AN INTRODUCTION TO BIOINFORMATIC PIPELINES FOR NGS

Ryan J. Schmidt, MD, PhD
Assistant Director, Clinical Genomics Laboratory
Center for Personalized Medicine
Department of Pathology and Laboratory Medicine
Children's Hospital Los Angeles
Assistant Professor of Clinical Pathology
Keck School of Medicine of USC

Pathology

biospecimen

↓

data

⇣

clinically useful knowledge

Molecular Diagnostics

biospecimen = nucleic acid typically extracted from cells or tissue

data = raw high-throughput sequencing reads

Bioinformatics

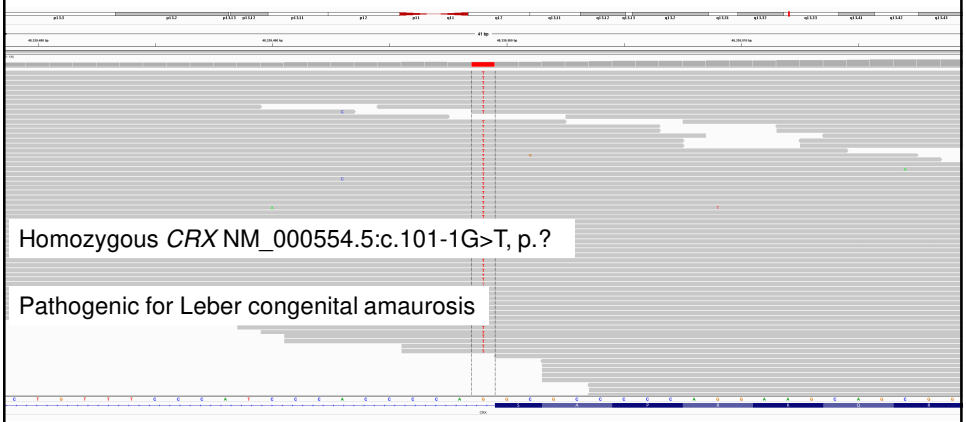
clinically useful knowledge = genetic variant(s) with interpreted clinical utility

Bioinformatics

"Bioinformatics is the discipline that conceptualizes biology in terms of macromolecules and then applies informatics techniques (applied math, computer science, and statistics) to understand and organize the information associated with these molecules, on a large scale."

Roy S 2018 *Journal of Molecular Diagnostics*

What is a variant?



- A difference from the reference
- The reference genome is a ~~lie~~ shortcut

Bioinformatic Pipelines

data = raw high-throughput sequencing or probing results

Primary

Secondary

Tertiary

clinically useful knowledge = genetic variant with interpreted clinical utility

Children's Hospital LOS ANGELES | USC University of Southern California

Workflow Management Software Tools

COMMON WORKFLOW LANGUAGE

docker

Workflow Language

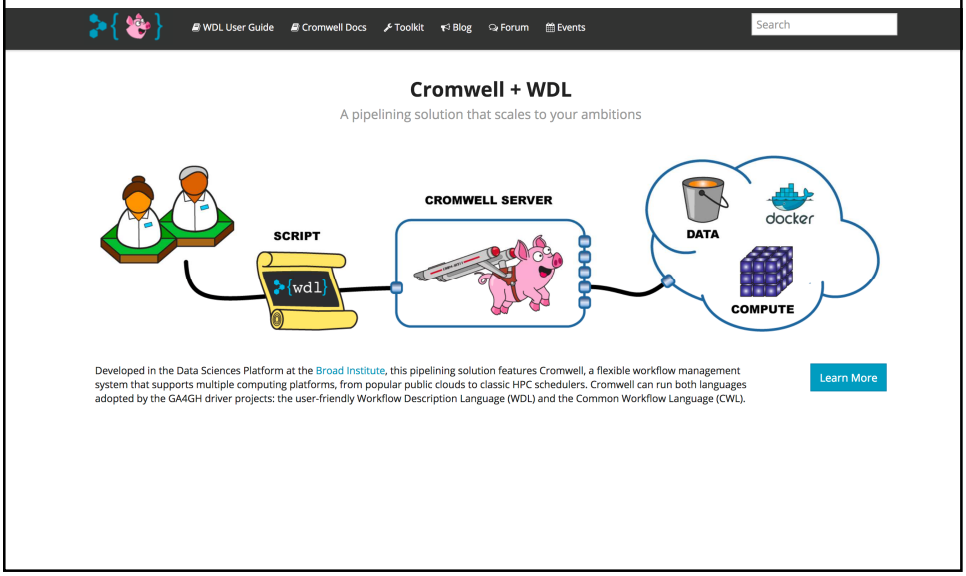
- Describes the workflow and the requirements for each step

