Variant calling with validated, scalable, community developed tools

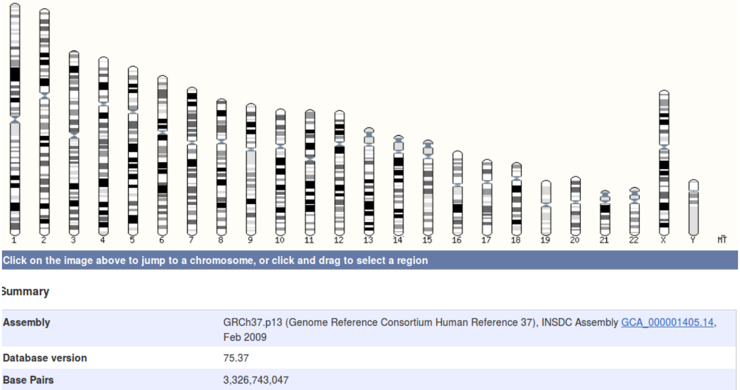
Brad Chapman
Bioinformatics Core, Harvard Chan School
https://bcb.io
http://j.mp/bcbiolinks

11 August 2016

Outline

- Overview of variant calling
- Motivate for using open source community resources
- bcbio validated variant calling
- Science
  - Human build 38 + HLA
  - Cancer calling of low frequency variants
  - Structural variation
- Practical calling example

Human whole genome sequencing



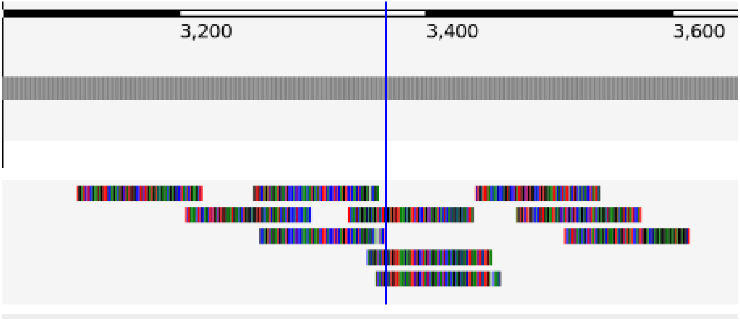
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

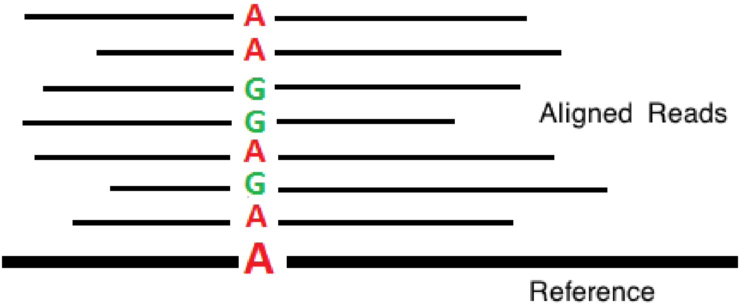
Assembly	GRCh37.p13 (Genome Reference Consortium Human Reference 37), INSDC Assembly GCA_000001405.14, Feb 2009
Database version	75.37
Base Pairs	3,326,743,047

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

High throughput sequencing

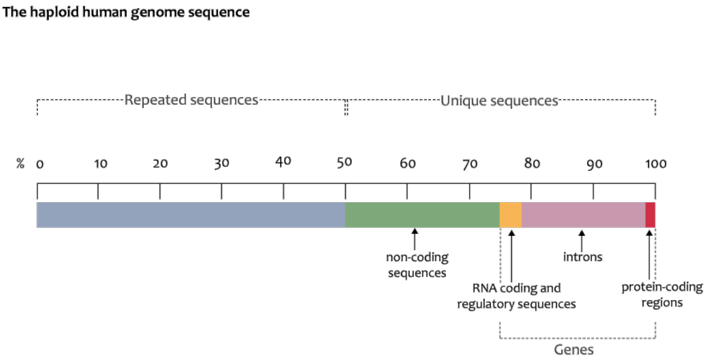


Variant calling



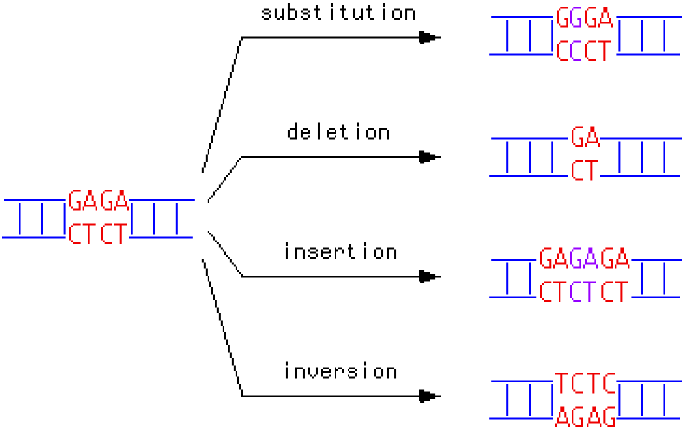
http://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

Scale: exome to whole genome



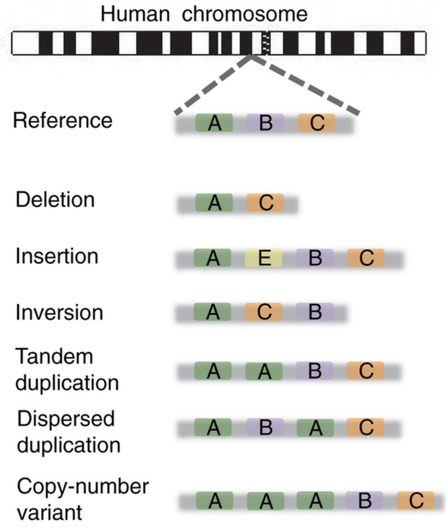
https://www.flickr.com/photos/119980645@N06/

