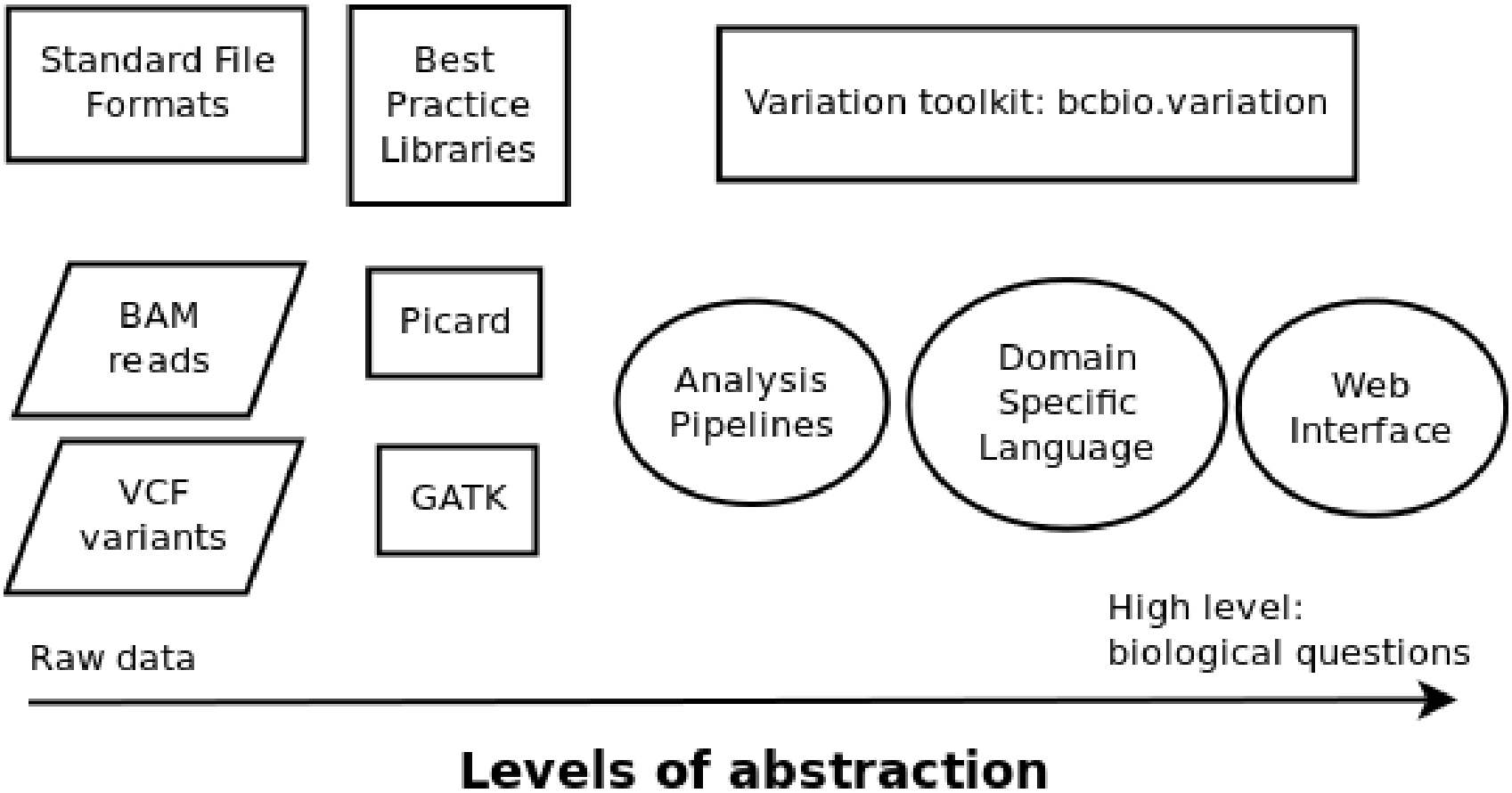
Toolkit for variation comparison and analysis

