Generating High Throughput Data and QC

Marylyn D Ritchie, PhD

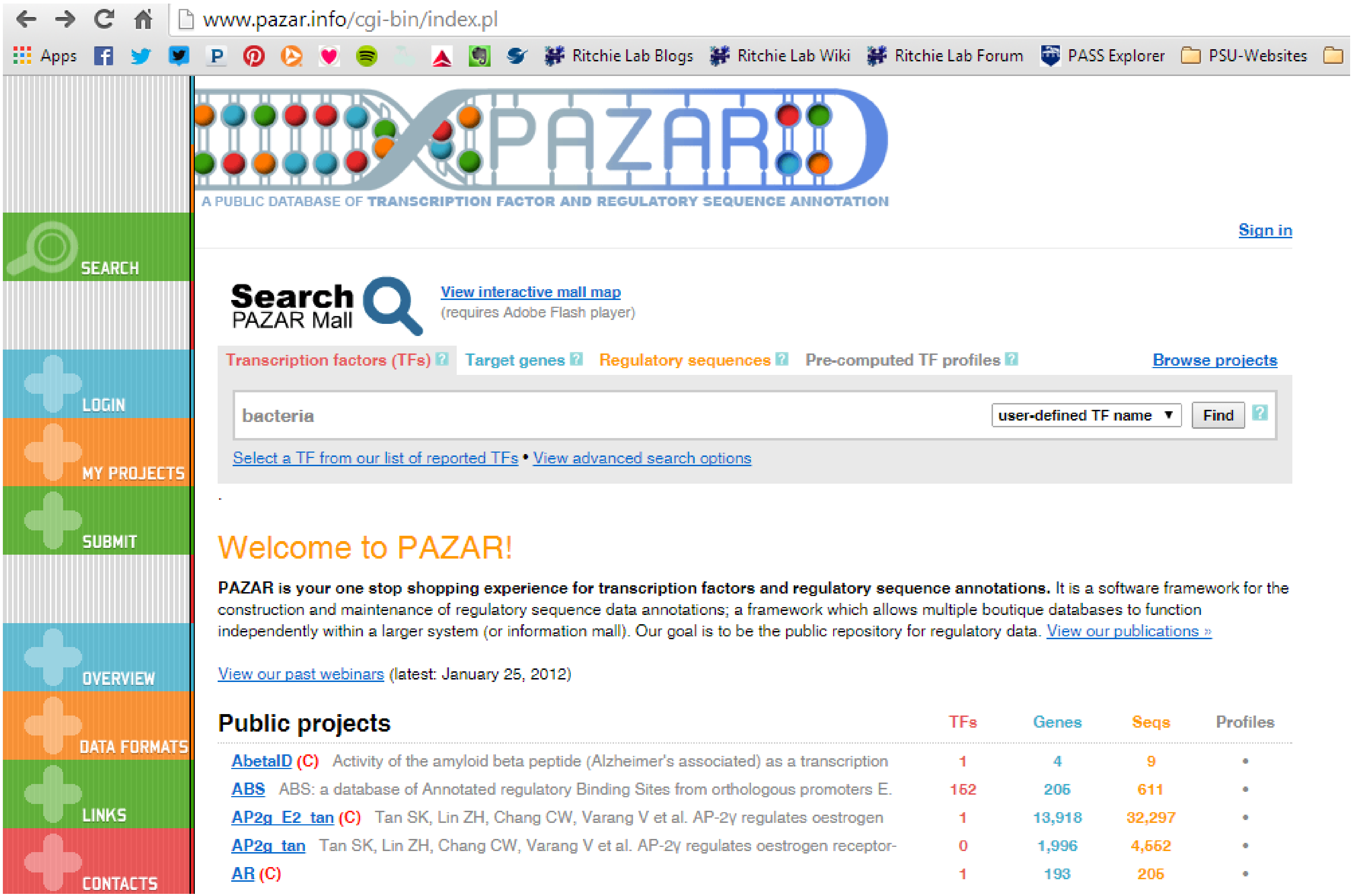
Professor, Biochemistry and Molecular Biology

Director, Center for Systems Genomics

The Pennsylvania State University

Center for Systems Genomics
Pennsylvania State University

Ritchie Lab
Pennsylvania State University



Welcome to PAZAR!

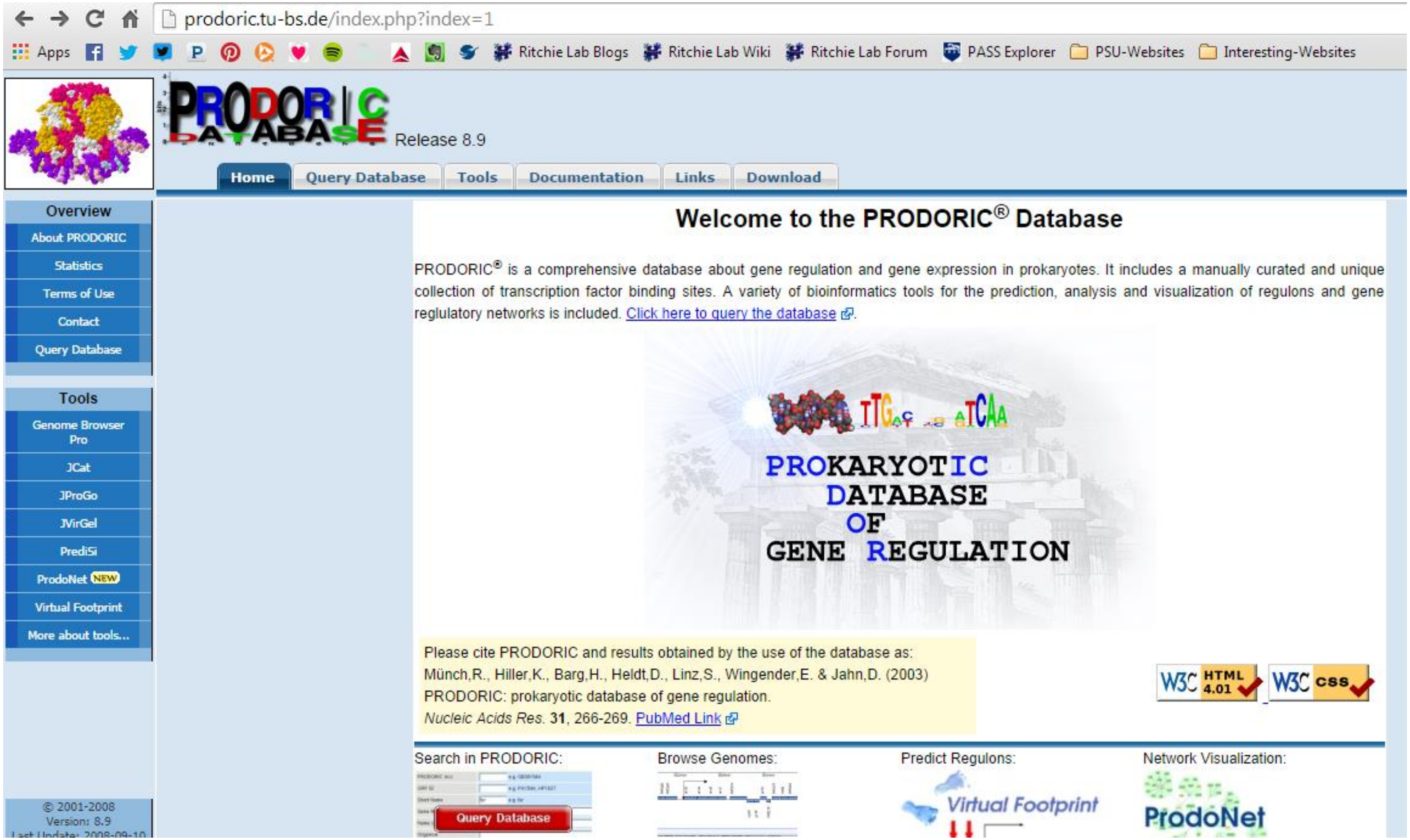
PAZAR is your one stop shopping experience for transcription factors and regulatory sequence annotations. It is a software framework for the construction and maintenance of regulatory sequence data annotations; a framework which allows multiple boutique databases to function independently within a larger system (or information mall). Our goal is to be the public repository for regulatory data. View our publications »

View our past webinars (latest: January 25, 2012)

Public projects

	TFs	Genes	Seqs	Profiles
AbetaID (C) Activity of the amyloid beta peptide (Alzheimer's associated) as a transcription	1	4	9	•
ABS ABS: a database of Annotated regulatory Binding Sites from orthologous promoters E.	152	206	611	•
AP2g_E2_tan (C) Tan SK, Lin ZH, Chang CW, Varang V et al. AP-2γ regulates oestrogen	1	13,918	32,297	•
AP2g_tan Tan SK, Lin ZH, Chang CW, Varang V et al. AP-2γ regulates oestrogen receptor-	0	1,996	4,552	•
AR (C)	1	193	205	•

Regulatory data for multiple species

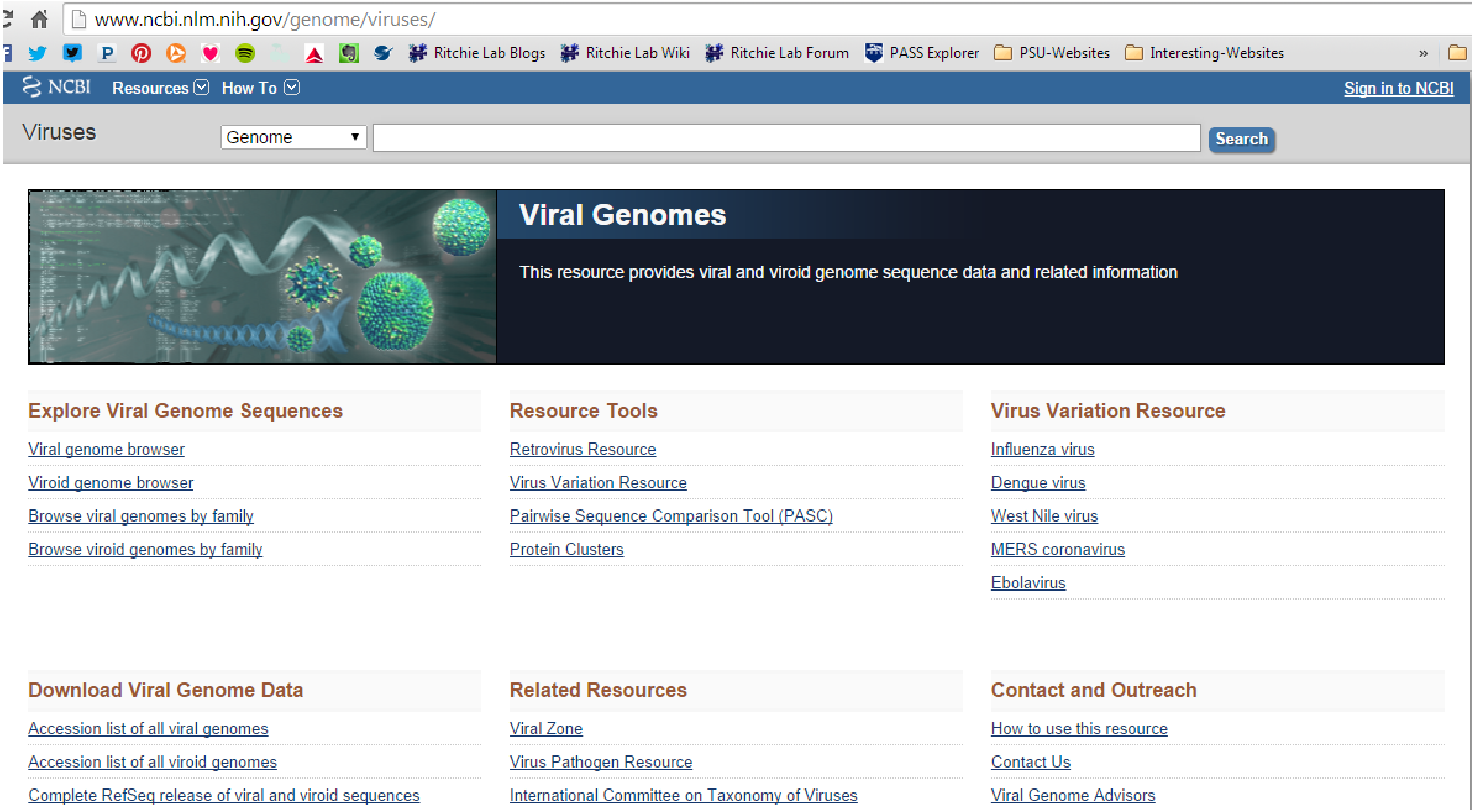


Welcome to the PRODORIC® Database

PRODORIC® is a comprehensive database about gene regulation and gene expression in prokaryotes. It includes a manually curated and unique collection of transcription factor binding sites. A variety of bioinformatics tools for the prediction, analysis and visualization of regulons and gene regulatory networks is included. Click here to query the database.

Please cite PRODORIC and results obtained by the use of the database as:
Münch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. & Jahn,D. (2003)
PRODORIC: prokaryotic database of gene regulation.
Nucleic Acids Res. **31**, 266-269. PubMed Link

Regulatory data for prokaryotic species



Viral Genomes

This resource provides viral and viroid genome sequence data and related information

Explore Viral Genome Sequences

- Viral genome browser
- Viroid genome browser
- Browse viral genomes by family
- Browse viroid genomes by family

Resource Tools

- Retrovirus Resource
- Virus Variation Resource
- Pairwise Sequence Comparison Tool (PASC)
- Protein Clusters

Virus Variation Resource

- Influenza virus
- Dengue virus
- West Nile virus
- MERS coronavirus
- Ebolavirus

Download Viral Genome Data

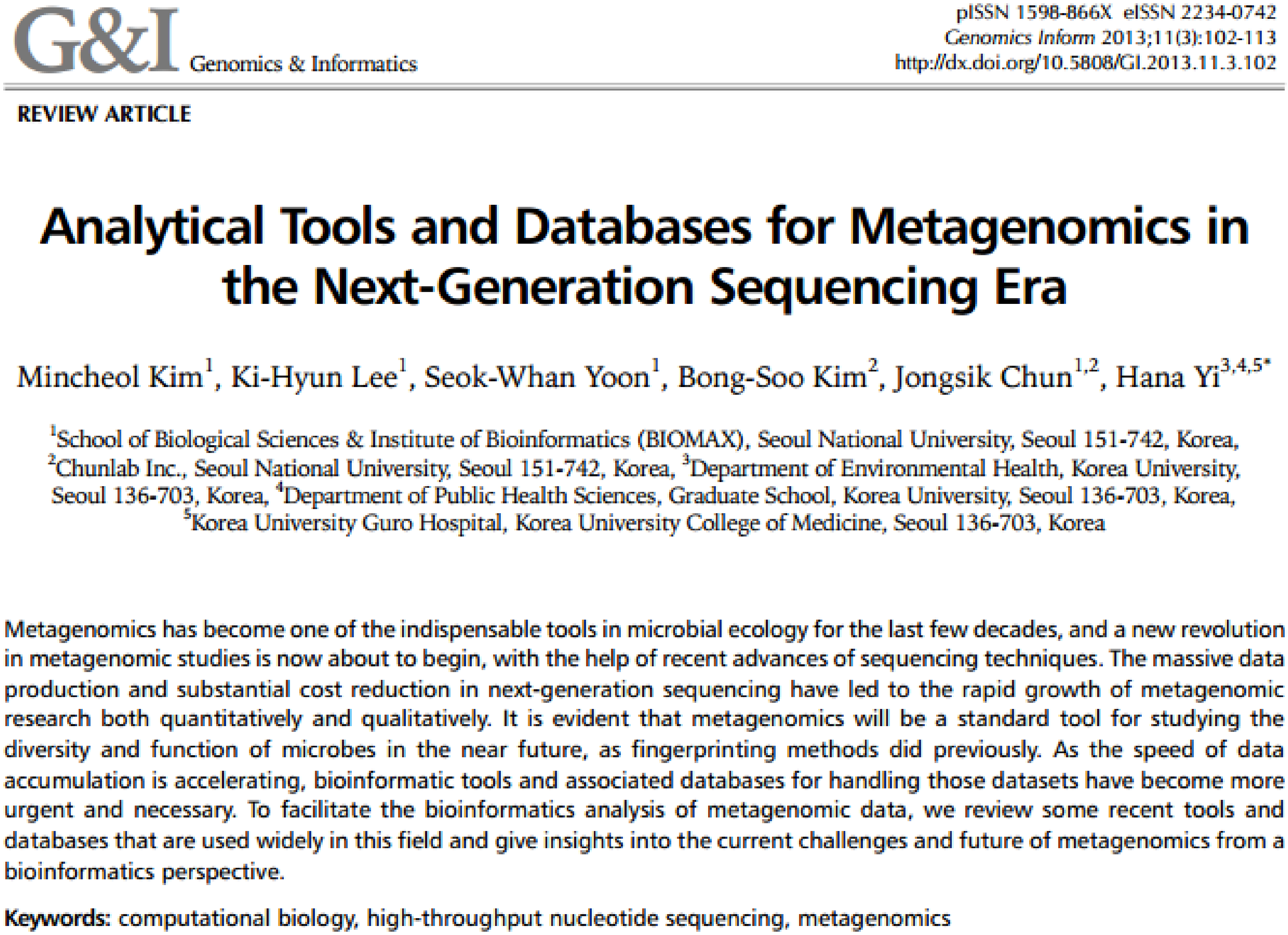
- Accession list of all viral genomes
- Accession list of all viroid genomes
- Complete RefSeq release of viral and viroid sequences

Related Resources

- Viral Zone
- Virus Pathogen Resource
- International Committee on Taxonomy of Viruses

Contact and Outreach

- How to use this resource
- Contact Us
- Viral Genome Advisors



G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2013;11(3):102-113
http://dx.doi.org/10.5808/GI.2013.11.3.102

REVIEW ARTICLE

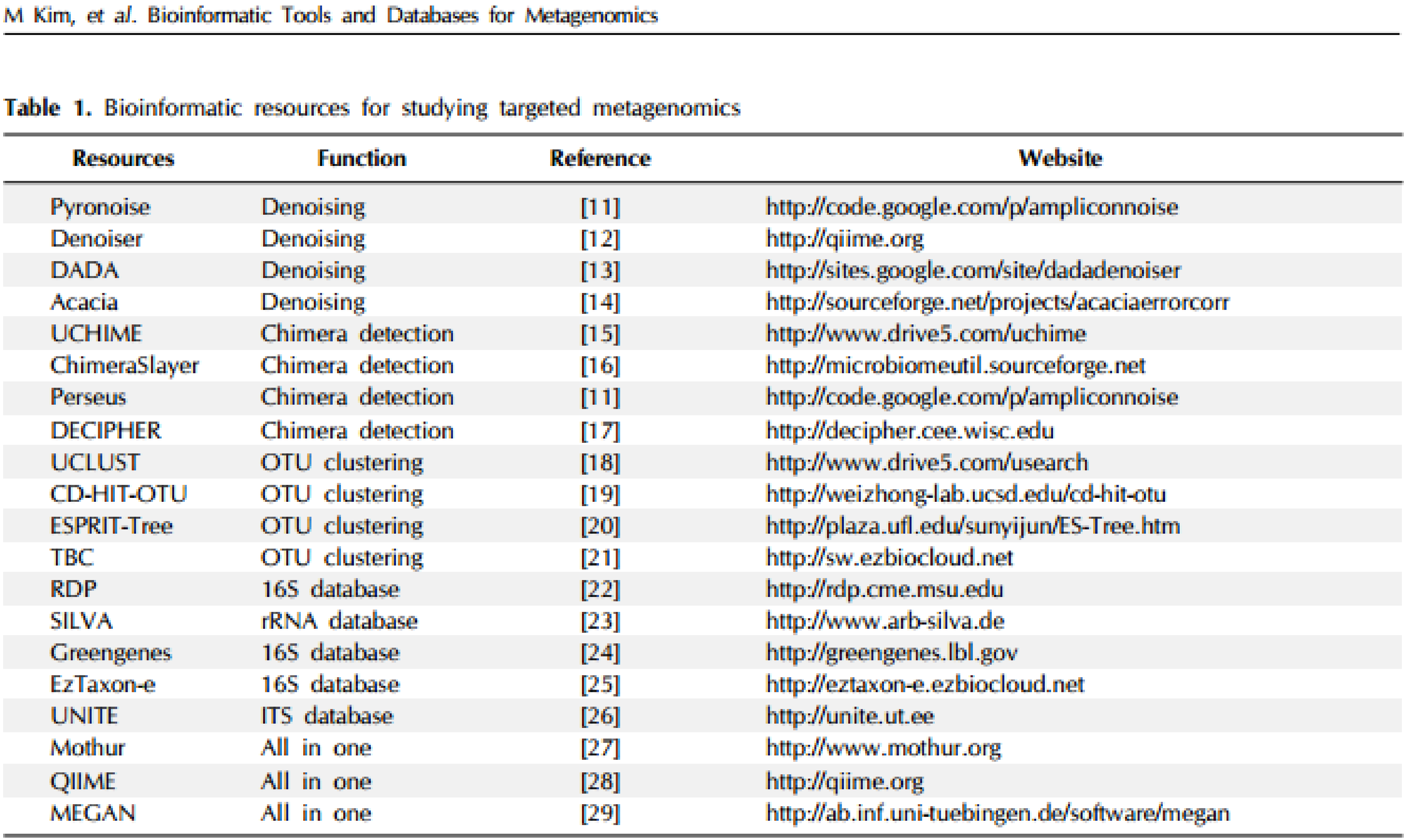
Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era

Mincheol Kim[1], Ki-Hyun Lee[1], Seok-Whan Yoon[1], Bong-Soo Kim[2], Jongsik Chun[1,2], Hana Yi[3,4,5*]

[1]School of Biological Sciences & Institute of Bioinformatics (BIOMAX), Seoul National University, Seoul 151-742, Korea,
[2]Chunlab Inc., Seoul National University, Seoul 151-742, Korea, [3]Department of Environmental Health, Korea University,
Seoul 136-703, Korea, [4]Department of Public Health Sciences, Graduate School, Korea University, Seoul 136-703, Korea,
[5]Korea University Guro Hospital, Korea University College of Medicine, Seoul 136-703, Korea

Metagenomics has become one of the indispensable tools in microbial ecology for the last few decades, and a new revolution in metagenomic studies is now about to begin, with the help of recent advances of sequencing techniques. The massive data production and substantial cost reduction in next-generation sequencing have led to the rapid growth of metagenomic research both quantitatively and qualitatively. It is evident that metagenomics will be a standard tool for studying the diversity and function of microbes in the near future, as fingerprinting methods did previously. As the speed of data accumulation is accelerating, bioinformatic tools and associated databases for handling those datasets have become more urgent and necessary. To facilitate the bioinformatics analysis of metagenomic data, we review some recent tools and databases that are used widely in this field and give insights into the current challenges and future of metagenomics from a bioinformatics perspective.

