# Variant calling with validated community developed tools

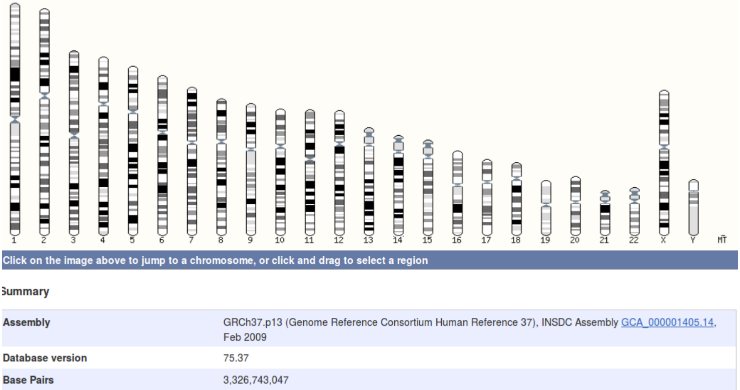
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

6 July 2017

# Outline

- Overview of variant calling tools
- Open source community resources
- bcbio validated variant analysis
- Science
  - Human build 38
  - GATK4 validation
  - Cancer calling of low frequency variants
  - Structural variation
- Practical calling example

# Human whole genome sequencing



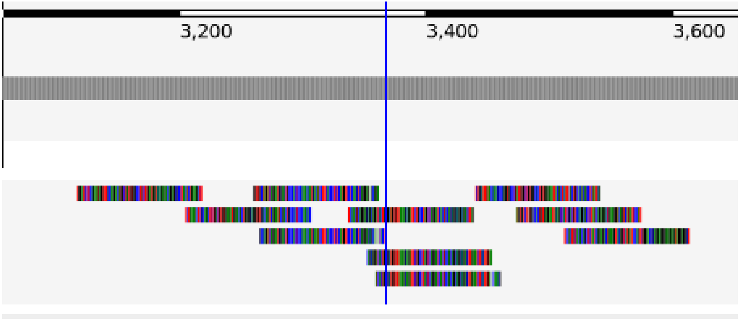
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

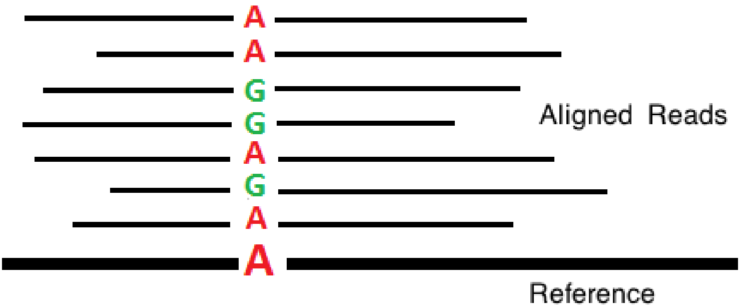
| | |
|---|---|
| **Assembly** | GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009 |
| **Database version** | 75.37 |
| **Base Pairs** | 3,326,743,047 |

http://ensembl.org/Homo_sapiens/Location/Genome

# High throughput sequencing

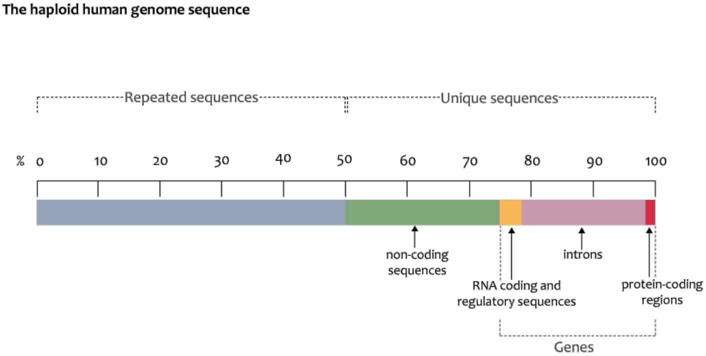


# Variant calling



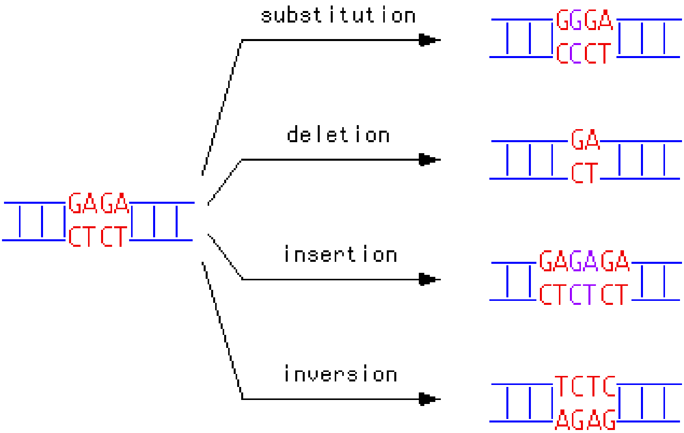
http://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

# Scale: exome to whole genome



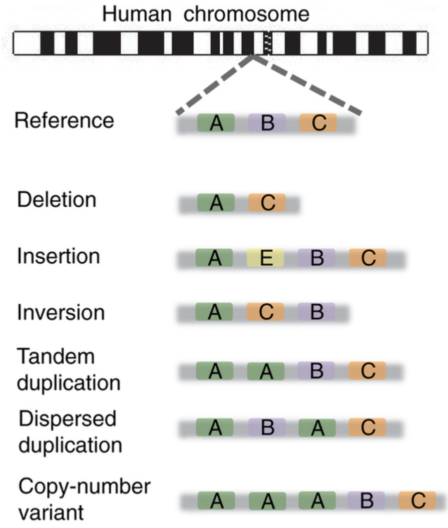
https://www.flickr.com/photos/119980645@N06/

# SNPs and Indels



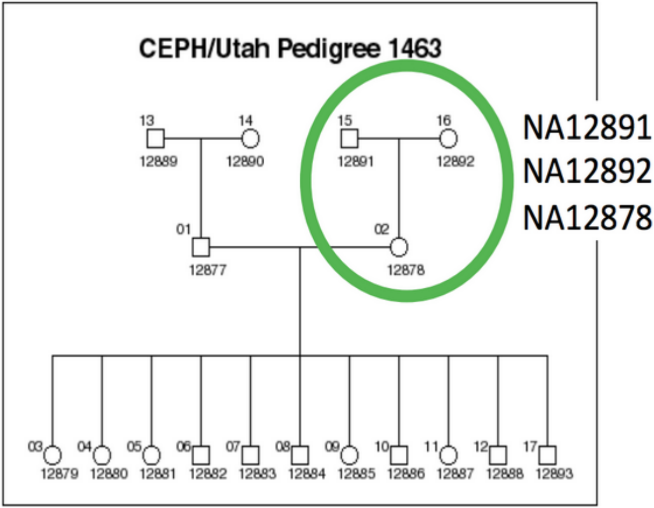
http://carolguze.com/text/442-2-mutations.shtml

# Structural variations



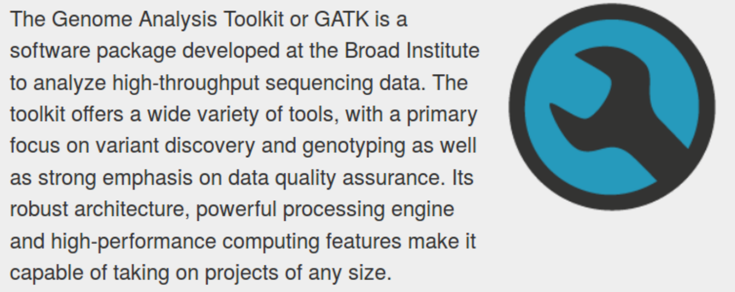
http://www.nature.com/nmeth/journal/v9/n2/full/nmeth.1858.html

# Germline population calling



http://blog.goldenhelix.com/grudy/the-state-of-ngs-variant-calling-dont-panic/

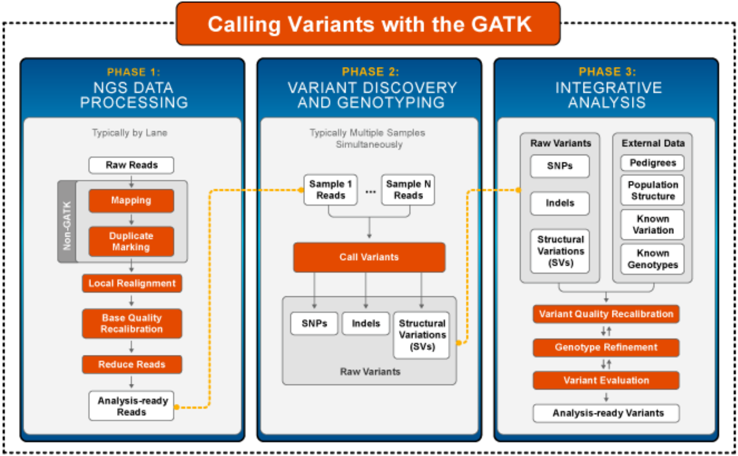
# Genome Analysis Toolkit (GATK)



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.