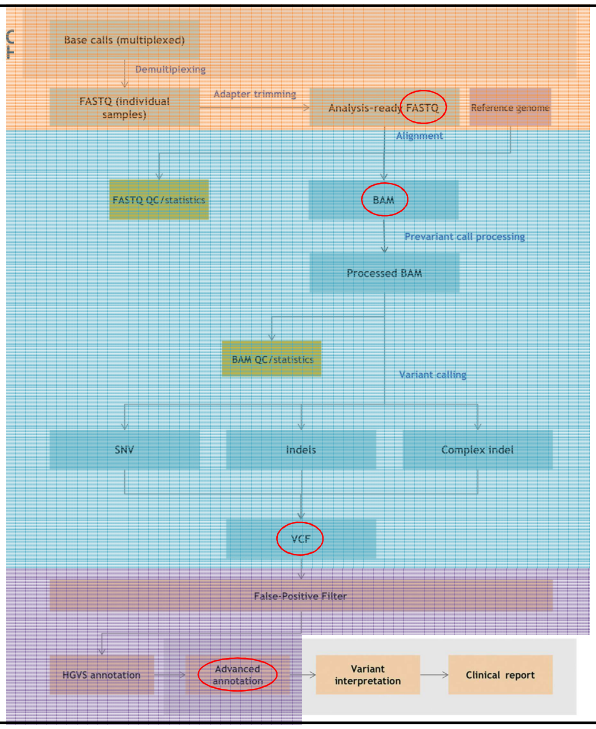
Containers

- Each step is run in a separate predictable environment
- Portable
  - Develop locally then deploy

Children's Hospital LOS ANGELES | USC University of Southern California

Workflow Management Software Tools





Primary

Secondary

Tertiary

Human Interpretation (semi-automated)

Figure 1 Next-generation sequencing (NGS) bioinformatics pipeline. The figure illustrates a bioinformatics pipeline and its components that are typically used for processing NGS data. The illustration may not represent nuances or additional algorithms that are specific to sequencing platforms. The components of the pipeline that overlap with the **gray shaded region** are out of scope of this guideline. BAM, binary alignment map; HGVS, Human Genome Variation Society; indel, insertion/deletion; QC, quality control; SNV, single-nucleotide variant; VCF, variant call format.

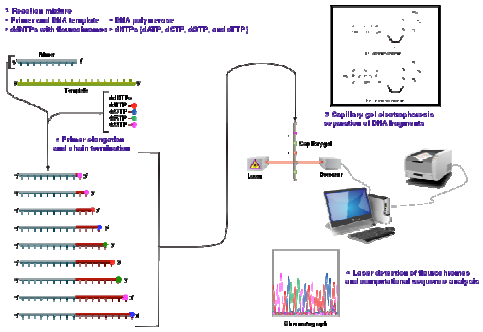
Roy S 2018 *JMD*

Children's Hospital LOS ANGELES | USC University of Southern California

Primary Analysis

- Converts detection signal to "raw" data for the bioinformatic pipeline
- Primarily occurs within the analyzer but we should be aware of it

Children's Hospital LOS ANGELES | USC University of Southern California



By Estevezj - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=23264166

Children's Hospital LOS ANGELES | USC University of Southern California

Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA; [2]Genome Sequencing Center, Washington University School of Medicine, Saint Louis, Missouri 63108 USA

The availability of massive amounts of DNA sequence information has begun to revolutionize the practice of biology. As a result, current large-scale sequencing output, while impressive, is not adequate to keep pace with growing demand and, in particular, is far short of what will be required to obtain the 3-billion-base human genome sequence by the target date of 2005. To reach this goal, improved automation will be essential, and it is particularly important that human involvement in sequence data processing be significantly reduced or eliminated. Progress in this respect will require both improved accuracy of the data processing software and reliable accuracy measures to reduce the need for human involvement in error correction and make human review more efficient. Here, we describe one step toward that goal: a base-calling program for automated sequencer traces, *phred*, with improved accuracy. *phred* appears to be the first base-calling program to achieve a lower error rate than the ABI software, averaging 40%–50% fewer errors in the data sets examined independent of position in read, machine running conditions, or sequencing chemistry.

Base-Calling of Automated Sequencer Traces Using *Phred*. II. Error Probabilities

Brent Ewing and Phil Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

Elimination of the data processing bottleneck in high-throughput sequencing will require both improved accuracy of data processing software and reliable measures of that accuracy. We have developed and implemented in our base-calling program *phred* the ability to estimate a probability of error for each base-call, as a function of certain parameters computed from the trace data. These error probabilities are shown here to be valid (correspond to actual error rates) and to have high power to discriminate correct base-calls from incorrect ones, for read data collected under several different chemistries and electrophoretic conditions. They play a critical role in our assembly program *phrap* and our finishing program *consed*.

