Automated pipeline for whole exome/genome sequencing analysis on Mendelian diseases

Yunfei Guo[1,2], Gholson J. Lyon [3], Kai Wang[1,2,4]
[1] Zilkha Neurogenetic Institute, [2] Department of Preventive Medicine, [4] Department of Psychiatry, Keck School of Medicine, University of Southern California, Los Angeles, CA ; [3] Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY

USC University of Southern California

Summary

The development of high-throughput sequencing technologies has dramatically changed the landscape of human genetics research on Mendelian diseases. Currently, there are approximately 3,659 known or suspected Mendelian diseases whose genetic basis remain elusive, based on the Online Mendelian Inheritance in Man (OMIM) database. Identifying causal genes and variants for these diseases will greatly improve disease diagnosis and facilitate the development of therapeutic strategies.

Compared to complex diseases with multiple causal genes and incomplete penetrance, Mendelian diseases may be easier to interrogate by sequencing a few cases and family members. Here we present a computational pipeline called **SeqMule** that performs a series of automated steps to help biologists and clinicians with limited bioinformatics skills identify candidate causal gene for Mendelian diseases. This pipeline perform QC assessments of raw sequencing reads, map single-end or paired-end reads to user-specified reference genome, remove duplicates where necessary, perform local realignments and base quality recalibration, generate SNPs and indels calls by multiple popular algorithms, then compile a set of consensus calls to improve calling accuracy while maintaining high sensitivity. Subsequently, we use a set of filtering and annotation procedures implemented in the ANNOVAR[1] package, including variant function, gene function, allele frequency, conservation scores, functional prediction scores, family information, combined information on multiple unrelated cases, gene-gene networks, phenotype similarity to known Mendelian diseases, to reduce the search space to a small ranked list of candidate genes.

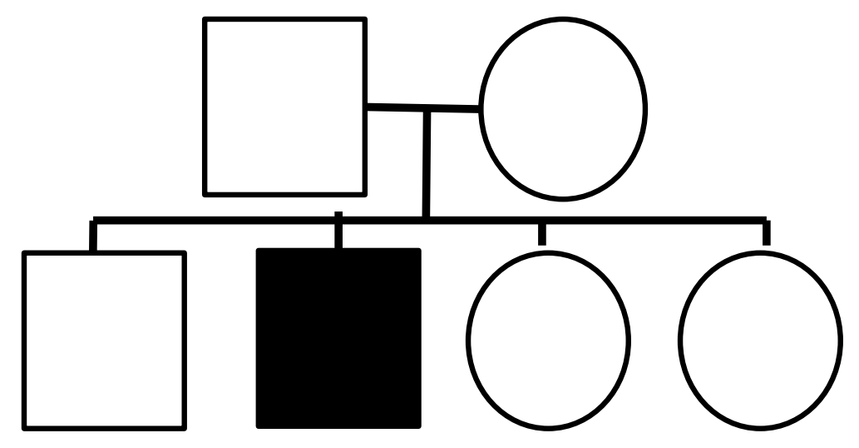
We tested the pipeline on whole-exome sequencing data sets on several Mendelian diseases or Mendelian forms of complex diseases. We demonstrated the efficiency of the analysis pipeline to generate candidate genes with minimum human intervention, and discussed the practical challenges in finding causal genes for Mendelian diseases.