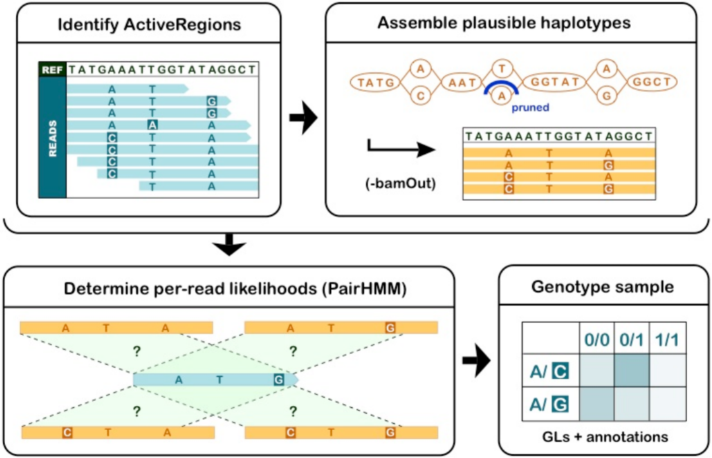
https://www.broadinstitute.org/gatk/

# GATK Best Practices



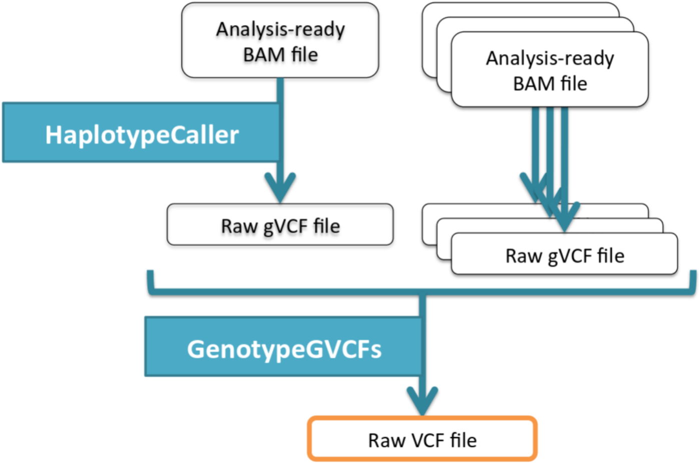
https://www.broadinstitute.org/gatk/guide/best-practices

# HaplotypeCaller



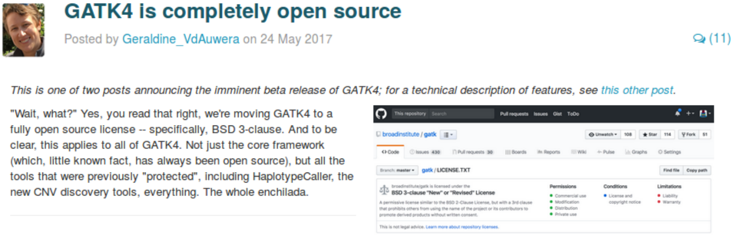
http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

# Joint calling on large populations



http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

# GATK4 now open source for all uses



## GATK4 is completely open source

Posted by Geraldine_VdAuwera on 24 May 2017 (11)

*This is one of two posts announcing the imminent beta release of GATK4; for a technical description of features, see this other post.*

"Wait, what?" Yes, you read that right, we're moving GATK4 to a fully open source license -- specifically, BSD 3-clause. And to be clear, this applies to all of GATK4. Not just the core framework (which, little known fact, has always been open source), but all the tools that were previously "protected", including HaplotypeCaller, the new CNV discovery tools, everything. The whole enchilada.

broadinstitute / gatk

gatk / LICENSE.TXT

broadinstitute/gatk is licensed under the

**BSD 3-clause "New" or "Revised" License**

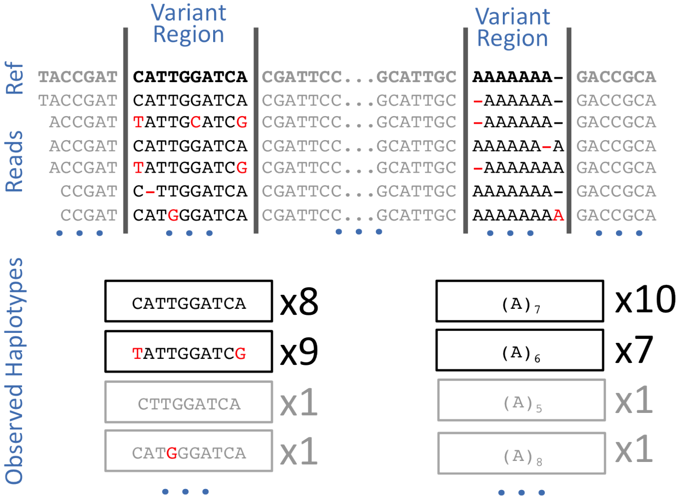
A permissive license similar to the BSD 2-Clause License, but with a 3rd clause that prohibits others from using the name of the project or its contributors to promote derived products without written consent.

| Permissions | Conditions | Limitations |
|---|---|---|
| ● Commercial use | ● License and copyright notice | ● Liability |
| ● Modification | | ● Warranty |
| ● Distribution | | |
| ● Private use | | |

This is not legal advice. Learn more about repository licenses.

https://software.broadinstitute.org/gatk/blog?id=9645

# FreeBayes



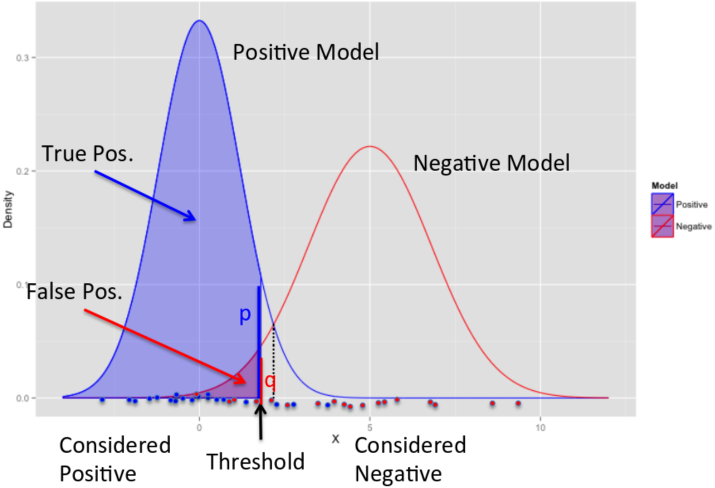
| | | Variant Region | | Variant Region | |
|---|---|---|---|---|---|
| Ref | TACCGAT | CATTGGATCA | CGATTCC...GCATTGC | AAAAAAA- | GACCGCA |
| Reads | TACCGAT | CATTGGATCA | CGATTCC...GCATTGC | -AAAAAA- | GACCGCA |
| | ACCGAT | TATTGCATCG | CGATTCC...GCATTGC | -AAAAAA- | GACCGCA |
| | ACCGAT | CATTGGATCA | CGATTCC...GCATTGC | AAAAAA-A | GACCGCA |
| | ACCGAT | TATTGGATCG | CGATTCC...GCATTGC | -AAAAAAA | GACCGCA |
| | CCGAT | C-TTGGATCA | CGATTCC...GCATTGC | AAAAAAA- | GACCGCA |
| | CCGAT | CATGGGATCA | CGATTCC...GCATTGC | AAAAAAAA | GACCGCA |
| | • • • | • • • | • • • | • • • | • • • |