Ewing B 1998 *Genome Biology*

Phred-scaled Quality Scores

An important technical aspect of our work is the use of log-transformed error probabilities rather than untransformed ones, which facilitates working with error rates in the range of most importance (very close to 0). Specifically, we define the quality value q assigned to a base-call to be

$$q = -10 \times \log_{10}(p)$$

where p is the estimated error probability for that base-call. Thus a base-call having a probability of 1/1000 of being incorrect is assigned a quality value of 30. Note that high quality values correspond to low error probabilities, and conversely.

FASTQ files

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

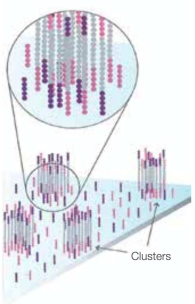
Base Call
+
Quality

Next-generation Sequencing

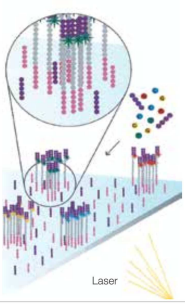
NGS = massively parallel short read sequencing-by-synthesis

Fluorescence (Illumina)	Ion Semiconductor (Ion Torrent)

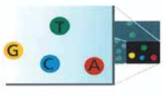
Illumina



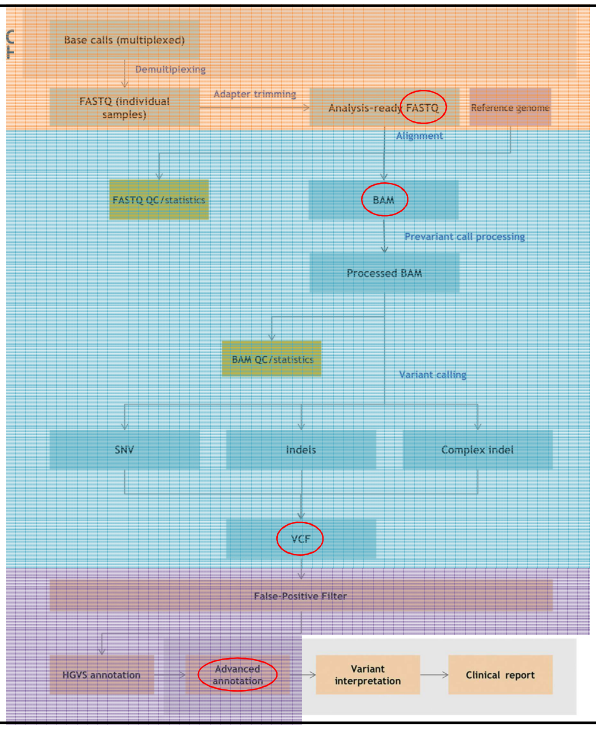
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.



FASTQ



Primary

Secondary

Figure 1 Next-generation sequencing (NGS) bioinformatics pipeline. The figure illustrates a bioinformatics pipeline and its components that are typically used for processing NGS data. The illustration may not represent nuances or additional algorithms that are specific to sequencing platforms. The components of the pipeline that overlap with the **gray shaded region** are out of scope of this guideline. BAM, binary alignment map; HGVS, Human Genome Variation Society; indel, insertion/deletion; QC, quality control; SNV, single-nucleotide variant; VCF, variant call format.

Tertiary

Human Interpretation (semi-automated)

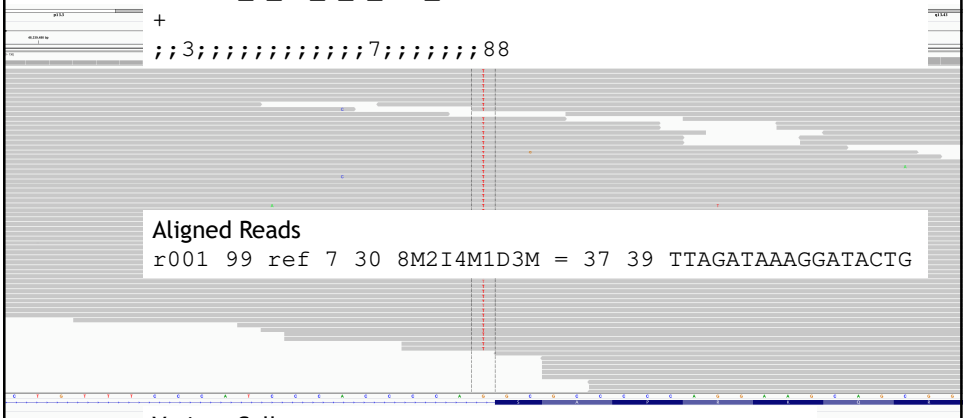
Roy S 2018 *JMD*

Children's Hospital LOS ANGELES | USC University of Southern California

Secondary Analysis

"Raw" Sequences

```
@EAS54_6_R1_2_1_413_324 CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```



Aligned Reads

```
r001  99  ref  7  30  8M2I4M1D3M  =  37  39  TTAGATAAAGGATACTG
```

Variant Call