Features

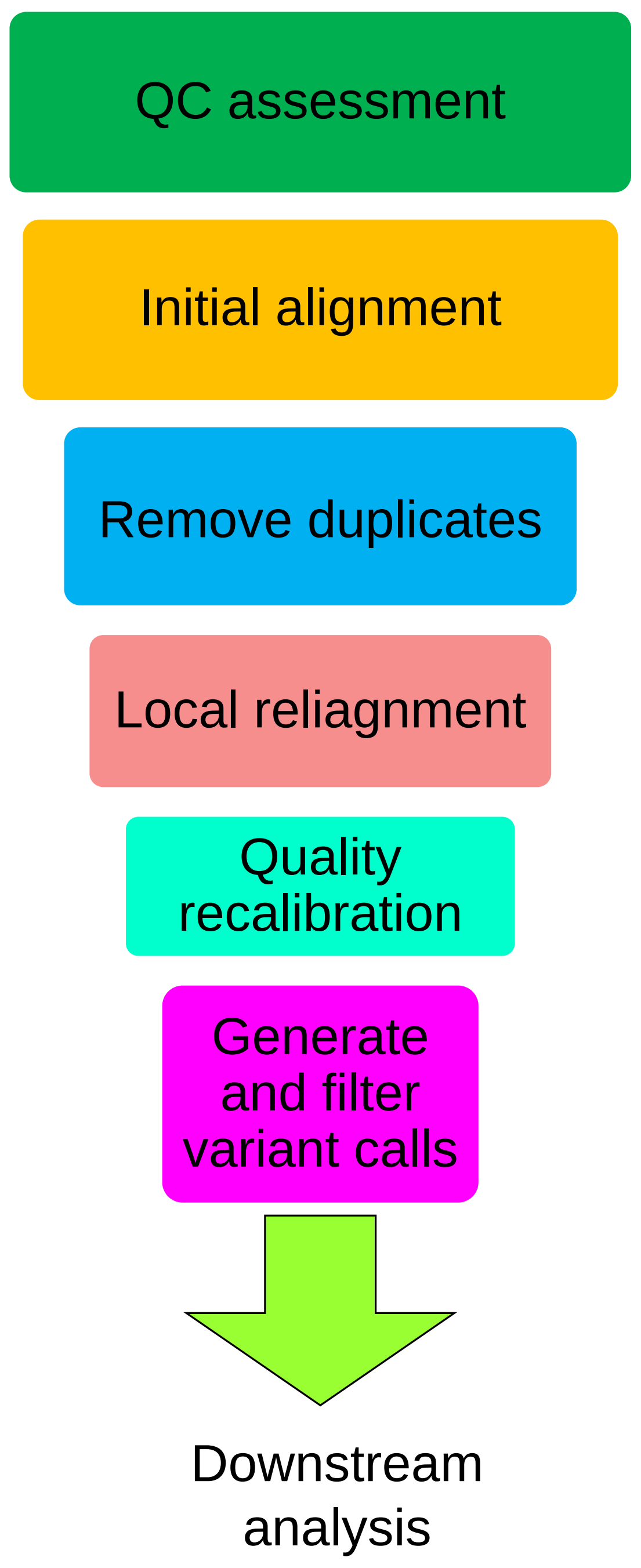
1. Simple commands for installing 10 popular NGS tools, and for downloading multiple databases. Most of the installation procedures in our pipeline was borrowed from the MAKER 2 pipeline developed by Cantarel et al.
2. Ability to start from FASTQ format data or from BAM alignment files
3. Easy customization of various sequence alignment and variant calling tools, and the ability to generate consensus calls from multiple algorithms