Observed Haplotypes

| | | | |
|---|---|---|---|
| CATTGGATCA | x8 | $(A)_7$ | x10 |
| TATTGGATCG | x9 | $(A)_6$ | x7 |
| CTTGGATCA | x1 | $(A)_5$ | x1 |
| CATGGGATCA | x1 | $(A)_8$ | x1 |
| • • • | | • • • | |

https://github.com/ekg/freebayes

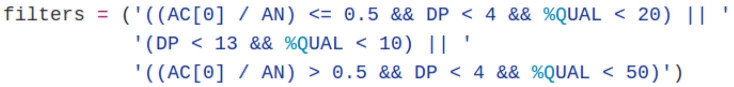
# Filtering – Variant Quality Score Recalibration



$$\text{VQSLOD}(x) = \text{Log}(p(x)/q(x))$$

http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

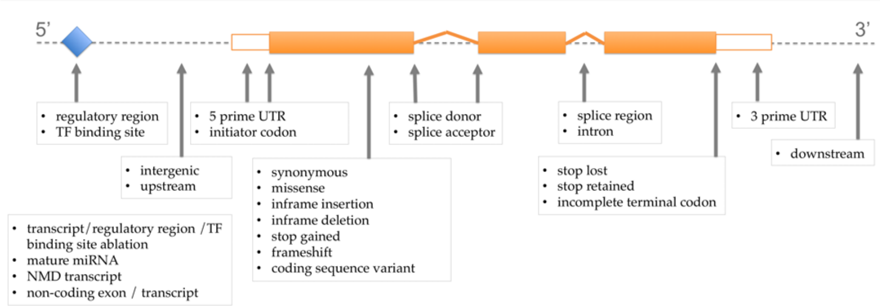
# Filtering – hard cutoffs



```
filters = ('((AC[0] / AN) <= 0.5 && DP < 4 && %QUAL < 20) || '
           '(DP < 13 && %QUAL < 10) || '
           '((AC[0] / AN) > 0.5 && DP < 4 && %QUAL < 50)')
```

http://bcb.io/2014/05/12/wgs-trio-variant-evaluation/

# Effects prediction



http://www.ensembl.org/info/genome/variation/predicted_data.html

# Tools for effects predictions

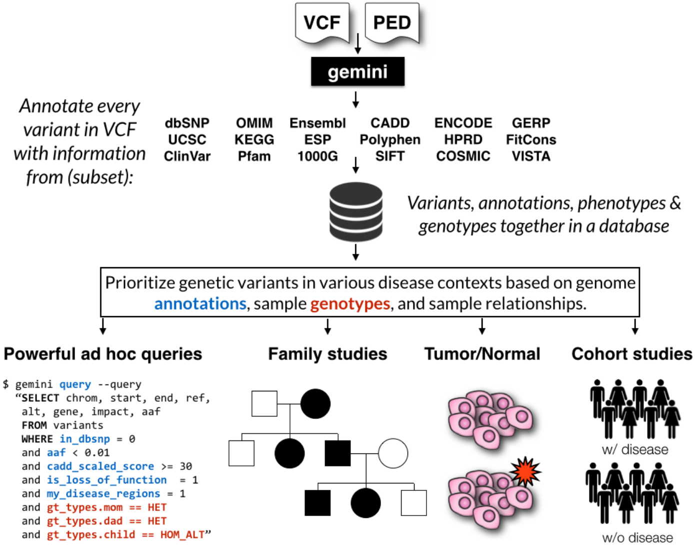
- snpEff

http://snpeff.sourceforge.net/

- Variant Effect Predictor (VEP) from Ensembl

http://www.ensembl.org/info/docs/tools/vep/index.html

# Annotation and analysis – GEMINI



VCF

PED

gemini

*Annotate every variant in VCF with information from (subset):*

dbSNP
UCSC
ClinVar

OMIM
KEGG
Pfam

Ensembl
ESP
1000G

CADD
Polyphen
SIFT

ENCODE
HPRD
COSMIC

GERP
FitCons
VISTA

*Variants, annotations, phenotypes & genotypes together in a database*

Prioritize genetic variants in various disease contexts based on genome **annotations**, sample **genotypes**, and sample relationships.

**Powerful ad hoc queries**

```
$ gemini query --query
  “SELECT chrom, start, end, ref,
   alt, gene, impact, aaf
   FROM variants
   WHERE in_dbsnp = 0
   and aaf < 0.01
   and cadd_scaled_score >= 30
   and is_loss_of_function  = 1
   and my_disease_regions = 1
   and gt_types.mom == HET
   and gt_types.dad == HET
   and gt_types.child == HOM_ALT”
```

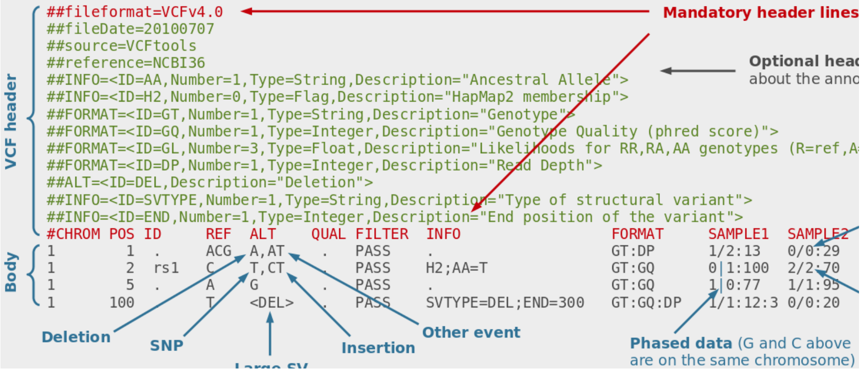
**Family studies**

**Tumor/Normal**

**Cohort studies**

https://github.com/arq5x/gemini

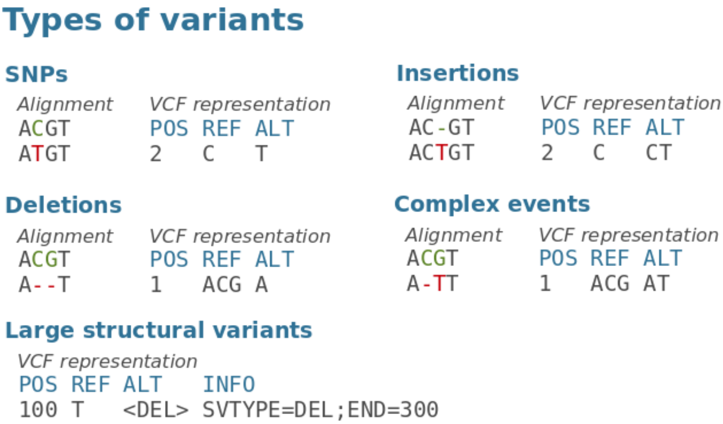
# VCF – overview



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS  ID    REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1    SAMPLE2
1        1  .     ACG  A,AT   .    PASS    .                    GT:DP      1/2:13     0/0:29
1        2  rs1   C    T,CT   .    PASS    H2;AA=T              GT:GQ      0|1:100    2/2:70
1        5  .     A    G      .    PASS    .                    GT:GQ      1|0:77     1/1:95
1      100  .     T    <DEL>  .    PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3   0/0:20
```

http://vcftools.sourceforge.net/VCF-poster.pdf

# VCF – representations



## Types of variants

### SNPs

```
Alignment     VCF representation
ACGT          POS REF ALT
ATGT          2   C   T
```

### Insertions

```
Alignment     VCF representation
AC-GT         POS REF ALT
ACTGT         2   C   CT
```

### Deletions

```
Alignment     VCF representation
ACGT          POS REF ALT
A--T          1   ACG A
```

### Complex events

```
Alignment     VCF representation
ACGT          POS REF ALT
A-TT          1   ACG AT
```

### Large structural variants