```
19  48339499  .  G  T
```

Children's Hospital LOS ANGELES | USC University of Southern California

Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

22 May 2018

The master version of this document can be found at `https://github.com/samtools/hts-specs`.
This printing is version b1ae9f9 from that repository, last modified on the date shown above.

1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '`@`', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

This specification is for version 1.6 of the SAM and BAM formats. Each SAM and BAM file may optionally specify the version being used via the `@HD VN` tag. For full version history see Appendix A.

Unless explicitly specified elsewhere, all fields are encoded using 7-bit US-ASCII [1] in using the POSIX / C locale. Regular expressions listed use the POSIX / IEEE Std 1003.1 extended syntax.

Children's Hospital LOS ANGELES | USC University of Southern California

The Variant Call Format (VCF) Version 4.2 Specification

25 Sep 2017

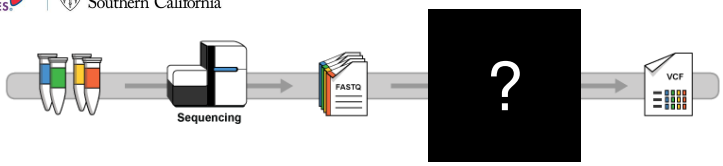
The master version of this document can be found at `https://github.com/samtools/hts-specs`.
This printing is version c8b9990 from that repository, last modified on the date shown above.

1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

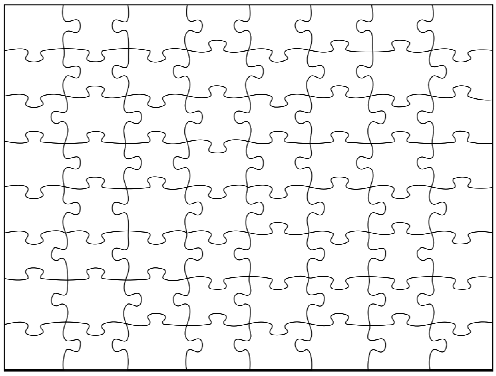
1.1 An example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##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      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

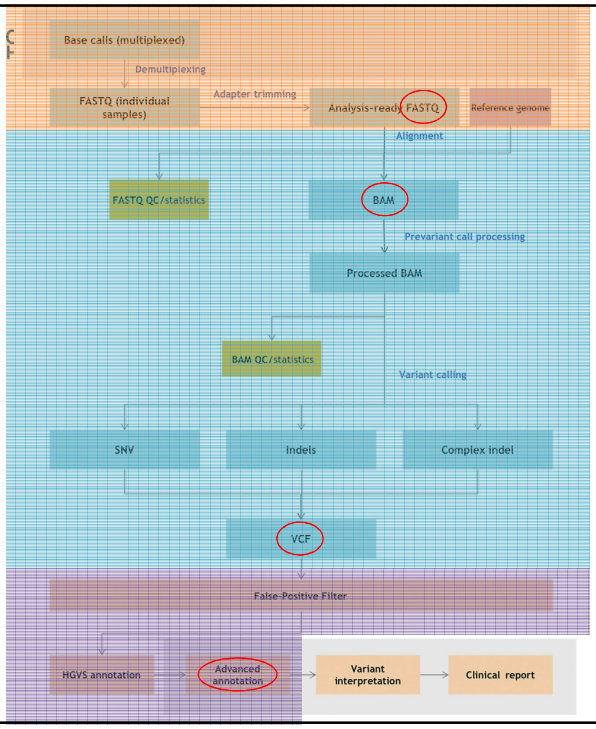


adapted from https://www.broadinstitute.org/gatk/guide/best-practices.php

Read Alignment Algorithm – ex. BWA-MEM



- Reads may not map uniquely
- Alignment algorithms are not generally "aware" of variants
  - Variants may impair alignment



Primary

Secondary