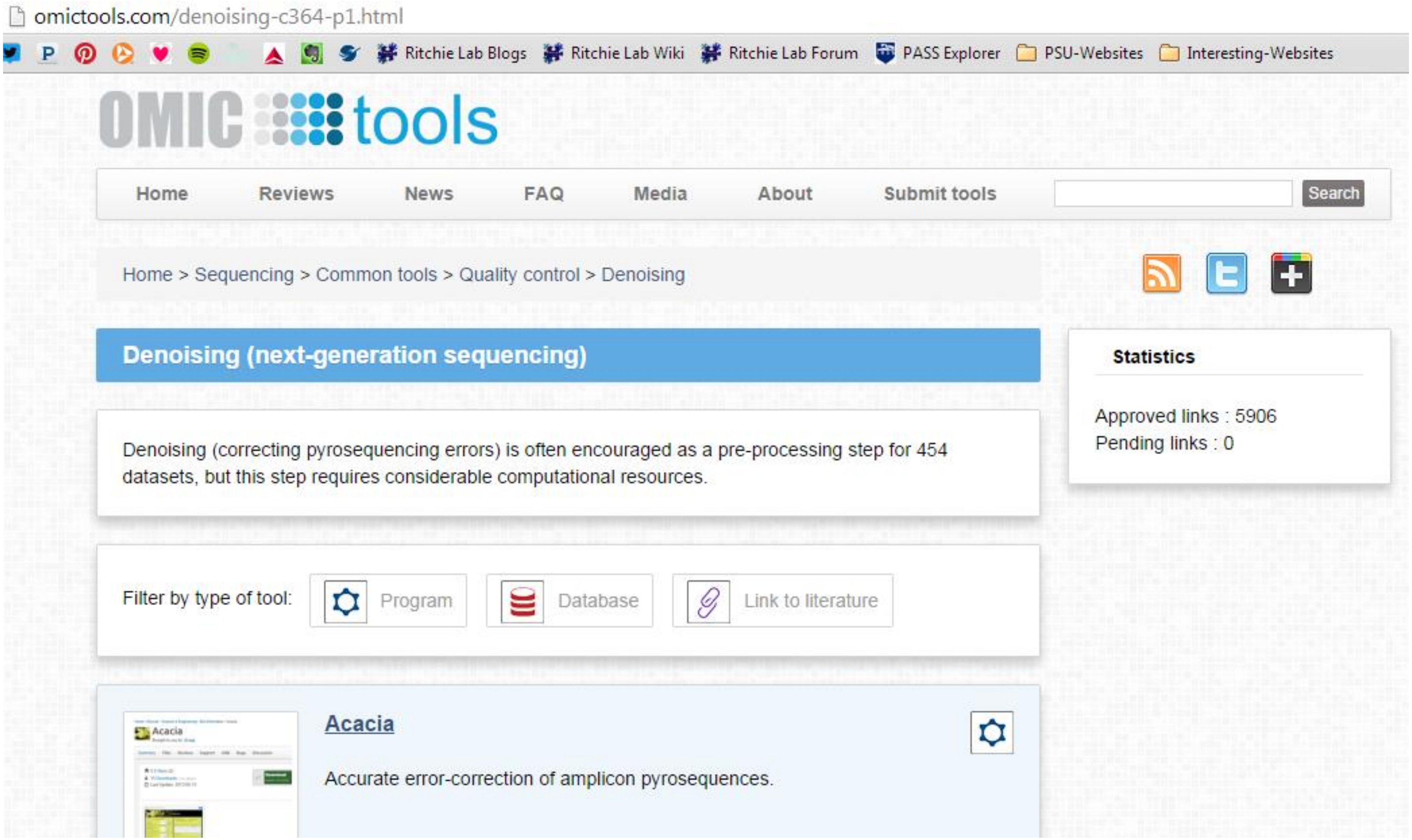
Keywords: computational biology, high-throughput nucleotide sequencing, metagenomics



M Kim, *et al*. Bioinformatic Tools and Databases for Metagenomics

Table 1. Bioinformatic resources for studying targeted metagenomics

Resources	Function	Reference	Website
Pyronoise	Denoising	[11]	http://code.google.com/p/ampliconnoise
Denoiser	Denoising	[12]	http://qiime.org
DADA	Denoising	[13]	http://sites.google.com/site/dadadenoiser
Acacia	Denoising	[14]	http://sourceforge.net/projects/acaciaerrorcorr
UCHIME	Chimera detection	[15]	http://www.drive5.com/uchime
ChimeraSlayer	Chimera detection	[16]	http://microbiomeutil.sourceforge.net
Perseus	Chimera detection	[11]	http://code.google.com/p/ampliconnoise
DECIPHER	Chimera detection	[17]	http://decipher.cee.wisc.edu
UCLUST	OTU clustering	[18]	http://www.drive5.com/usearch
CD-HIT-OTU	OTU clustering	[19]	http://weizhong-lab.ucsd.edu/cd-hit-otu
ESPRIT-Tree	OTU clustering	[20]	http://plaza.ufl.edu/sunyijun/ES-Tree.htm
TBC	OTU clustering	[21]	http://sw.ezbiocloud.net
RDP	16S database	[22]	http://rdp.cme.msu.edu
SILVA	rRNA database	[23]	http://www.arb-silva.de
Greengenes	16S database	[24]	http://greengenes.lbl.gov
EzTaxon-e	16S database	[25]	http://eztaxon-e.ezbiocloud.net
UNITE	ITS database	[26]	http://unite.ut.ee
Mothur	All in one	[27]	http://www.mothur.org
QIIME	All in one	[28]	http://qiime.org
MEGAN	All in one	[29]	http://ab.inf.uni-tuebingen.de/software/megan



Home > Sequencing > Common tools > Quality control > Denoising

Denoising (next-generation sequencing)

Denoising (correcting pyrosequencing errors) is often encouraged as a pre-processing step for 454 datasets, but this step requires considerable computational resources.

Statistics

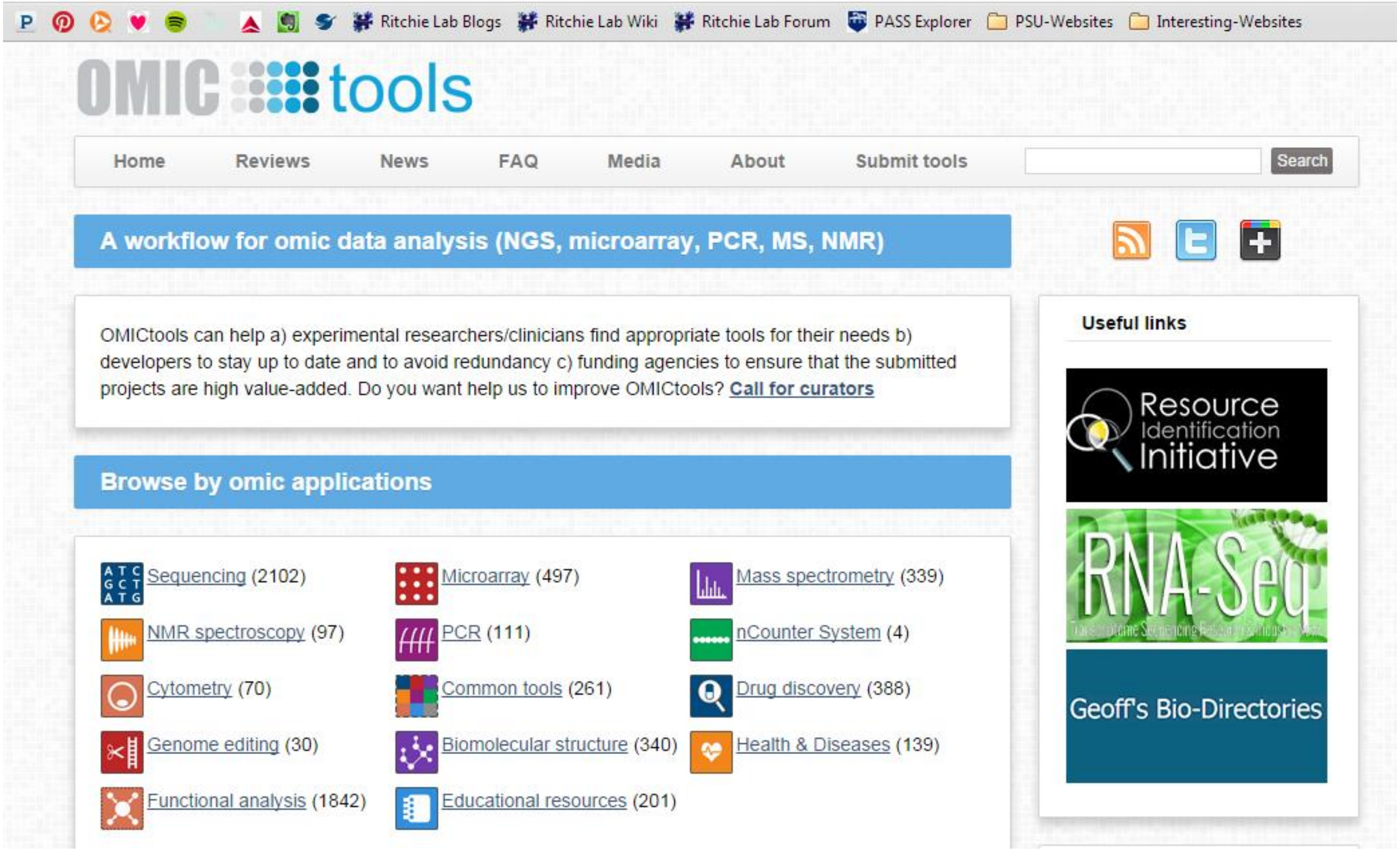
Approved links : 5906
Pending links : 0

Filter by type of tool: Program Database Link to literature

Acacia

Accurate error-correction of amplicon pyrosequences.

omictools.com



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)

Generating High Throughput Data and QC

Marylyn D Ritchie, PhD

Professor, Biochemistry and Molecular Biology

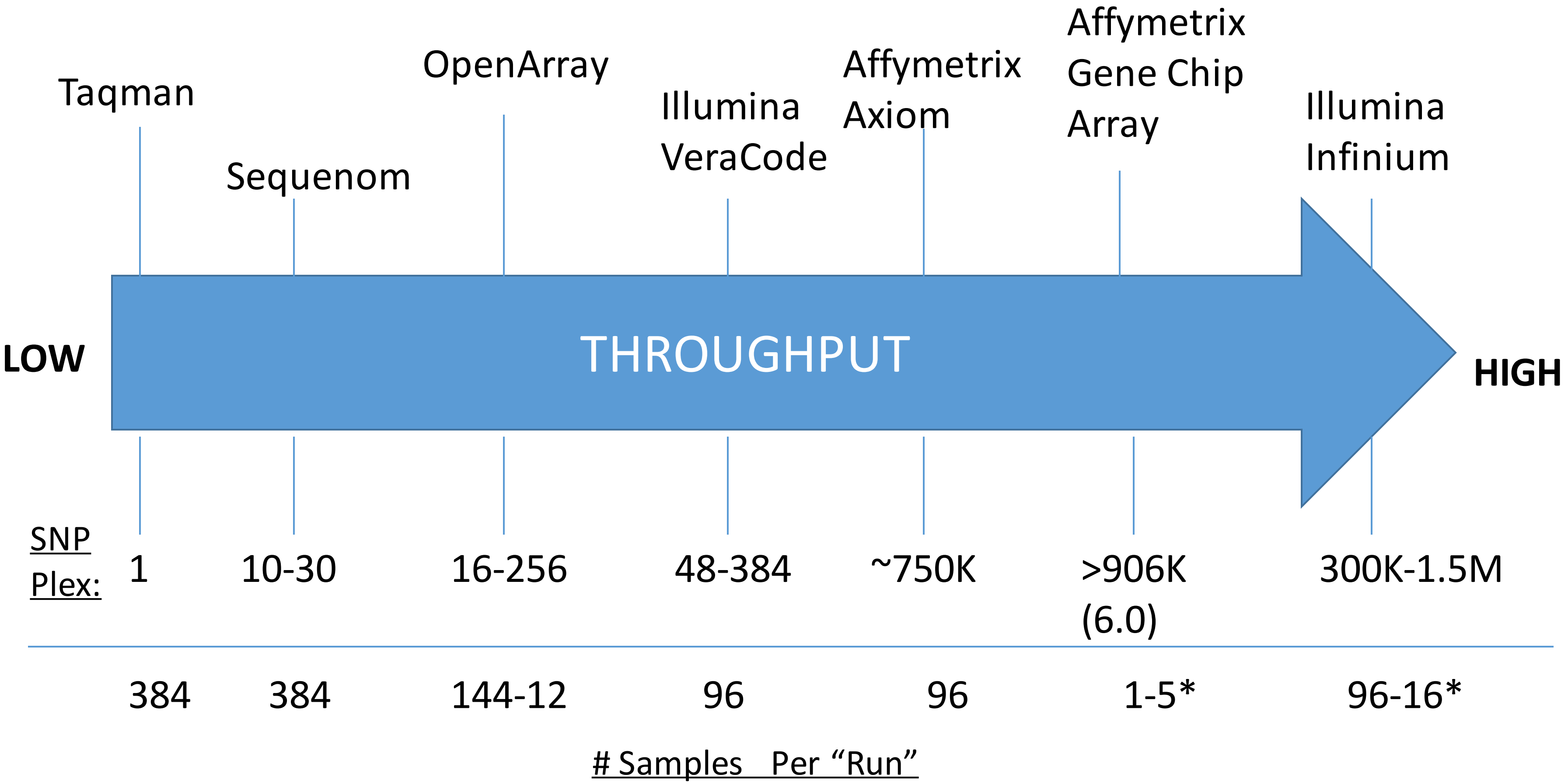
Director, Center for Systems Genomics

The Pennsylvania State University

Center for Systems Genomics
Pennsylvania State University

Ritchie Lab
Pennsylvania State University

Options for Genotyping SNPs



LOW — THROUGHPUT — HIGH

	Taqman	Sequenom	OpenArray	Illumina VeraCode	Affymetrix Axiom	Affymetrix Gene Chip Array	Illumina Infinium
SNP Plex:	1	10-30	16-256	48-384	~750K	>906K (6.0)	300K-1.5M
# Samples Per "Run"	384	384	144-12	96	96	1-5*	96-16*

Genotyping Platforms

	Assay Type	Technology Basis	Throughput/person	Multiplexing (# SNPs)	Application

Ragoussis, J. Genotyping Technologies for Genetic Research. Annu. Rev. Genom. Hum. Genet 2009. 10:117-133.

Genotyping Platforms

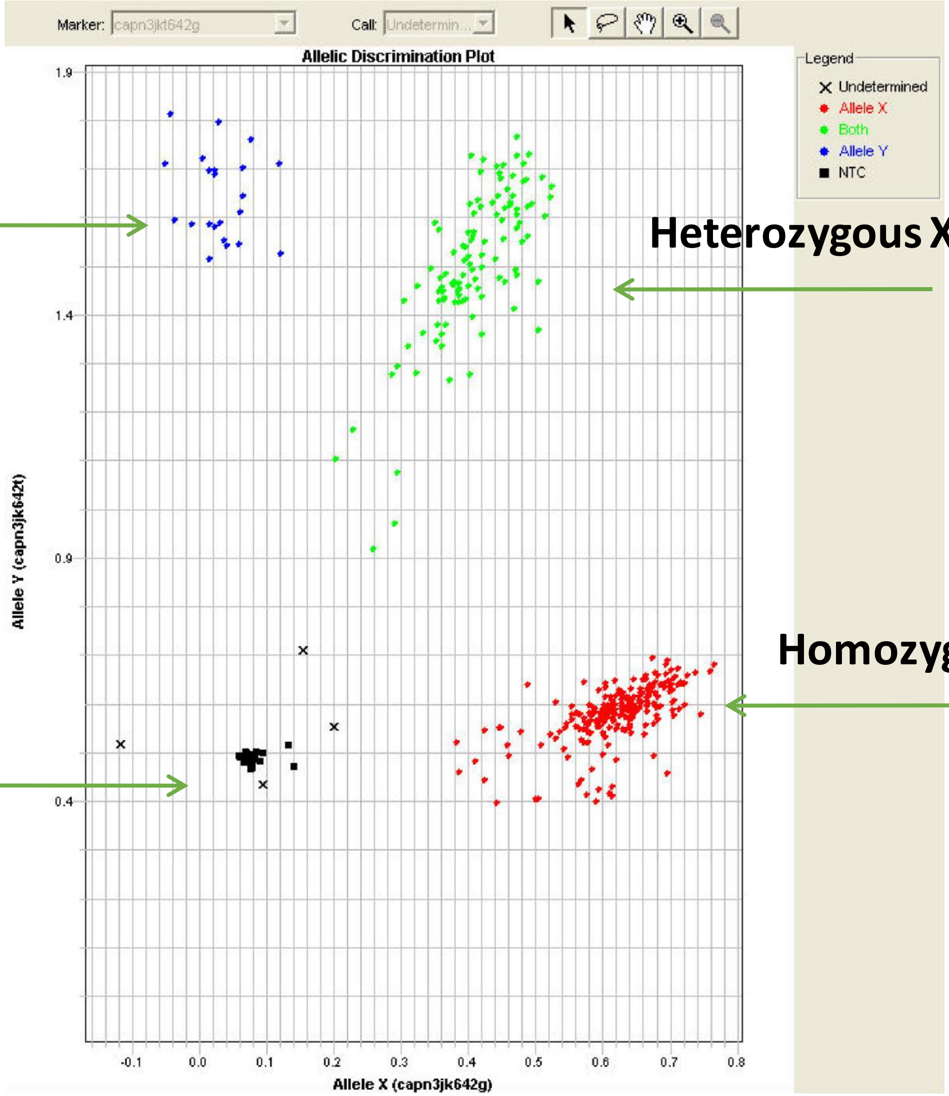
	Assay Type	Technology Basis	Throughput/person	Multiplexing (# SNPs)	Application
TaqMan / OpenArray	5' exonuclease/PCR	TaqMan probes	384-1536 samples/day	64-256	Medium custom SNP density; medium-large sample size
SNPlex	OLA/PCR	Capillary electrophoresis	1536 samples/ 3 days	24-48plex	Medium custom SNP density; large sample size
iPlex	Primer extension	MALDI-TOF Mass spec	3840 samples/ 2.5 days	12-40 plex	Medium custom SNP density; large sample size
Goldengate	Primer extension/ ligation	Bead Array	172 samples/ 3 days	384-1536	High custom or off-shelf SNP density; medium-large sample size
GeneChip	Hybridization	Oligonucleotide array	96 samples/ 5 days	10,000 – 1.8M	WGA studies; off-shelf assays; small-large sample size
Infinium II	Hybridization/Primer extension and ligation	Bead Array	32-128 samples/5 days	6,000-1.2M	WGA studies; very high density custom SNP studies; small-large sample size

Ragoussis, J. Genotyping Technologies for Genetic Research. Annu. Rev. Genom. Hum. Genet 2009. 10:117-133.

TaqMan/OpenArray

- 5’ nuclease assay
- Single tube/well
- Real-time PCR required (ABI 7900HT)
- Detects fluorescence
- Advantages
  - 1 reaction
  - several million validated assays available off-the-shelf
- OpenArray
  - Multiplexed TaqMan
  - 64-256 SNPs at one time on 12-48 samples

TaqMan/OpenArray


Marker: capn3jkt642g

Call: Undetermin...