```
VCF representation
POS REF ALT   INFO
100 T   <DEL> SVTYPE=DEL;END=300
```

http://vcftools.sourceforge.net/VCF-poster.pdf

# Learning to read VCFs

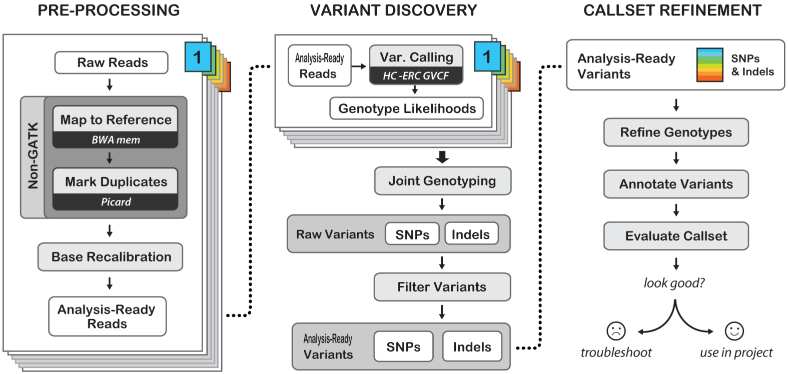
- Step by step guide from Broad

https://www.broadinstitute.org/gatk/guide/article?id=1268

- Specification

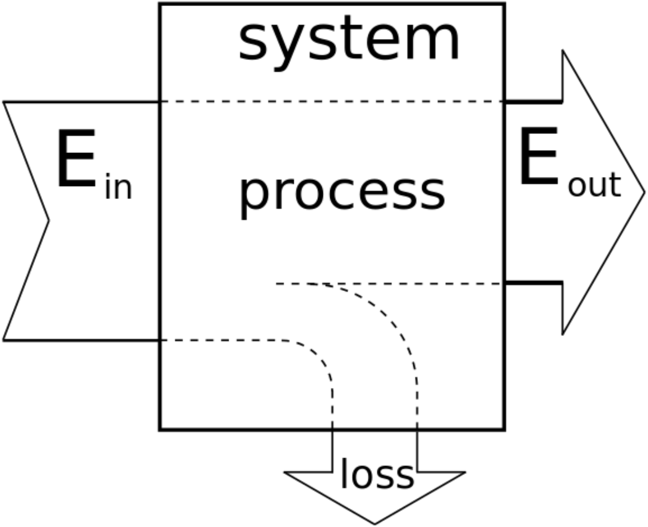
http://samtools.github.io/hts-specs/

# You want to build a variant calling pipeline



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/best-practices/

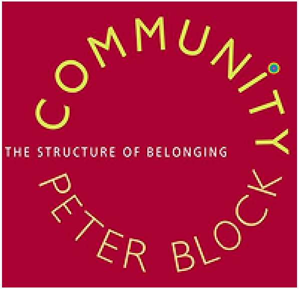


https://commons.wikimedia.org/wiki/File:Efficiency_diagram_by_Zureks.svg

# Barriers to implementing yourself

- Changing tools
- Feature support burden
- Validation

# Build open source communities



COMMUNITY

THE STRUCTURE OF BELONGING

PETER BLOCK

O|B|F

Page Discussion

BOSC 2017

Main Page

Projects

News

BOSC

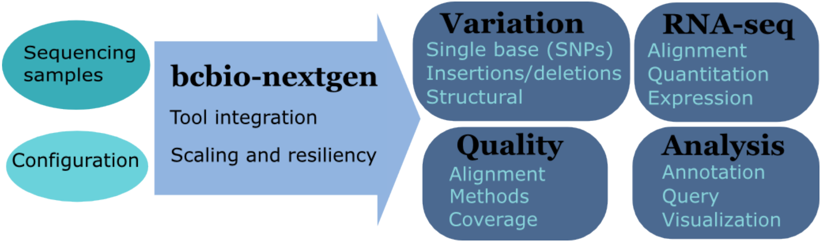
OBF Board

Join

http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

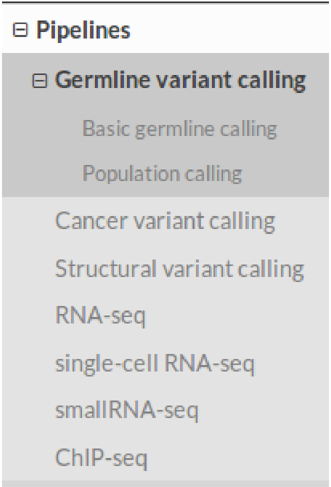
http://www.open-bio.org/wiki/BOSC_2017

# Overview



https://github.com/chapmanb/bcbio-nextgen

# Supported analysis types



⊟ **Pipelines**

- ⊟ **Germline variant calling**
  - Basic germline calling
  - Population calling
- Cancer variant calling
- Structural variant calling
- RNA-seq
- single-cell RNA-seq
- smallRNA-seq
- ChIP-seq

https://bcbio-nextgen.readthedocs.org/en/latest/contents/pipelines.html

# We made a pipeline – so what?

> *There have been a number of previous efforts to create publicly available analysis pipelines for high throughput sequencing data. Examples include Omics-Pipe, bcbio-nextgen, TREVA and NGSane. These pipelines offer a comprehensive, automated process that can analyse raw sequencing reads and produce annotated variant calls. However, the main audience for these pipelines is the research community. Consequently, there are many features required by clinical pipelines that these examples do not fully address. Other groups have focused on improving specific features of clinical pipelines. The Churchill pipeline uses specialised techniques to achieve high performance, while maintaining reproducibility and accuracy. However it is not freely available to clinical centres and it does not try to improve broader clinical aspects such as detailed quality assurance reports, robustness, reports and specialised variant filtering. The Mercury pipeline offers a comprehensive system that addresses many clinical needs: it uses an automated workflow system (Valence) to ensure robustness, abstract computational resources and simplify customisation of the pipeline. Mercury also includes detailed coverage reports provided by ExCID, and supports compliance with US privacy laws (HIPAA) when run on DNANexus, a cloud computing platform specialised for biomedical users. Mercury offers a comprehensive solution for clinical users, however it does not achieve our desired level of transparency, modularity and simplicity in the pipeline specification and design. Further, Mercury does not perform specialised variant filtering and prioritisation that is specifically tuned to the needs of clinical users.*

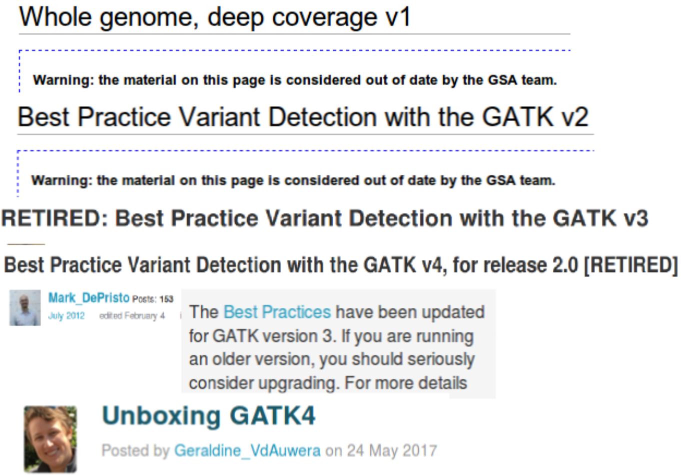
http://www.genomemedicine.com/content/7/1/68

# Sustainability

A piece of software is being sustained if people are using it, fixing it, and improving it rather than replacing it.

http://software-carpentry.org/blog/2014/08/sustainability.html

# Complex, rapidly changing baseline functionality



Whole genome, deep coverage v1

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v2

Warning: the material on this page is considered out of date by the GSA team.

RETIRED: Best Practice Variant Detection with the GATK v3

Best Practice Variant Detection with the GATK v4, for release 2.0 [RETIRED]

Mark_DePristo Posts: 153
July 2012 edited February 4

The Best Practices have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

Unboxing GATK4

Posted by Geraldine_VdAuwera on 24 May 2017

# Feature support burden

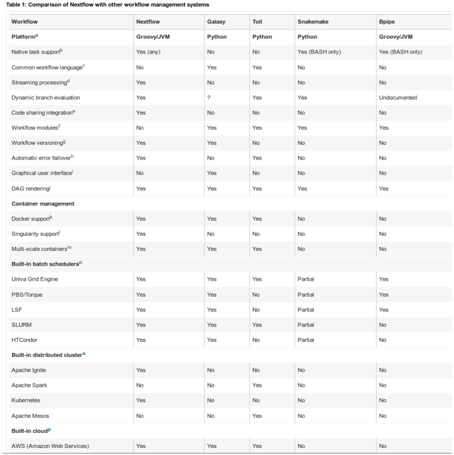