SNPs and Indels



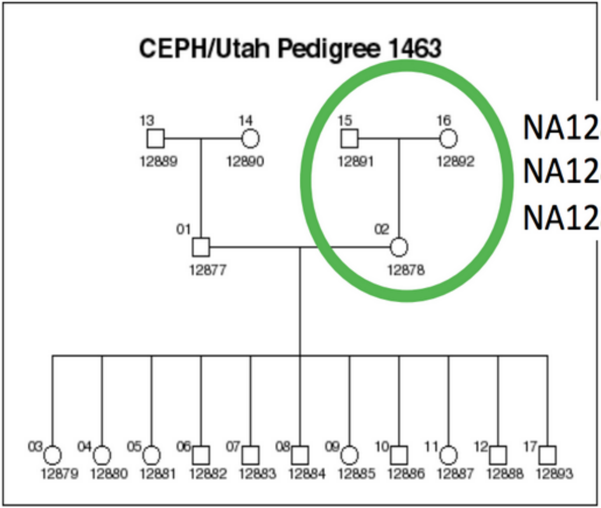
http://carolguze.com/text/442-2-mutations.shtml

Structural variations



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

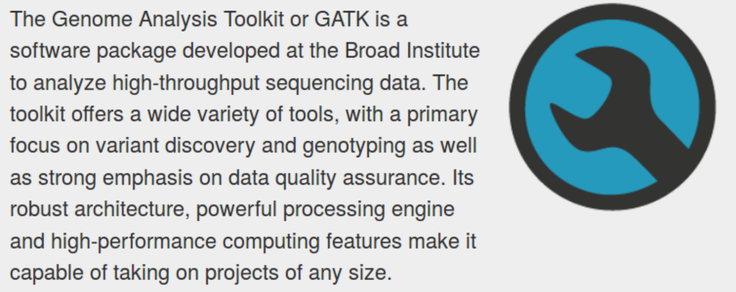
Germline population calling



NA12891
NA12892
NA12878

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

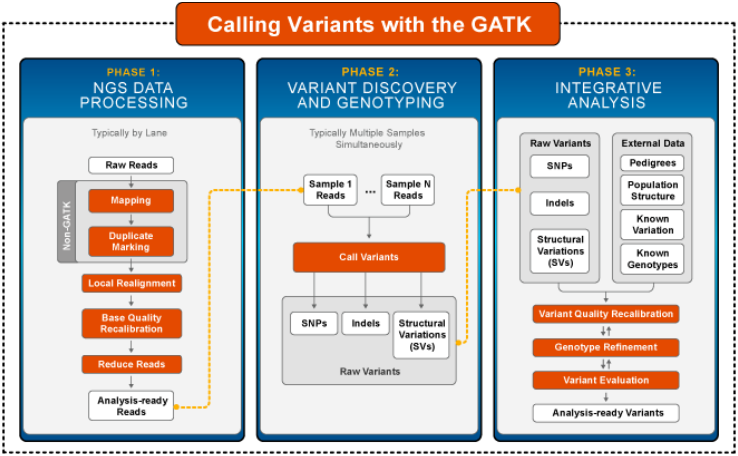
Genome Analysis Toolkit (GATK)



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.

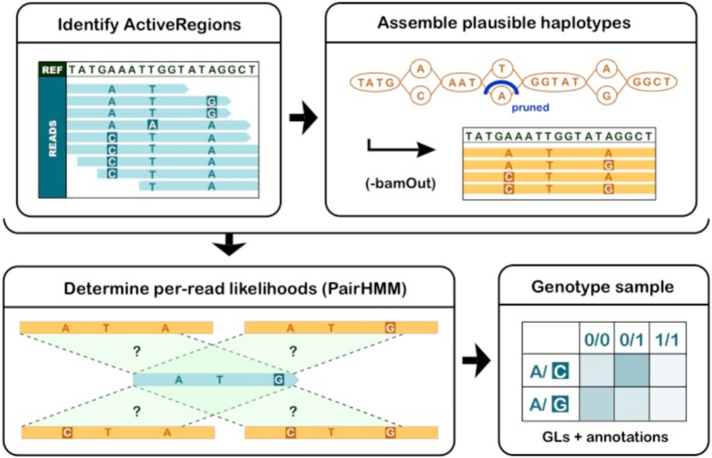
`https://www.broadinstitute.org/gatk/`

GATK Best Practices



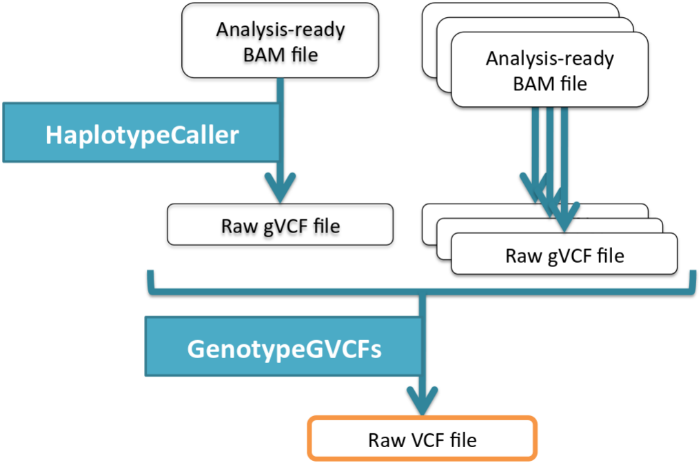
https://www.broadinstitute.org/gatk/guide/best-practices

HaplotypeCaller



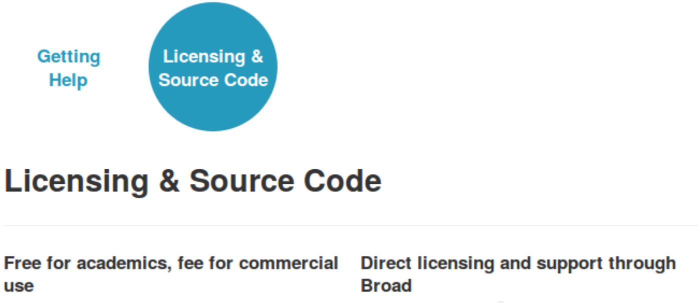
http://gatkforums.broadinstitute.org/discussion/5464/workshop-presentations-2015-uk-4-20-24