Brad Chapman, Bioinformatics Core at Harvard School of Public Health

Bioinformatics Open Source Conference, 13 July 2012

Variation

```
1    1156131 rs2887286    T    C    1714.07 PASS
AB=0;ABP=0;AC=1;AC1=2;AF=1.0;AF1=1;AN=1;AO=56;
BVAR;CIGAR=1X;DB;DP=59;DP4=0,0,27,26;DPRA=0;Dels=0.00;
EPP=6.88793;EPPR=0;FQ=-187;FS=0.000;GC=55.45;HRun=0;
HaplotypeScore=0.0000;LEN=1;MEANALT=2;MQ=70.00;MQ0=0;
MQM=70;MQMR=0;NBQ=27.76;NS=1;NUMALT=1;ODDS=308.76;
PAIRED=0.928571;QD=30.61;RO=0;RPP=12.937;RUN=1;
SAP=3.16541;SNPEFF_EFFECT=INTRON;SNPEFF_GENE_NAME=SDF4;
SNPEFF_FUNCTIONAL_CLASS=NONE;SNPEFF_IMPACT=MODIFIER;
SNPEFF_GENE_BIOTYPE=protein_coding;
SNPEFF_TRANSCRIPT_ID=ENST00000263741;
Samples=NA19239_illumina;TYPE=snp;VDB=0.0416;
VQSLOD=16.7453;XAI=0.00018797;XAM=0.00154116
GT:AO:DP:GL:GQ:QA:QR:RO
1:56:58:-141.12,-3.61:99:1527:0:0
```



Levels of abstraction

Answer biological questions; help people

by ERIK CORONA • JULY 9, 2012

The goal of this blog is to catalog the process of finding the genetic cause of an inherited disease. Stacy was diagnosed with CMT4 (Charcot-Marie-Tooth Type 4) in the 1980s following a nerve biopsy which showed the characteristic onion-bulb myelin sheath around peripheral nerves and after measuring nerve conduction velocities along the length of her leg. There are many different causes of CMT, many known and many unknown. Stacy recently had her exome sequenced. My goal is to find the genetic basis of this disease. My name is Erik Corona and I am PhD student in the Stanford Medical school studying the genetics of complex disease. I will be making all results public and I will also post all code and/or scripts I use to generate these results. I will be posting screenshots of candidate structural variants and making every step along the way transparent and clear. This blog will represent a detailed case study of how we now have the technology and ability to use sequence data to find the genetic cause of a rare inherited disease. I welcome all comments, advice, and suggestions.

http://cmtproject.blogspot.com/

Solutions

Comparisons

- Multiple callers: GATK, FreeBayes, samtools
- Multiple technologies: Illumina, SOLiD, IonTorrent

Identify real variants

- Summarize associated metrics
- Remove false positives

Scale

- Millions of variants
- Thousands of samples

http://genomics.xprize.org/

Clinical grade genome

- 98 percent genome coverage
- 1 error per million bases (SNPs + small indels)
- Full haplotype phasing
- Structural variations

Sequencing for patients

My son Bertrand has a new genetic disorder.

Patient 0.

To find it, a team of scientists at Duke University used whole-exome sequencing (a protein-focused variant of whole-genome sequencing) on me, my wife and my son.

We discovered that my son inherited two *different* (thus-far-unique) mutations in the same gene--the NGLY1 gene--which encodes the enzyme N-glycanase 1. Consequently, he cannot make this enzyme.

My son is the only human being known to lack this enzyme.

Below, I'm documenting our journey to the unlikeliest of diagnoses.

This is a story about the kind of hope that only science can provide.

http://matt.might.net/articles/my-sons-killer/

Technology overview

- Clojure
  http://clojure.org/
- Genome Analysis Toolkit
  http://www.broadinstitute.org/gsa/wiki/index.php/Home_Page
- GenomeSpace
  http://www.genomespace.org/

https://github.com/chapmanb/bcbio.variation

Why Clojure?

- Hosted on Java Virtual Machine
  - Interoperability with existing libraries: GATK, GenomeSpace
  - Wonderful build, deployment and testing tools
- Functional and immutable
  - Easier to write correct code
  - Small functions: concise and refactorable
- Community
  - Smart people working on hard problems
- Ecosystem
  - Multiple backends: ClojureScript

Example analysis pipeline

- Variant files from two different callers
  - GATK UnifiedGenotyper
  - FreeBayes: https://github.com/ekg/freebayes
- Compare, identifying:
  - Identical variants
  - Different variants in each caller
  - Metrics to help discriminate

High level description

```
experiments:
  - sample: Test1
    ref: test/data/GRCh37.fa
    intervals: test/data/target-regions.bed
    align: test/data/aligned-reads.bam
    calls:
      - name: gatk
        file: test/data/gatk-calls.vcf
      - name: freebayes
        file: test/data/freebayes-calls.vcf
        prep: true
        annotate: true
        filters:
          - QD < 2.0
          - MQRankSum < -12.5
```

