI)	Choose variants to be included in bin using Bayesian probability instead of functional prediction algorithm

Dispersion tests



SCHOOL OF PUBLIC HEALTH
Powerful ideas for a healthier world

A to Z Index

Search

GO

People • Calendar • myHSPH • Email • Newsletters  

⚙ SNP-set (Sequence) Kernel Association Test (SKAT)

SKAT Home

SKAT Download

MetaSKAT

[Home](#) > SNP-set (Sequence) Kernel Association Test (SKAT)

SKAT



Test for association between a set of rare (and common) variants and continuous/dichotomous phenotypes using kernel machine methods

SKAT is a SNP-set (e.g., a gene or a region) level test for association between a set of rare (or common) variants and dichotomous or quantitative phenotypes. SKAT aggregates individual score test statistics of SNPs in a SNP set and efficiently computes SNP-set level p-values, e.g. a gene or a region level p-value, while adjusting for covariates, such as principal components to account for population stratification. SKAT

FEATURES

Obesity

Global progress to reverse obesity epidemic 'unacceptably slow'

Genetics

Genes affecting allergies, asthma identified

PENNSSTATE



<http://www.hsph.harvard.edu/skat/>

Why collapse based on genomic location?

Collapse based on biology

more on this in the next lecture with Sarah Pendergrass

Which method should you use?

- Many statistical tests available for rare variants
 - burden
 - dispersion
- Many ways to consider collapsing the variants
- Perform simulation studies similar to your own study

SimRare

- Developed by Biao Li in Leal lab
- Simulates realistic population demographic and phenotypic models
- Can simulate and evaluate association methods on one platform
- Uses forward-time simulation
- Can simulate case-control and quantitative traits
- The phenotypic effect can be detrimental, protective, or non-causal

SimRare GUI

Form

**SimRare: Simulator of Generating Sequence-based Data
for Testing Rare Variants association Methods**

Need simulation of rare variants (srv) ? ----> [Click Here](#)

Need analyze *.ped file or load R script? ----> [Click Here](#)

How do you like to establish phenotype-genotype associations? Case-control population attributable risk model of a sample

☐ Use Saved Input Parameters? [Load File](#) ☒ Use Default Input Parameters?

>>>> Initialization <<<<

[Load .maf File](#)

[Load .sel File](#)

[Load .pos File](#)

>>>> Simulation Parameter Settings <<<<

Proportion of Detrimental RVs # Cases # Unphenotyped

Proportion of Protective RVs # Controls Population Size

☐ Fixed Effect Model Prevalence Odds Ratios for Common Mutations

☐ Variable Effect Model Odds Ratios for Protective Mutations

Odds Ratios for Detrimental Mutations

Mode of Inheritance ☐ Constant Parameters? QT Coefficient for Common Variants

Attributable Risk for Detrimental Mutations QT Coefficients for Causal Variants

Attributable Risk for Protective Mutations

Percentage of Causal RVs QTL Cutoffs ☐ Mark Case-Control?

☐ Allelic Heterogeneity? Proportion of Heterogeneous Cases

>>>> Mimic Genotyping Process <<<<

Proportion of Missing Detrimental RVs

Proportion of Missing Protective RVs

Proportion of Missing Non-causal Mutations

Proportion of Missing Synonymous Mutations

☐ Mark Missing RVs?

>>>> Save Simulation Results <<<<

☐ Is Syno Trimmed? [Set Path for Output File](#)

☐ Is CV Trimmed? Output File Name

☐ Is Ped Written? ☐ Save Input Parameters?

>>>> Show Data <<<<

☐ Print Genotypes? ☐ Print Phenotypes?

☐ Print Minor Allele Frequencies?

[Help](#) [Cancel](#) [Run!](#)

SimRare GUI

Form

Simulator of Rare Variants

☐ Use Saved Input Parameters? ☒ Use Default Input Parameters?

Basics

Gene Length Mode ☐ Fixed ☒ Random

Random Gene Length within a region

2500 5000

Output Files Name (prefix)

MySimuRV

Number of Replicates

Demographic Model

Effective Population Sizes

8100, 8100, 7900, 9000

Numbers of Generations per Stage

500, 10, 370

Genetic Forces

Mutation Model ☐ infinite_sites ☒ finite_sites

Mutation Rate

1.8e-8

Multi-locus Selection Model

additive

Selection Coefficient Distribution Model

Boyko_2008_European

☐ Need Customize Selection Coefficient?

Customized Selection Coefficient

☐ Need Recombination?

Recombination Rate

Screen Output

Screen Output Mode minimum

Detailed Screen Output Interval per Stage

100

Optional File Output

☐ Save Genotype for Replicates? ☐

☐ Save Statistics for Replicates? ☐

Set Path to Output Files

Path

☐ Save Input Parameters?