Joint calling on large populations



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

Support and availability



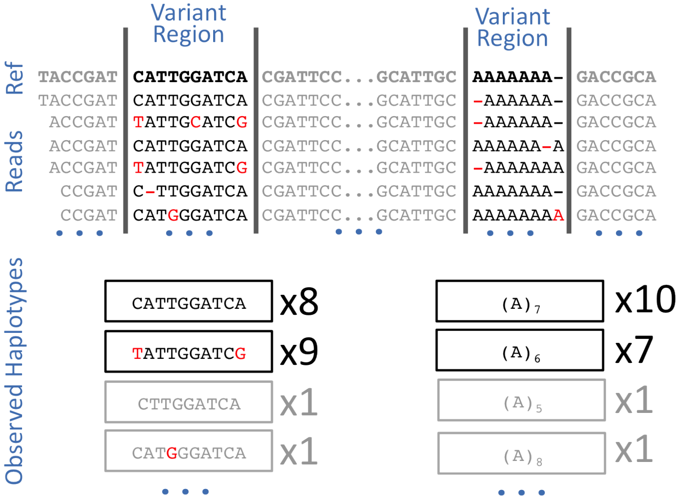
Licensing & Source Code

Free for academics, fee for commercial use

Direct licensing and support through Broad

https://github.com/broadgsa

FreeBayes



Variant Region

Variant Region

Ref

Reads

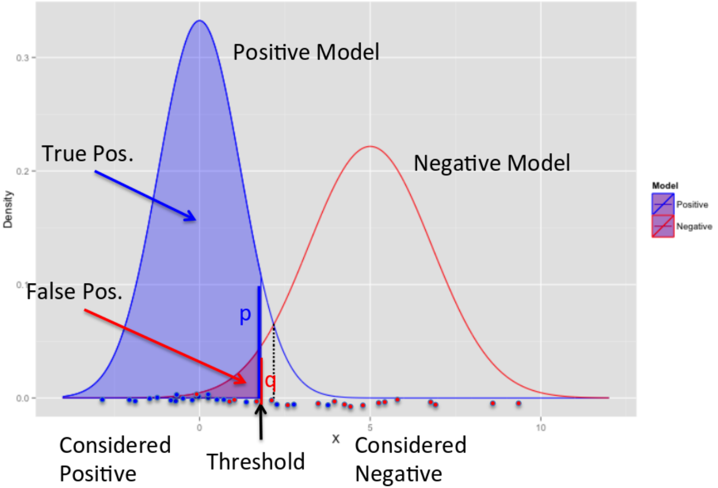
```
TACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
TACCGAT CATTGGATCA CGATTCC...GCATTGC -AAAAAA- GACCGCA
 ACCGAT TATTGCATCG CGATTCC...GCATTGC -AAAAAA- GACCGCA
 ACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAA-A GACCGCA
 ACCGAT TATTGGATCG CGATTCC...GCATTGC -AAAAAAA GACCGCA
  CCGAT C-TTGGATCA CGATTCC...GCATTGC 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

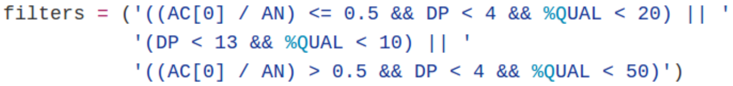
Filtering – Variant Quality Score Recalibration



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

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

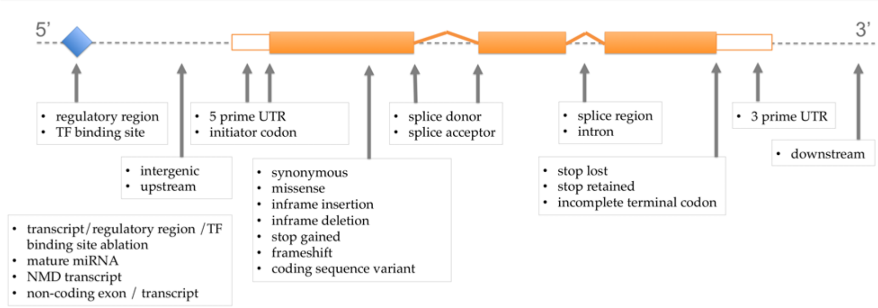
Filtering – hard cutoffs



```
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



- regulatory region
- TF binding site

- intergenic
- upstream

- transcript/regulatory region /TF binding site ablation
- mature miRNA
- NMD transcript
- non-coding exon / transcript

- 5 prime UTR
- initiator codon

- synonymous
- missense
- inframe insertion
- inframe deletion
- stop gained
- frameshift
- coding sequence variant