Simple to run

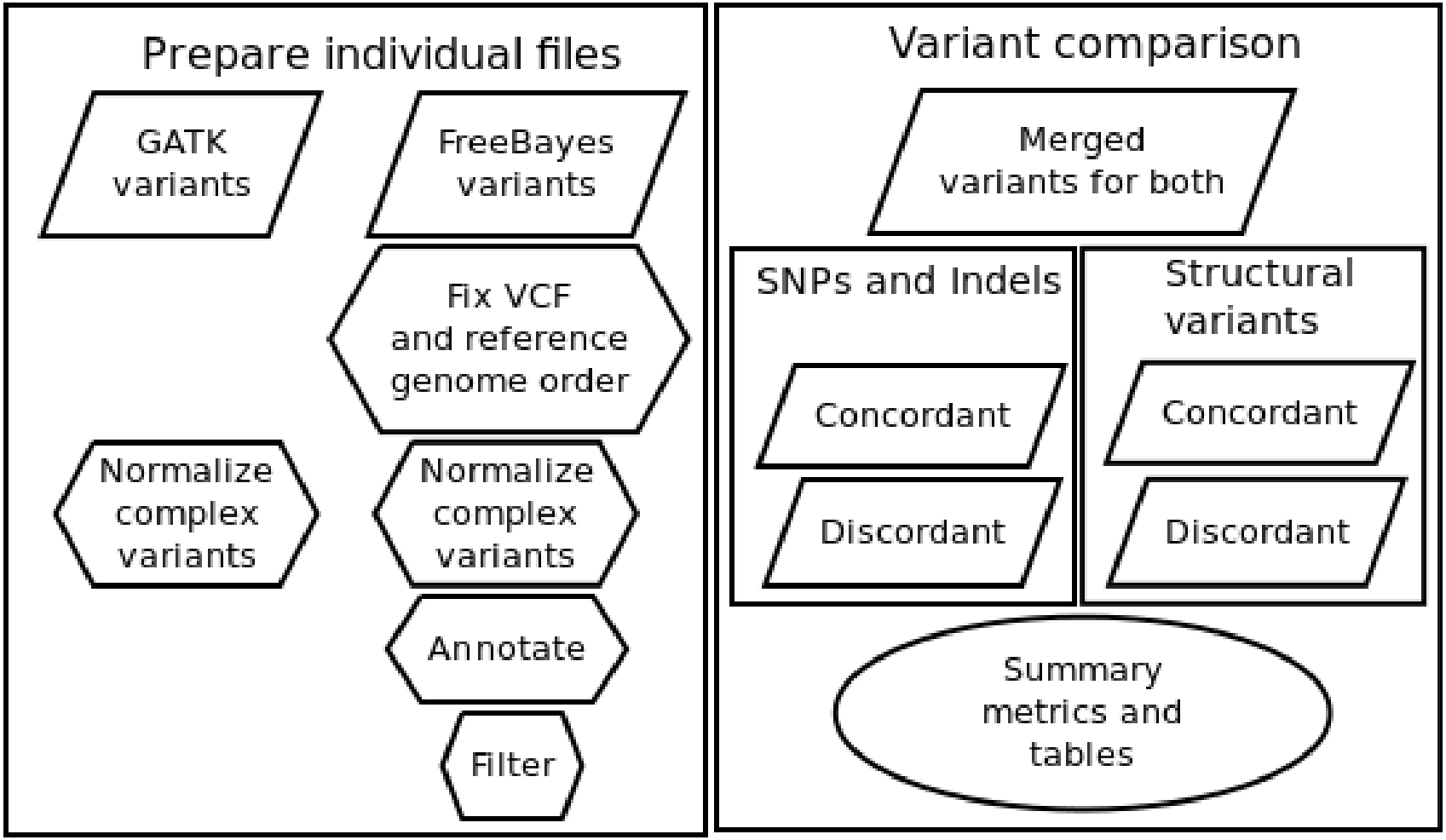
With a new blank machine, get Java and automated Clojure build tool:

```
$ sudo apt-get install openjdk-7-jdk openjdk-7-jre
$ wget https://raw.github.com/technomancy/leiningen/\
      preview/bin/lein
$ chmod 755 lein && sudo mv lein /usr/local/bin
```

Get code and dependencies, then run:

```
$ git pull git://github.com/chapmanb/bcbio.variation.git
$ cd bcbio.variation
$ lein deps
$ lein variant-compare comparison_description.yaml
```

What happened?