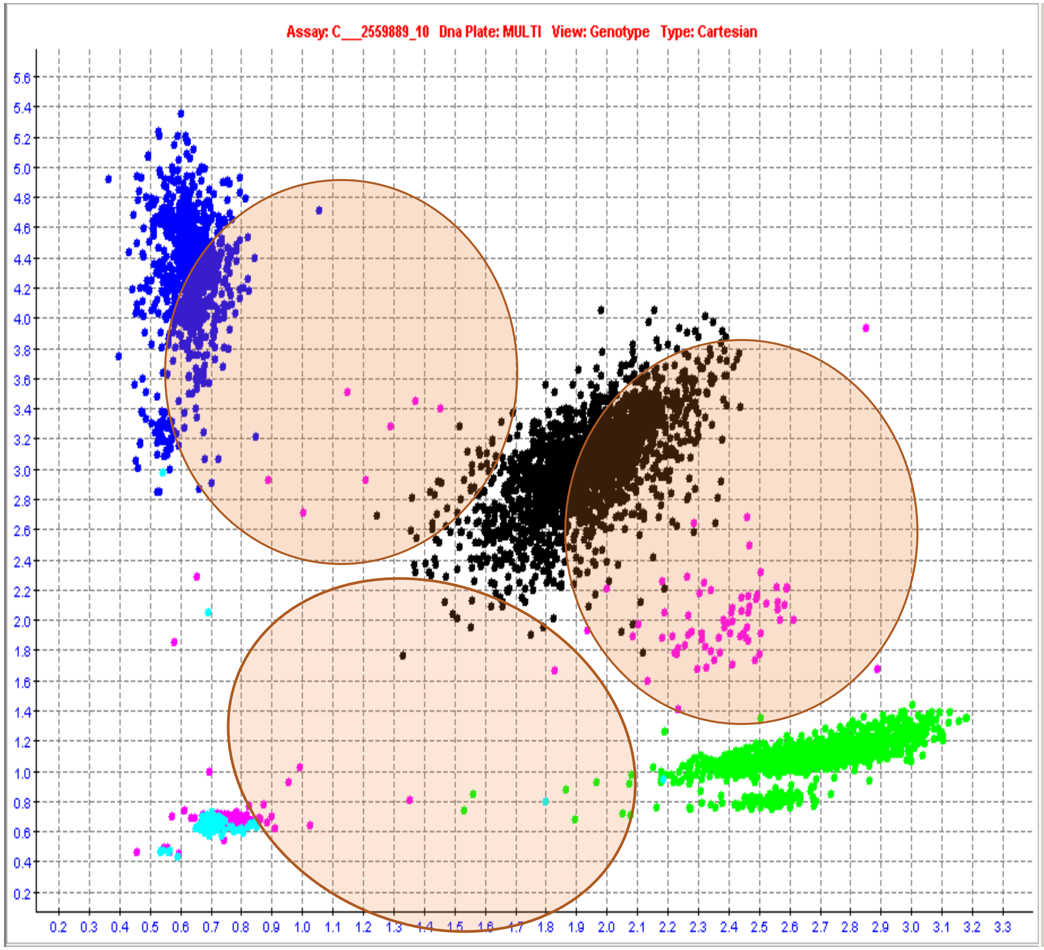
Homozygous YY

Heterozygous XY

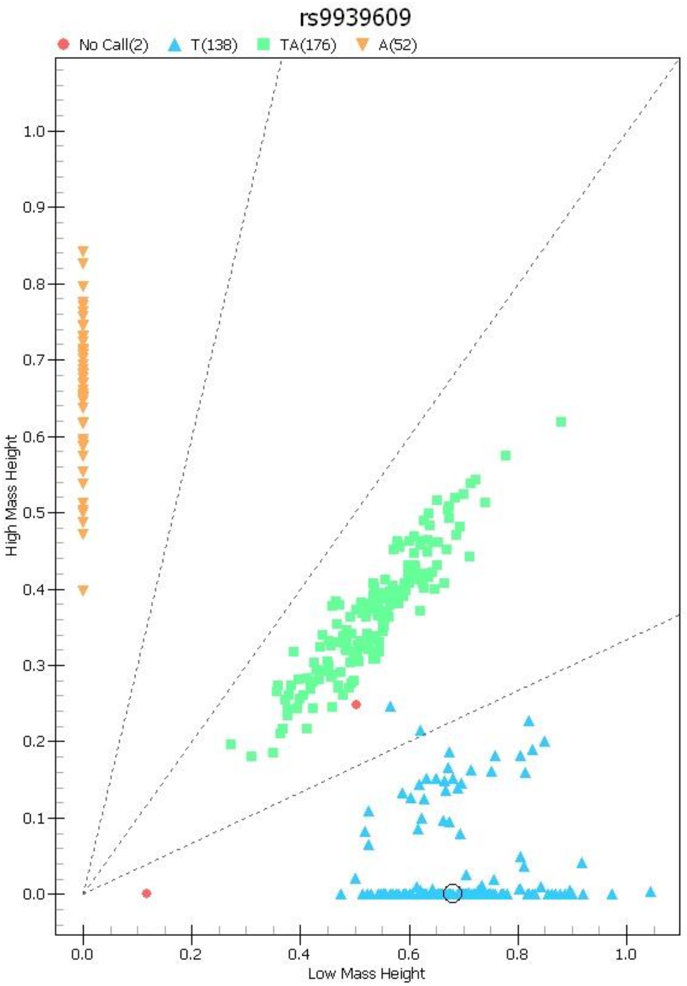
Homozygous XX

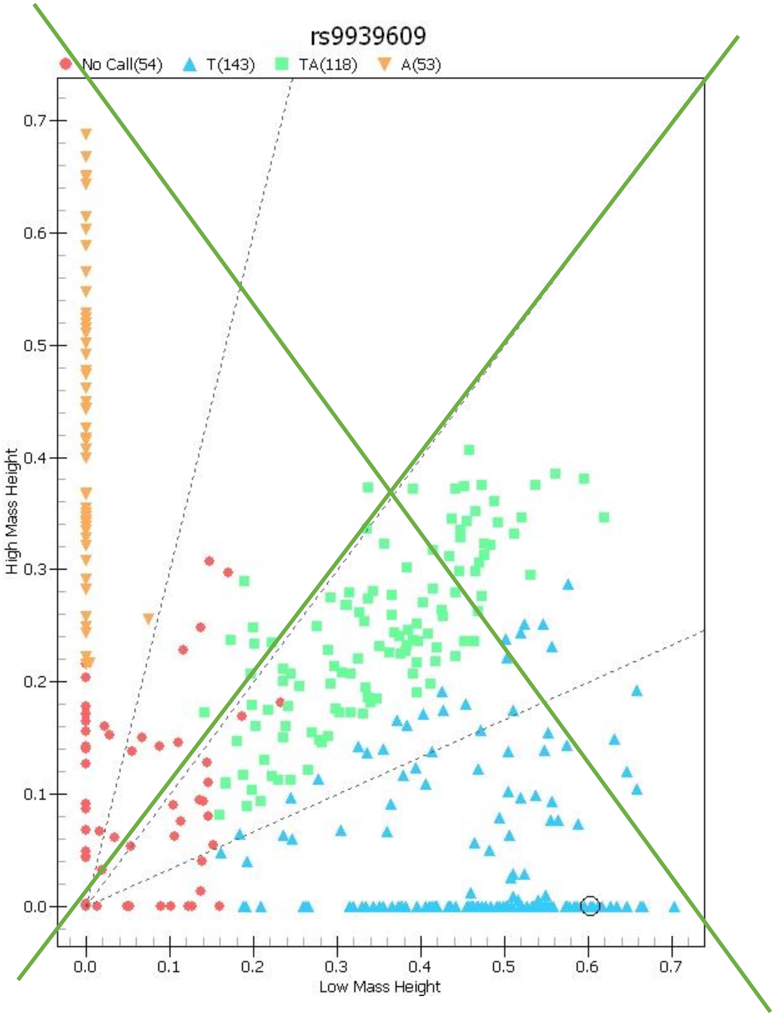
Undetermined

TaqMan/OpenArray

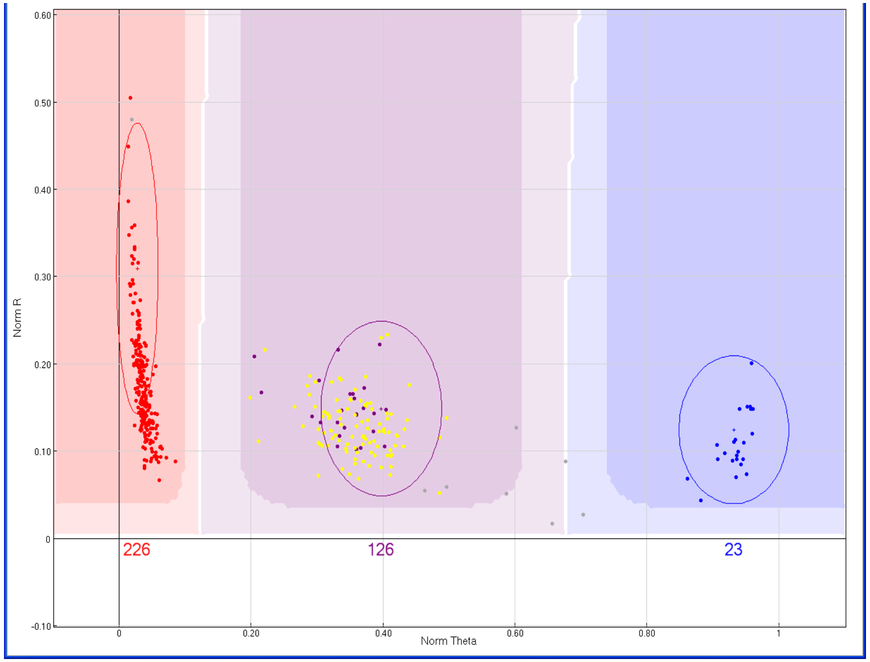


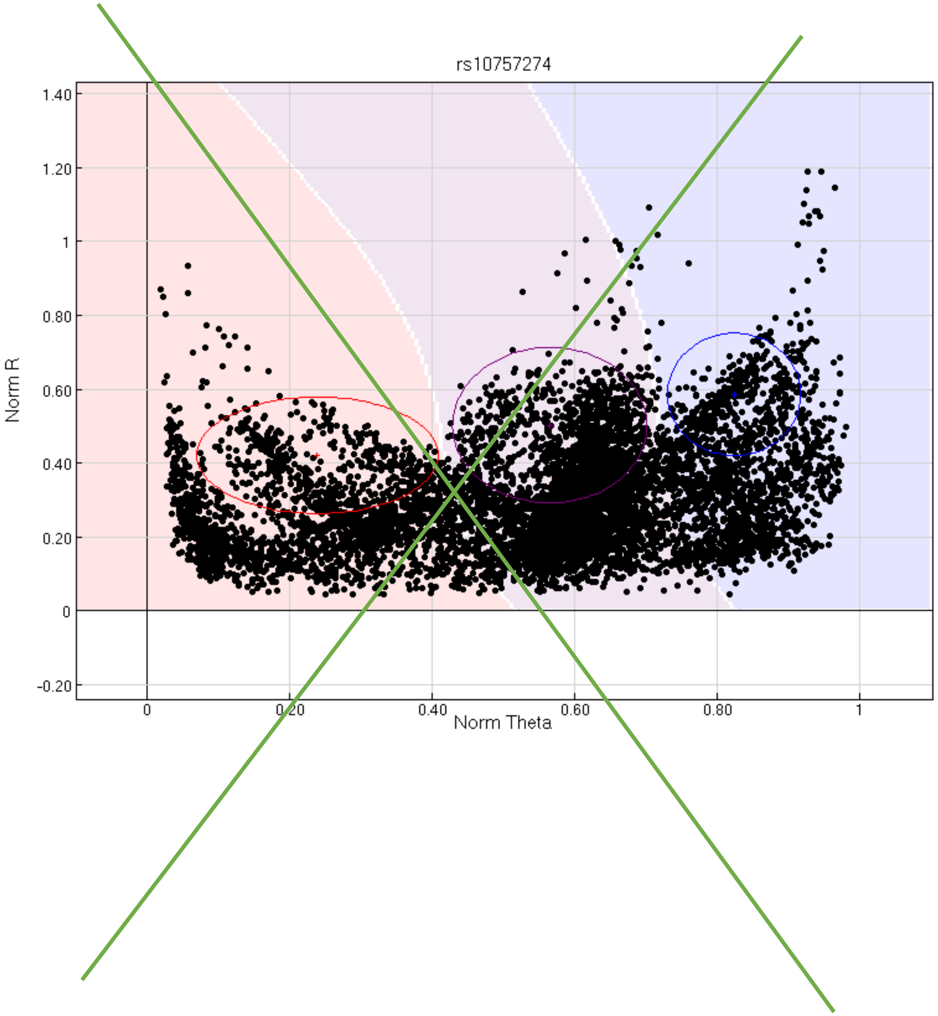
iPlex - Sequenom





Illumina Goldengate

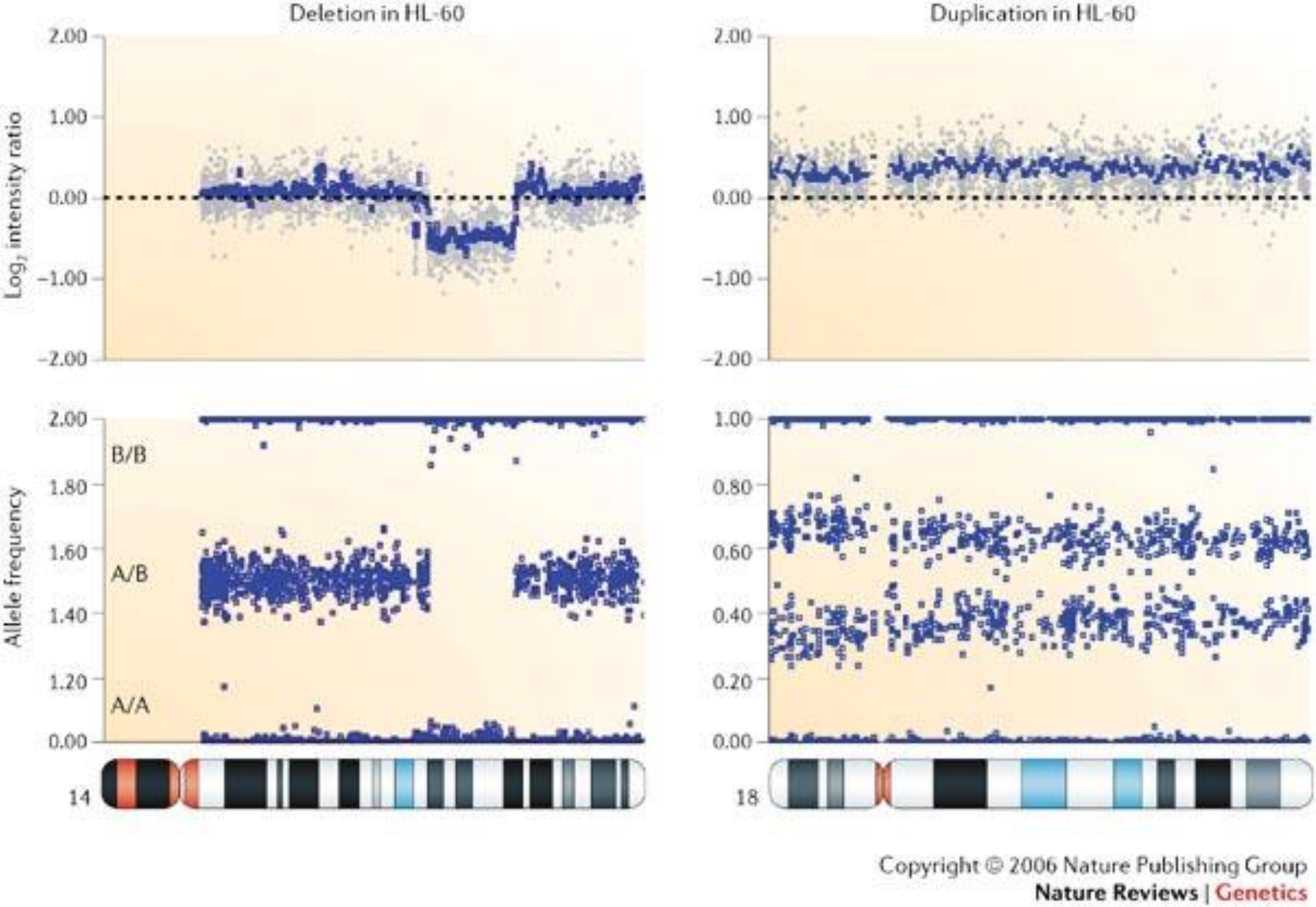




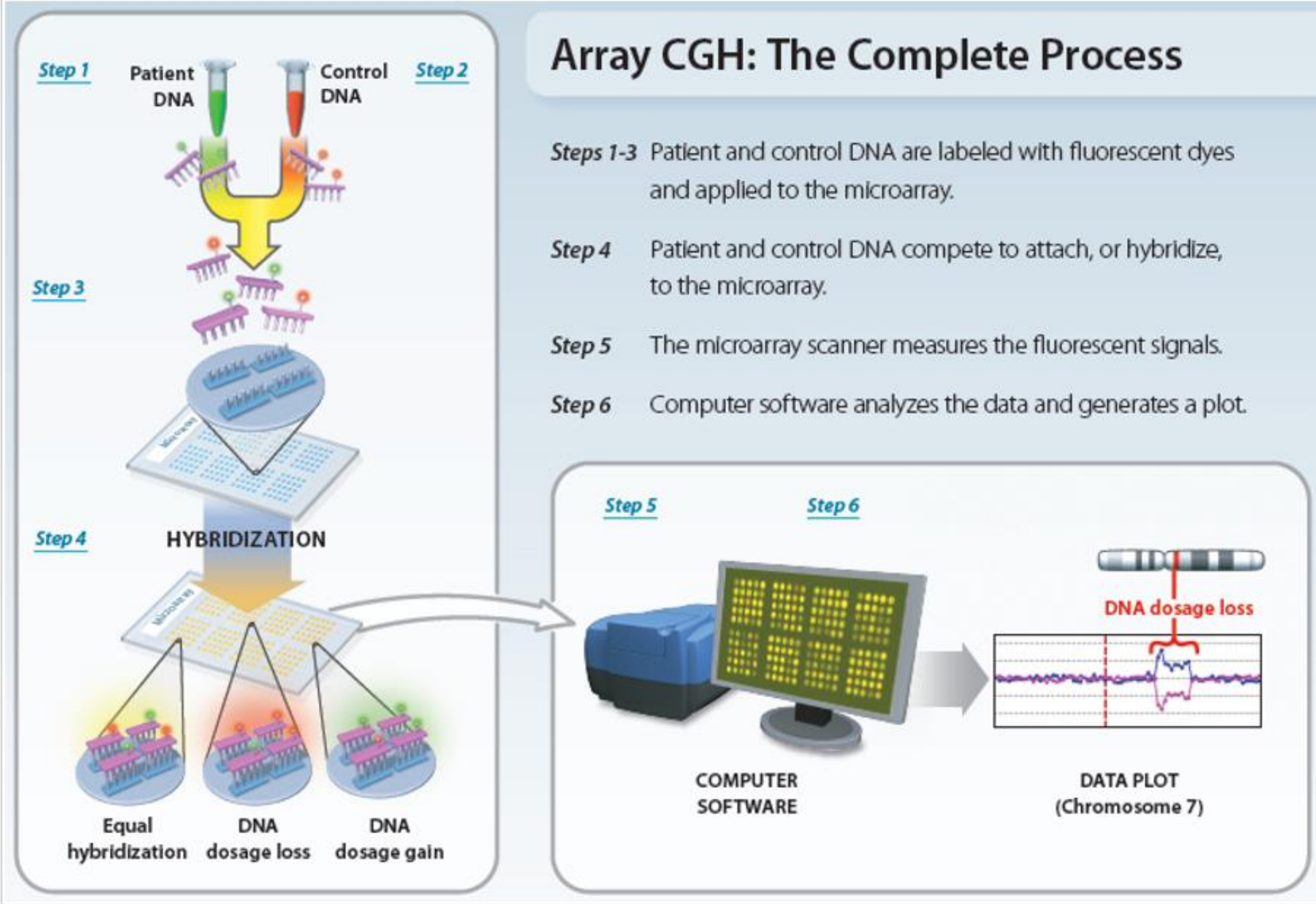
Hardware for Genotyping

ay Genotyping System, see the feature article,
com/taqmanopenarray

Figure 1. TaqMan® OpenArray™ Genotyping System.



Nature Reviews Genetics 7, 632-644 (August 2006)



Array CGH: The Complete Process

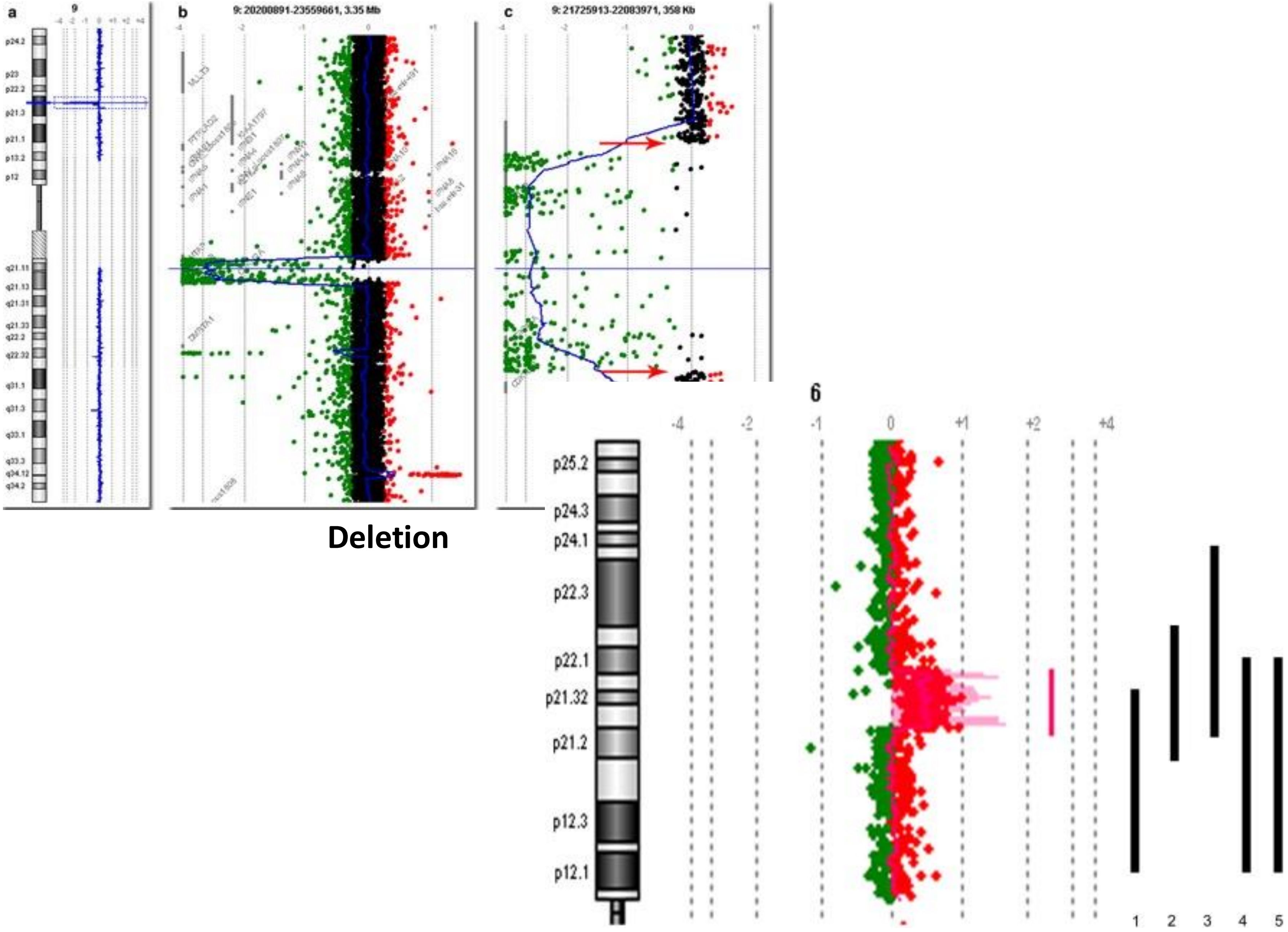
Steps 1-3 Patient and control DNA are labeled with fluorescent dyes and applied to the microarray.