Figure 1 Next-generation sequencing (NGS) bioinformatics pipeline. The figure illustrates a bioinformatics pipeline and its components that are typically used for processing NGS data. The illustration may not represent nuances or additional algorithms that are specific to sequencing platforms. The components of the pipeline that overlap with the **gray shaded region** are out of scope of this guideline. BAM, binary alignment map; HGVS, Human Genome Variation Society; indel, insertion/deletion; QC, quality control; SNV, single-nucleotide variant; VCF, variant call format.

Tertiary

Human Interpretation (semi-automated)

Roy S 2018 *JMD*

Tertiary Analysis

- Goals
  - Add meaning to the variants identified
  - Exclude specific variants from further review
  - Prioritize specific variants for review

Tertiary Analysis

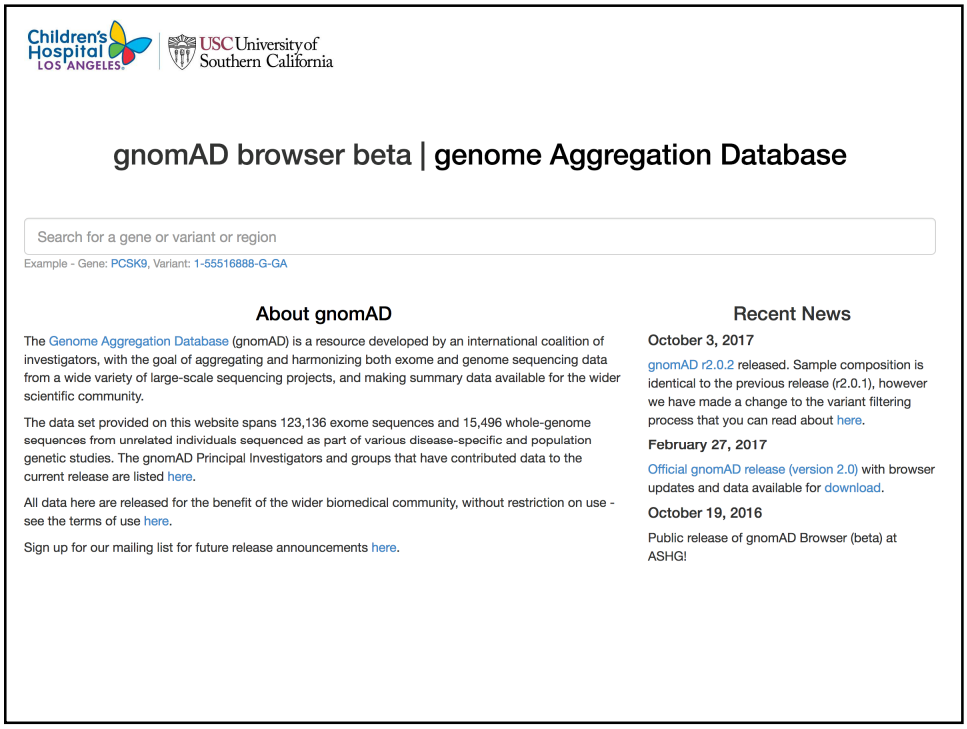
- Methods
  - Annotation
  - Filtering
    - “hard” vs. “soft”
  - Prioritization

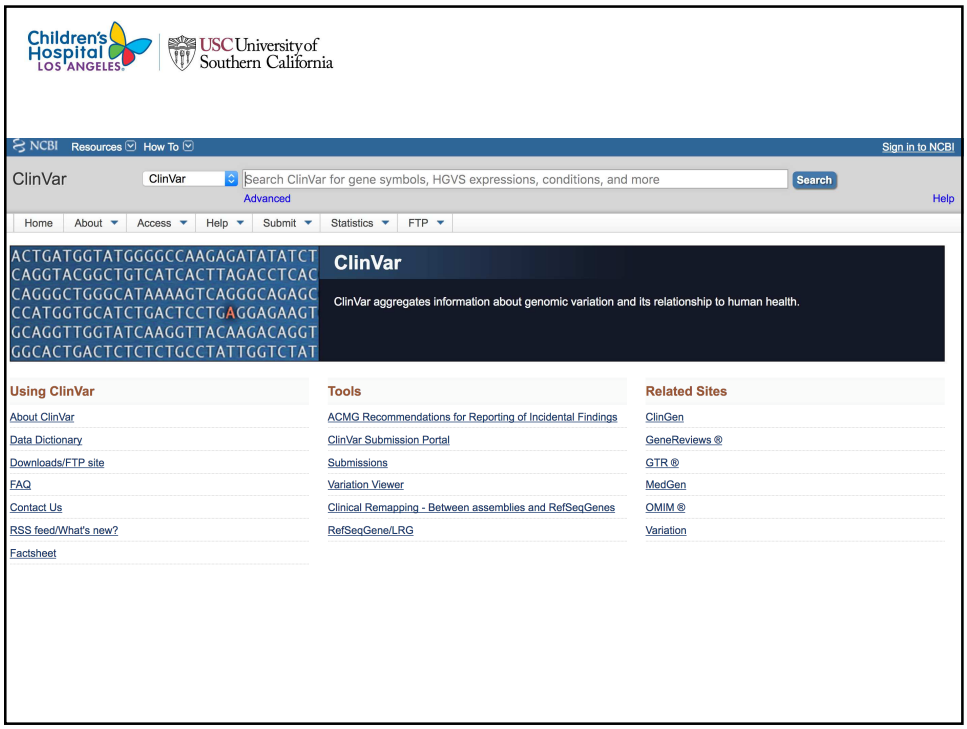
Tertiary Analysis

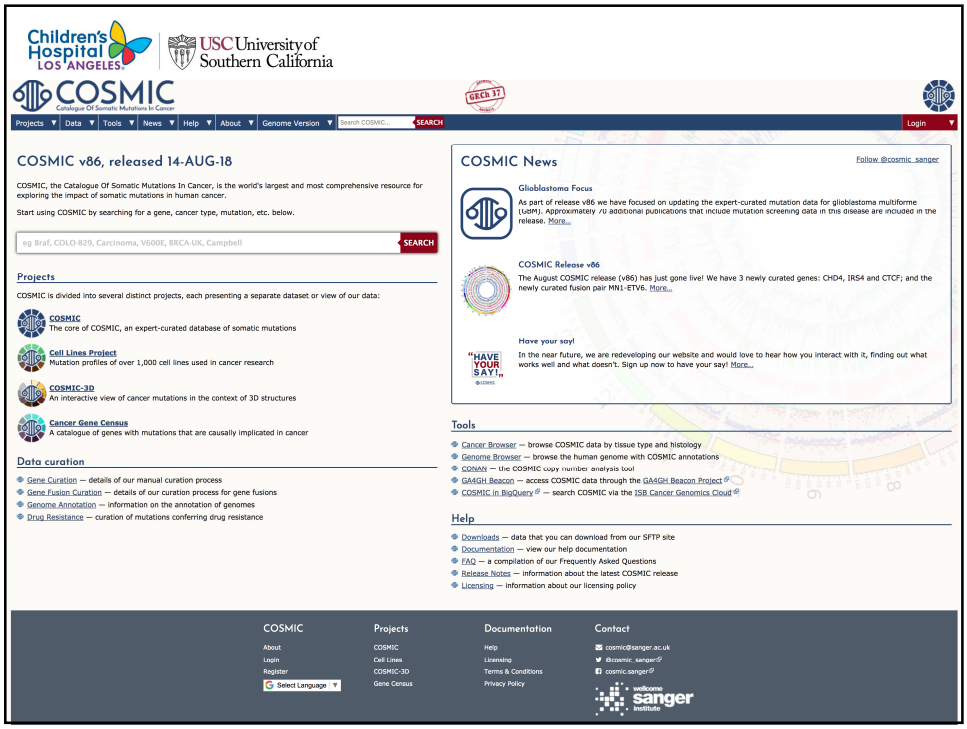
- Filtering to Identify Somatic Mutations
  - Remove germline variants
    - Tumor Only Workflow
      - uses population frequency to remove common germline variants in the population
    - Paired Tumor/Normal Workflow
      - directly subtracts all germline variants for the tested individual

Annotations

- HGVS Nomenclature
  - Effect on mRNA and protein
- Consequence
- Population Frequency
- Presence in Disease Databases
- Associations in Literature
- Computational Prediction
  - Splicing
  - Missense
- Artifact/Internal Frequency Counts


Children's Hospital LOS ANGELES
USC University of Southern California
gnomAD browser beta | genome Aggregation Database
Search for a gene or variant or region
Example - Gene: PCSK9, Variant: 1-55516888-G-GA
About gnomAD
The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.
The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.
All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use here.
Sign up for our mailing list for future release announcements here.