- splice donor
- splice acceptor

- splice region
- intron

- stop lost
- stop retained
- incomplete terminal codon

- 3 prime UTR

- downstream

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

Tools for effects predictions

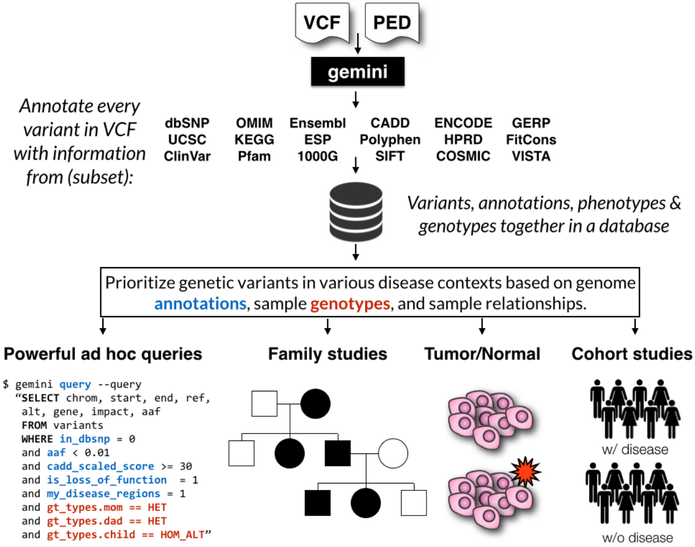
- snpEff

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”
```

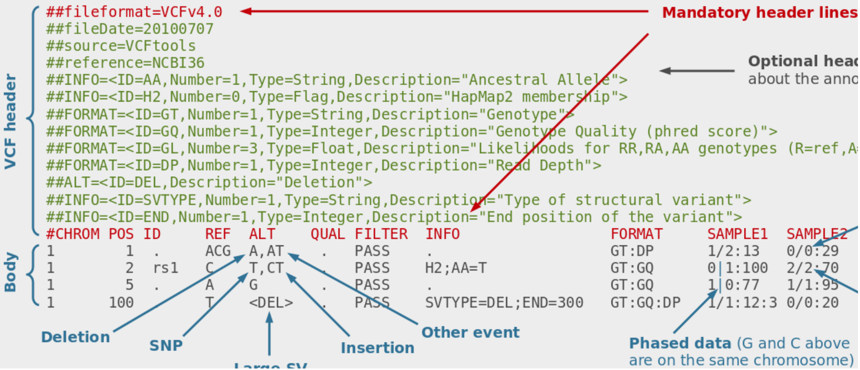
Family studies

Tumor/Normal

Cohort studies

https://github.com/arq5x/gemini

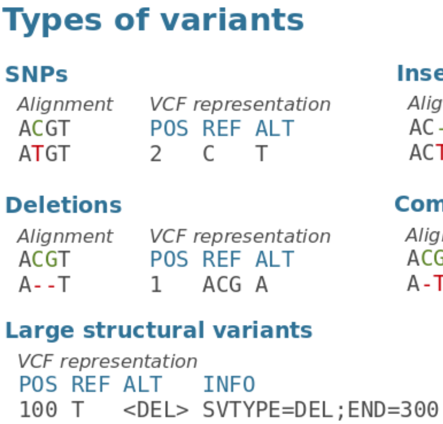
VCF – overview



```
##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

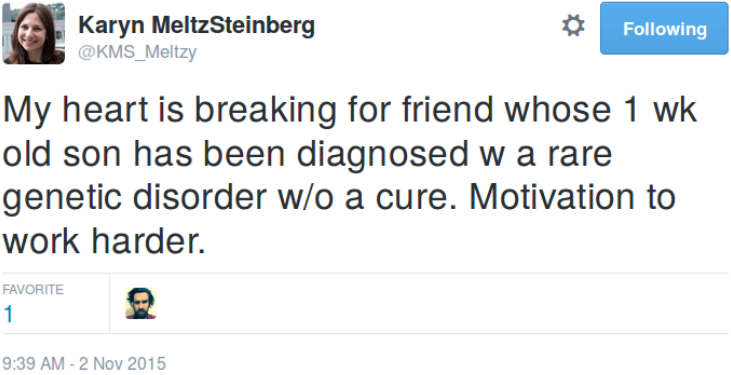
- Step by step guide from Broad

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

- Specification

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

We need to do science faster



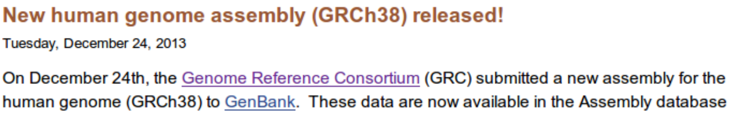
Karyn MeltzSteinberg
@KMS_Meltzy

My heart is breaking for friend whose 1 wk old son has been diagnosed w a rare genetic disorder w/o a cure. Motivation to work harder.

9:39 AM - 2 Nov 2015

https://twitter.com/KMS_Meltzy/status/661206070308794368

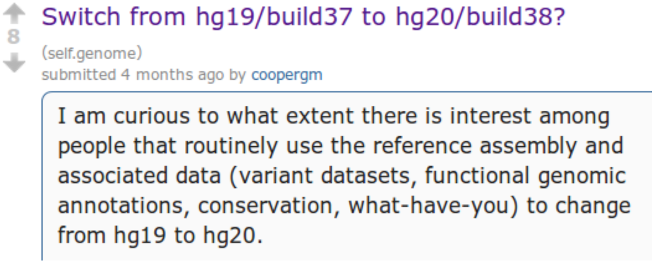
We need to incorporate improvements faster



New human genome assembly (GRCh38) released!

Tuesday, December 24, 2013

On December 24th, the Genome Reference Consortium (GRC) submitted a new assembly for the human genome (GRCh38) to GenBank. These data are now available in the Assembly database



Switch from hg19/build37 to hg20/build38?

(self.genome)

submitted 4 months ago by coopergm

I am curious to what extent there is interest among people that routinely use the reference assembly and associated data (variant datasets, functional genomic annotations, conservation, what-have-you) to change from hg19 to hg20.

https://www.reddit.com/r/genome/comments/3b3s3t/switch_from_hg19build37_to_hg20build38/

Daily bioinformatics work

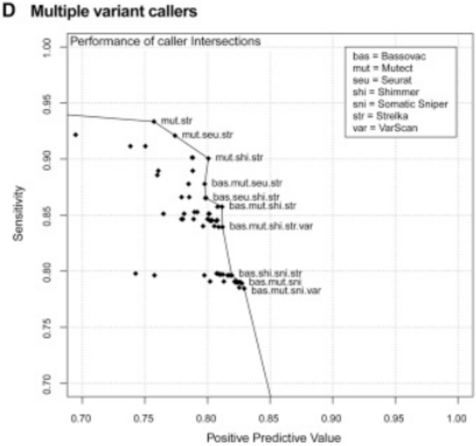
- Install tools
- Put tools together
- Test and validate
- Scale
- Improve algorithms
- Read literature
- Do biology

Standard analyses not routine

> *Four major genome centers predicted single-nucleotide variants (SNVs) for The Cancer Genome Atlas (TCGA) lung cancer samples, but only 31.0% (1,667/5,380) of SNVs were identified by all four.*

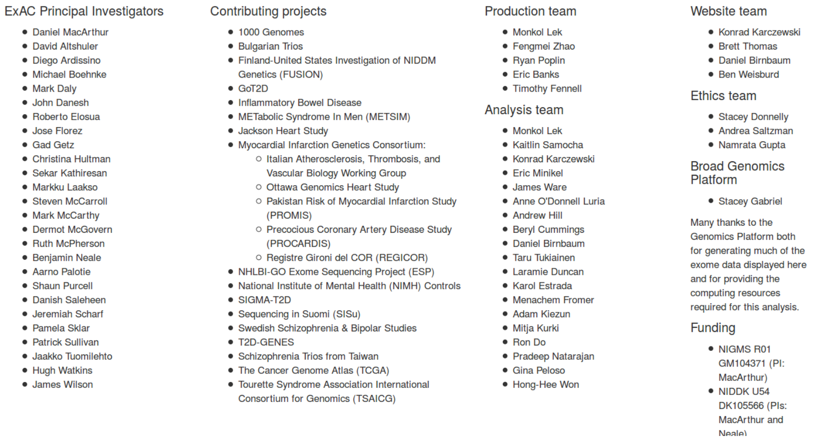
http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.3407.html