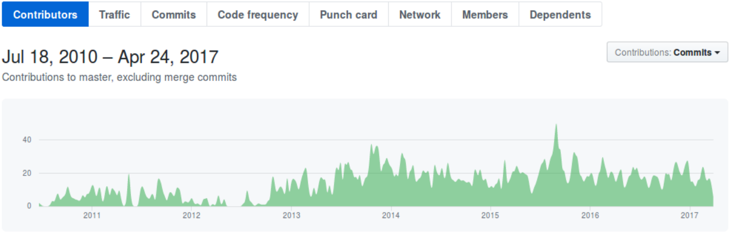


**Table 1: Comparison of Nextflow with other workflow management systems**

| **Workflow** | **Nextflow** | **Galaxy** | **Toil** | **Snakemake** | **Bpipe** |
|---|---|---|---|---|---|
| **Platform**[a] | **Groovy/JVM** | **Python** | **Python** | **Python** | **Groovy/JVM** |
| Native task support[b] | Yes (any) | No | No | Yes (BASH only) | Yes (BASH only) |
| Common workflow language[c] | No | Yes | Yes | No | No |
| Streaming processing[d] | Yes | No | No | No | No |
| Dynamic branch evaluation | Yes | ? | Yes | Yes | Undocumented |
| Code sharing integration[e] | Yes | No | No | No | No |
| Workflow modules[f] | No | Yes | Yes | Yes | Yes |
| Workflow versioning[g] | Yes | Yes | No | No | No |
| Automatic error failover[h] | Yes | No | Yes | No | No |
| Graphical user interface[i] | No | Yes | No | No | No |
| DAG rendering[j] | Yes | Yes | Yes | Yes | Yes |
| **Container management** | | | | | |
| Docker support[k] | Yes | Yes | Yes | No | No |
| Singularity support[l] | Yes | No | No | No | No |
| Multi-scale containers[m] | Yes | Yes | Yes | No | No |
| **Built-in batch schedulers**[n] | | | | | |
| Univa Grid Engine | Yes | Yes | Yes | Partial | Yes |
| PBS/Torque | Yes | Yes | No | Partial | Yes |
| LSF | Yes | Yes | No | Partial | Yes |
| SLURM | Yes | Yes | Yes | Partial | No |
| HTCondor | Yes | Yes | No | Partial | No |
| **Built-in distributed cluster**[o] | | | | | |
| Apache Ignite | Yes | No | No | No | No |
| Apache Spark | No | No | Yes | No | No |
| Kubernetes | Yes | No | No | No | No |
| Apache Mesos | No | No | Yes | No | No |
| **Built-in cloud**[p] | | | | | |
| AWS (Amazon Web Services) | Yes | Yes | Yes | No | No |

http://www.nature.com/nbt/journal/v35/n4/full/nbt.3820.html

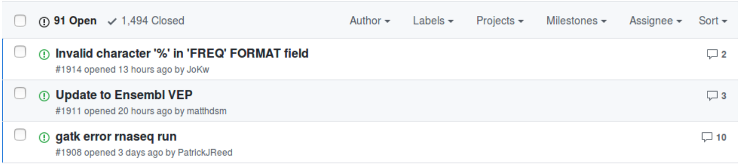
# Community: sustainability and support



Contributions: **Commits**

Jul 18, 2010 – Apr 24, 2017

Contributions to master, excluding merge commits



□ ⓘ **91 Open** ✓ 1,494 Closed

□ ⓘ **Invalid character '%' in 'FREQ' FORMAT field** 2
#1914 opened 13 hours ago by JoKw

□ ⓘ **Update to Ensembl VEP** 3
#1911 opened 20 hours ago by matthdsm

□ ⓘ **gatk error rnaseq run** 10
#1908 opened 3 days ago by PatrickJReed

https://github.com/chapmanb/bcbio-nextgen

# Value of validation

- Integration tests for pipelines
- Unbiased algorithm comparisons
- Baseline for improving methods

# Reference materials

Genome in a Bottle Consortium

Global Alliance for Genomics & Health

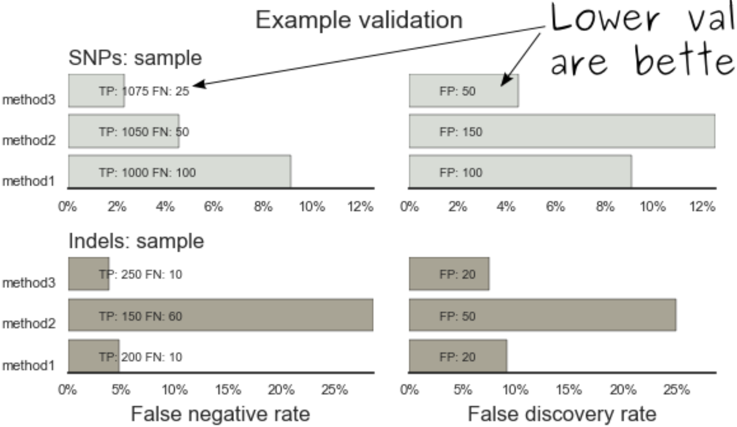
ICGC-TCGA DREAM Mutation Calling challenge

http://www.genomeinabottle.org/
http://ga4gh.org/#/benchmarking-team
https://www.synapse.org/#!Synapse:syn312572

# GATK and Genome in a Bottle

- Collaboration with GATK methods development
- Compare HaplotypeCaller to other methods
- Germline validation
- Genome in a Bottle reference materials
  - NA12878 – Caucasian
  - NA24385 – Ashkenazim Jewish
  - NA24631 – Chinese

# Validation graphs

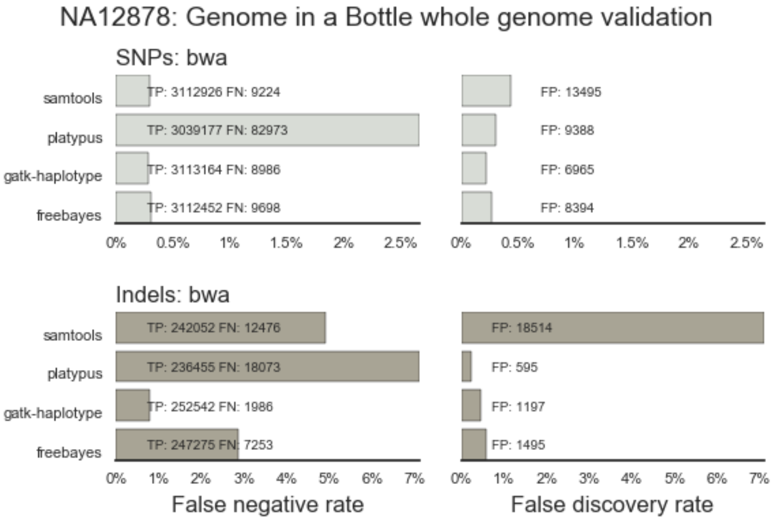


Lower values are better

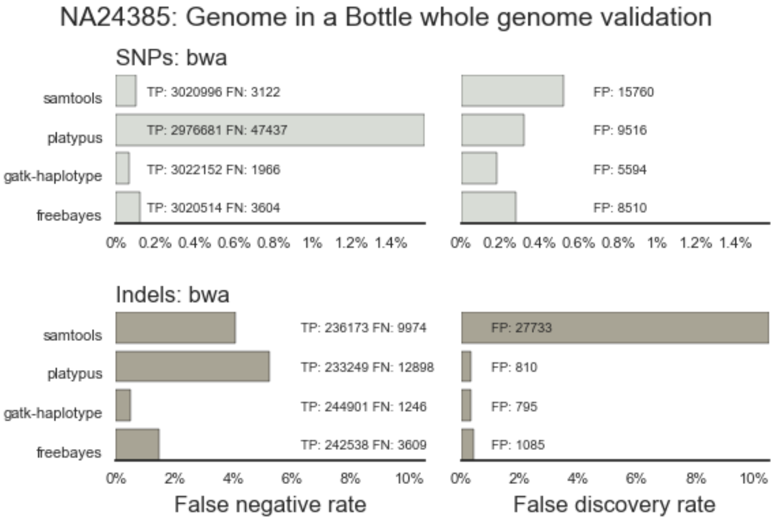
Sensitivity

precision

# NA12878


NA12878: Genome in a Bottle whole genome validation
SNPs: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 3112926 FN: 9224
TP: 3039177 FN: 82973
TP: 3113164 FN: 8986
TP: 3112452 FN: 9698
FP: 13495
FP: 9388
FP: 6965
FP: 8394
0%
0.5%
1%
1.5%
2%
2.5%
Indels: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 242052 FN: 12476
TP: 236455 FN: 18073
TP: 252542 FN: 1986
TP: 247275 FN: 7253
FP: 18514
FP: 595
FP: 1197
FP: 1495
0%
1%
2%
3%
4%
5%
6%
7%
False negative rate
False discovery rate

# NA24385


NA24385: Genome in a Bottle whole genome validation
SNPs: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 3020996 FN: 3122
TP: 2976681 FN: 47437
TP: 3022152 FN: 1966
TP: 3020514 FN: 3604
FP: 15760
FP: 9516
FP: 5594
FP: 8510
0% 0.2% 0.4% 0.6% 0.8% 1% 1.2% 1.4%
0% 0.2% 0.4% 0.6% 0.8% 1% 1.2% 1.4%
Indels: bwa
samtools
platypus
gatk-haplotype
freebayes
TP: 236173 FN: 9974
TP: 233249 FN: 12898
TP: 244901 FN: 1246
TP: 242538 FN: 3609
FP: 27733