Step 4 Patient and control DNA compete to attach, or hybridize, to the microarray.

Step 5 The microarray scanner measures the fluorescent signals.

Step 6 Computer software analyzes the data and generates a plot.

Figure 1 : Diagram of the microarray-based comparative genomic hybridization (aCGH) process

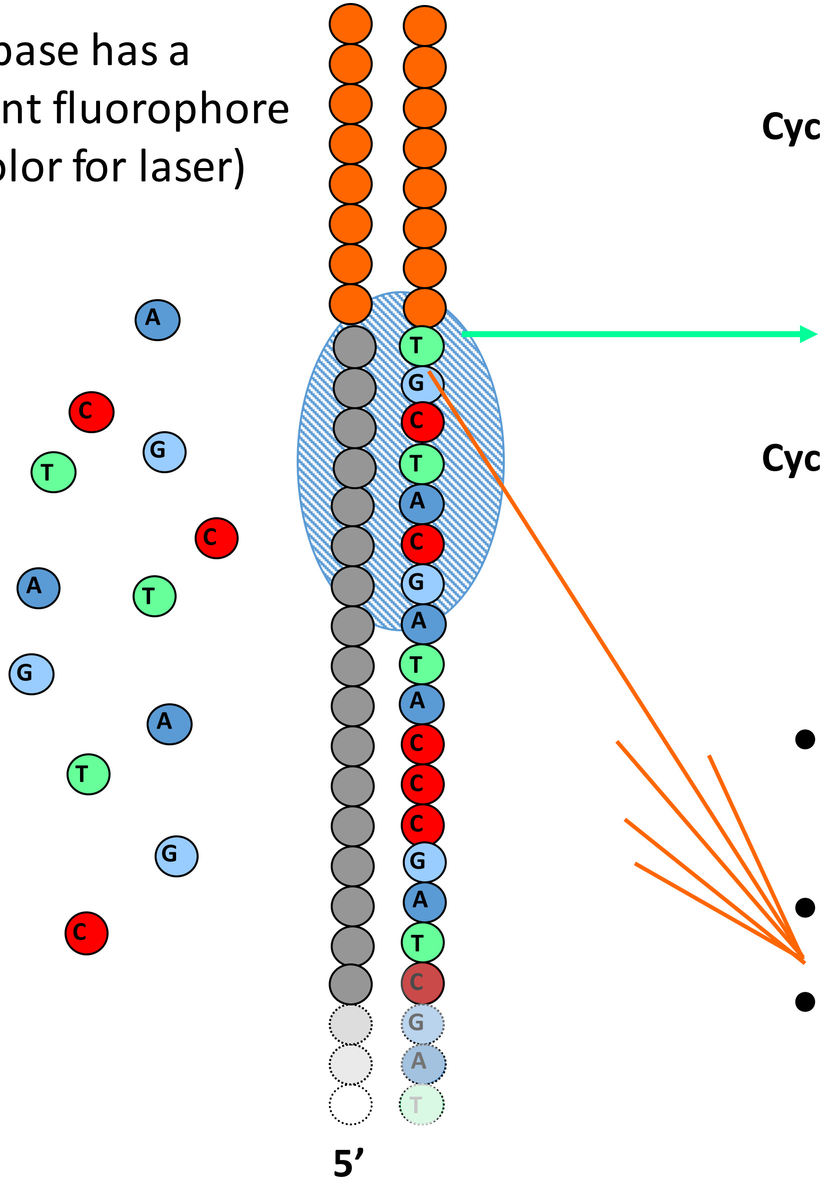


Deletion

Duplication

Sequencing by Synthesis: Reverse Terminator Chain Sequencing

Every base has a
different fluorophore
(diff color for laser)



Cycle 1: **Add sequencing reagents**

First base incorporated

Remove unincorporated bases

Detect signal

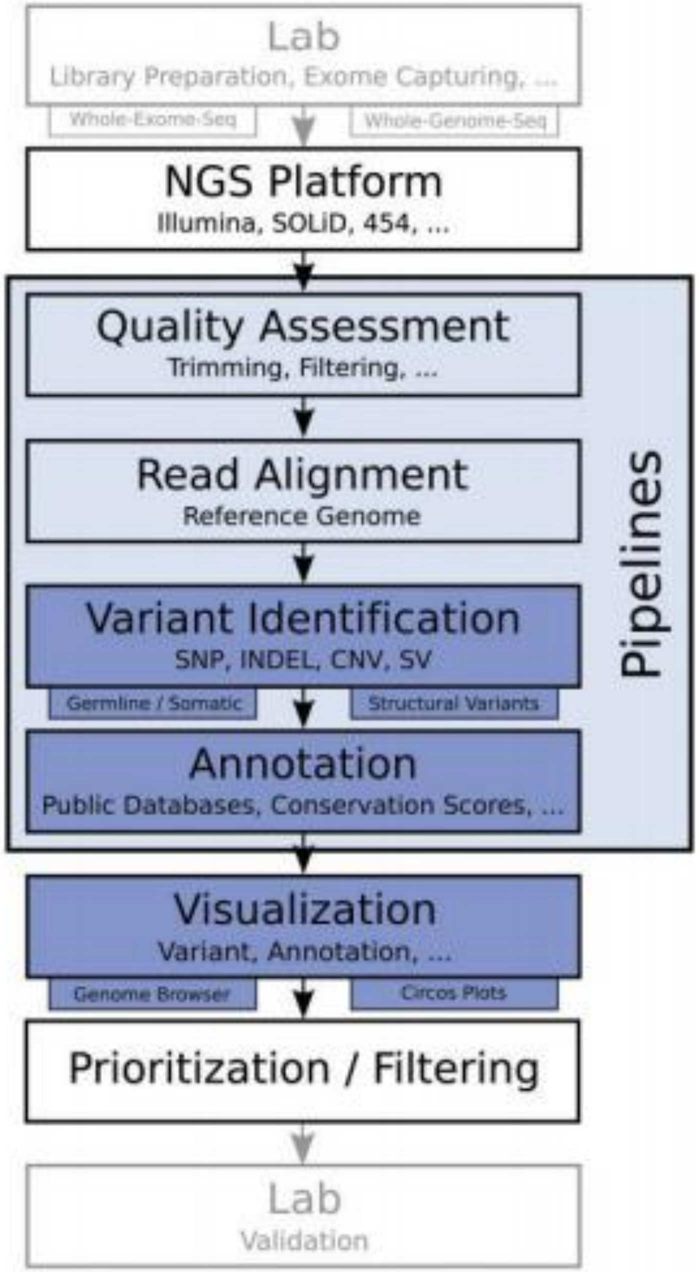
Cycle 2-n: **Add sequencing reagents and repeat**

- All four labeled nucleotides in one reaction
- Base-by-base sequencing
- Polymerase can only extend by one base

Slide courtesy of Holli Dilks

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



Lab

Library Preparation, Exome Capturing, ...

Whole-Exome-Seq

Whole-Genome-Seq

NGS Platform

Illumina, SOLiD, 454, ...

Quality Assessment

Trimming, Filtering, ...

Read Alignment

Reference Genome

Variant Identification

SNP, INDEL, CNV, SV

Germline / Somatic

Structural Variants

Annotation

Public Databases, Conservation Scores, ...

Pipelines

Visualization

Variant, Annotation, ...

Genome Browser

Circos Plots

Prioritization / Filtering

Lab

Validation

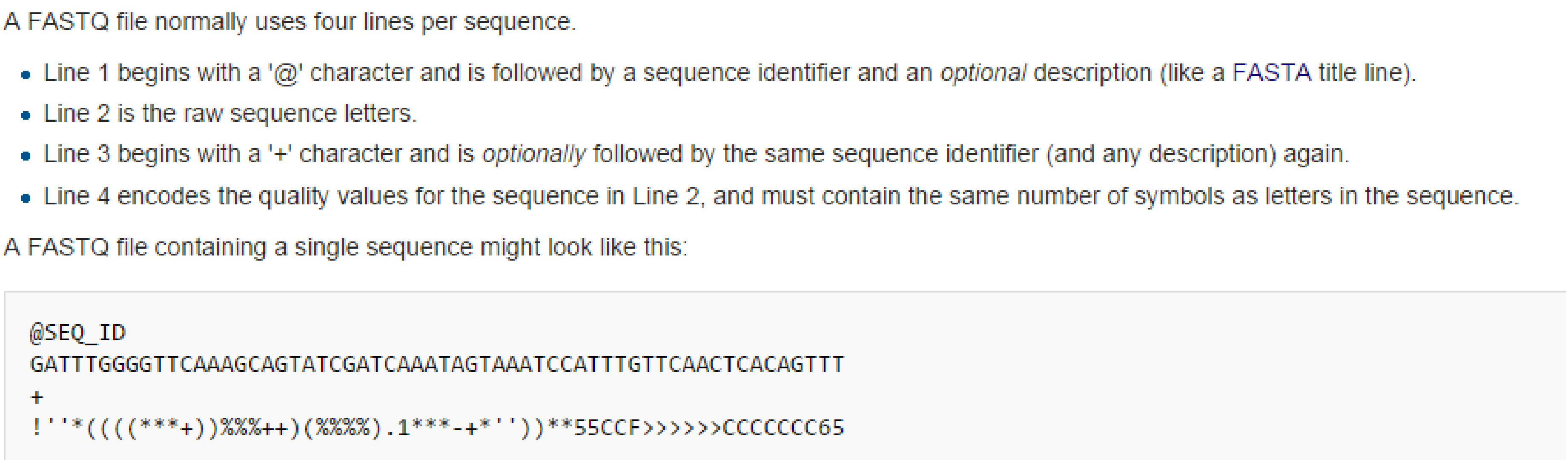
Quality assessment

- 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.

Pabinger et a. 2013 Briefings in Bioinformatics

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
TileQC [11]	Lin, Mac	Eland output	Graphs	Illumina	yes	no	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
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

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

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

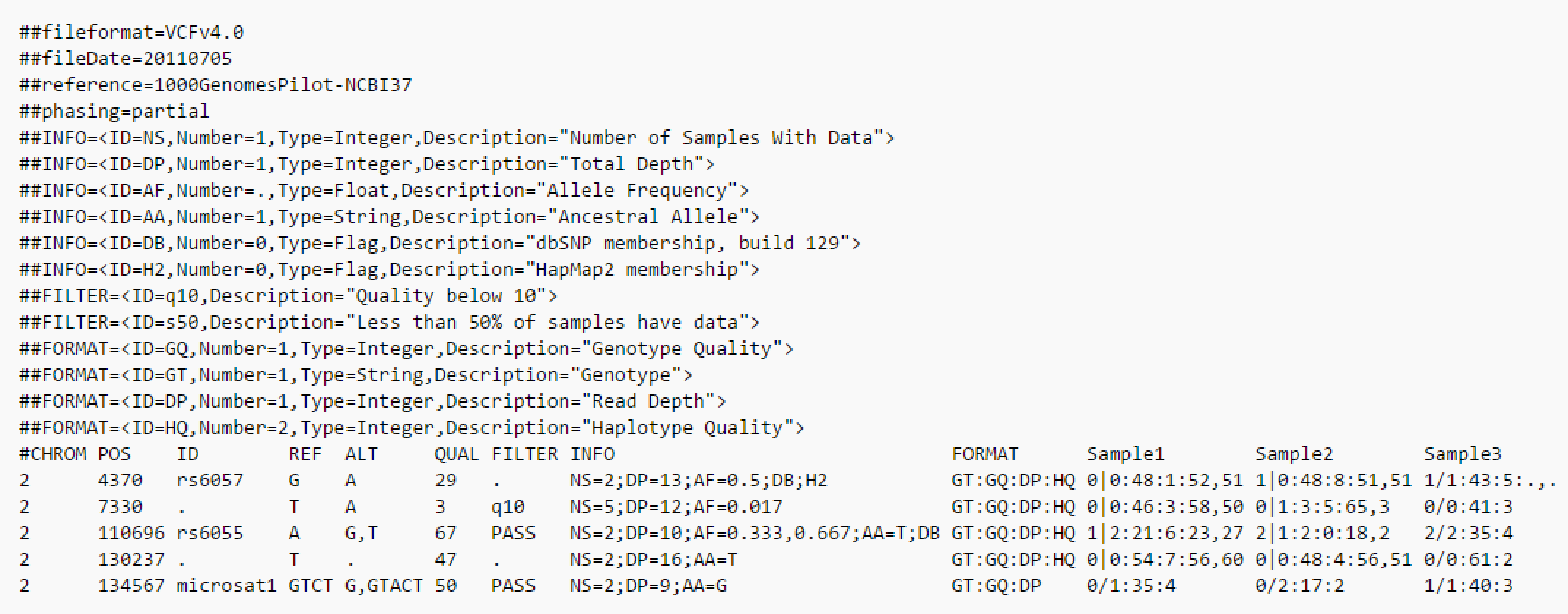
Step 2: Variant Calling

- 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

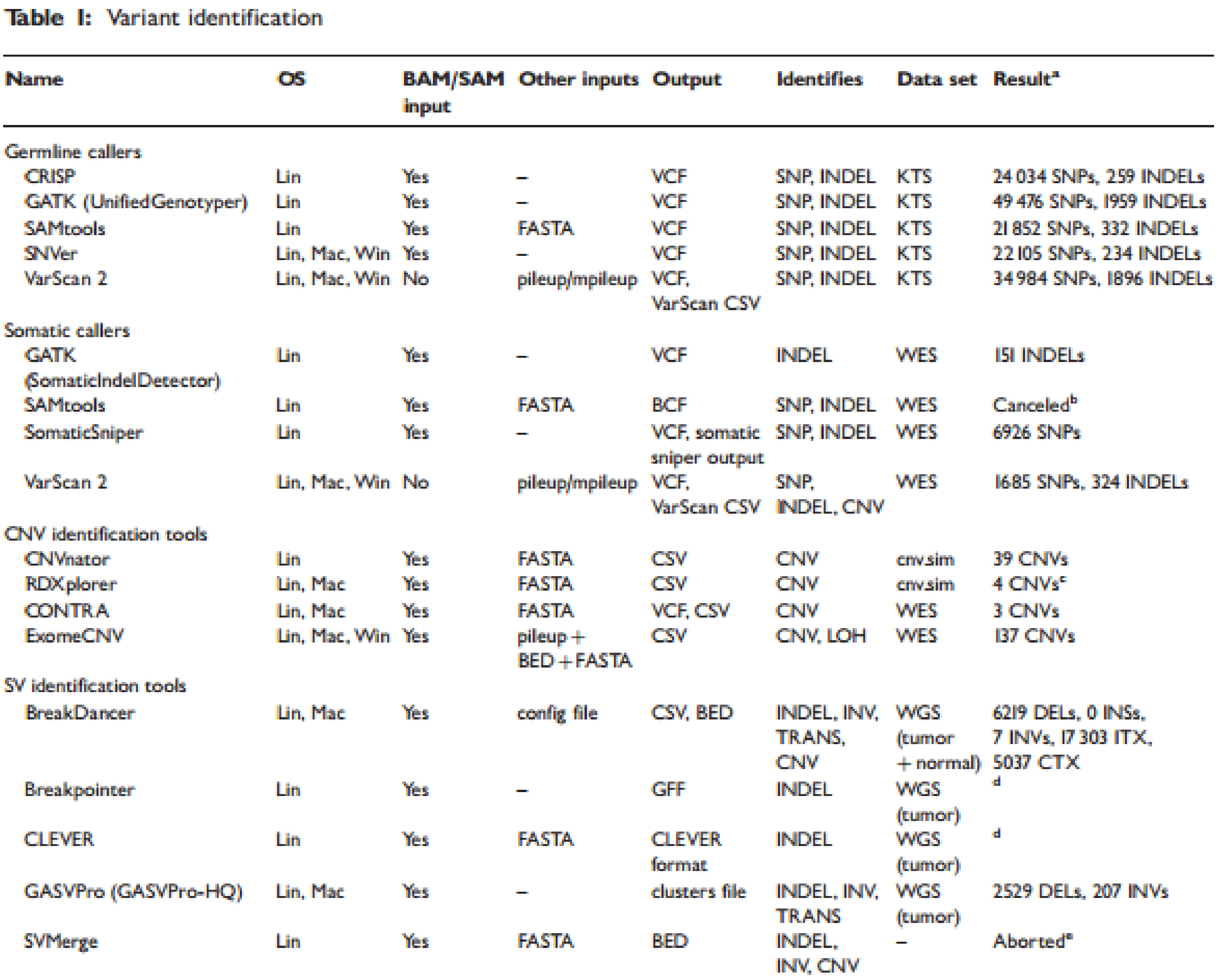


Table 1: Variant identification

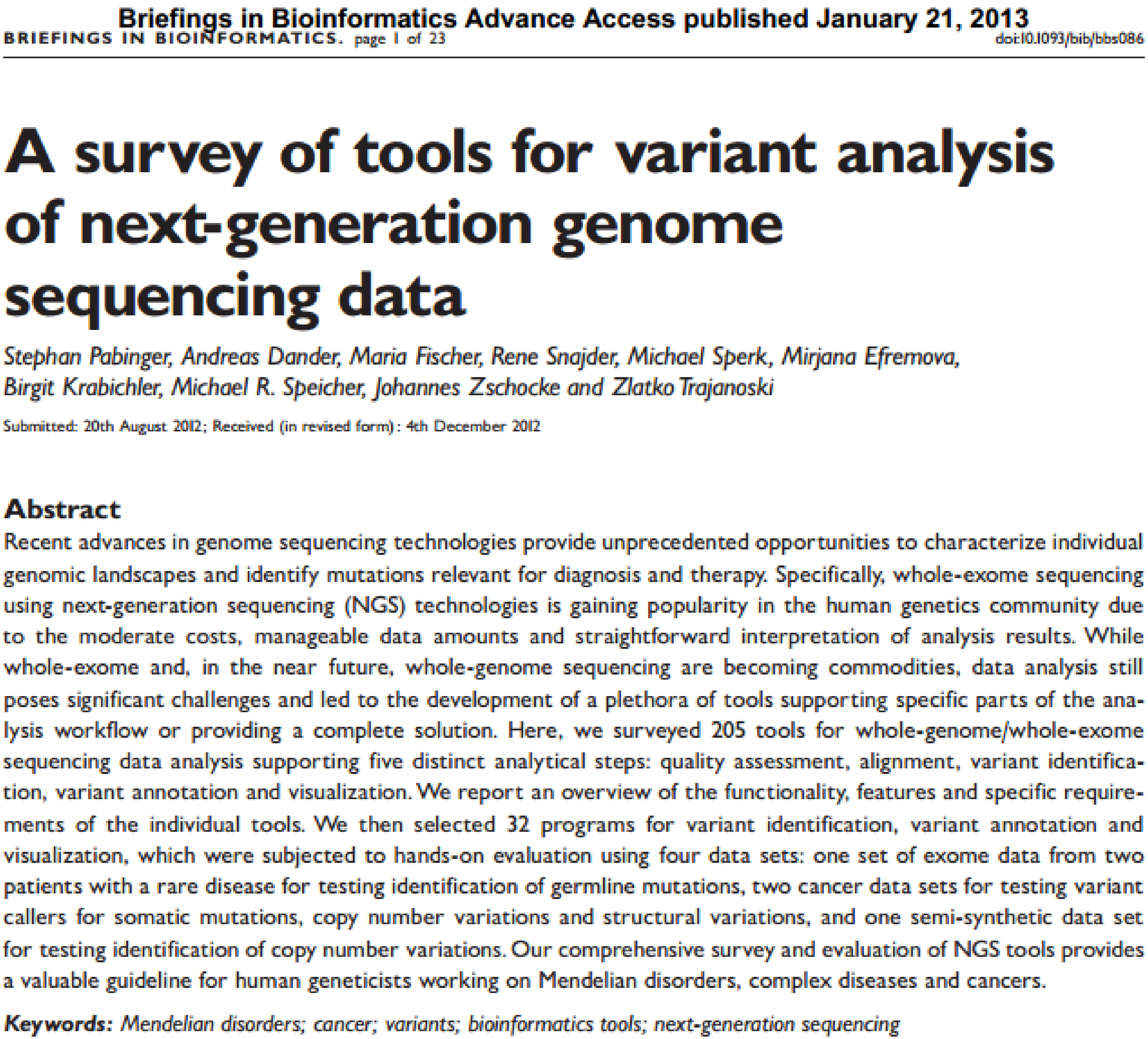
Name	OS	BAM/SAM input	Other inputs	Output	Identifies	Data set	Result[a]
Germline callers							
CRISP	Lin	Yes	–	VCF	SNP, INDEL	KTS	24 034 SNPs, 259 INDELs
GATK (UnifiedGenotyper)	Lin	Yes	–	VCF	SNP, INDEL	KTS	49 476 SNPs, 1959 INDELs
SAMtools	Lin	Yes	FASTA	VCF	SNP, INDEL	KTS	21 852 SNPs, 332 INDELs
SNVer	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 INSs, 7 INVs, 17 303 ITX, 5037 CTX
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