Combining analyses = better results



http://www.cell.com/cell-systems/abstract/S2405-4712%2815%2900113-1

Working together produces great things



ExAC Principal Investigators

- Daniel MacArthur
- David Altshuler
- Diego Ardissino
- Michael Boehnke
- Mark Daly
- John Danesh
- Roberto Elosua
- Jose Florez
- Gad Getz
- Christina Hultman
- Sekar Kathiresan
- Markku Laakso
- Steven McCarroll
- Mark McCarthy
- Dermot McGovern
- Ruth McPherson
- Benjamin Neale
- Aarno Palotie
- Shaun Purcell
- Danish Saleheen
- Jeremiah Scharf
- Pamela Sklar
- Patrick Sullivan
- Jaakko Tuomilehto
- Hugh Watkins
- James Wilson

Contributing projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
    - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
    - Ottawa Genomics Heart Study
    - Pakistan Risk of Myocardial Infarction Study (PROMIS)
    - Precocious Coronary Artery Disease Study (PROCARDIS)
    - Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP)
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

Production team

- Monkol Lek
- Fengmei Zhao
- Ryan Poplin
- Eric Banks
- Timothy Fennell

Analysis team

- Monkol Lek
- Kaitlin Samocha
- Konrad Karczewski
- Eric Minikel
- James Ware
- Anne O'Donnell Luria
- Andrew Hill
- Beryl Cummings
- Daniel Birnbaum
- Taru Tukiainen
- Laramie Duncan
- Karol Estrada
- Menachem Fromer
- Adam Kiezun
- Mitja Kurki
- Ron Do
- Pradeep Natarajan
- Gina Peloso
- Hong-Hee Won

Website team

- Konrad Karczewski
- Brett Thomas
- Daniel Birnbaum
- Ben Weisburd

Ethics team

- Stacey Donnelly
- Andrea Saltzman
- Namrata Gupta

Broad Genomics Platform

- Stacey Gabriel

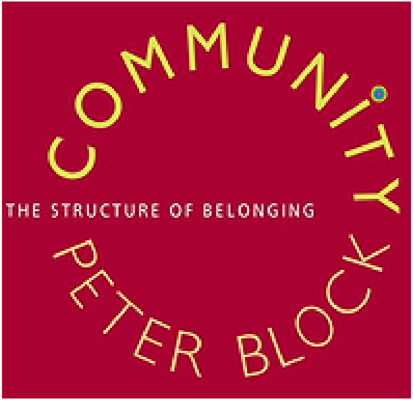
Many thanks to the Genomics Platform both for generating much of the exome data displayed here and for providing the computing resources required for this analysis.

Funding

- NIGMS R01 GM104371 (PI: MacArthur)
- NIDDK U54 DK105566 (PIs: MacArthur and Neale)

http://exac.broadinstitute.org/about

Solution



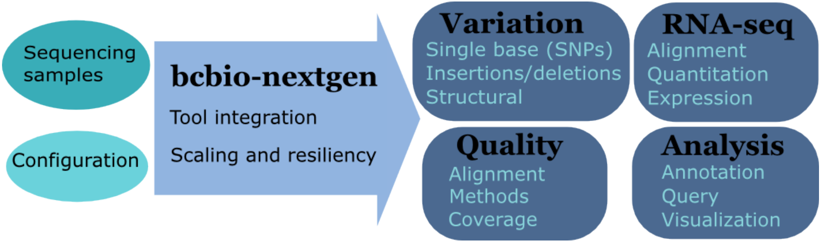
http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

Large scale infrastructure development

- Shared problems – academic, industry, startups
- Community developed analyses
- Validation
- Scaling
- Supporting a community of users

White box software

Overview



Sequencing samples

Configuration

bcbio-nextgen

Tool integration

Scaling and resiliency

Variation

Single base (SNPs)

Insertions/deletions

Structural

RNA-seq

Alignment

Quantitation

Expression

Quality

Alignment

Methods

Coverage

Analysis

Annotation

Query

Visualization

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

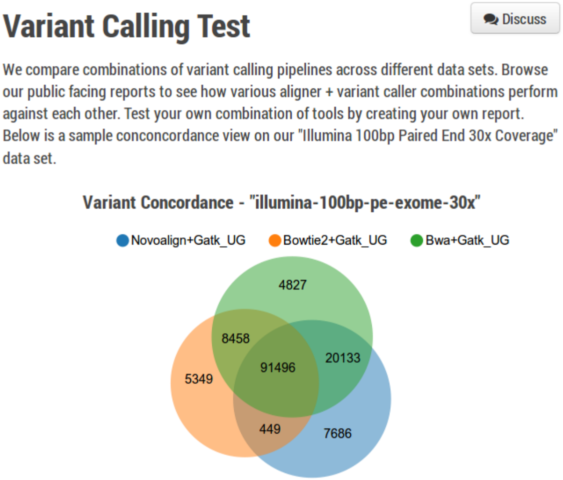
Uses

- Aligners: bwa, novoalign, bowtie2, HISAT2
- Variantion: FreeBayes, GATK, VarDict, MuTecT2, Scalpel, SnpEff, VEP, GEMINI, Lumpy, Manta, CNVkit, WHAM
- RNA-seq: Tophat, STAR, Cufflinks, Sailfish
- Quality control: fastqc, bamtools, Qualimap
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib, vt

Provides

- Community – collected set of expertise
- Installation of tools and data
- Tool integration
- Validation – outputs + automated evaluation
- Scaling

Quality differences between methods



Variant Calling Test

Discuss

We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample conconcordance view on our "Illumina 100bp Paired End 30x Coverage" data set.