Example of real data analysis

Idiopathic hemolytic anemia



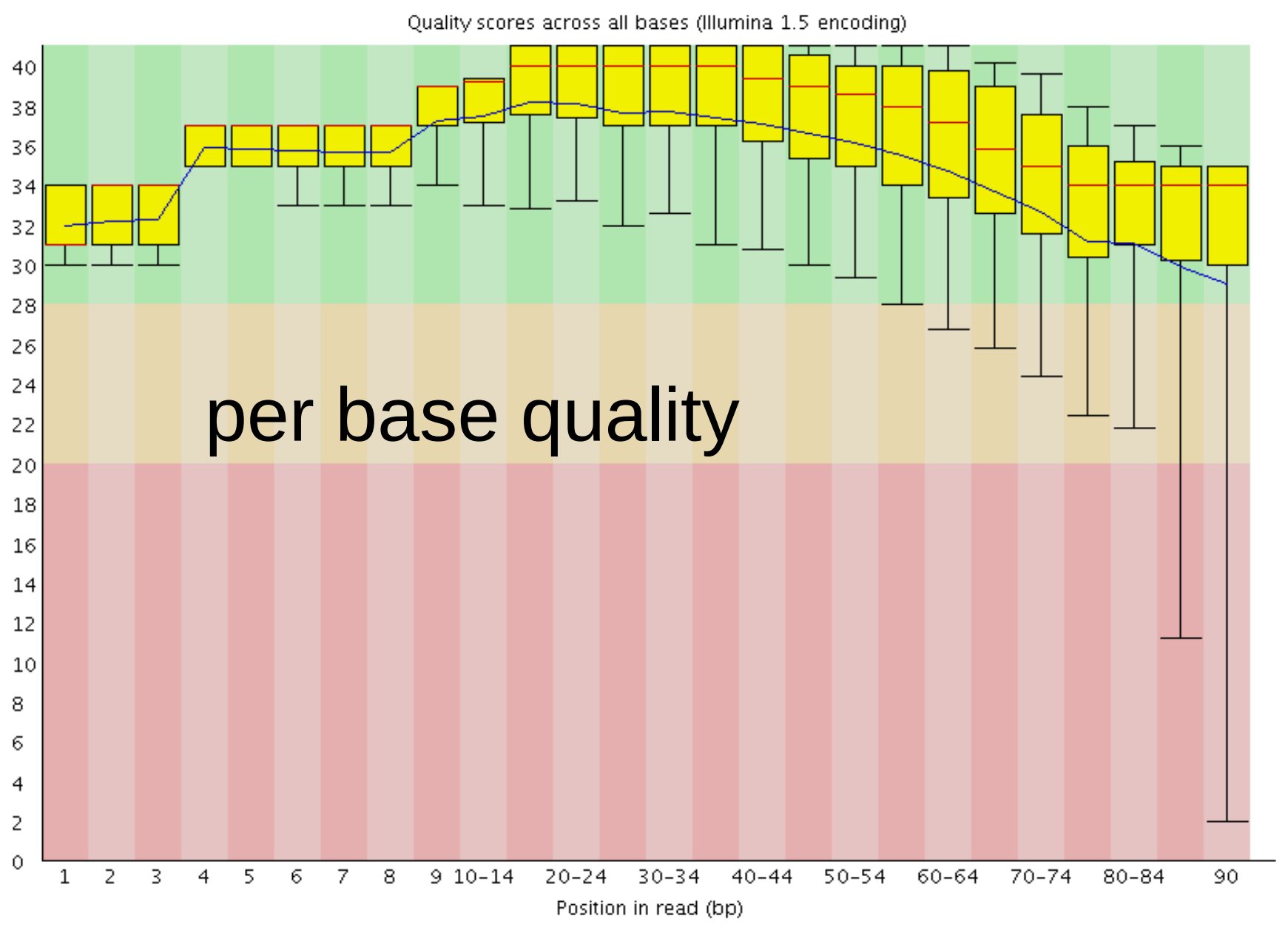
Possible inheritance model
1.X-linked mode of inheritance
2.Autosomal recessive
(compound heterozygous or homozygous)
3.De novo dominant mutation