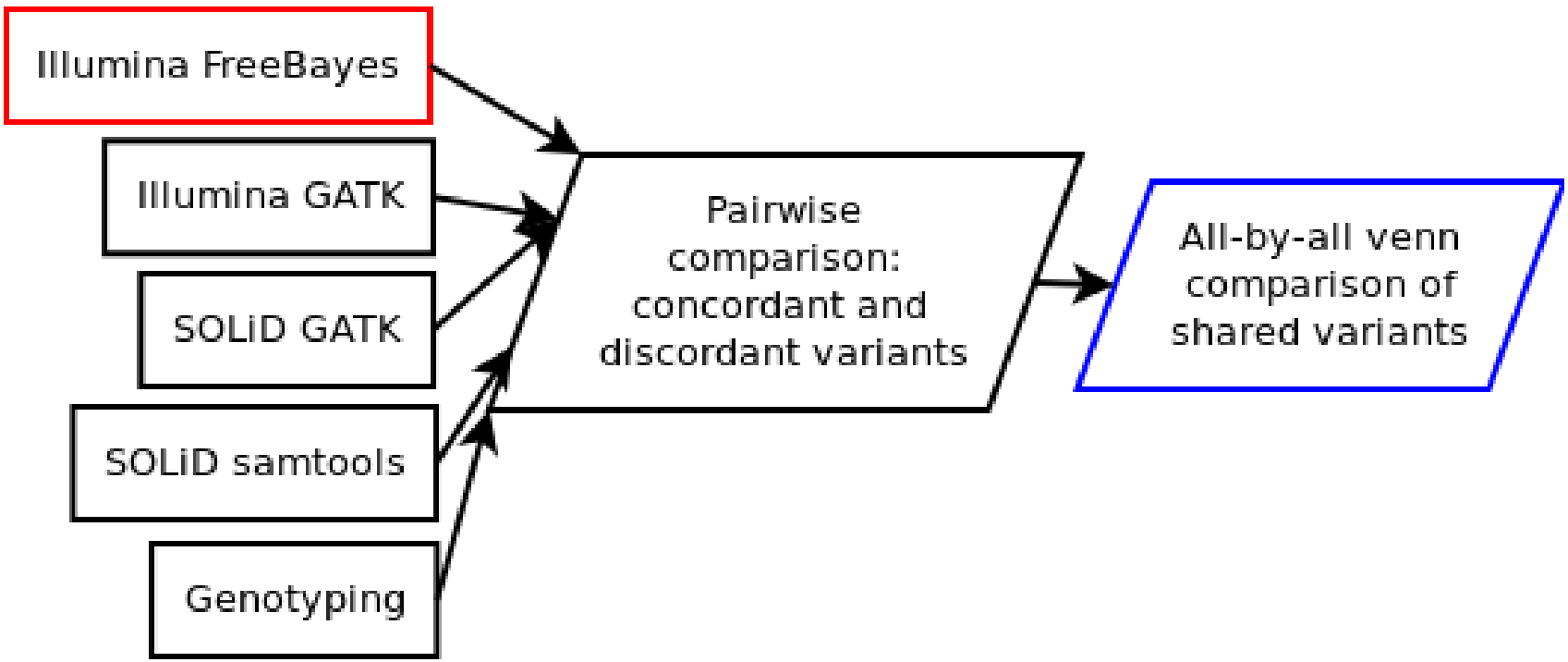


Establishing true variants

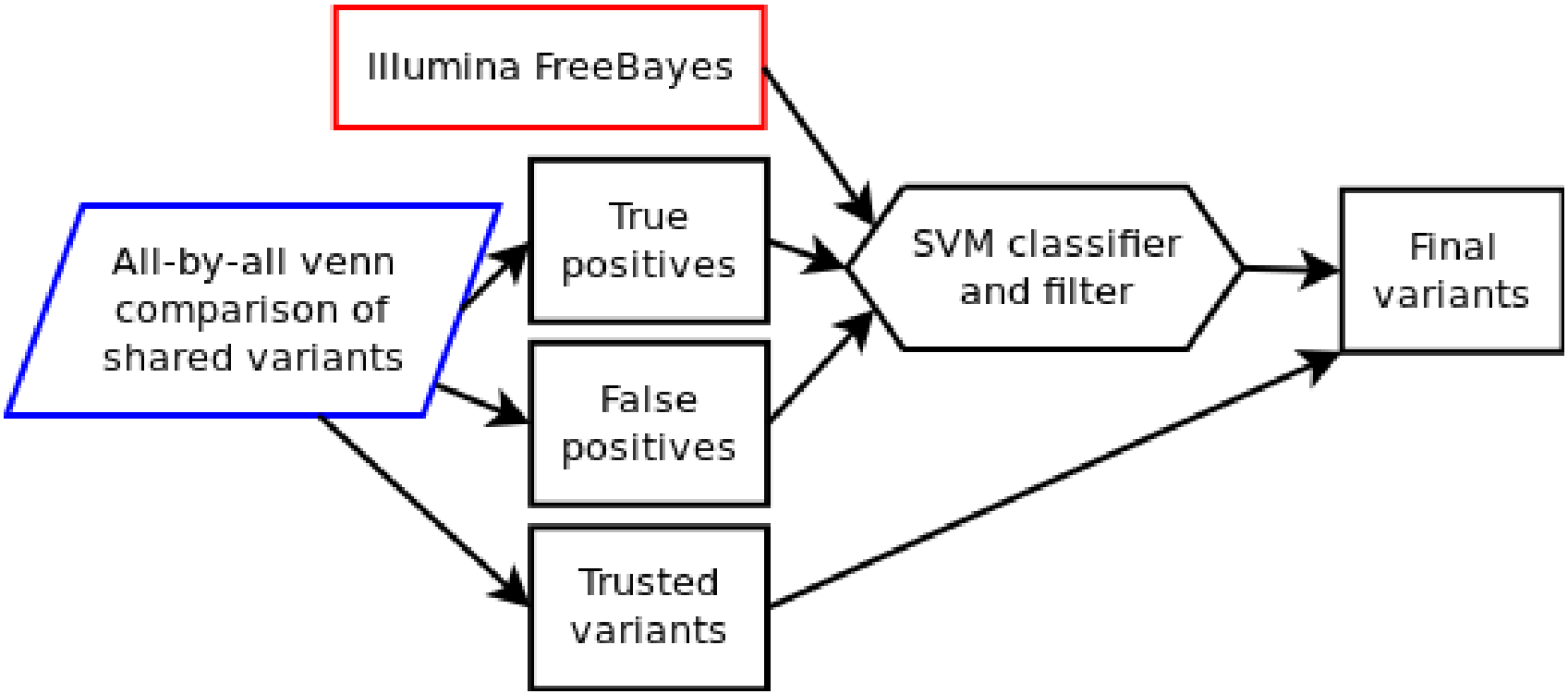
X Prize: haploid gold standard genome

- Public genomes from HapMap/1000 genomes
  - NA12878: Caucasian female from Utah.
  - NA19239: Yoruba male from Ibadan, Nigeria.
- Multiple technologies
  - Illumina
  - SOLiD
  - IonTorrent
- Multiple callers
  - GATK
  - FreeBayes
  - samtools mpileup