Evaluating Power

- Challenges for sequencing studies
 - What is the correct model?
 - What are the allele frequencies going to be?
 - How many variants should/can be in a bin?
 - What is the expected effect size?
 - Challenge: more variants in a bin vs mitigating true signal
 - How can we even be sure about the direction of effect?

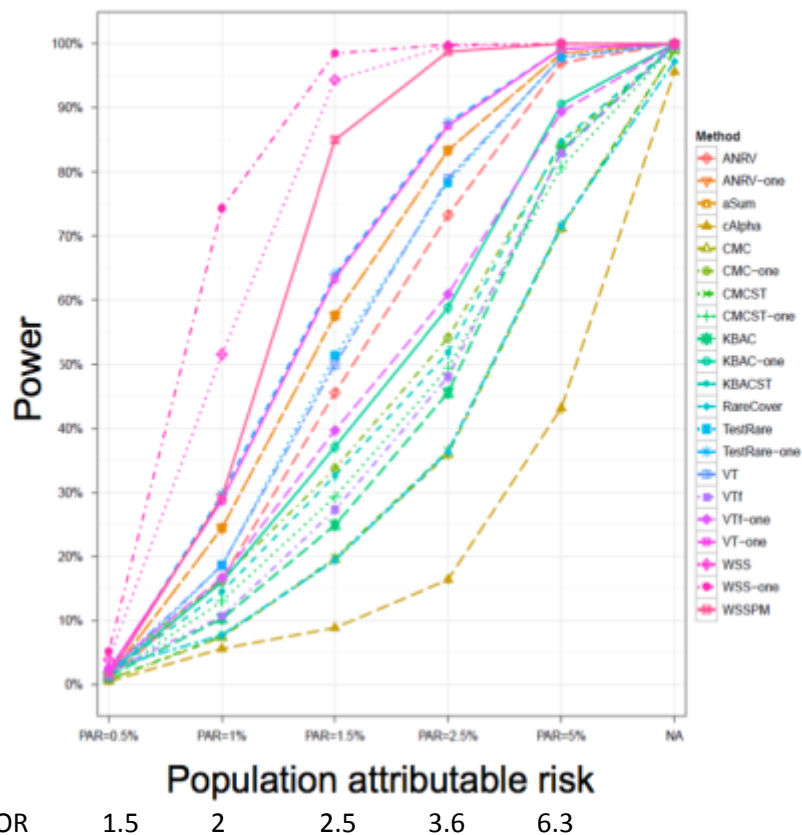
Evaluating Power

In order to assess power for BioBin:

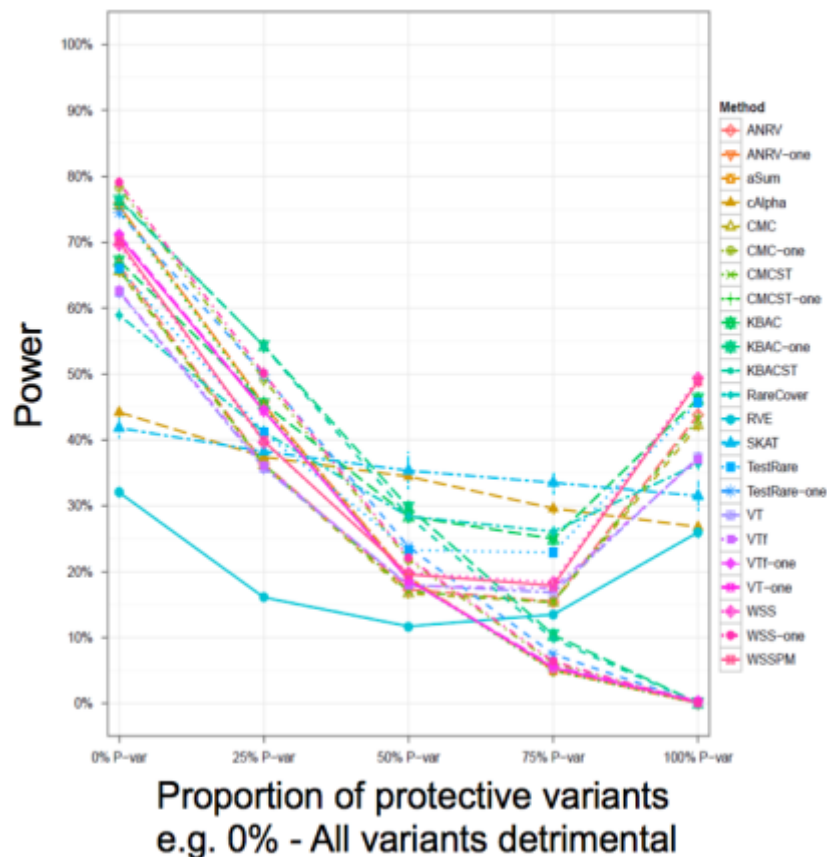
- Further develop the method
- Simulate sequence data
 - Consider opposing directions of effect (variants with neutral effect)
 - Various genetic models (recessive, additive, dominant)
 - Simulate different examples of testable features (pathways, genes, introns, etc)
- Test BioBin using simulated data to evaluate power