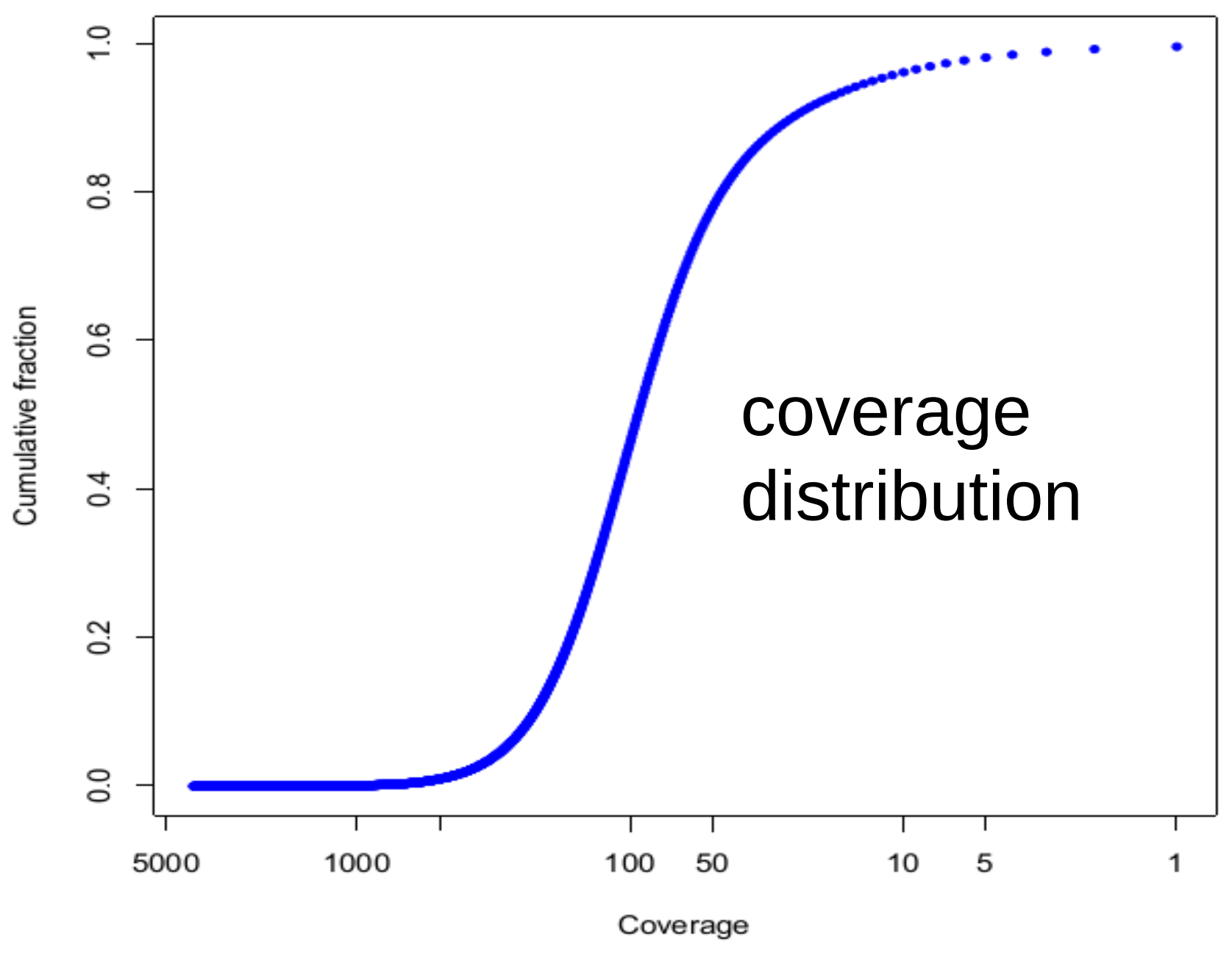


WORKFLOW	TOOLS AVAILABLE
QC assessment	FastQC
Initial alignment	BWA, Bowtie, Bowtie2, SOAPaligner
Remove duplicates	Picard Tools
Local reliagnment	GATK
Quality recalibration	GATK
Generate and filter variant calls	SAMtools, VarScan, GATK, SOAPsnp
Downstream analysis	ANNOVAR