FP: 810
FP: 795
FP: 1085
0% 2% 4% 6% 8% 10%
0% 2% 4% 6% 8% 10%
False negative rate
False discovery rate

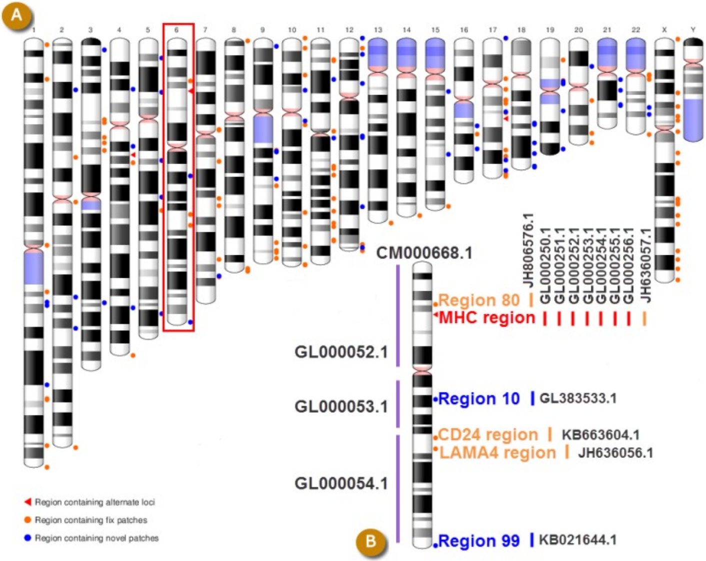
## Validation results

- Good performance for GATK HaplotypeCaller
- Other good performing callers: FreeBayes
- Consistency across diverse samples
- Identify potential problem areas for tuning
  - samtools Indel false positive rates
  - Platypus SNP sensitivity
- PrecisionFDA: https://precision.fda.gov/

# Outline: Science

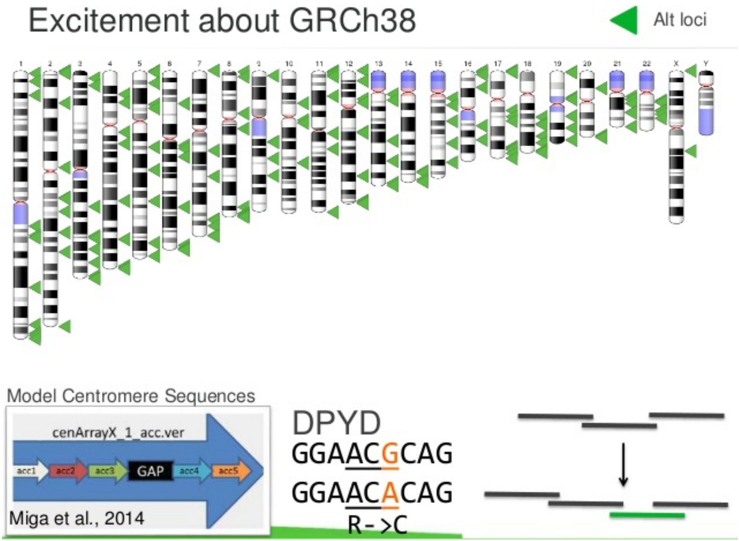
- **Human build 38**
- GATK4 validation
- Low frequency somatic calling
- Structural variation

# GRCh37/hg19



http://www.ncbi.nlm.nih.gov/books/NBK153600/?report=reader

# GRCh38 – graph based, many more alternative loci



Excitement about GRCh38

Alt loci

Model Centromere Sequences

Miga et al., 2014

DPYD

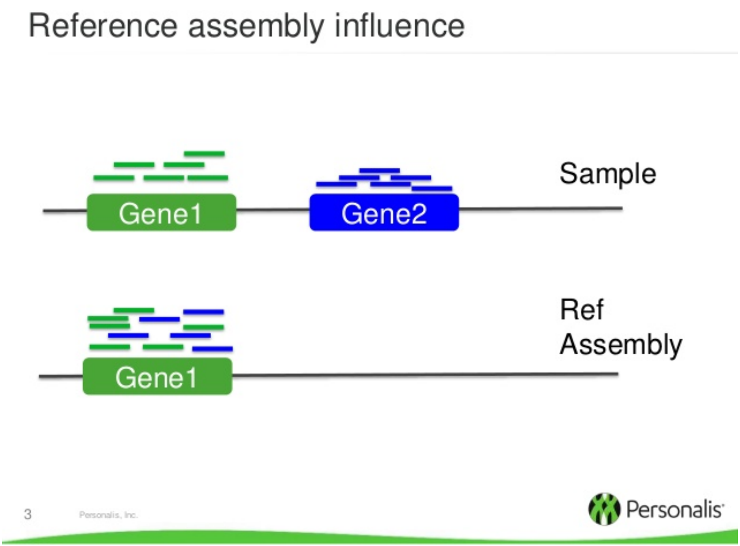
GGAACGCAG

GGAACACAG

R->C

http://www.slideshare.net/GenomeRef/transitioning-to-grch38

# GRCh38 – advantage for variant calling



## Reference assembly influence

Gene1

Gene2

Sample

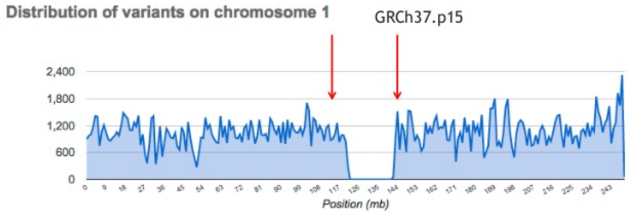
Gene1

Ref Assembly

Personalis

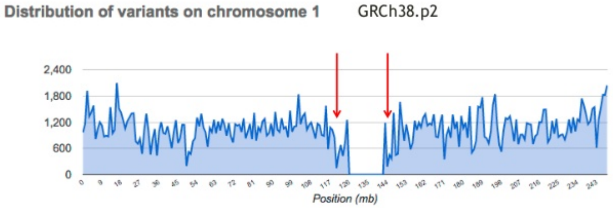
http://www.slideshare.net/GenomeRef/transitioning-to-grch38

# Avoiding collapsed repeats



Distribution of variants on chromosome 1

GRCh37.p15



Distribution of variants on chromosome 1

GRCh38.p2

http://www.slideshare.net/kmsteinberg/

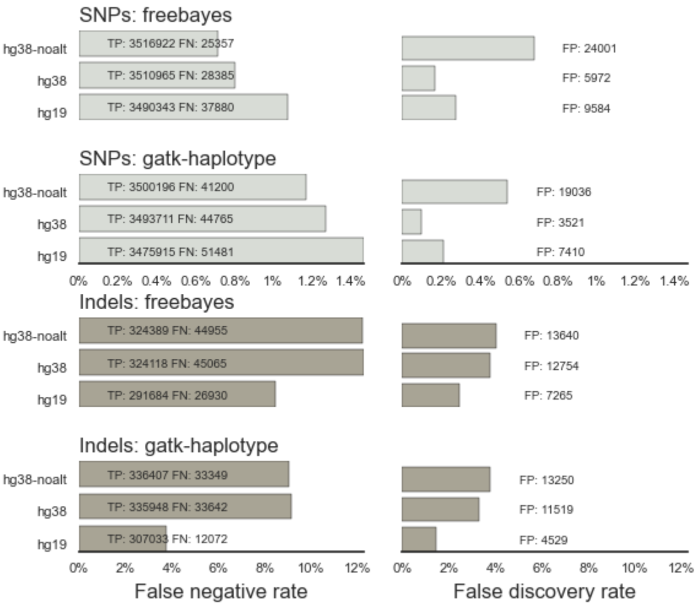
the-importance-of-high-quality-reference-genome-assemblies-to-personal-and-medical-genomics

# Comparison

- Build 37 and 38
- Validation sets: Genome in a Bottle, Illumina Platinum Genomes
- 38 builds: with/without alternative alleles
- Variant callers: FreeBayes, GATK HaplotypeCaller

http://bcb.io/2015/09/17/hg38-validation/

# hg19/hg38 comparison: NA12878 Platinum Genomes



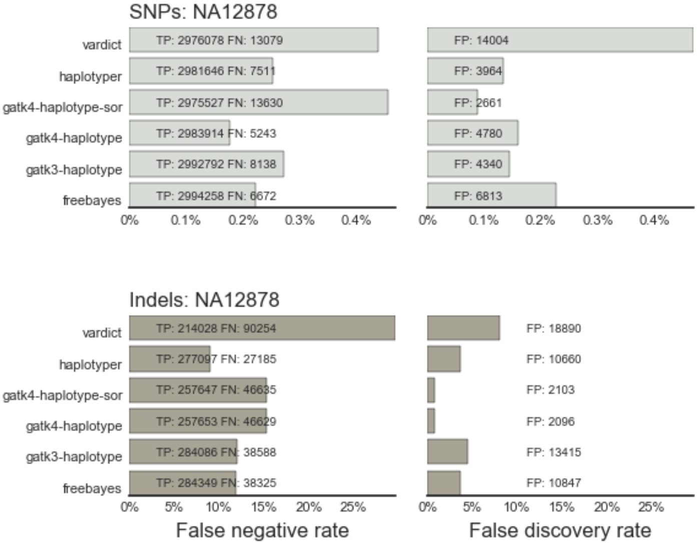
# Small variant results

- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

# Outline: science

- Human build 38
- **GATK4 validation**
- Low frequency somatic calling
- Structural variation

# NA12878 hg38: GATK4



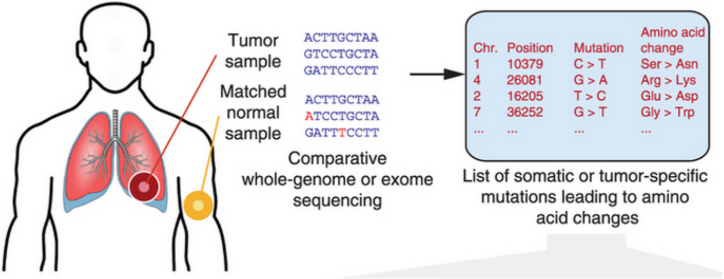
# GATK4 validation results

- Comparable sensitivity and specificity to GATK3
- Removed a recommended filter
  - Strand Odds Ratio (SOR) – strand bias
  - Improves sensitivity
  - ~6000 TPs vs ~2000 FPs
- Indels in GATK need additional tuning
  - Sensitivity/specificity tradeoff
  - ~26k TPs vs ~11k FPs

# Outline: science

- Human build 38
- GATK4 validation
- **Low frequency somatic calling**
- Structural variation

# Cancer somatic calling

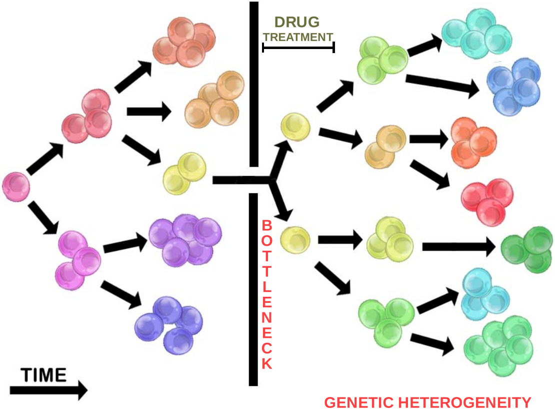


| Chr. | Position | Mutation | Amino acid change |
|---|---|---|---|
| 1 | 10379 | C > T | Ser > Asn |
| 4 | 26081 | G > A | Arg > Lys |
| 2 | 16205 | T > C | Glu > Asp |
| 7 | 36252 | G > T | Gly > Trp |
| ... | ... | ... | ... |

List of somatic or tumor-specific mutations leading to amino acid changes

http://www.nature.com/nmeth/journal/v10/n8/fig_tab/nmeth.2562_F1.html

# Cancer heterogeneity



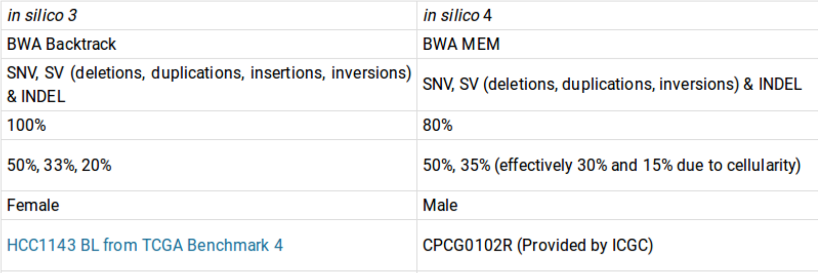
http://en.wikipedia.org/wiki/Tumour_heterogeneity

# VarDict

- AstraZeneca
- Germline + Cancer calling
- SNP + Insertion/Deletions
- Whole genome + exome
- Also works on deep targeted data

https://github.com/AstraZeneca-NGS/VarDictJava
http://nar.oxfordjournals.org/content/early/2016/04/07/nar.gkw227.full

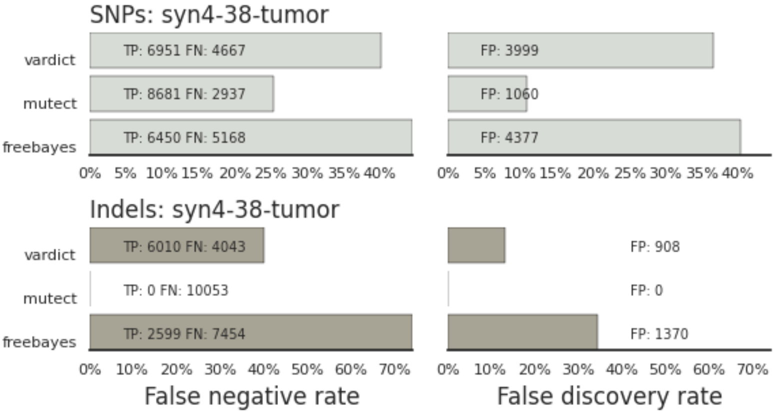
# DREAM synthetic dataset 4



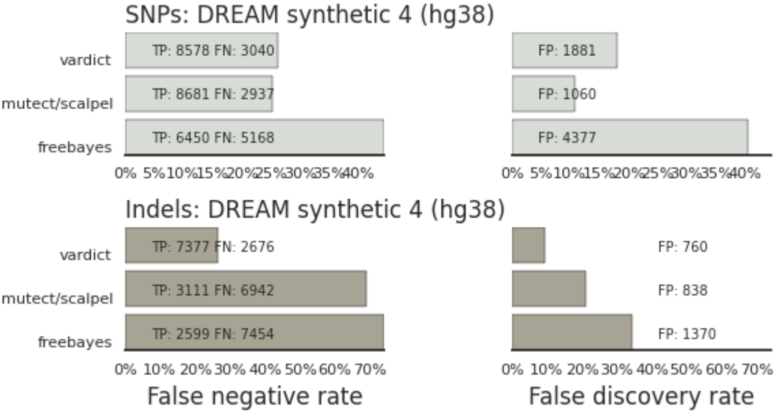
| *in silico 3* | *in silico* 4 |
|---|---|
| BWA Backtrack | BWA MEM |
| SNV, SV (deletions, duplications, insertions, inversions) & INDEL | SNV, SV (deletions, duplications, inversions) & INDEL |
| 100% | 80% |
| 50%, 33%, 20% | 50%, 35% (effectively 30% and 15% due to cellularity) |
| Female | Male |
| HCC1143 BL from TCGA Benchmark 4 | CPCG0102R (Provided by ICGC) |

https://www.synapse.org/#!Synapse:syn312572/wiki/62018

# VarDict sensivitity/precision before



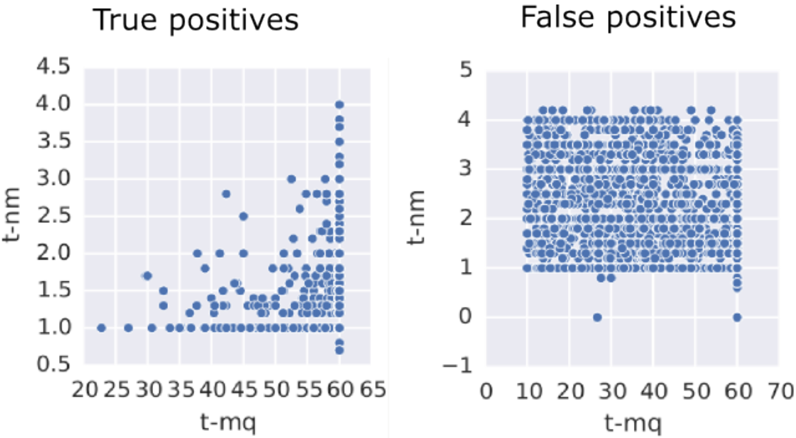
# VarDict sensivitity/precision after



# How? Filter summary