Briefings in Bioinformatics Advance Access published January 21, 2013

doi:10.1093/bib/bbs086

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

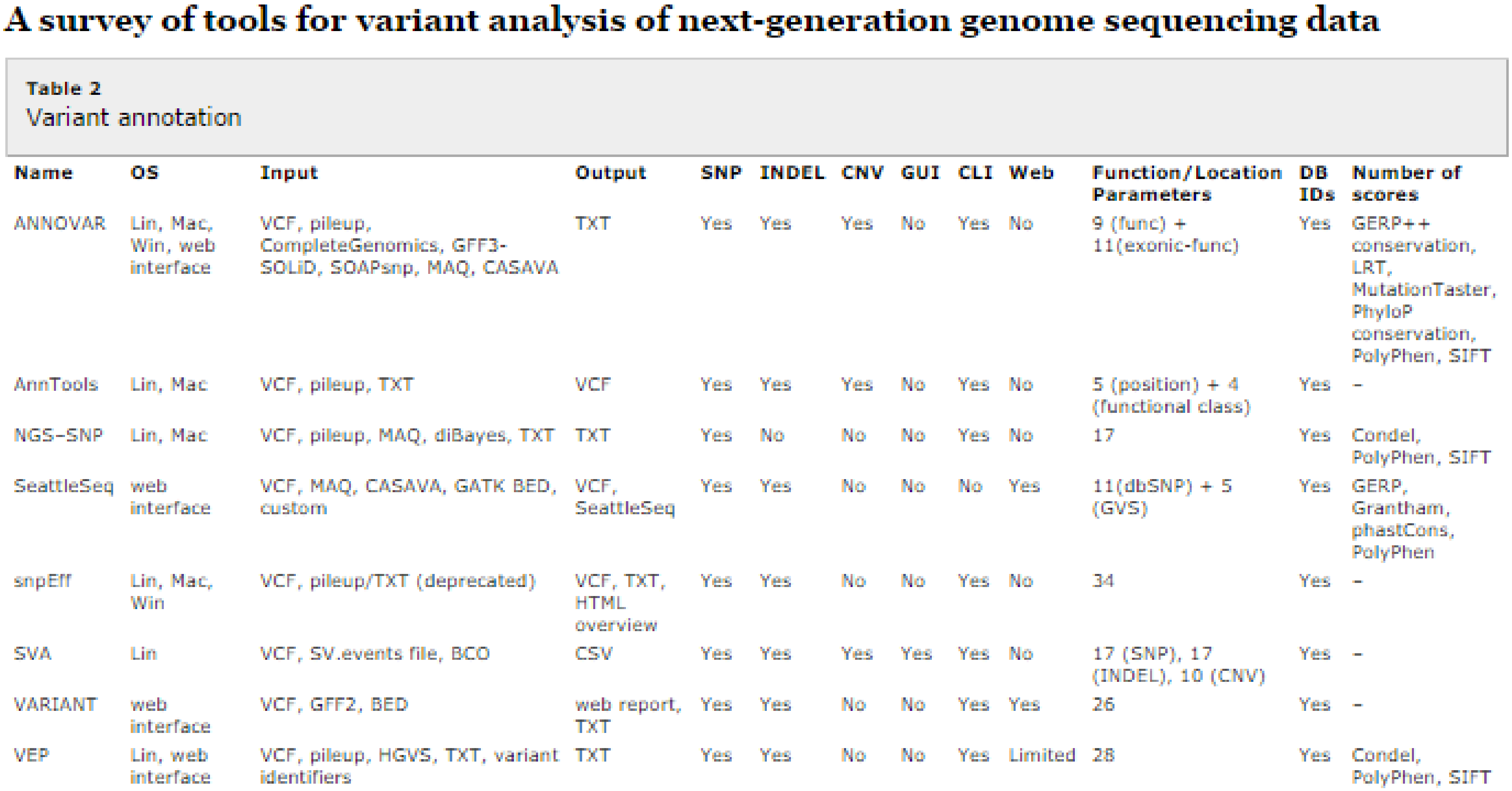
Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski

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

Data Management

- Generating 300,000-1,000,000 SNPs on 1,000-5,000 individuals means 300 Million-5 Billion genotypes.
- Then there’s all the clinical data you have to match with the genotypes (age, smoking status, BMI, etc.)
- This is way beyond Excel. Can your computer handle it?

Data Management

- Most files stored in binary compressed format
  - This means you cannot open them and look at it on the screen
- Need to rely on scripts and computer programs to work with the data
- Led to an influx in jobs in bioinformatics

Quality control analysis

- Two different types of QA/QC performed
  - QA in the lab where genotyping is done
  - QC in the lab where data analysis is underway
- Each checking for different things
  - With some overlap
- Important to ensure data integrity
- Without QC, can lead to spurious results
  - Type I errors and Type II errors

Quality control analysis

- VERY different QA pipelines in genotyping labs for research and clinical use
    - CLIA: Clinical Laboratory Improvement Amendments
    - CLIA: United States federal regulatory standards that apply to all clinical laboratory testing performed on humans in the United States, except clinical trials and basic research.

Quality control analysis

- Primary differences between CLIA and research lab genotyping
  - Sample tracking
  - Assay validation
  - Security
  - Equipment validation/calibration
  - SOPs (standard operating procedures)
    - With verification
  - <u>COST</u>

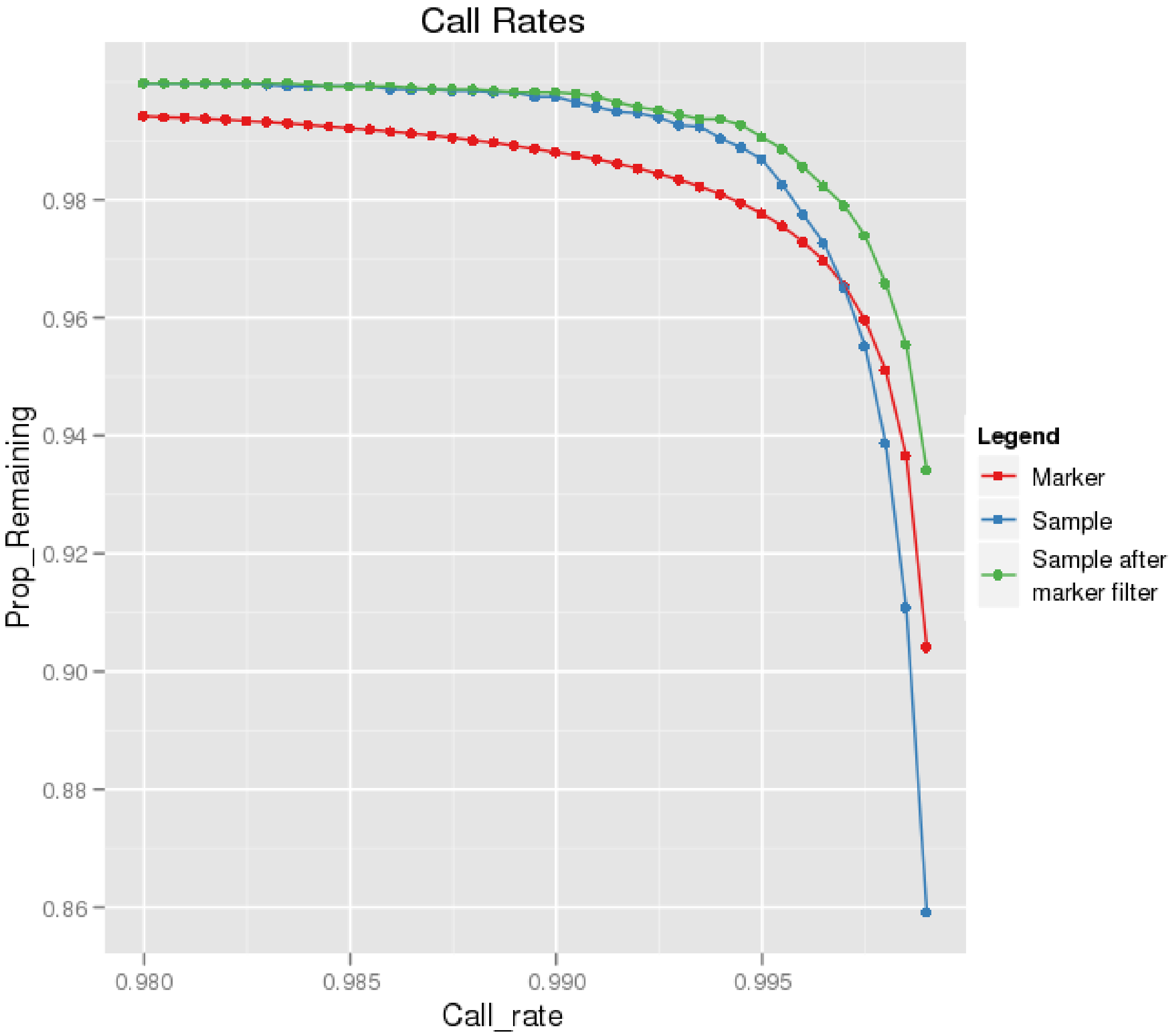
Quality control analysis

- Differences between CLIA and research plays a role in
  - What variants go into clinical practice
  - Timeline for variants being used in clinic

Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.

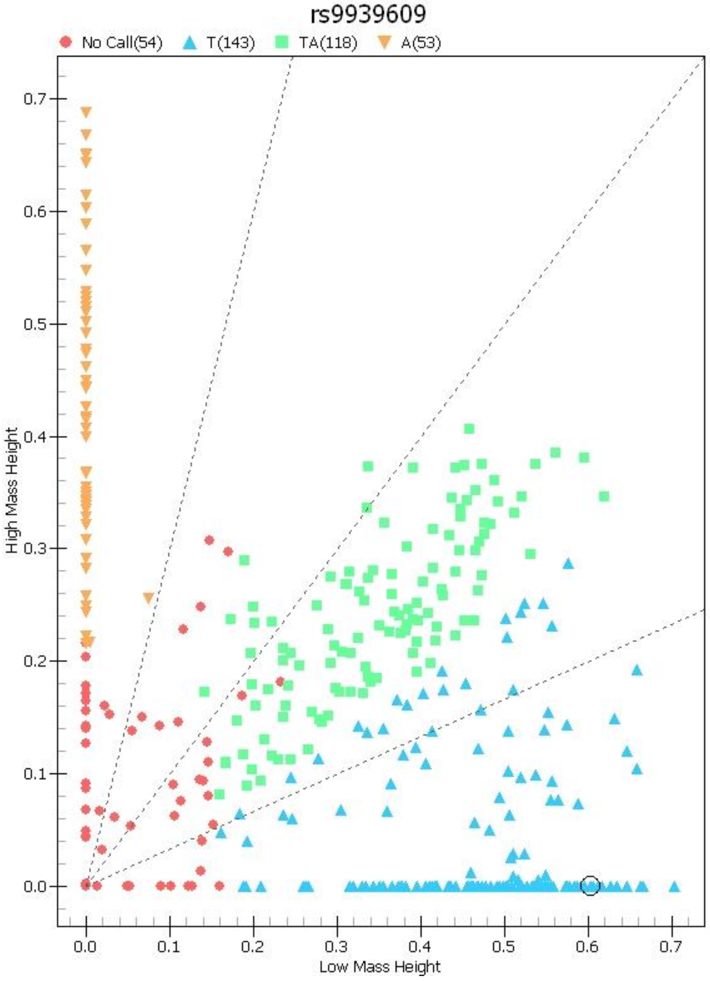
Marker and Sample Call Rate



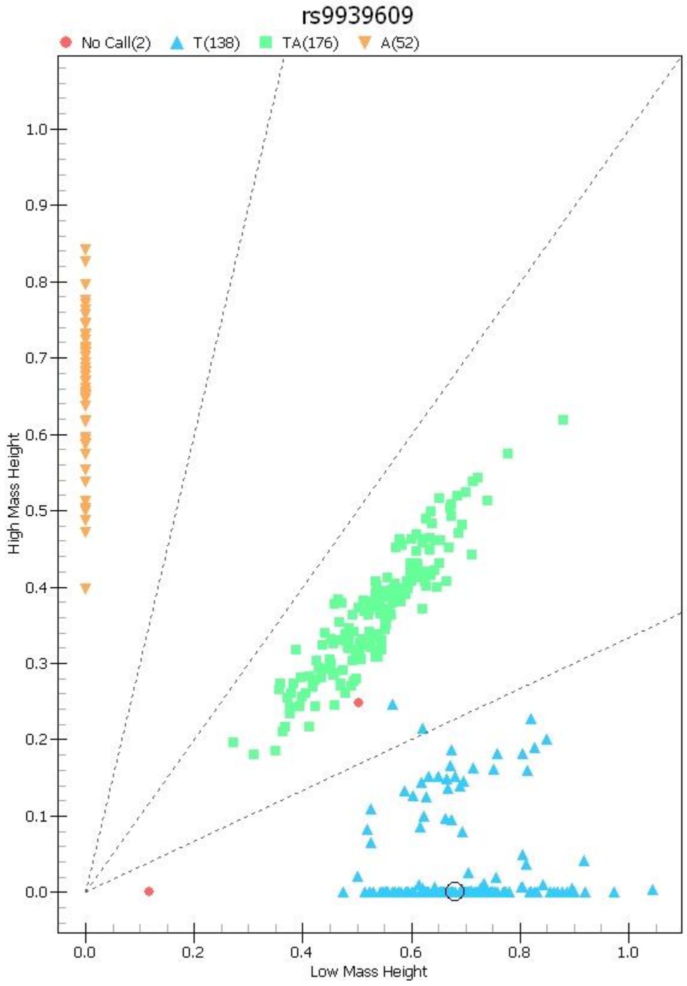
Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate

Genotyping Failures



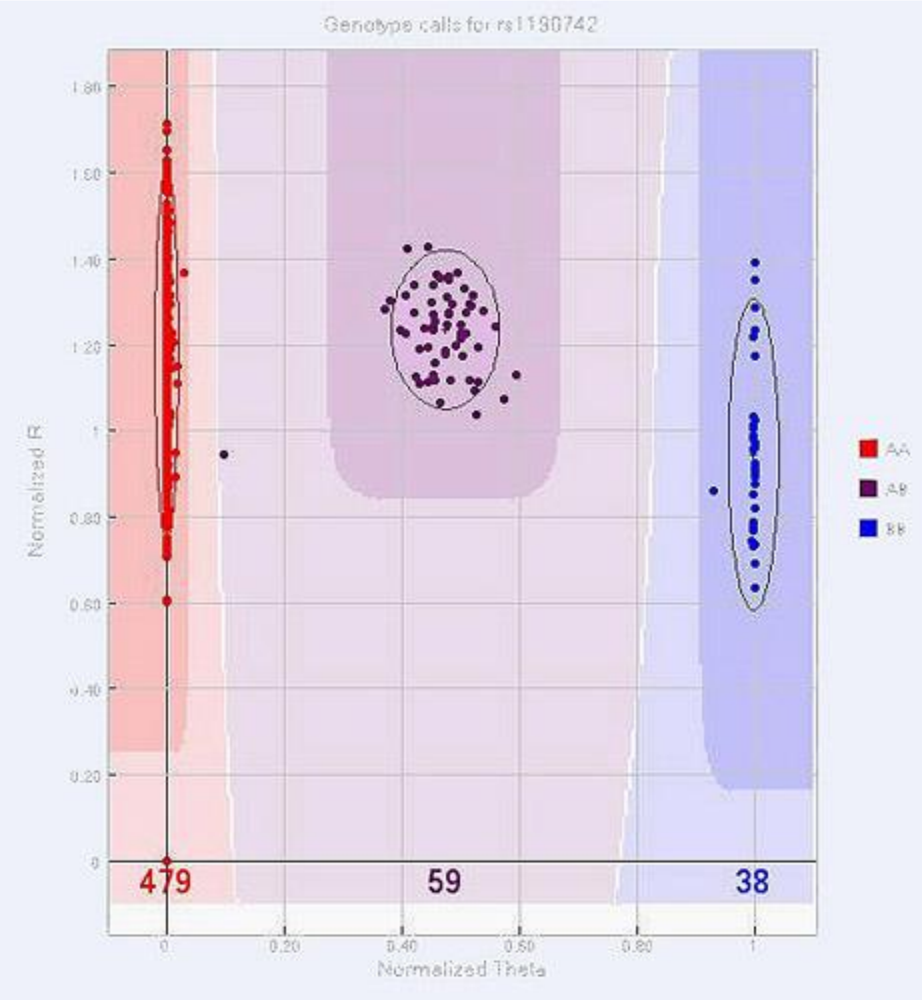
NHANES III
(Failed)



NHANES 99-02
(OK)

Courtesy of Dana Crawford

Genotyping success



Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.

Female: pedsex=2, SNPsex=2
Male: pedsex=1, SNPsex=1

Sex Concordance Check

emerge_id	Pedsex	SNPsex	PLINK_F	Note
16230834	2	0	0.4746	CIDR comment after review of B allele freq and Log R ratio plots for all chromosomes: This sample has large loss-of-heterozygosity (LOH) blocks on X (and other autosomes). The sample is definitely female (2 X chromosomes by intensities).
16228083	2	0	0.2654	Same as above
16231930	2	0	0.4376	Same as above
16233764	2	0	0.2603	Same as above
16221112	2	0	0.2048	XX/XO mosaic not caught by initial check completed by CIDR
16222319	2	0	0.7452	Annotation by CIDR at data release: Appears to be XX/XO mosaic
16228204	2	1	1	Annotation by CIDR at data release: Appears to be XX/XO mosaic
16233113	1	0	0.4752	Annotation by CIDR at data release: Appears to be XXY
16214881	1	2	0.136	Annotation by CIDR at data release: Appears to be XXY/XY mosaic

- Female: pedsex=2, SNPsex=2
- Male: pedsex= 1, SNPsex=1
- A male call is made if the F (actual X chromosome inbreeding estimate) is more than 0.8; a female call is made if the F is less than 0.2.

Sex Concordance

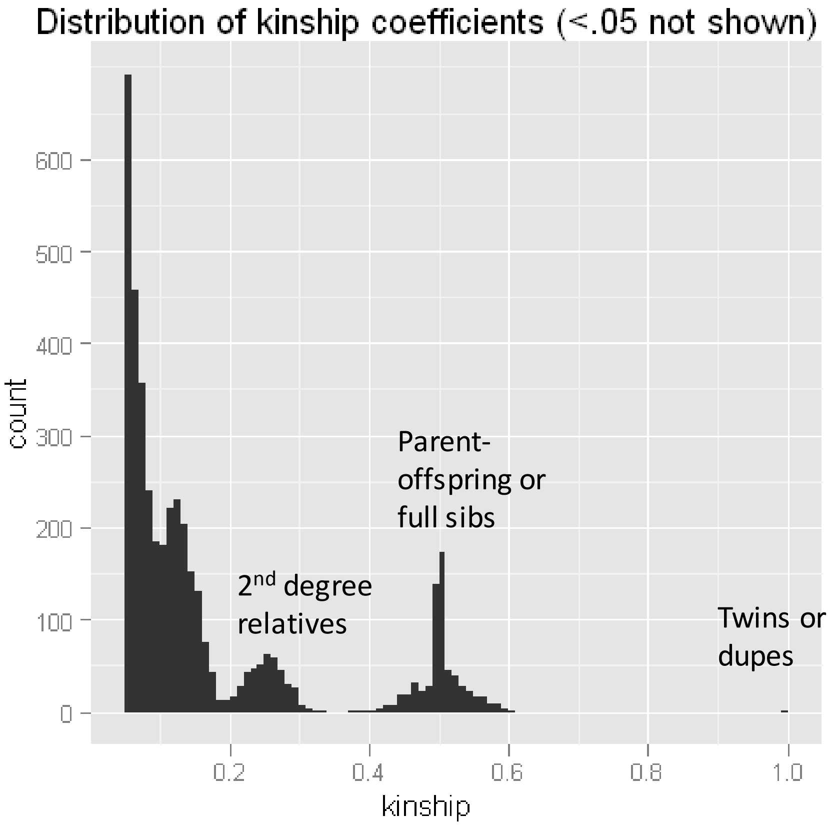
- Check sex chromosome markers for two reasons
  1. To identify and sex chromosome anomalies
  2. To identify and sample mix-ups
     - Phenotype = male, genotype = female or vice versa
     - Can be indicative of sample mix-up

Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)

Sample Relatedness

Z0	Z1	Z2	Kinship	Relationship
0.0	0.0	1.0	1.0	MZ twin or duplicate
0.0	1.0	0.0	0.50	Parent-offspring
0.25	0.50	0.25	0.50	Full siblings
0.50	0.50	0.0	0.25	Half siblings
0.75	0.25	0.0	0.125	Cousins
1.0	0.0	0.0	0.0	Unrelated

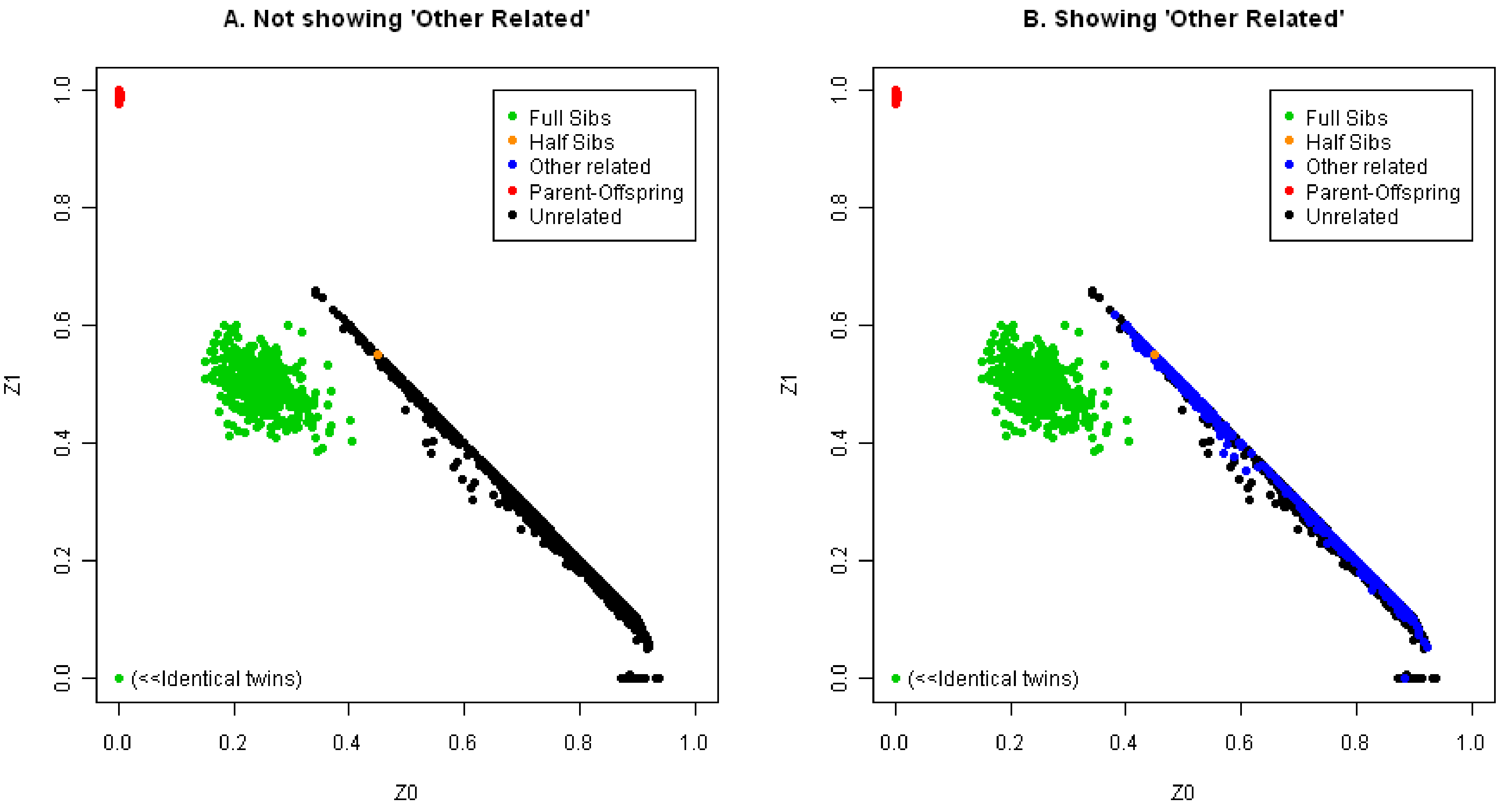


2^{nd} degree relatives

Parent-offspring or full sibs

Twins or dupes

Sample Relatedness



Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)
Mendelian Inheritance Errors	For trio/family data, can identify problem samples and families. Can estimate error rate.