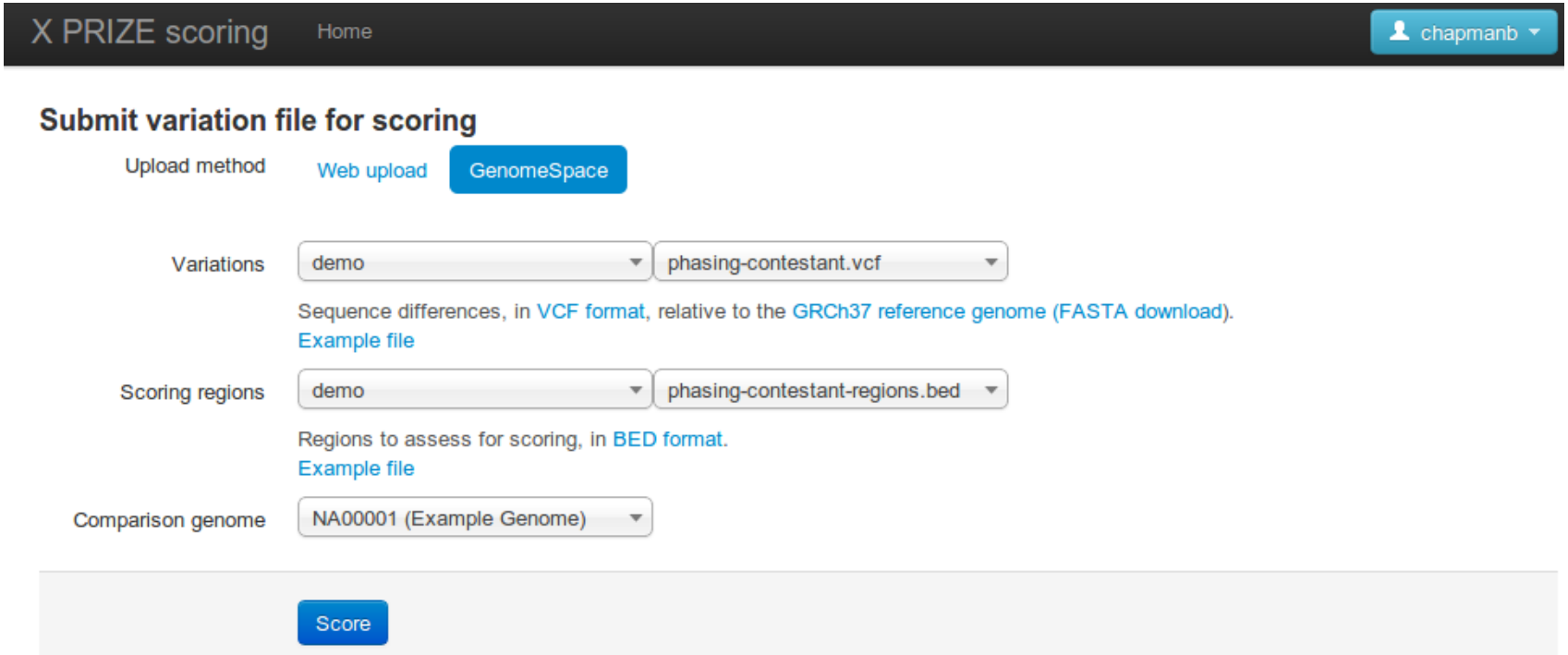
True variant workflow – comparisons



True variant workflow – finalize

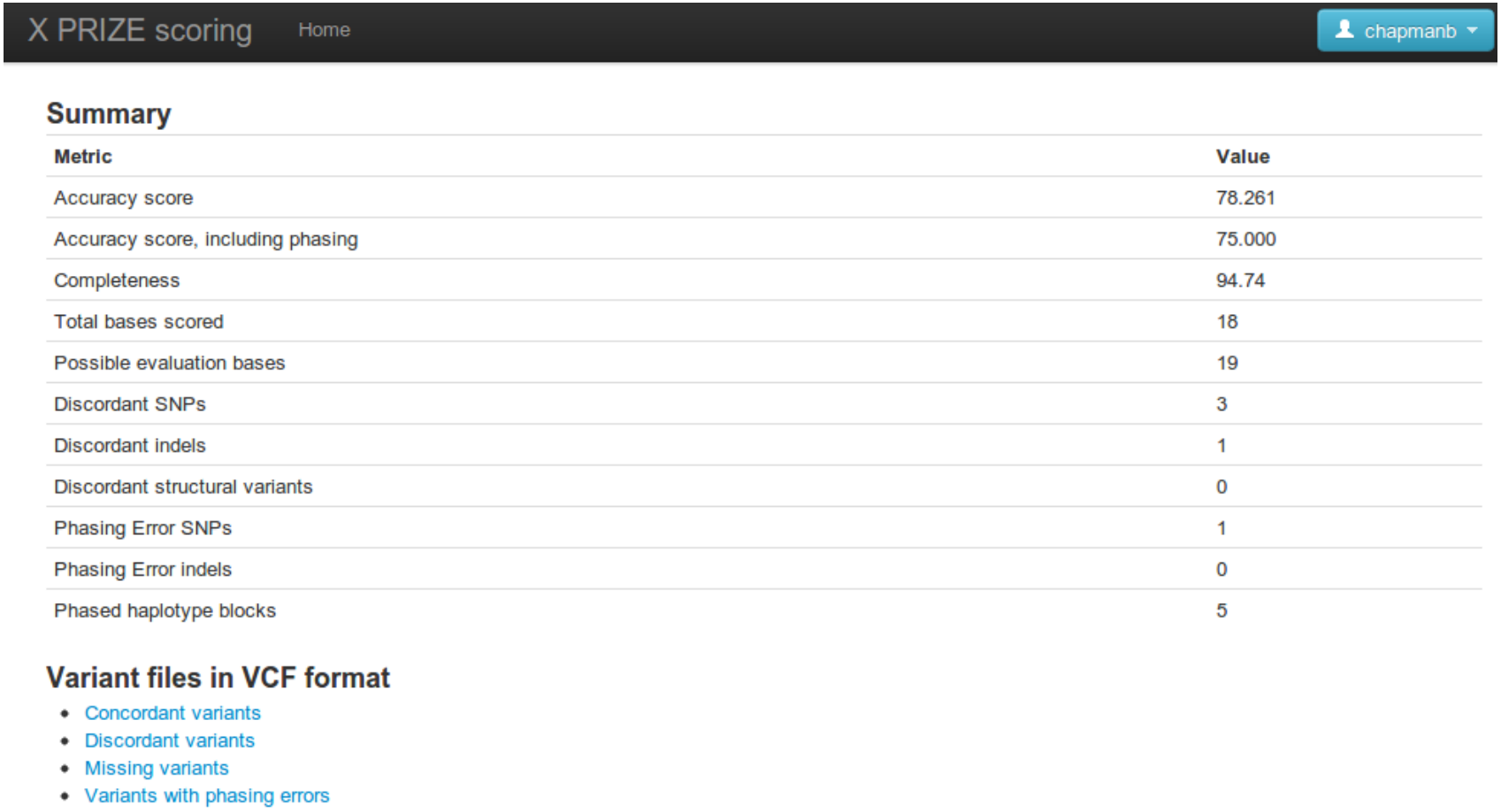


Comparison web interface

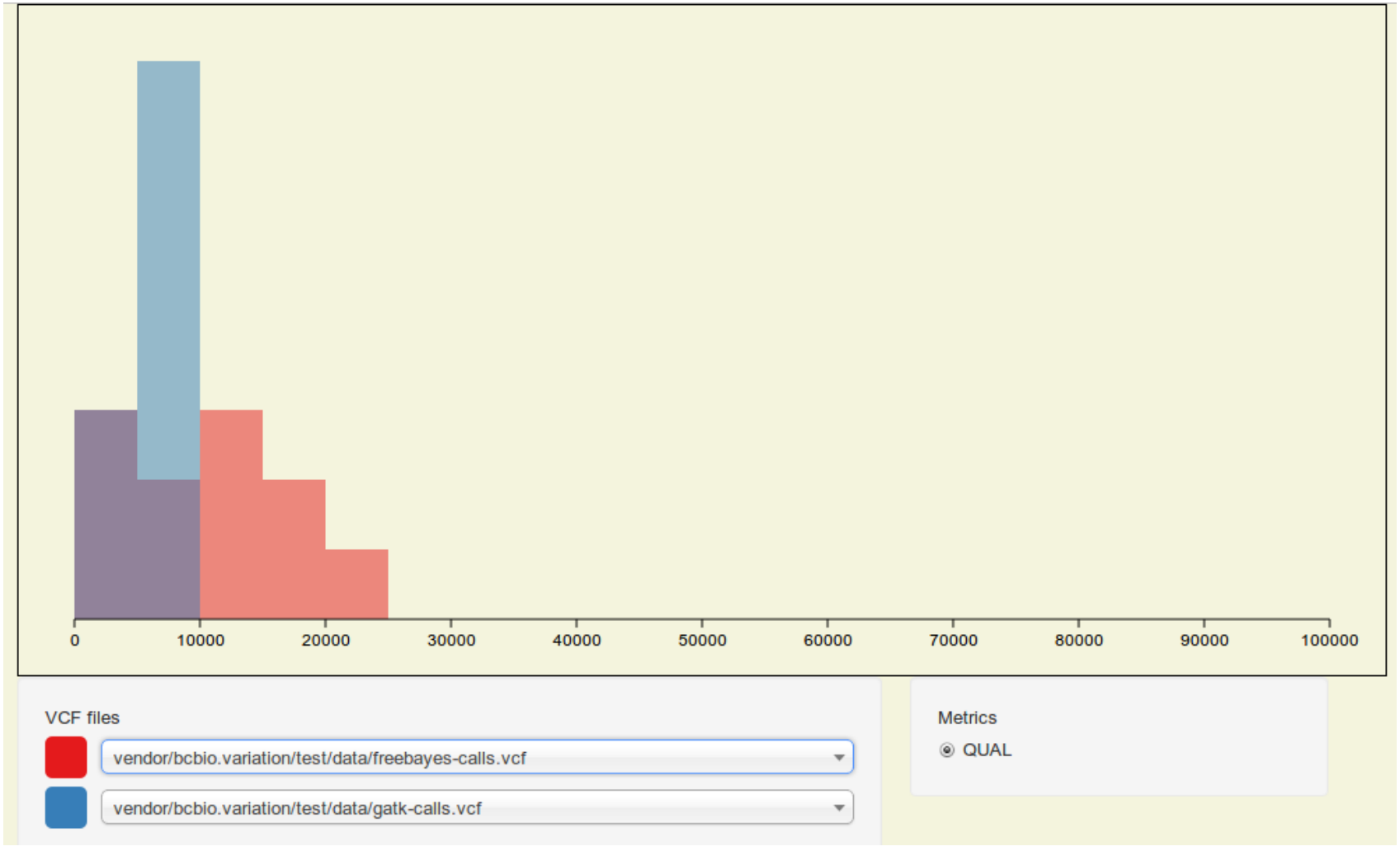


https://validationprotocol.org/