Variant Concordance - "illumina-100bp-pe-exome-30x"

http://www.bioplanet.com/gcat

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.*

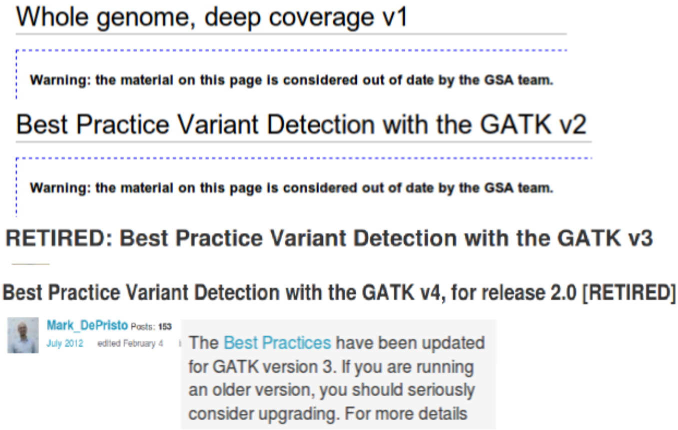
http://www.genomemedicine.com/content/7/1/68

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.

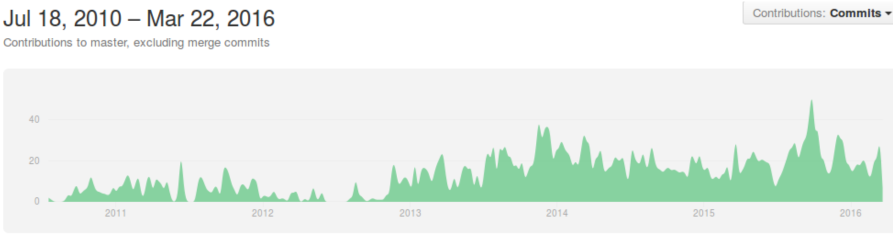
RETIRED: Best Practice Variant Detection with the GATK v3

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

Community: sustainability

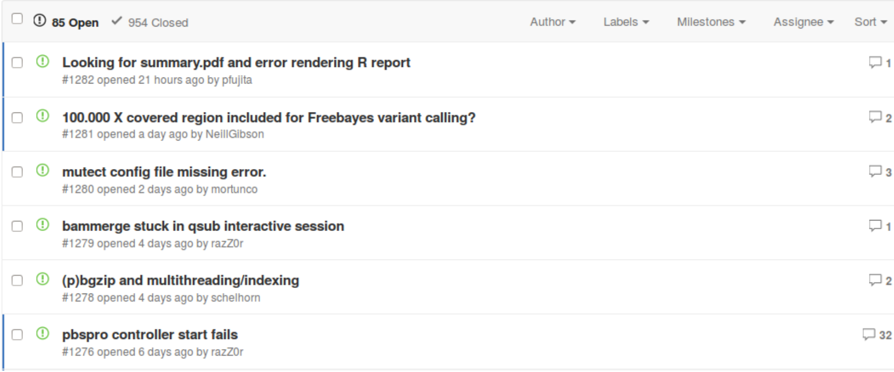


Jul 18, 2010 – Mar 22, 2016

Contributions to master, excluding merge commits

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

Community: support



□ **Looking for summary.pdf and error rendering R report** 1
#1282 opened 21 hours ago by pfujita

□ **100.000 X covered region included for Freebayes variant calling?** 2
#1281 opened a day ago by NeillGibson

□ **mutect config file missing error.** 3
#1280 opened 2 days ago by mortunco

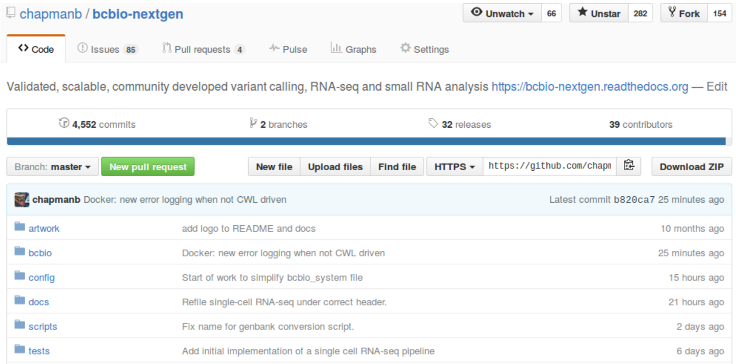
□ **bammerge stuck in qsub interactive session** 1
#1279 opened 4 days ago by razZ0r

□ **(p)bgzip and multithreading/indexing** 2
#1278 opened 4 days ago by schelhorn

□ **pbspro controller start fails** 32
#1276 opened 6 days ago by razZ0r

https://bcbio-nextgen.readthedocs.org

Community: contribution



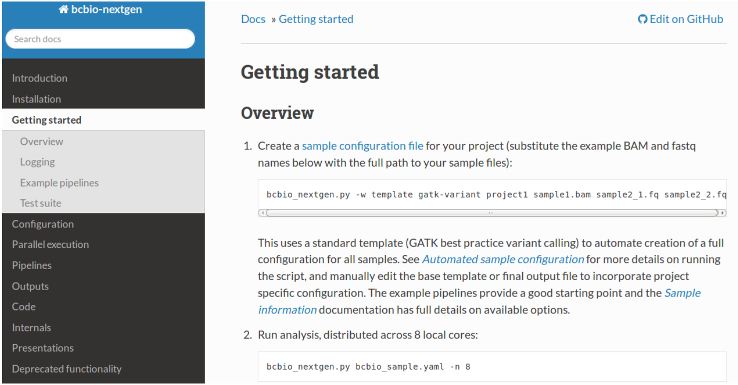
Validated, scalable, community developed variant calling, RNA-seq and small RNA analysis https://bcbio-nextgen.readthedocs.org — Edit

chapmanb Docker: new error logging when not CWL driven — Latest commit b820ca7 25 minutes ago

artwork	add logo to README and docs	10 months ago
bcbio	Docker: new error logging when not CWL driven	25 minutes ago
config	Start of work to simplify bcbio_system file	15 hours ago
docs	Refile single-cell RNA-seq under correct header.	21 hours ago
scripts	Fix name for genbank conversion script.	2 days ago
tests	Add initial implementation of a single cell RNA-seq pipeline	6 days ago

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

Community: documentation



Getting started

Overview

1. Create a sample configuration file for your project (substitute the example BAM and fastq names below with the full path to your sample files):

```
bcbio_nextgen.py -w template gatk-variant project1 sample1.bam sample2_1.fq sample2_2.fq
```