Mendelian Inheritance Errors

- Typically HapMap trios are plated and genotyped in addition to study samples
- Allows for an additional QC step

Number Mendelian Errors	Number SNPs pre QC	Number SNPs post marker QC
0	558821	552346
1	1519	1353
2	97	64
3	5	1

Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)
Mendelian Inheritance Errors	For trio/family data, can identify problem samples and families. Can estimate error rate.
Replicate concordance	Check for consistent genotype calls in duplicate samples

Replicate Concordance

emerge	Samp1	samp2	discordant	total	concordance_rate
16231453	A	B	171	558882	0.99969
16223704	A	B	137	557783	0.99975
16216270	A	B	133	559711	0.99976
16230108	A	B	69	559341	0.99987
16224359	A	B	67	558868	0.99988
16234120	A	B	43	560202	0.99992
16232463	A	B	42	560355	0.99992
16234233	A	B	33	560384	0.99994
16216349	A	B	30	559345	0.99994
16215309	A	B	12	560041	0.99997
16224779	A	B	7	560412	0.99998
16231724	A	B	5	560427	0.99999
16233841	A	B	4	560519	0.99999
16221647	A	B	2	560457	0.99999
16230404	A	B	2	560309	0.99999
16226433	A	B	2	560500	0.99999
16234367	A	B	2	560373	0.99999
16224635	A	B	1	560560	0.99999
16219214	A	B	1	560535	0.99999
16231219	A	B	1	560547	0.99999
16220060	A	B	0	560580	1

Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)
Mendelian Inheritance Errors	For trio/family data, can identify problem samples and families. Can estimate error rate.
Replicate concordance	Check for consistent genotype calls in duplicate samples
Batch effects	Check for genotyping call differences due to plate

Batch Effects

- Evidence that associations can result due to allele frequency difference due to plate effects
- Careful consideration when creating plate maps
    - Plate cases and controls together
    - Randomize by race, gender, age, BMI, others…
- After genotyping look for plate effects
    - MAF differences by plate
    - Call rate by plate
    - Association tests (one plate versus all others)

Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)
Mendelian Inheritance Errors	For trio/family data, can identify problem samples and families. Can estimate error rate.
Replicate concordance	Check for consistent genotype calls in duplicate samples
Batch effects	Check for genotyping call differences due to plate
Hardy-Weinberg Equilibrium	Violation across all sample groups may indicate error, but can also be a good test of association

Hardy Weinberg Equilibrium

All individuals

threshhold	below	exp_below	excess_below
0.05	37690	28022	9668
0.01	12774	5604	7170
0.001	4766	560	4206
1.00E-04	2949	56	2893
1.00E-05	2337	5	2332
1.00E-06	2004	0	2004
1.00E-07	1785	0	1785

All cases

threshold	below	exp_below	excess_below
0.05	34646	28022	6624
0.01	10843	5604	5239
0.001	3642	560	3082
1.00E-04	2194	56	2138
1.00E-05	1792	5	1787
1.00E-06	1563	0	1563
1.00E-07	1394	0	1394

All controls

threshold	below	exp_below	excess_below
0.05	30557	28022	2535
0.01	8859	5604	3255
0.001	2614	560	2054
1.00E-04	1517	56	1461
1.00E-05	1180	5	1175
1.00E-06	982	0	982
1.00E-07	860	0	860

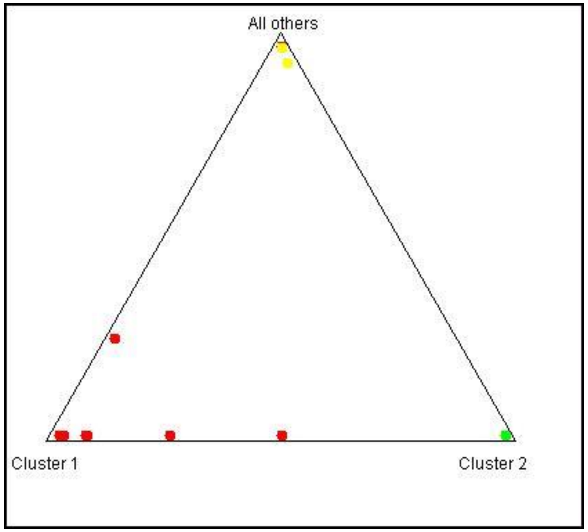
Quality control analysis

Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)
Mendelian Inheritance Errors	For trio/family data, can identify problem samples and families. Can estimate error rate.
Replicate concordance	Check for consistent genotype calls in duplicate samples
Batch effects	Check for genotyping call differences due to plate
Hardy-Weinberg Equilibrium	Violation across all sample groups may indicate error, but can also be a good test of association
Population Stratification	Check for population substructure using the genome-wide data

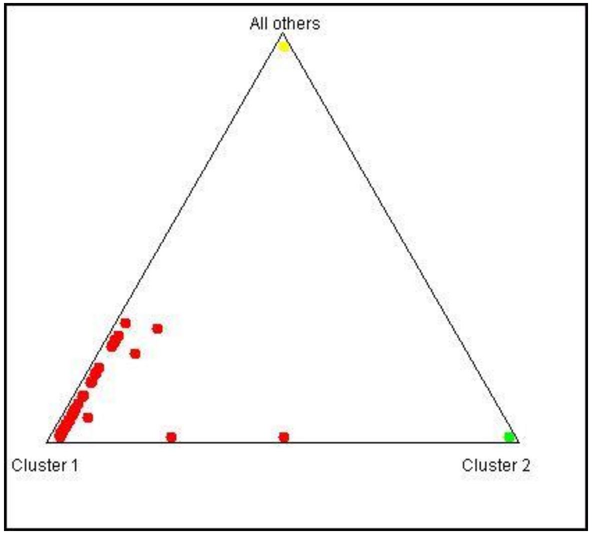
Population Stratification

STRUCTURE plot (CEU+Marshfield=Red, CHB=Green, YRI=Yellow)

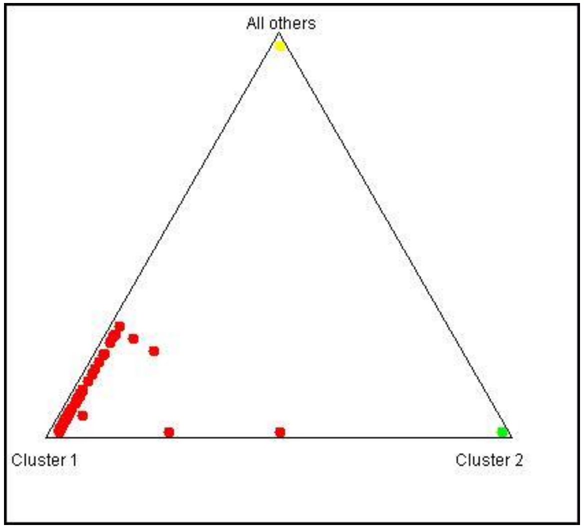
k=3



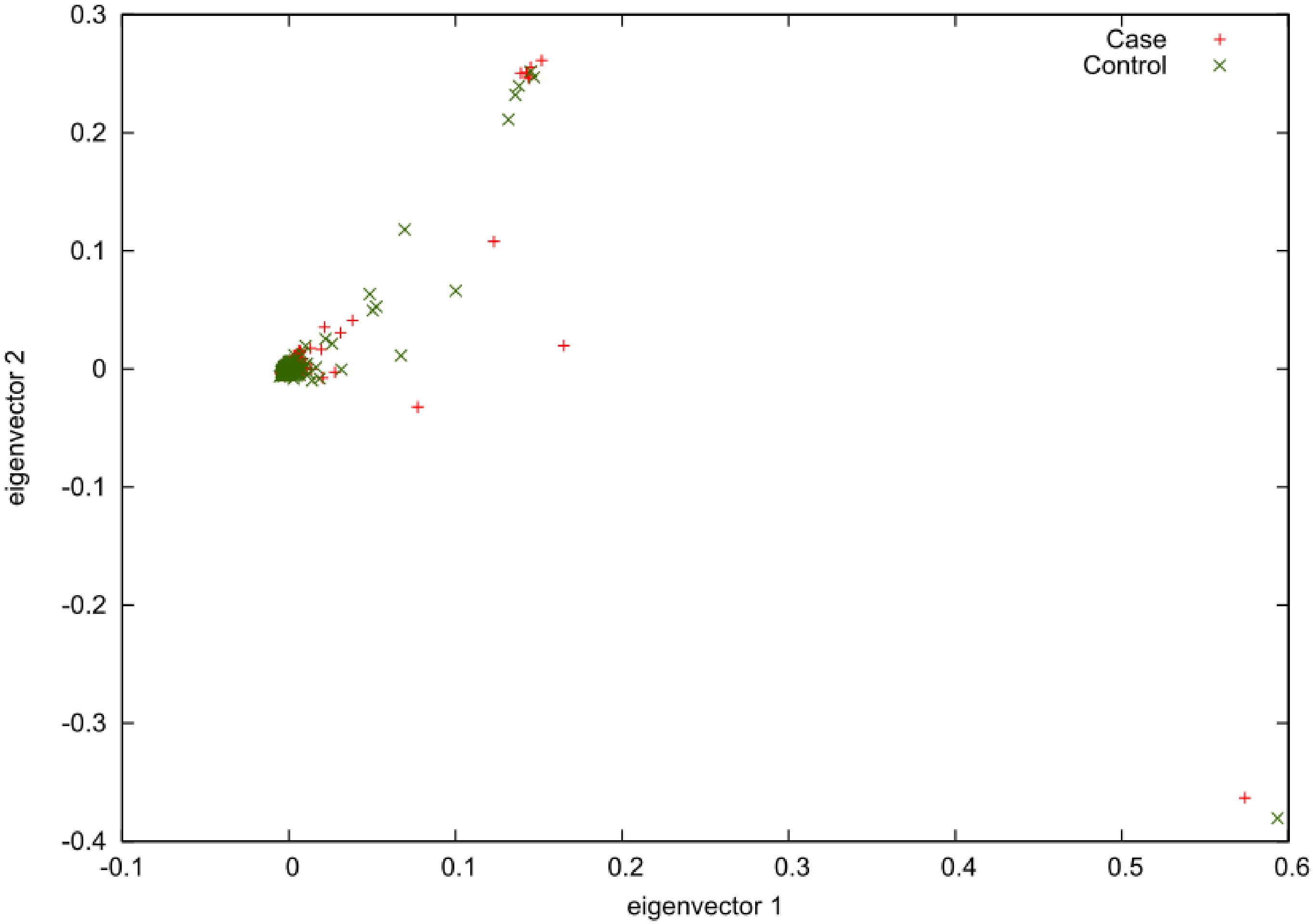
k=4

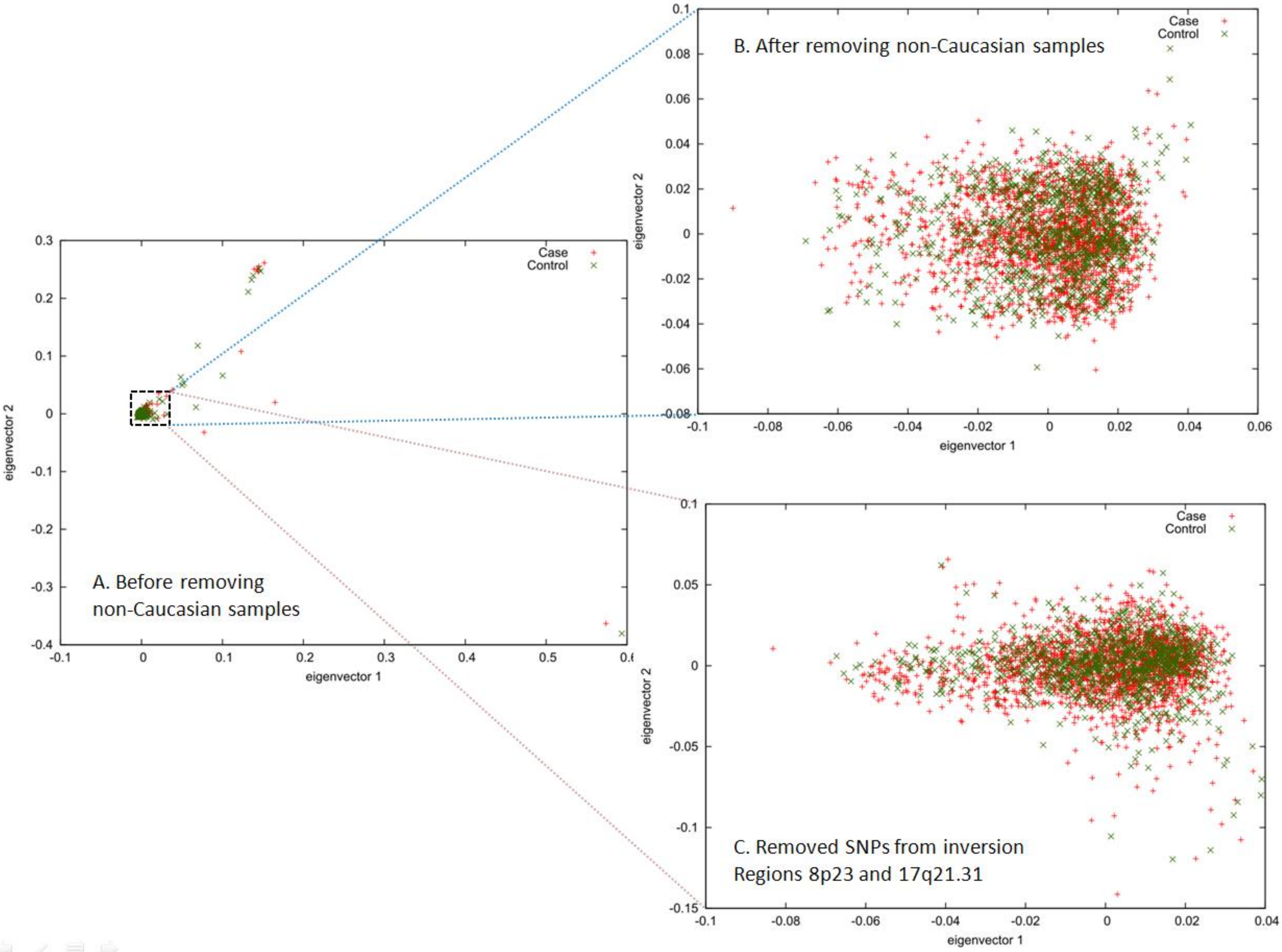


k=5



Population Stratification





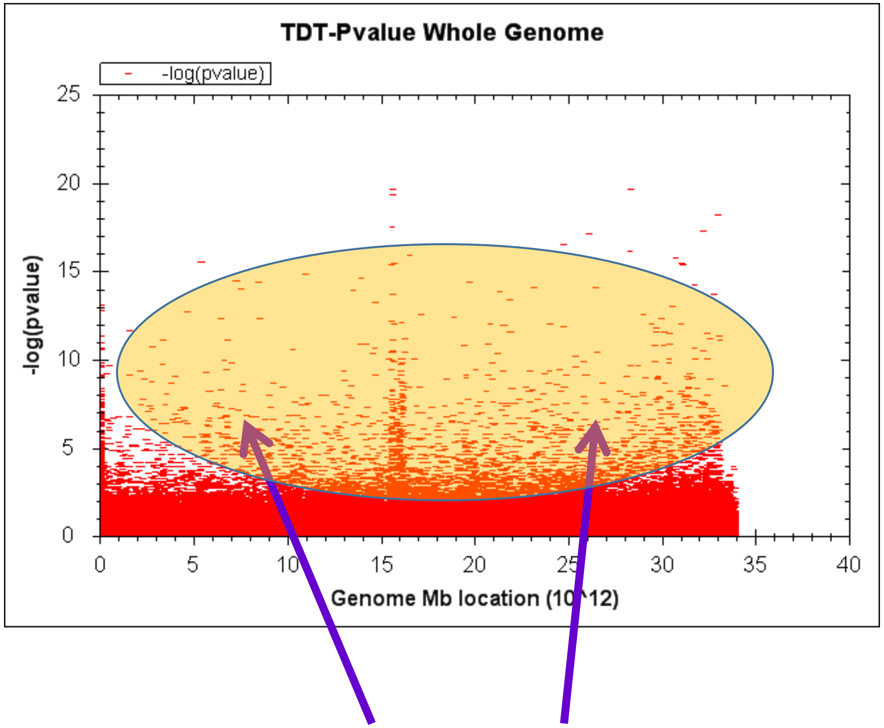
A. Before removing non-Caucasian samples

B. After removing non-Caucasian samples

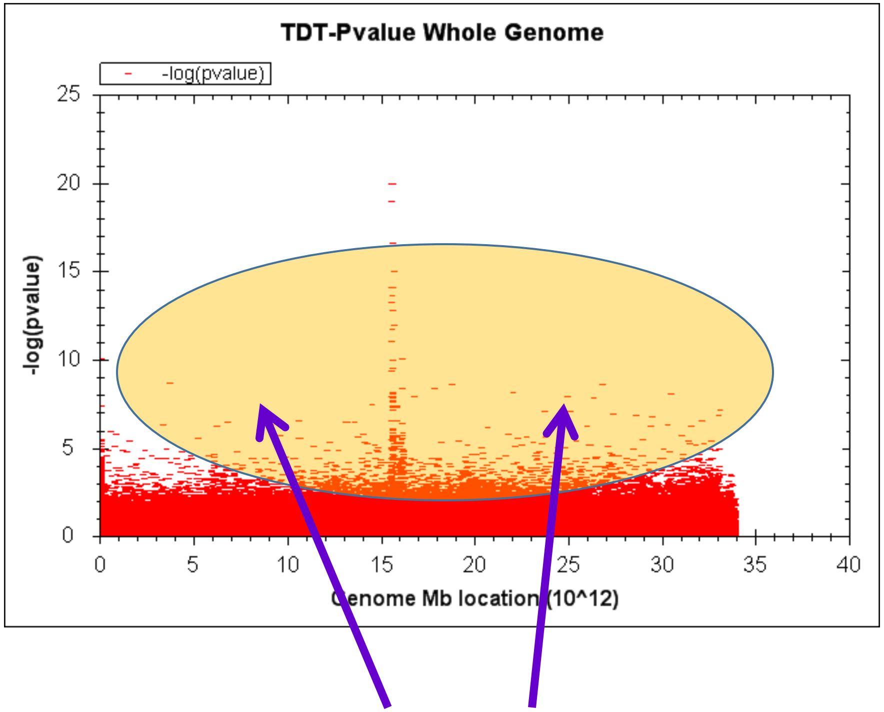
C. Removed SNPs from inversion Regions 8p23 and 17q21.31