SimRare Power Analyses

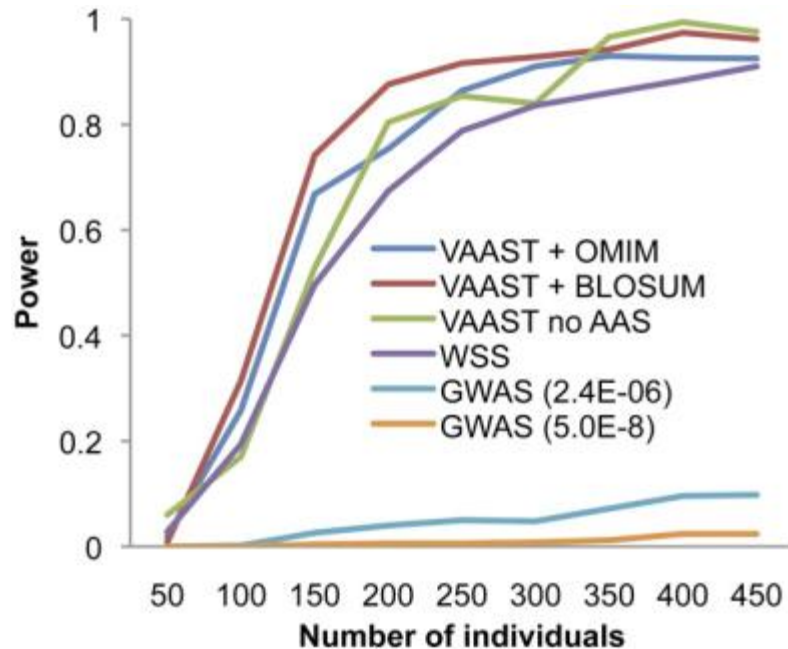
Population attributable risk model,
Recessive, N=800:800



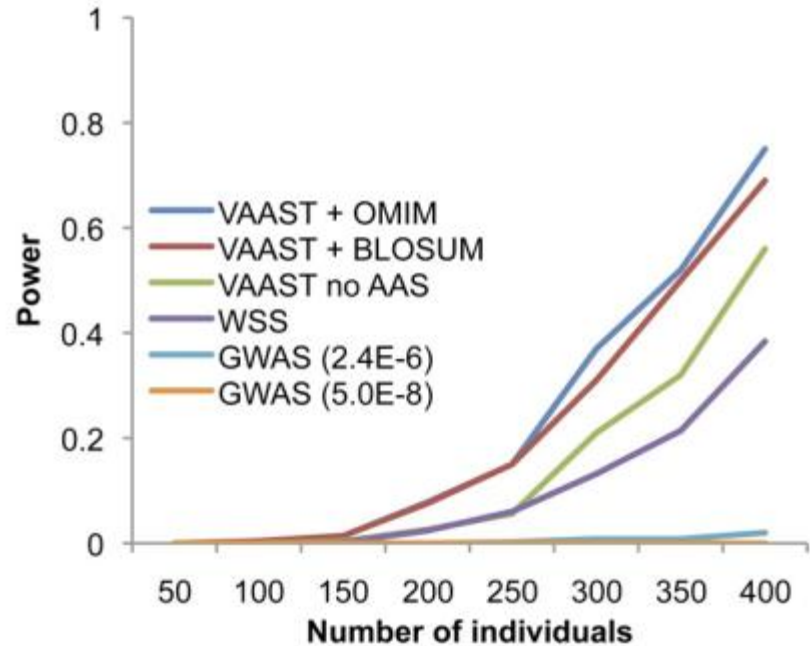
Impact of protective variants, fixed effect model
OR = 2 vs 0.5. Prevalence 1%. N=1000:1000



VAAST Power Analyses



A. *NOD2*. Crohn Disease.



B. *LPL*. Hypertriglyceridemia

Statistical power as a function of number of target genomes for two common disease genes.

Summary

- Alignment and base calling tools are important to determine downstream data quality
- Quality assessment is critical
- Analysis tools for WES and WGS continue to be developed
- Optimal strategies have not yet been determined

Questions???