Recent News
October 3, 2017
gnomAD r2.0.2 released. Sample composition is identical to the previous release (r2.0.1), however we have made a change to the variant filtering process that you can read about here.
February 27, 2017
Official gnomAD release (version 2.0) with browser updates and data available for download.
October 19, 2016
Public release of gnomAD Browser (beta) at ASHG!

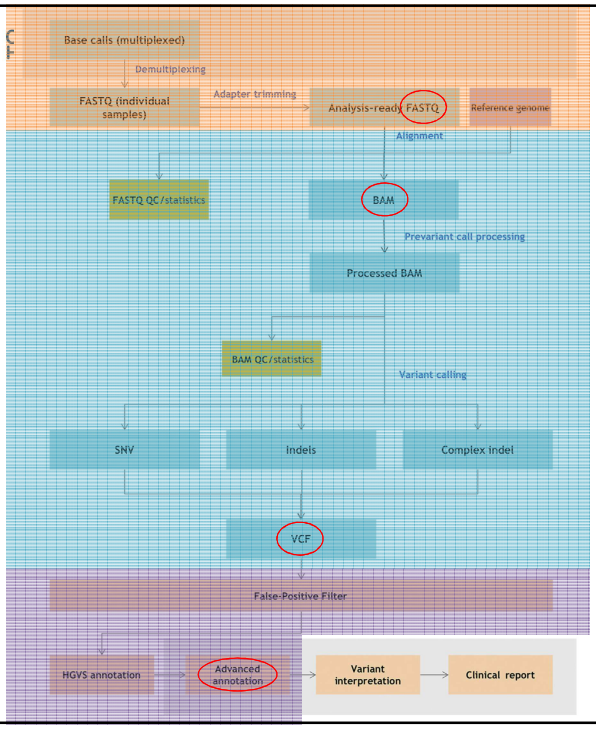

Children's Hospital LOS ANGELES
USC University of Southern California
NCBI Resources How To
Sign in to NCBI
ClinVar
ClinVar
Search ClinVar for gene symbols, HGVS expressions, conditions, and more
Search
Advanced
Help
Home About Access Help Submit Statistics FTP
ACTGATGGTATGGGGCCAAGAGATATATCT
CAGGTACGGCTGTCATCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCCTGAGGAGAAGT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTCTGCCTATTGGTCTAT
ClinVar
ClinVar aggregates information about genomic variation and its relationship to human health.
Using ClinVar
About ClinVar
Data Dictionary
Downloads/FTP site
FAQ
Contact Us
RSS feed/What's new?
Factsheet
Tools
ACMG Recommendations for Reporting of Incidental Findings
ClinVar Submission Portal
Submissions
Variation Viewer
Clinical Remapping - Between assemblies and RefSeqGenes
RefSeqGene/LRG
Related Sites
ClinGen
GeneReviews ®
GTR ®
MedGen
OMIM ®
Variation



Children's Hospital LOS ANGELES | USC University of Southern California

Filtering

- Region of Interest
- Population Frequency
- Quality Scores
- Sequencing Depth



Primary

Secondary

Tertiary

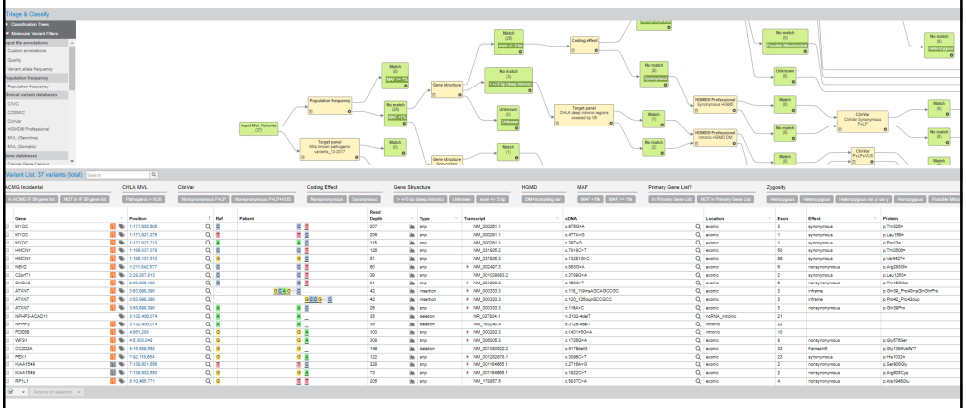
Human Interpretation (semi-automated)