Quality Control Analysis

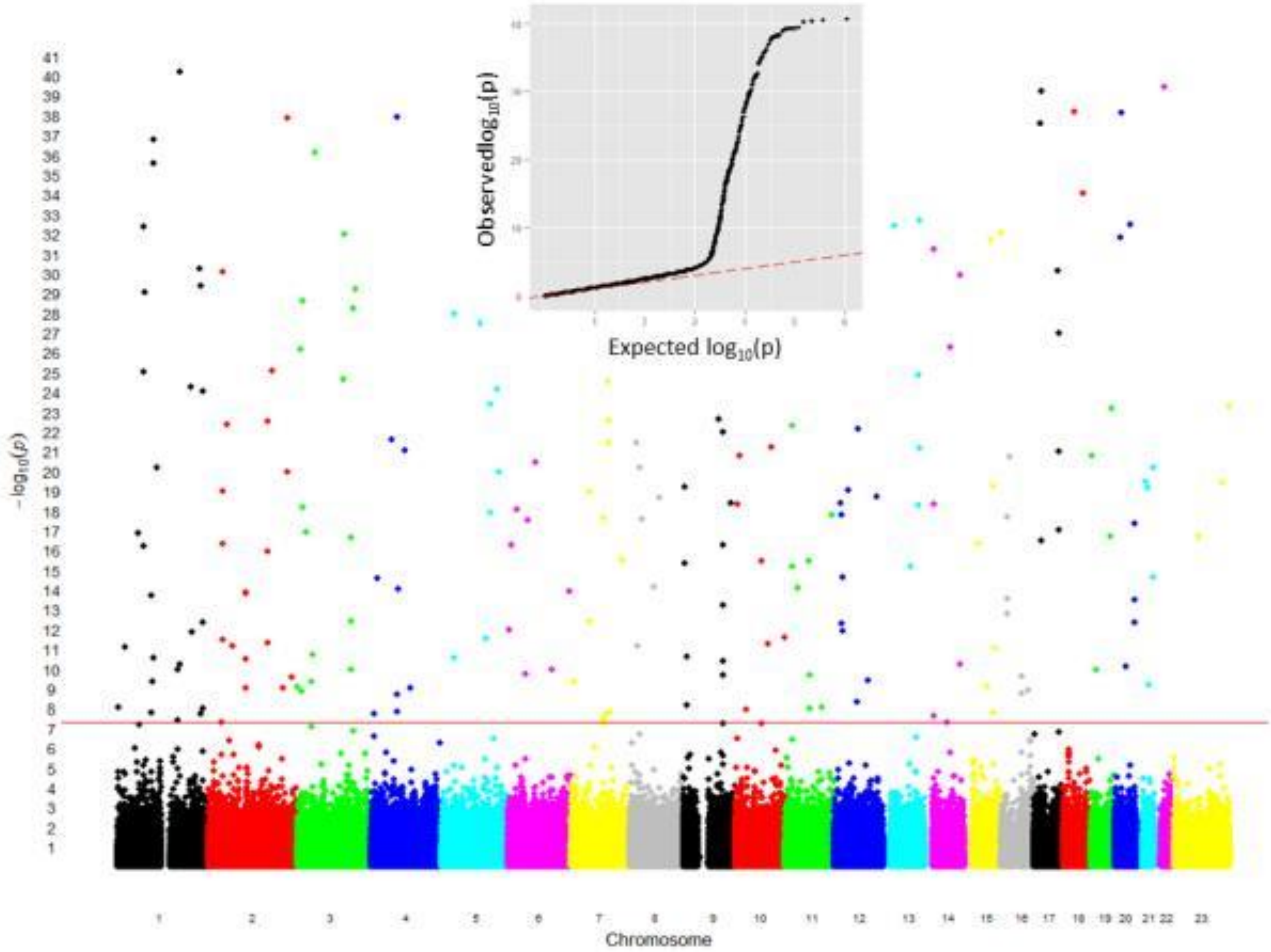
Pre-QC Thresholds



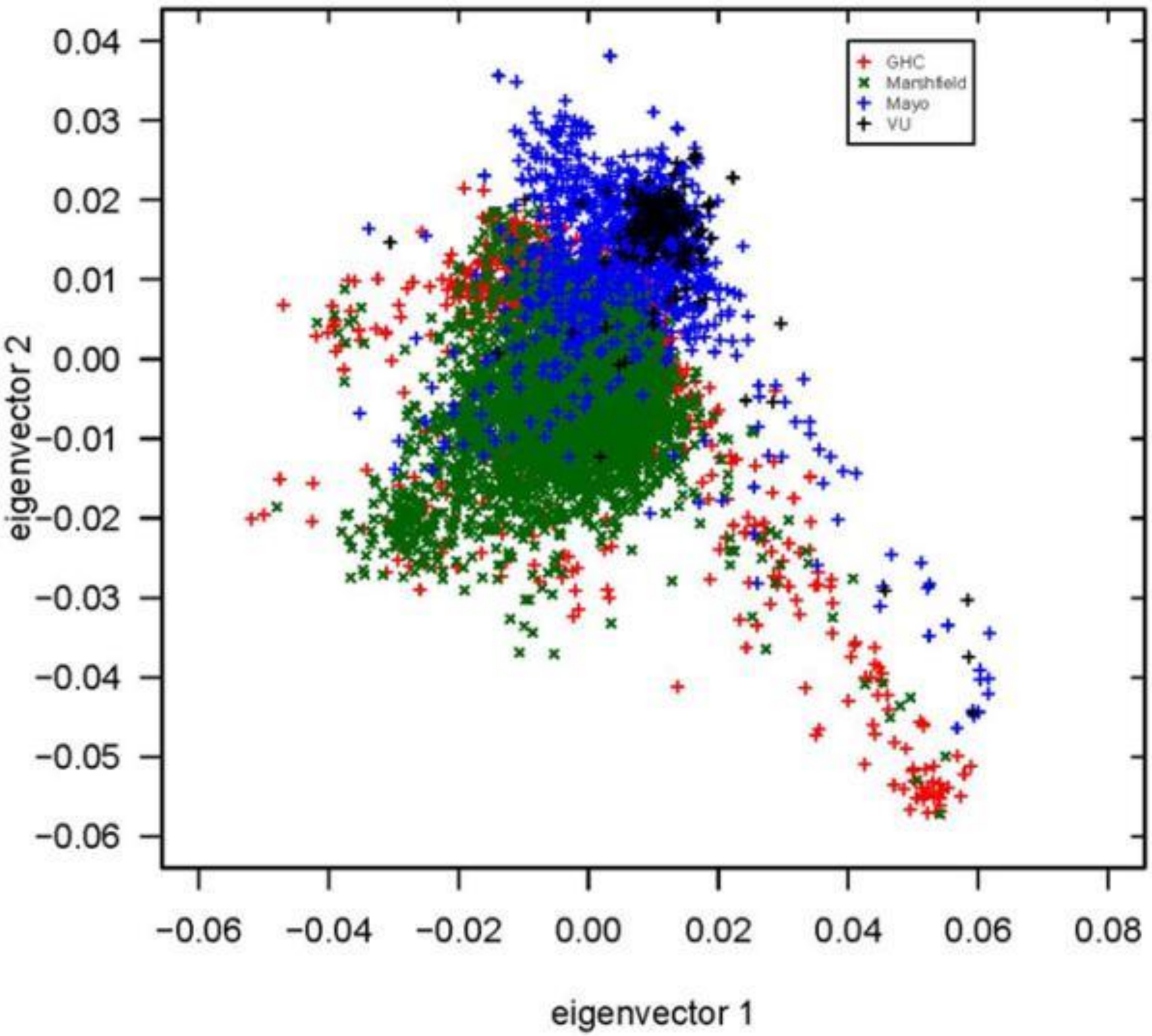
Post-QC Thresholds



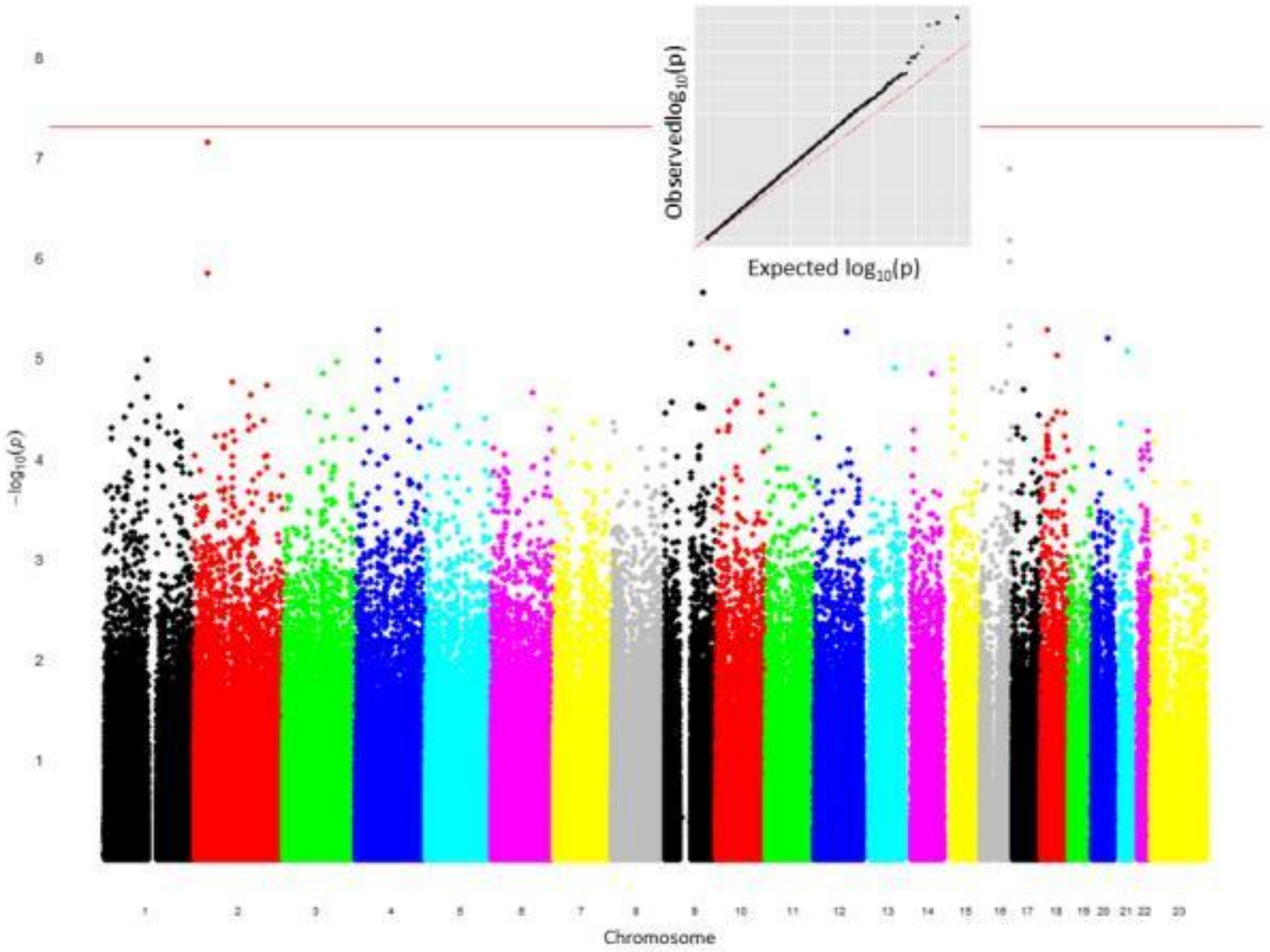
Many false positives disappear after QC



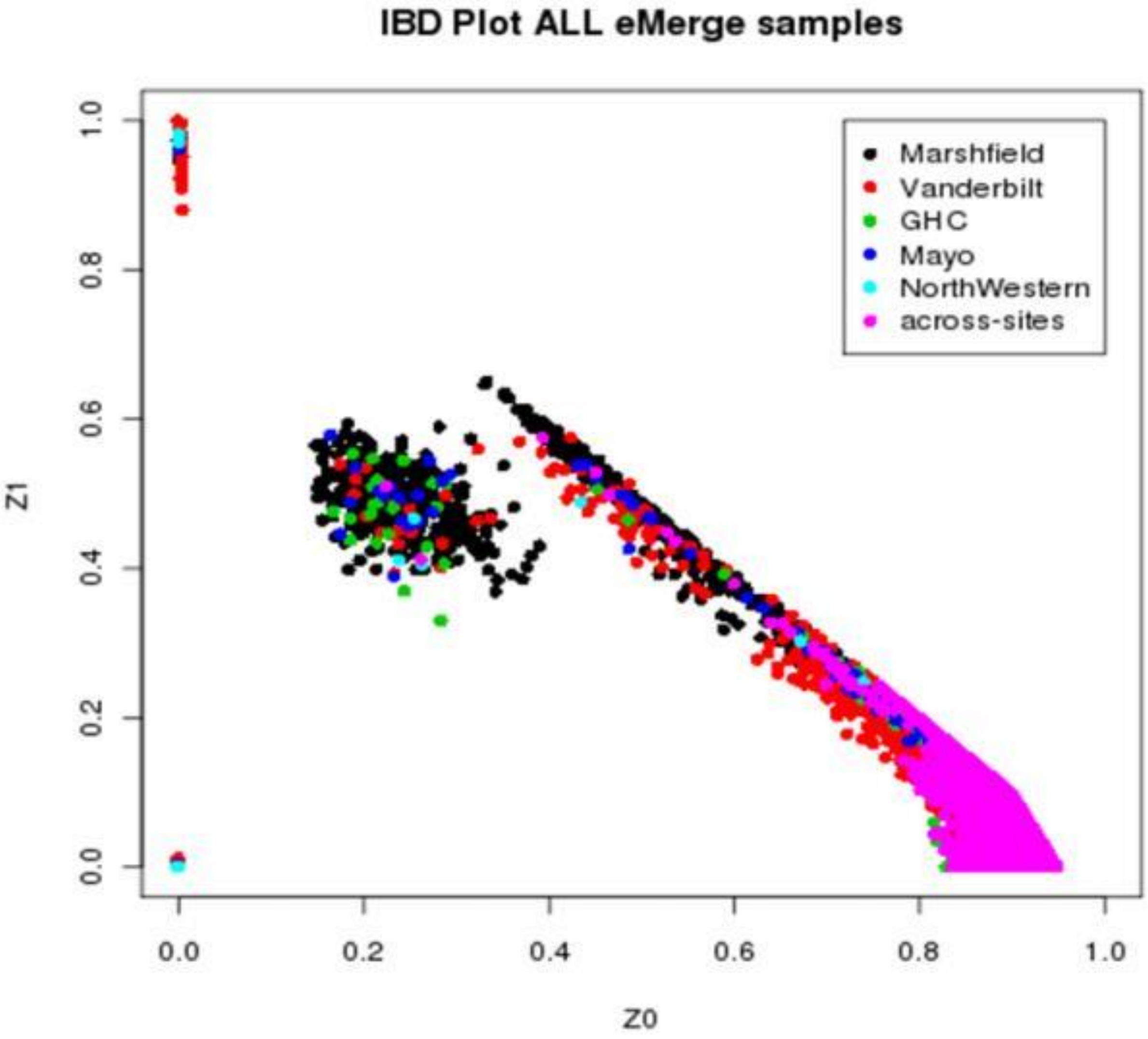
Zuvich et al. Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genet Epidemiol. 2011 December; 35(8): 887–898.



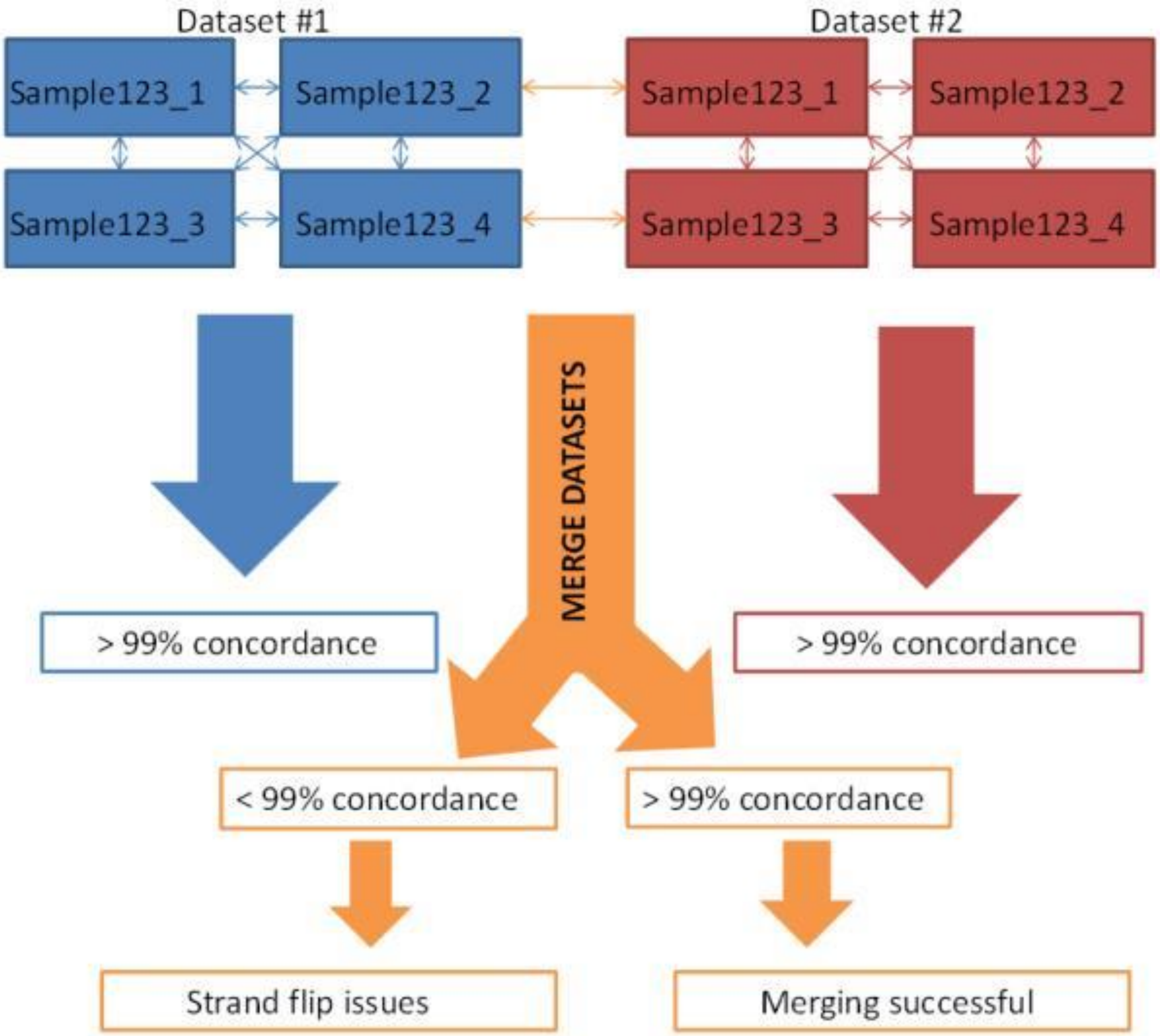
Zuvich et al. Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genet Epidemiol. 2011 December; 35(8): 887–898.



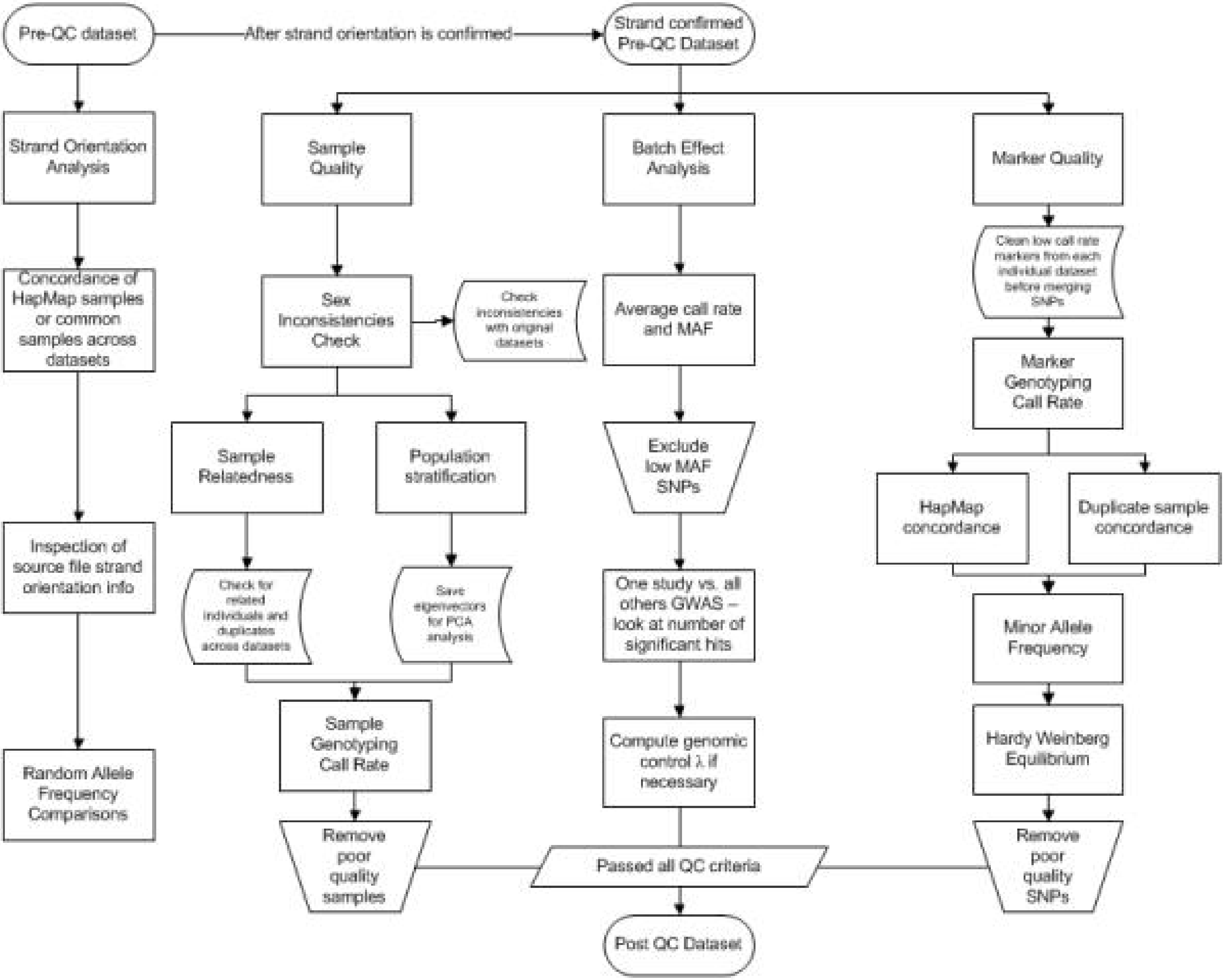
Zuvich et al. Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genet Epidemiol. 2011 December; 35(8): 887–898.