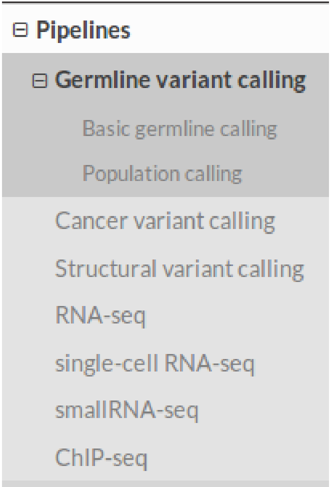
   This uses a standard template (GATK best practice variant calling) to automate creation of a full configuration for all samples. See *Automated sample configuration* for more details on running the script, and manually edit the base template or final output file to incorporate project specific configuration. The example pipelines provide a good starting point and the *Sample information* documentation has full details on available options.

2. Run analysis, distributed across 8 local cores:

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

https://bcbio-nextgen.readthedocs.org

Supported analysis types



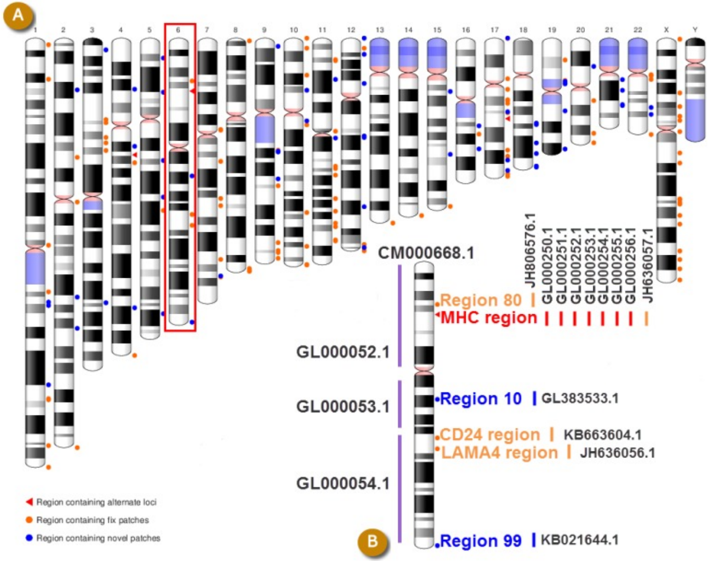
- ⊟ **Pipelines**
  - ⊟ **Germline variant calling**
    - Basic germline calling
    - Population calling
  - Cancer variant calling
  - Structural variant calling
  - RNA-seq
  - single-cell RNA-seq
  - smallRNA-seq
  - ChIP-seq

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

Outline: Science

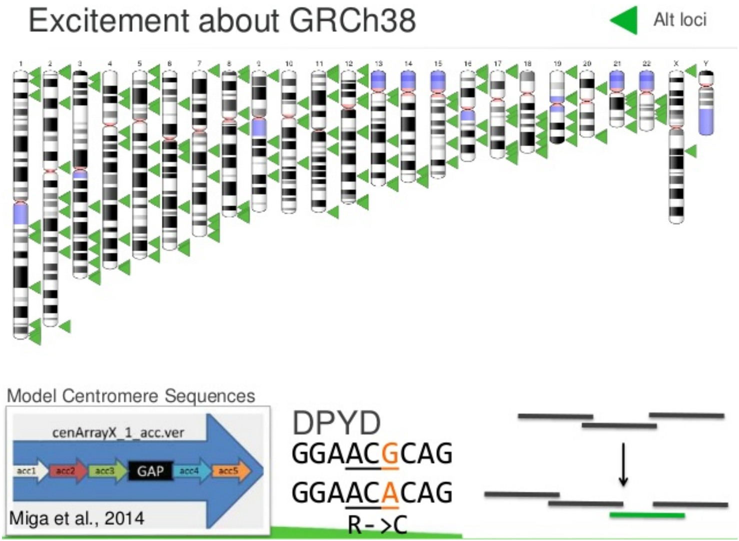
- **Human build 38 + HLA**
- Low frequency somatic calling
- Structural variation