Figure 1 Next-generation sequencing (NGS) bioinformatics pipeline. The figure illustrates a bioinformatics pipeline and its components that are typically used for processing NGS data. The illustration may not represent nuances or additional algorithms that are specific to sequencing platforms. The components of the pipeline that overlap with the **gray shaded region** are out of scope of this guideline. BAM, binary alignment map; HGVS, Human Genome Variation Society; indel, insertion/deletion; QC, quality control; SNV, single-nucleotide variant; VCF, variant call format.

Roy S 2018 *JMD*

Children's Hospital LOS ANGELES | USC University of Southern California

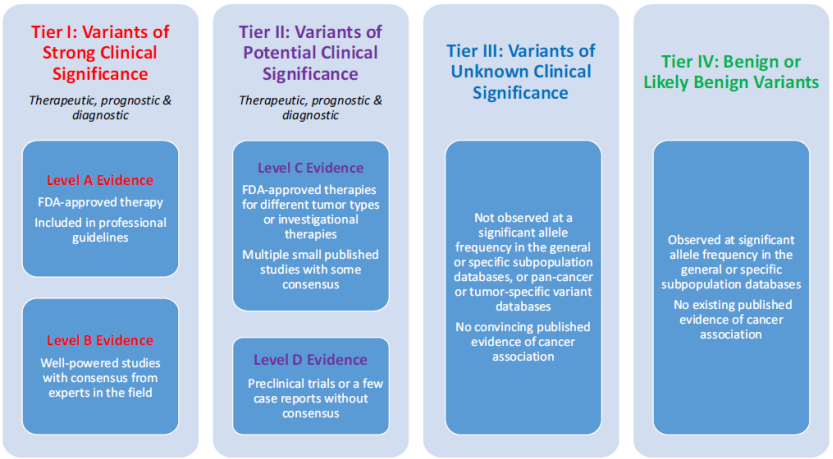
Human Interpretation



Interpretation

- Two Steps (semi-automated process)
  - Variant review and classification
  - Reporting
- Differs for different types of testing
  - Germline sequencing
  - Germline microarray
  - Tumor sequencing
  - Tumor microarray

Classification and Tiering



Li M 2017 *JMD*

Table 3 Categories of Clinical and/or Experimental Evidence

Category	Therapeutic	Diagnosis	Prognosis
Level A	1. Biomarkers that predict response or resistance to FDA-approved therapies for a specific type of tumor 2. Biomarkers included in professional guidelines that predict response or resistance to therapies for a specific type of tumor	Biomarkers included in professional guidelines as diagnostic for a specific type of tumor	Biomarkers included in professional guidelines as prognostic for a specific type of tumor
Level B	Biomarkers that predict response or resistance to therapies for a specific type of tumor based on well-powered studies with consensus from experts in the field	Biomarkers of diagnostic significance for a specific type of tumor based on well-powered studies with consensus from experts in the field	Biomarkers of prognostic significance for a specific type of tumor based on well-powered studies with consensus from experts in the field
Level C	1. Biomarkers that predict response or resistance to therapies approved by the FDA or professional societies for a different type of tumor 2. Biomarkers that serve as inclusion criteria for clinical trials	Biomarkers of diagnostic significance based on the results of multiple small studies	Biomarkers of prognostic significance based on the results of multiple small studies