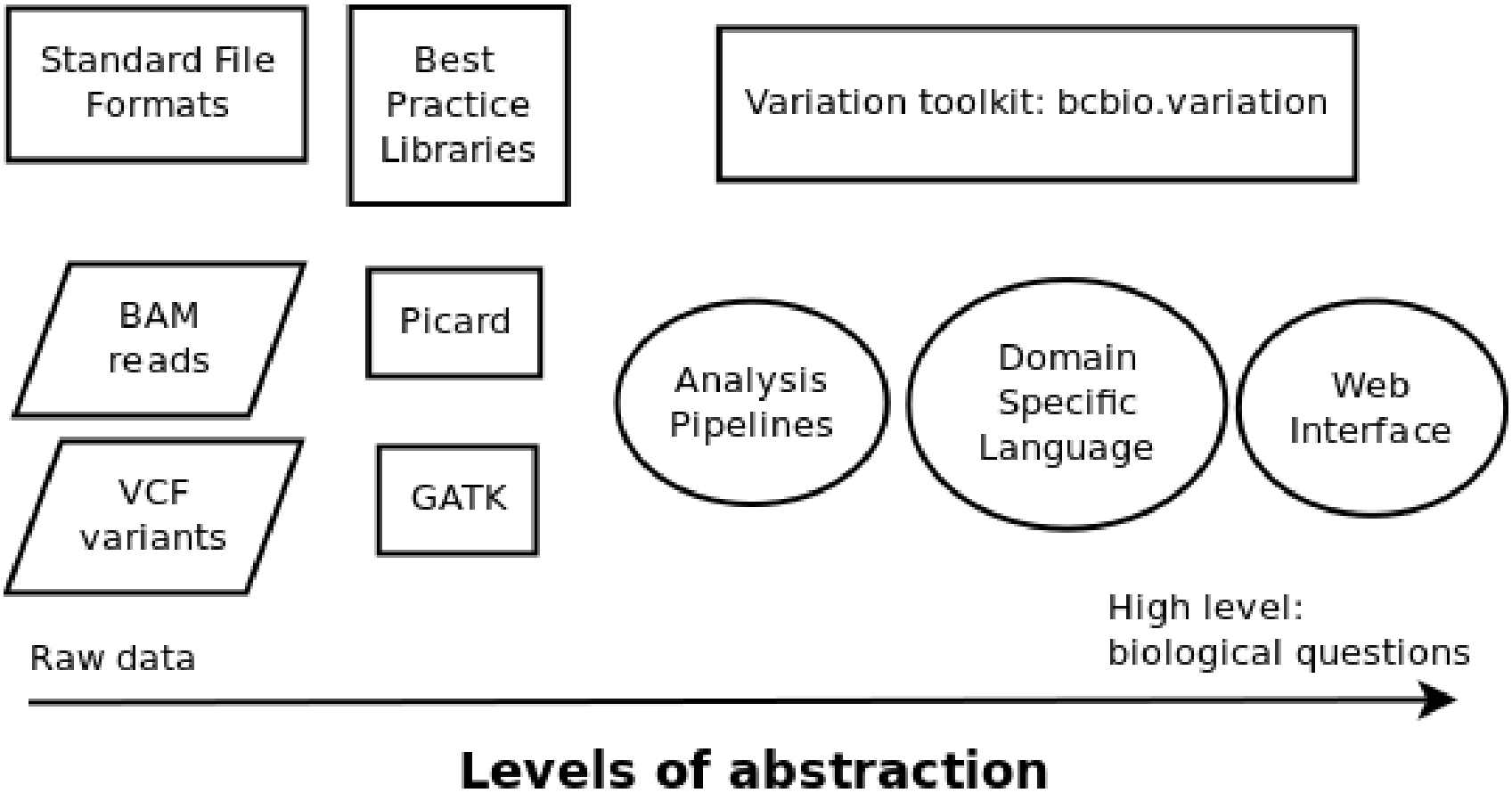
Summary web interface



Next web steps – metrics



http://keminglabs.com/



Levels of abstraction

Answering biological questions

- Establish set of true variants
  - Understand boundaries of certainty
  - Make patient decisions
- Comparison architecture
  - Cancer: tumor/normal pairs
  - Mendelian inherited diseases: father/mother/child
- Annotate with known data
  - Gemini: https://github.com/arq5x/gemini
  - 1000 genomes frequencies
  - Mappability
  - Clinical information