```
((AF * DP < 6) &&
 ((MQ < 55.0 && NM > 1.0) ||
  (MQ < 60.0 && NM > 2.0) ||
  (DP < 10) ||
  (QUAL < 45)))
```

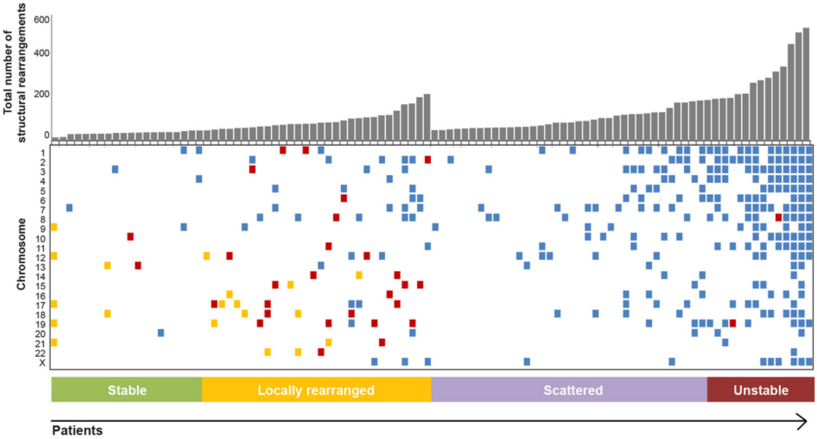
# Example filter: mapping quality and number of mismatches



# Outline: science

- Human build 38
- GATK4 validation
- Low frequency somatic calling
- **Structural variation**

# Structural variants critical in cancer

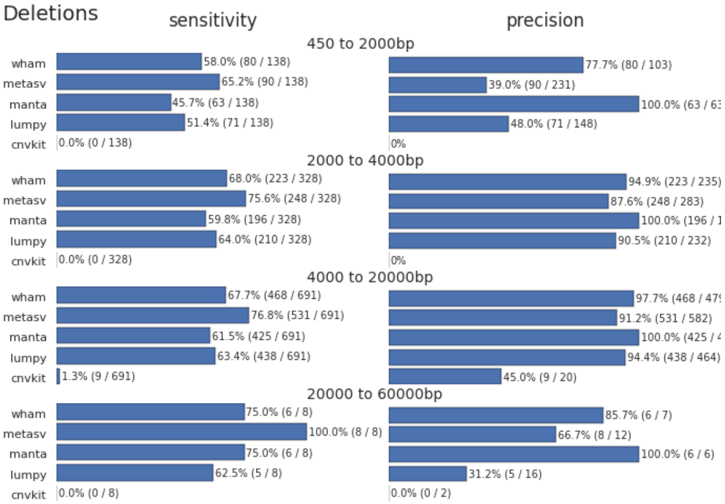


http://www.nature.com/nature/journal/v518/n7540/full/nature14169.html

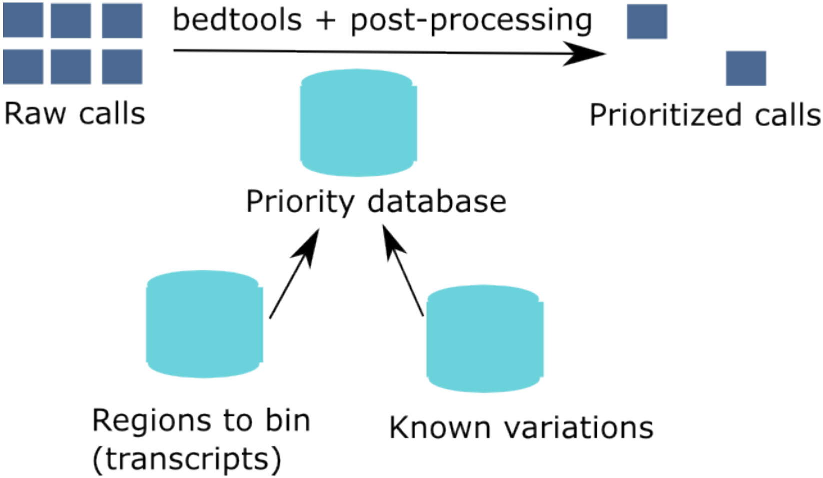
# Improvements in speed, sensitivity and precision

- Manta: https://github.com/Illumina/manta
- CNVkit: https://github.com/etal/cnvkit
- Lumpy: https://github.com/arq5x/lumpy-sv
- WHAM: https://github.com/zeeev/wham
- MetaSV: https://github.com/bioinform/metasv

# Results: Somatic deletions

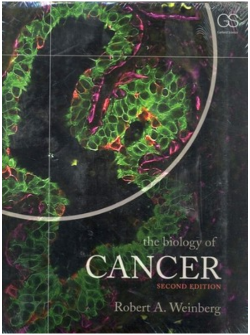


# Prioritize in previously known regions



# Public cancer variant databases

- CIViC: https://civic.genome.wustl.edu
- IntOGen: http://www.intogen.org



http://www.amazon.com/The-Biology-Cancer-Robert-Weinberg/dp/0815340761

# Practical overview

- Small dataset – single chromosome, exome
- Cancer sample from DREAM synthetic dataset 3
- Call against build 38

https://www.synapse.org/#!Synapse:syn312572/wiki/58893

# Demonstrates

- Somatic tumor/normal samples
- SNP and indel calling at lower frequency
- Structural variant detection
- Prioritization with CIViC
- HLA typing

# bcbio configuration file