Zuvich et al. Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genet Epidemiol. 2011 December; 35(8): 887–898.

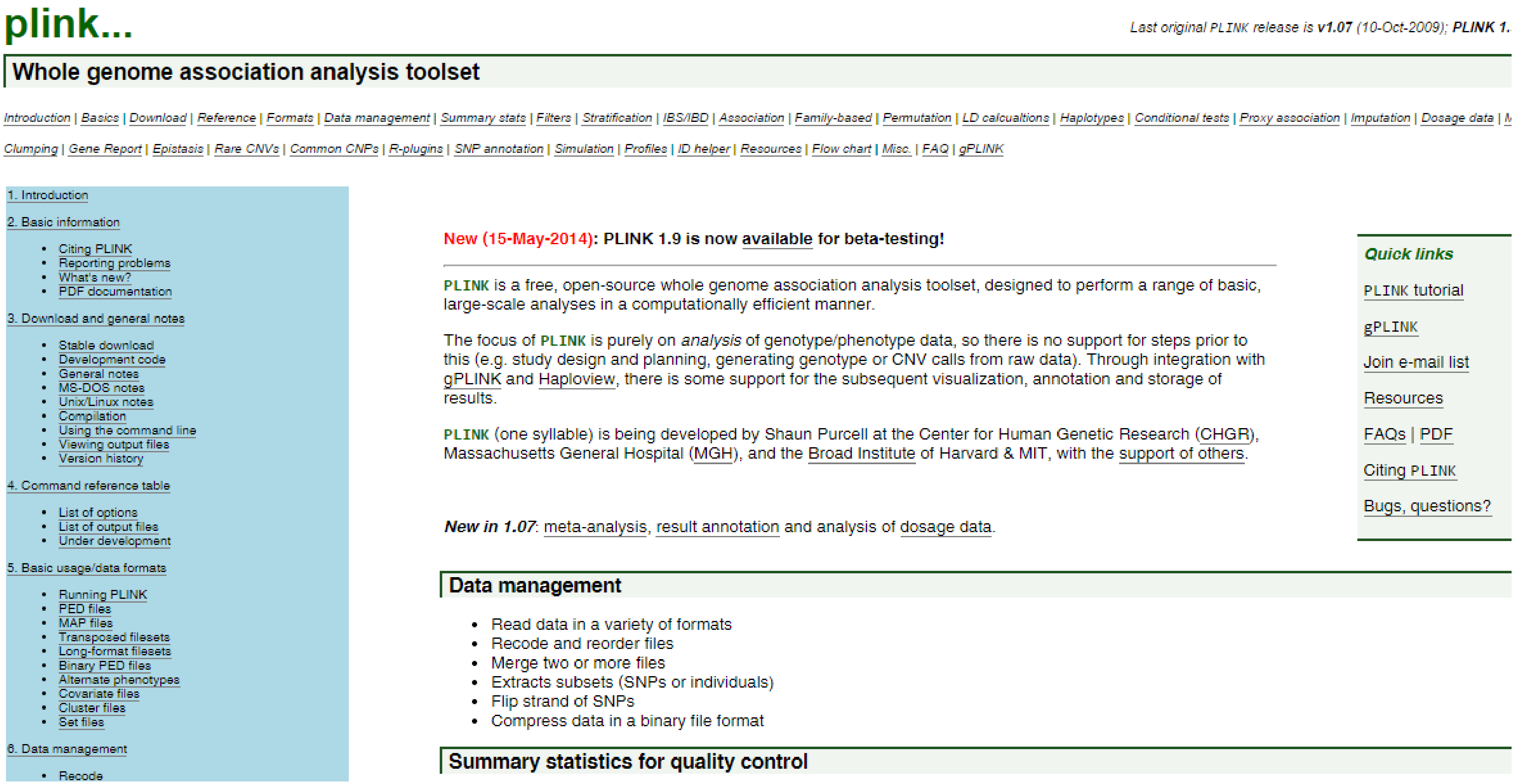


Zuvich et al. Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genet Epidemiol. 2011 December; 35(8): 887–898.


Pre-QC dataset
After strand orientation is confirmed
Strand confirmed Pre-QC Dataset
Strand Orientation Analysis
Concordance of HapMap samples or common samples across datasets
Inspection of source file strand orientation info
Random Allele Frequency Comparisons
Sample Quality
Sex Inconsistencies Check
Check inconsistencies with original datasets
Sample Relatedness
Population stratification
Check for related individuals and duplicates across datasets
Save eigenvectors for PCA analysis
Sample Genotyping Call Rate
Remove poor quality samples
Batch Effect Analysis
Average call rate and MAF
Exclude low MAF SNPs
One study vs. all others GWAS – look at number of significant hits
Compute genomic control λ if necessary
Passed all QC criteria
Post QC Dataset
Marker Quality
Clean low call rate markers from each individual dataset before merging SNPs
Marker Genotyping Call Rate
HapMap concordance
Duplicate sample concordance
Minor Allele Frequency
Hardy Weinberg Equilibrium
Remove poor quality SNPs

Zuvich et al. Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genet Epidemiol. 2011 December; 35(8): 887–898.

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 calcualtions | 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 filesets
- Long-format filesets
- 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.

Quick links

PLINK tutorial

gPLINK

Join e-mail list

Resources

FAQs | PDF

Citing PLINK

Bugs, questions?

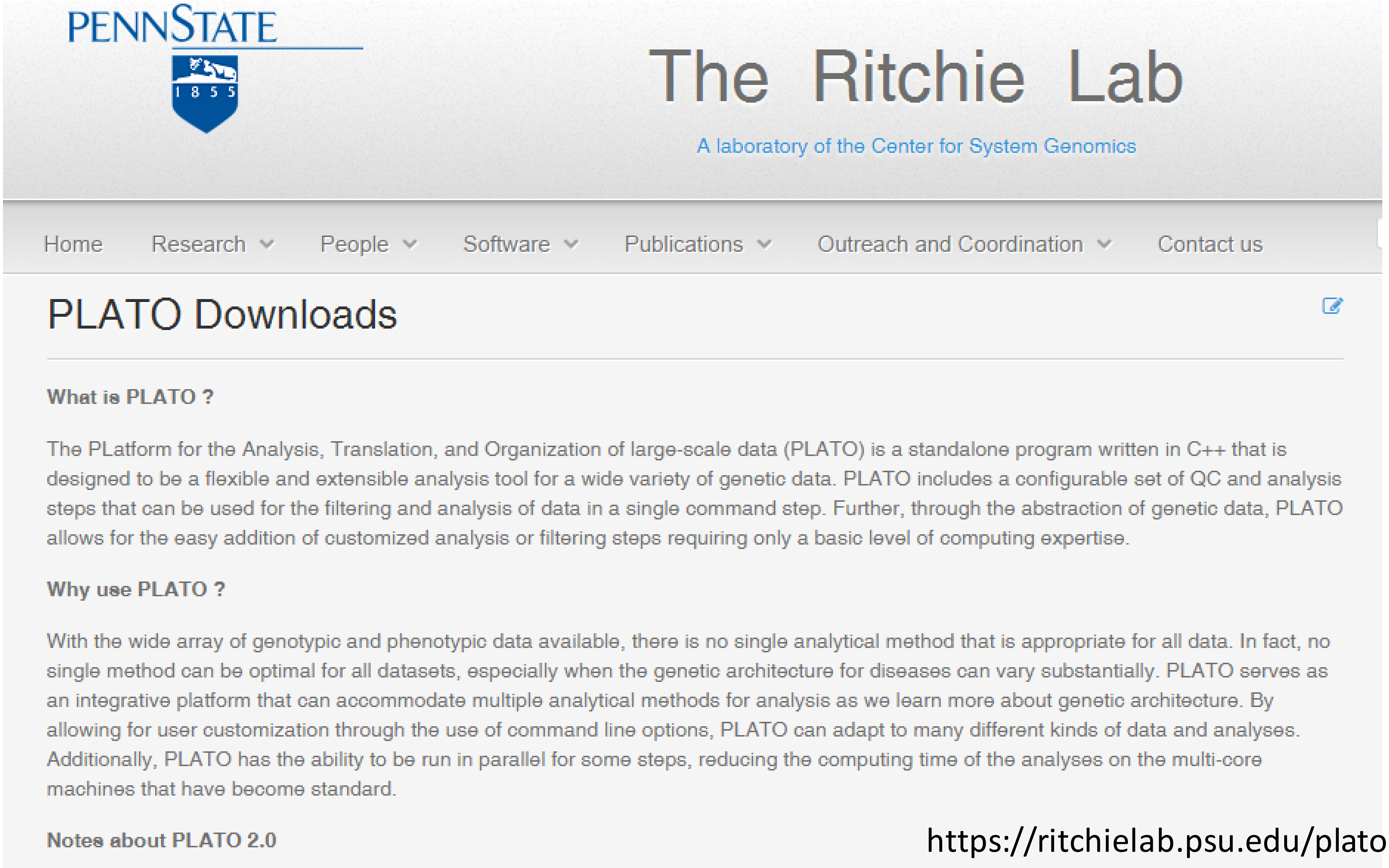
Data management

- 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

http://pngu.mgh.harvard.edu/~purcell/plink/

Software for SNP QC



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.

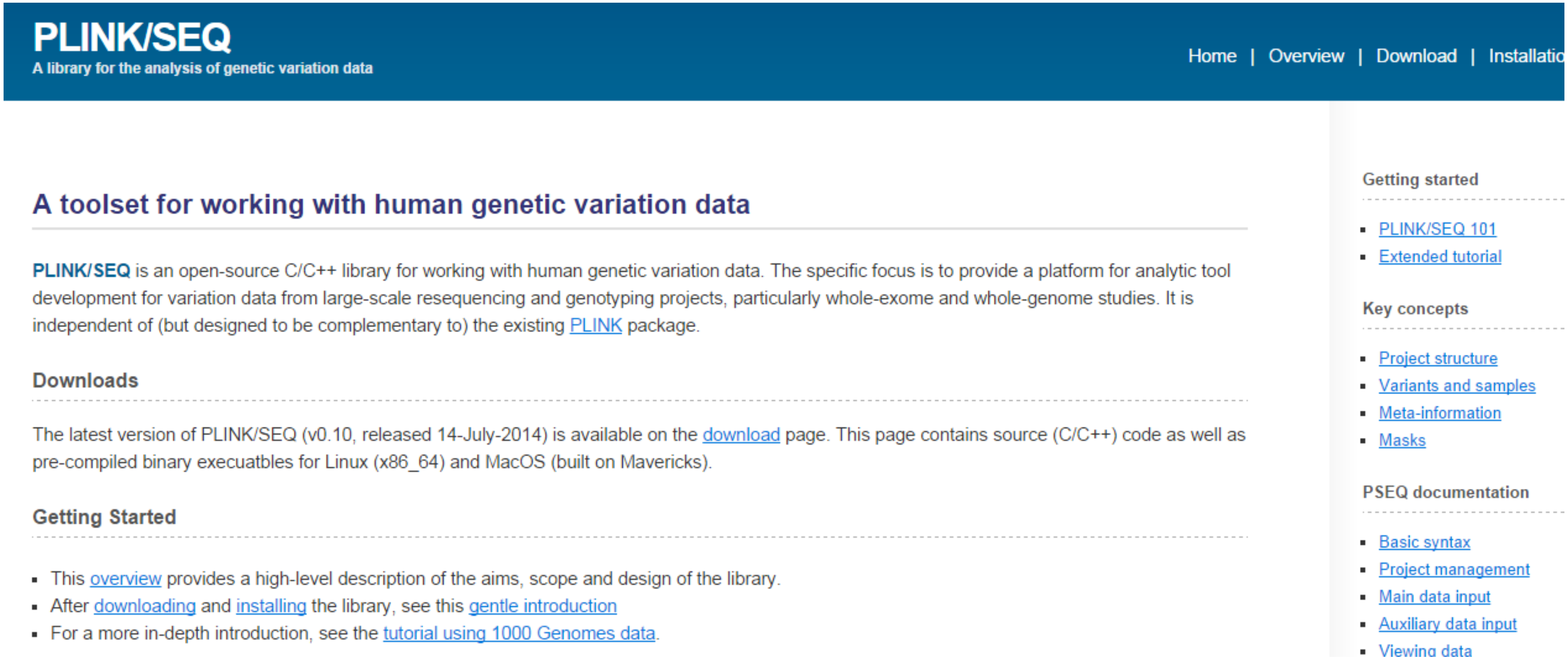
Why use PLATO ?

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 | Installatio

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

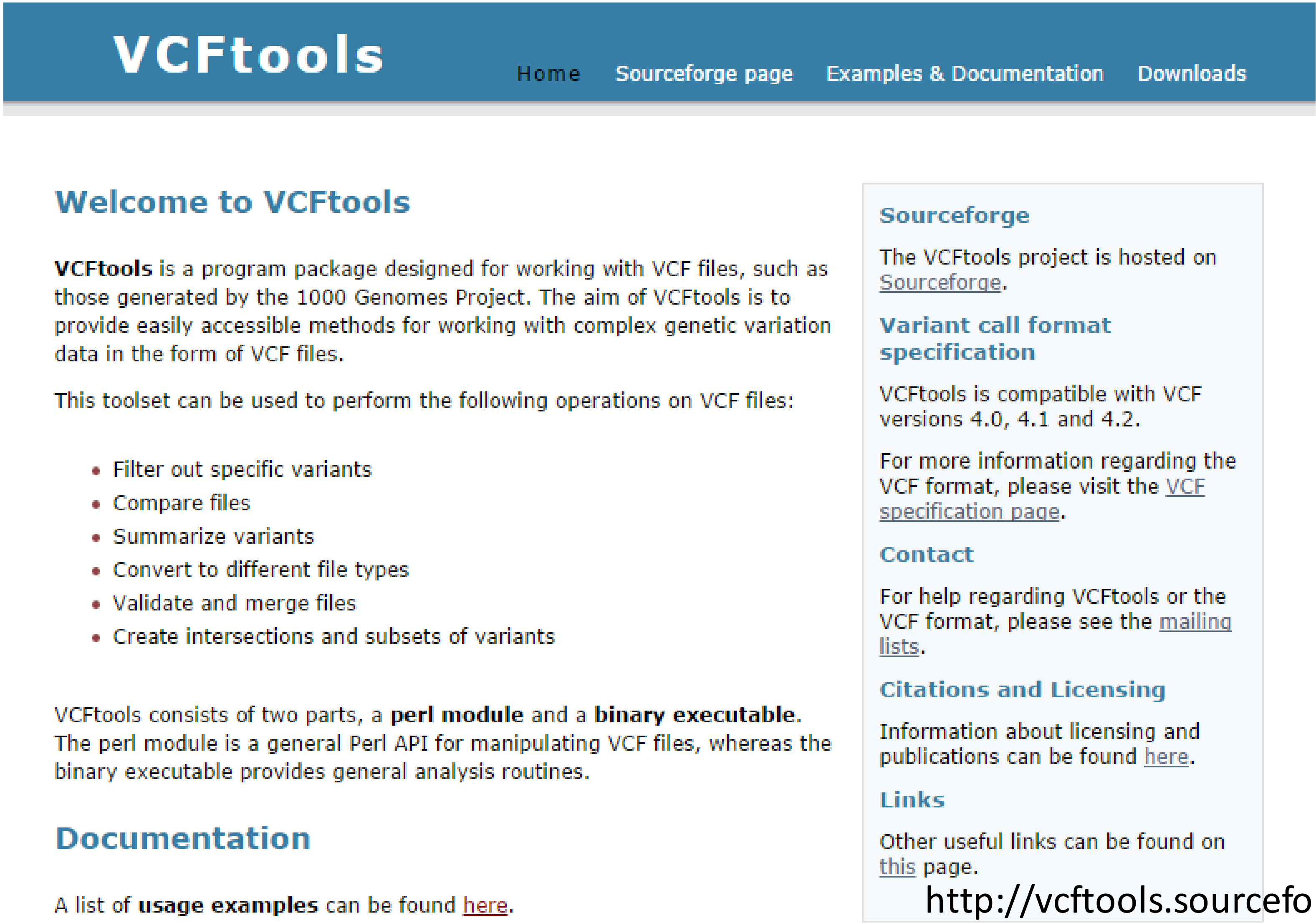
The latest version of PLINK/SEQ (v0.10, released 14-July-2014) is available on the download page. This page contains source (C/C++) code as well as pre-compiled binary execuatbles for Linux (x86_64) and MacOS (built on Mavericks).

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.

https://atgu.mgh.harvard.edu/plinkseq/

Software for Sequence QC



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

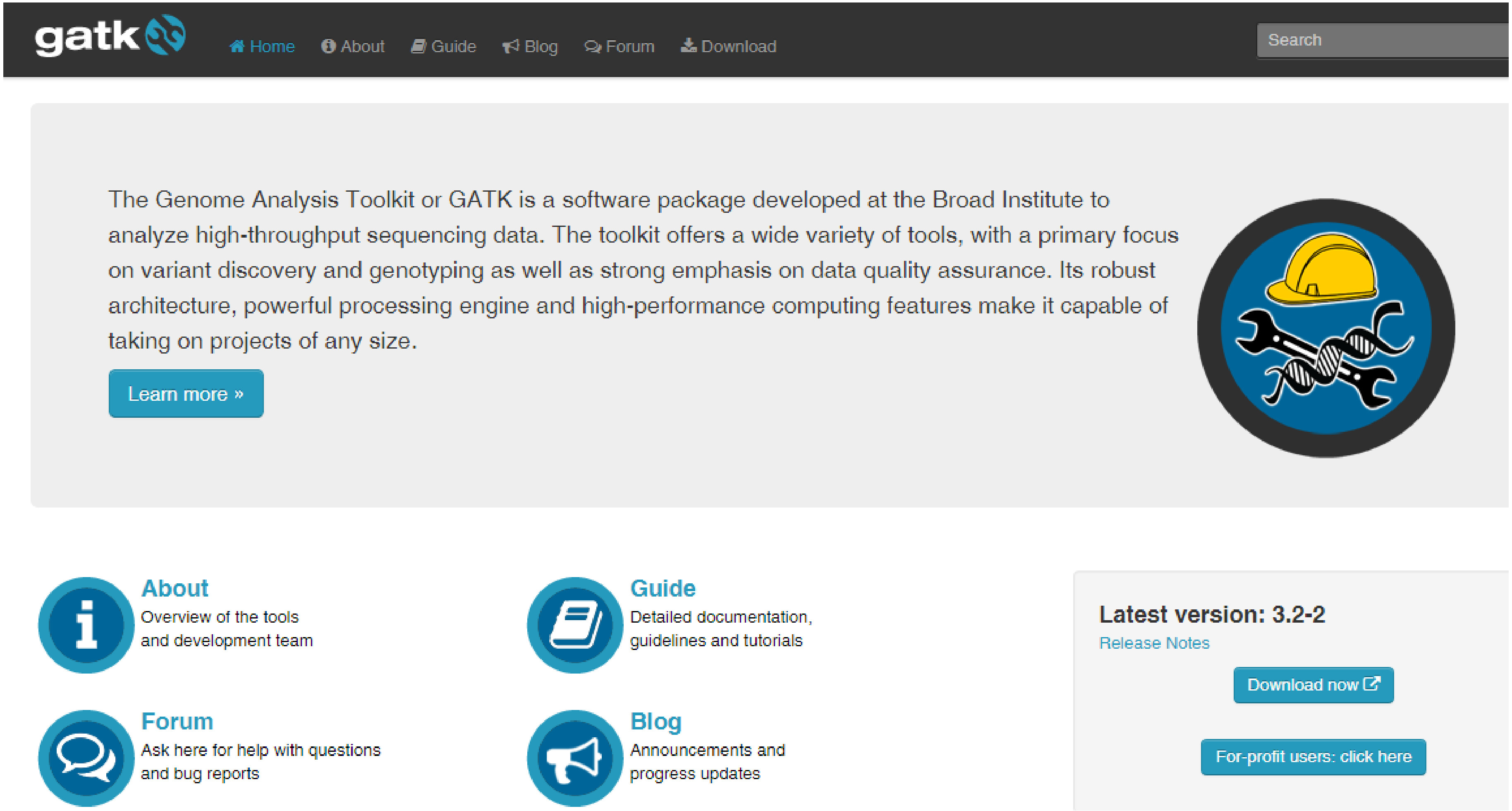
Information about licensing and publications can be found here.

Links

Other useful links can be found on this page.

http://vcftools.sourceforge.net/

Software for Sequence QC



The Genome Analysis Toolkit or GATK is a software package developed at the Broad Institute to analyze high-throughput sequencing data. The toolkit 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.

About
Overview of the tools and development team

Guide
Detailed documentation, guidelines and tutorials

Forum
Ask here for help with questions and bug reports

Blog
Announcements and progress updates

Latest version: 3.2-2

Release Notes

https://www.broadinstitute.org/gatk/

Questions???