Results panel

Quality Control

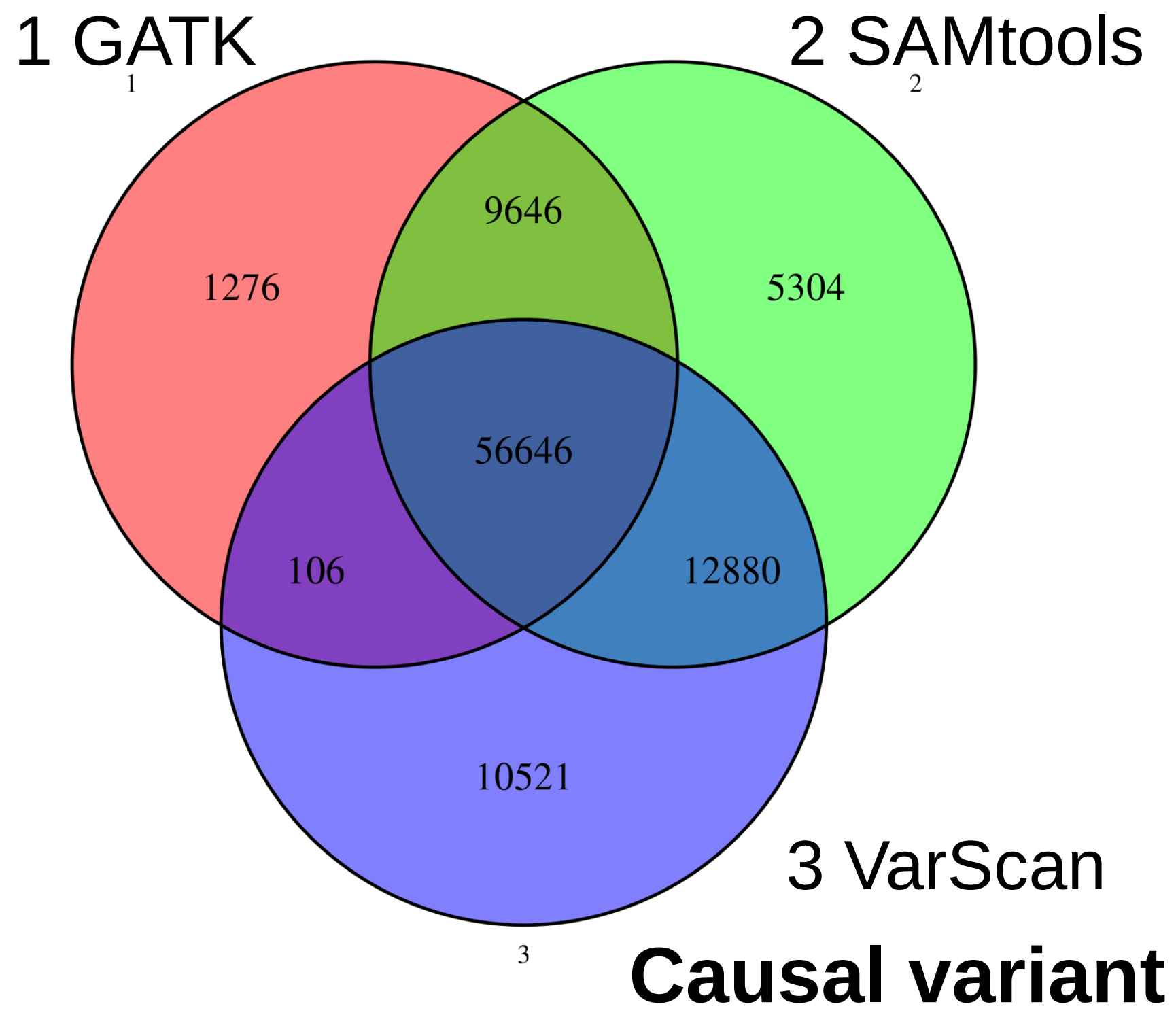




Alignment performance

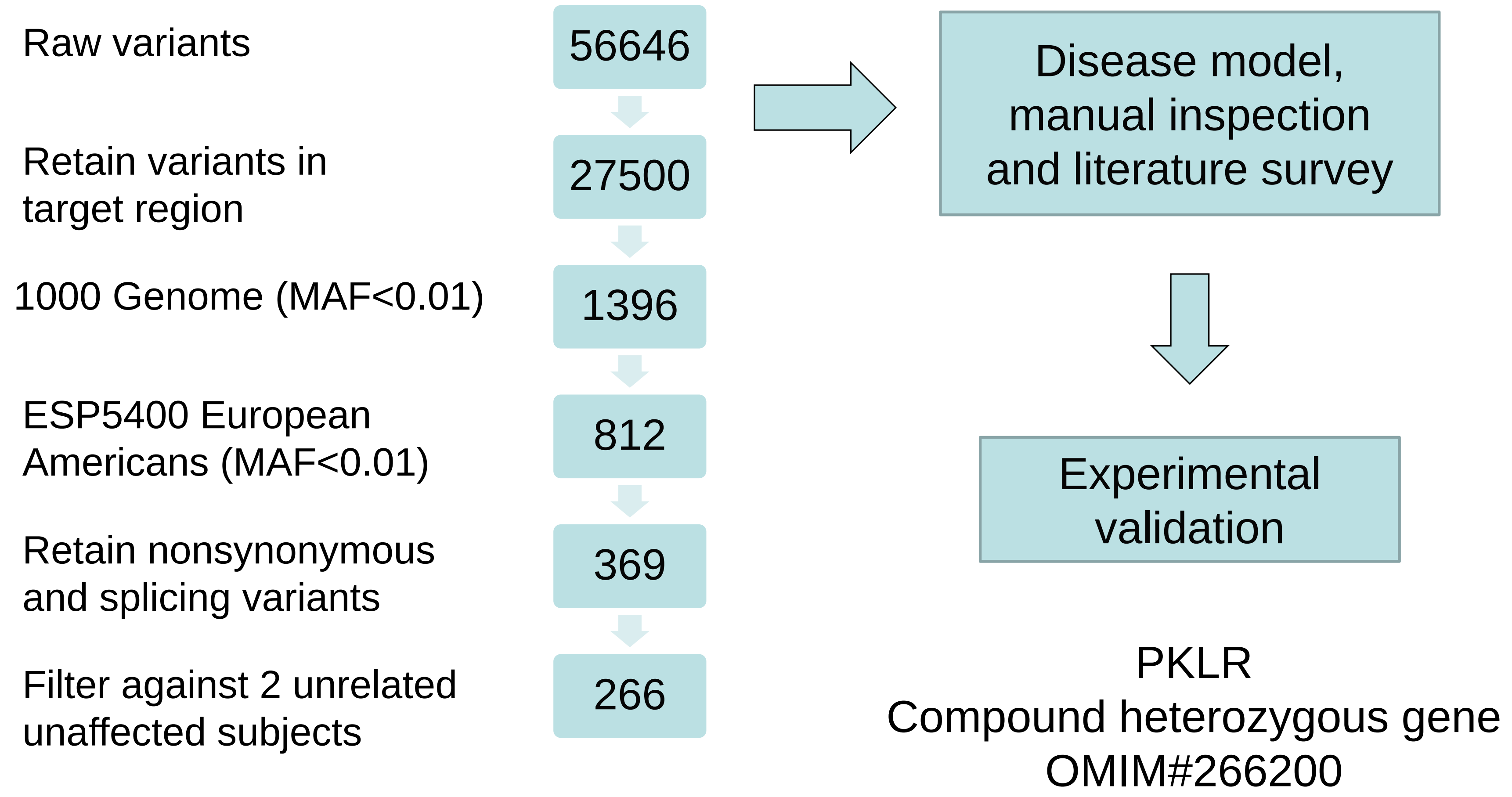
Fraction of reads mapped to target region: **44.3%**
Average coverage on target region:**115x**
Fraction of reads with coverage > 20x: **92.2%**
Fraction of reads with coverage > 10x: **96.1%**

Variant calling



Venn Diagram showing variants overlaping among different callers

Causal variant identification



Raw variants	56646
Retain variants in target region	27500
1000 Genome (MAF<0.01)	1396
ESP5400 European Americans (MAF<0.01)	812
Retain nonsynonymous and splicing variants	369
Filter against 2 unrelated unaffected subjects	266

Disease model, manual inspection and literature survey

Experimental validation

PKLR
Compound heterozygous gene
OMIM#266200

Future directions

1. Currently SeqMule runs analysis script in a non-parallel manner, an it takes only one sample at a time, while some users might want to do multi-sample analysis. These limitations will be addressed in future releases.
2. Additional alignment and variant calling algorithms such as SNVer, GNUMAP, ComB, SOAP3 etc. will be integrated.
3. Web version of SeqMule will be developed which enables a user-friendly graphical interface for data analysis.
4. SeqMule will be extended to generate RNA-Seq or ChIP-Seq data analysis script.

Download (beta version)

System requirement: Linux x86-64
http://bionas.usc.edu/seqmule

References

1. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38:e164, 2010
2. Cantarel, BL et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, 18 188-96, 2008