```yaml
---
details:
  - analysis: variant2
    genome_build: hg38
    algorithm:
      aligner: bwa
      mark_duplicates: true
      recalibrate: false
      realign: false
      variantcaller: [vardict, mutect, freebayes]
      ensemble:
        numpass: 2
      svcaller: [lumpy, manta]
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html

# bcbio template file – CSV

```
samplename,description,batch,phenotype,sex,variant_regions
sample1,ERR256785,batch1,normal,female,/path/to/regions.bed
sample2,ERR256786,batch1,tumor,,/path/to/regions.bed
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html#automated-sample-configuration

# Template to full configuration

```
bcbio_nextgen.py -w template \
   tumor-paired.yaml project1.csv \
   sample1.bam sample2_1.fq sample2_2.fq
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/configuration.html#automated-sample-configuration

# Running bcbio

```
bcbio_nextgen.py project1.yaml -n 8
```

https://bcbio-nextgen.readthedocs.org/en/latest/contents/testing.html

# AWS example configuration and output

https://bcbio-nextgen.readthedocs.org/en/latest/contents/teaching.html

- Pre-downloaded and analysis run
- AMI (ami-5e84fe34)

# Summary

- Overview of variant calling tools
- Open source community resources
- bcbio validated variant analysis
- Science
  - Human build 38
  - GATK4 validation
  - Cancer calling of low frequency variants
  - Structural variation
- Practical calling example

http://bcb.io