Level D	Biomarkers that show plausible therapeutic significance based on preclinical studies	Biomarkers that may assist disease diagnosis themselves or along with other biomarkers based on small studies or a few case reports	Biomarkers that may assist disease prognosis themselves or along with other biomarkers based on small studies or a few case reports

FDA, Food and Drug Administration.

Li M 2017 *JMD*

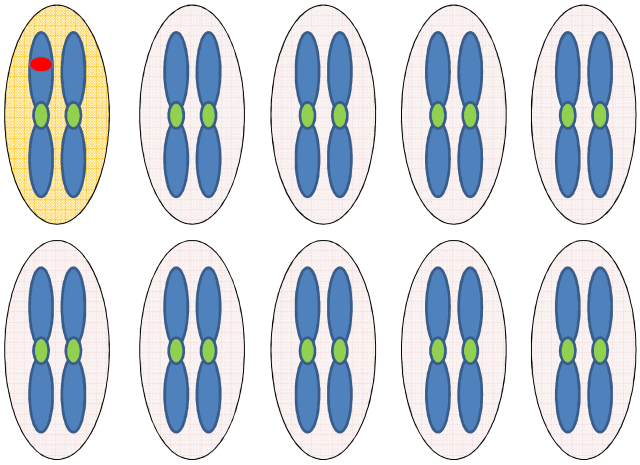
Children's Hospital LOS ANGELES | USC University of Southern California

Important QC Values

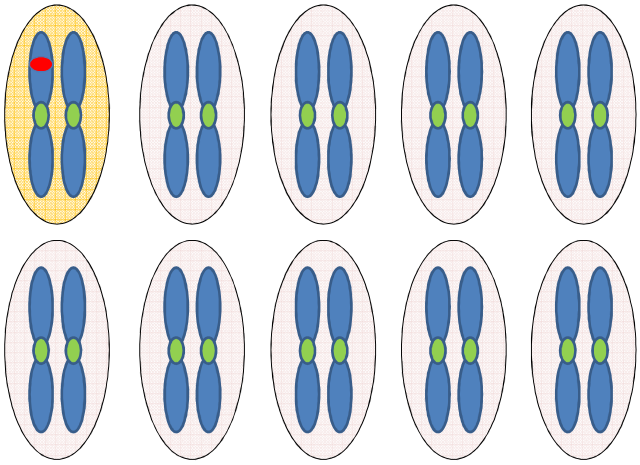
- Mean Read Length
- Mean Depth of Coverage
- Number and/or percentage of regions under 250X coverage

Variant Allele Frequency/Fraction (VAF)

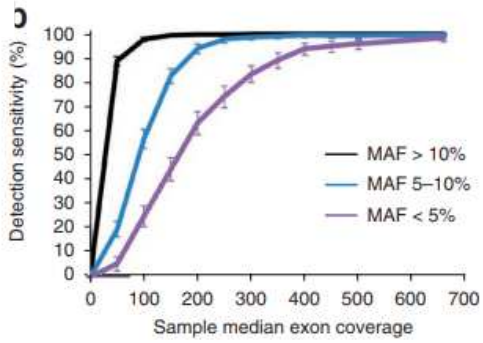
What is the VAF of this mutation?



Conventional NGS can detect Mutations present in ~ 1/10 cells or ~ 1/20 (5%) DNA molecules



Sensitivity Requires Depth



- Detection sensitivity is a function of depth of coverage
- High depth of coverage (>250x) is needed to avoid false negatives

Frampton G 2013 *Nat Biotechnology*

Main Points

- Bioinformatic pipelines convert raw sequencing data into interpretable information
- NGS pipelines are divided into primary, secondary, and tertiary analysis with corresponding intermediate file types
  - FASTQ – raw reads
  - BAM – aligned reads
  - VCF – called variants
- Somatic mutation detection requires high sequencing depth and removal of germline variants

Questions or comments?

Ryan J. Schmidt, MD, PhD | rschmidt@chla.usc.edu
Assistant Director, Clinical Genomics Laboratory
Center for Personalized Medicine
Department of Pathology and Laboratory Medicine
Children's Hospital Los Angeles
Assistant Professor of Clinical Pathology
Keck School of Medicine of USC