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

cenArrayX_1_acc.ver

Miga et al., 2014

DPYD

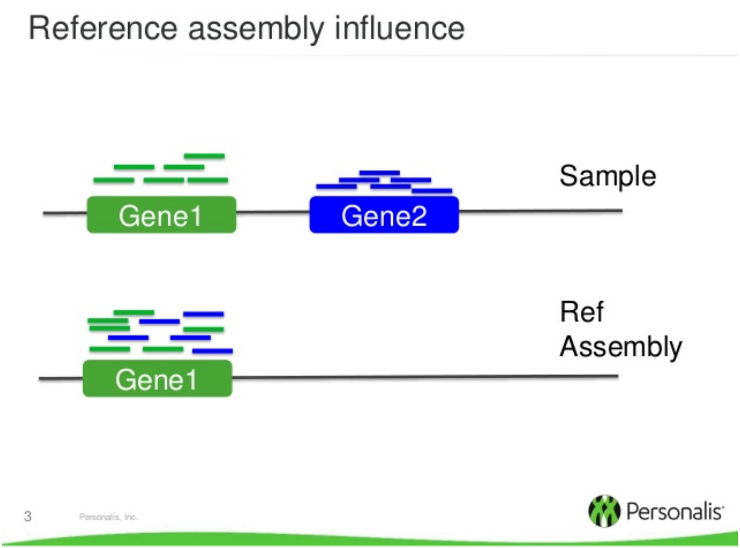
GGAACGCAG

GGAACACAG

R->C

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

GRCh38 – advantage for variant calling



Reference assembly influence

Gene1

Gene2

Sample

Gene1

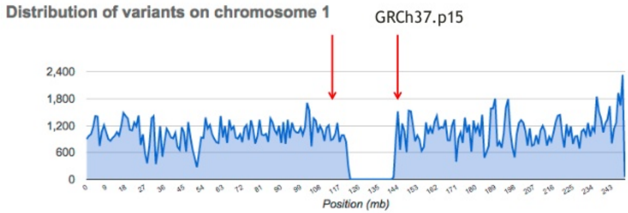
Ref Assembly

3 Personalis, Inc.

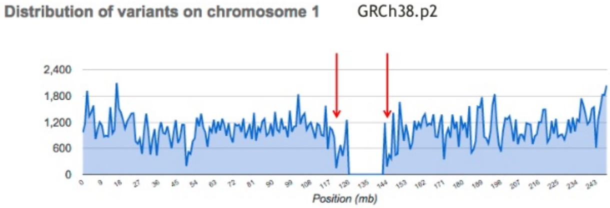
Personalis

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

Avoiding collapsed repeats



GRCh37.p15



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/

Reference materials

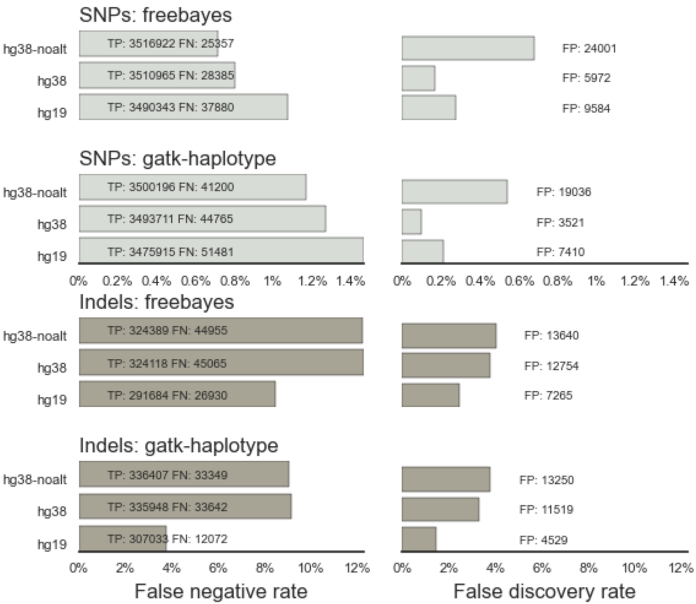
Genome in a Bottle Consortium

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

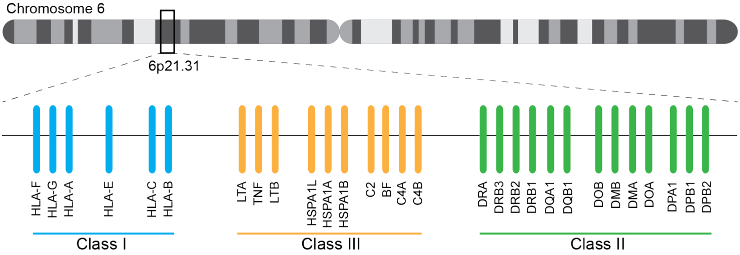
hg19/hg38 comparison: NA12878 Platinum Genomes



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

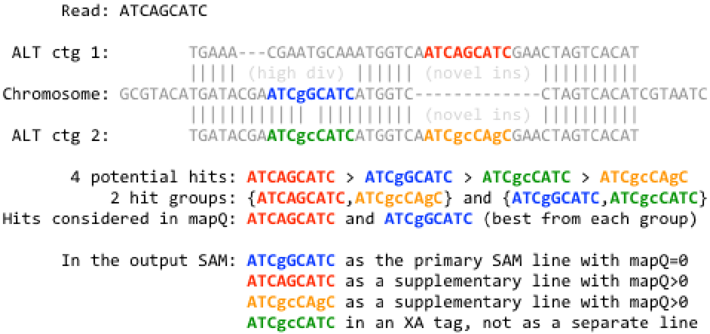
Major histocompatibility complex (MHC) – HLAs



http://www.ebi.ac.uk/ipd/imgt/hla/

http://sciscogenetics.com/technology/human-leukocyte-antigen-complex/

Alignment: bwa alternative allele support



```
      Read: ATCAGCATC

 ALT ctg 1:        TGAAA---CGAATGCAAATGGTCAATCAGCATCGAACTAGTCACAT
                   ||||| (high div) |||||| (novel ins) ||||||||||
Chromosome: GCGTACATGATACGAATCgGCATCATGGTC------------CTAGTCACATCGTAATC
                   |||||||||||| |||||||||| (novel ins) ||||||||||
 ALT ctg 2:        TGATACGAATCgcCATCATGGTCAATCgcCAgCGAACTAGTCACAT

       4 potential hits: ATCAGCATC > ATCgGCATC > ATCgcCATC > ATCgcCAgC
           2 hit groups: {ATCAGCATC,ATCgcCAgC} and {ATCgGCATC,ATCgcCATC}
Hits considered in mapQ: ATCAGCATC and ATCgGCATC (best from each group)

      In the output SAM: ATCgGCATC as the primary SAM line with mapQ=0
                         ATCAGCATC as a supplementary line with mapQ>0
                         ATCgcCAgC as a supplementary line with mapQ>0
                         ATCgcCATC in an XA tag, not as a separate line
```

https://github.com/lh3/bwa/blob/master/README-alt.md

HLA typing

- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa mem extracts HLA reads
- Map reads only to HLA sequences
- OptiType: Call HLA types

https://github.com/lh3/bwa/blob/master/README-alt.md#hla-typing

https://github.com/FRED-2/OptiType

Validations

- Omixon example data
- Exome (1000 genomes) and deep targeted data
- P-group resolution
- HLA type I calls (A, B, C)
- Great results across exome and targeted

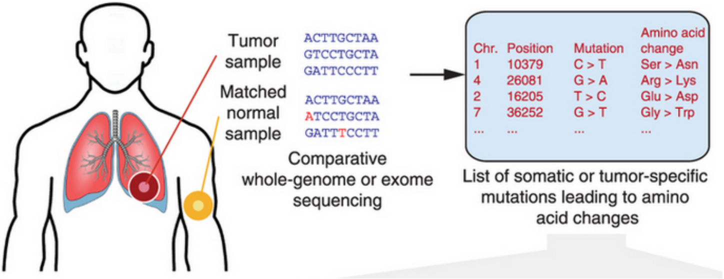
http://www.omixon.com/hla-typing-example-data/

https://gist.github.com/chapmanb/8f994618a7fc5e88f893

Outline: science

- Human build 38 + HLA
- **Low frequency somatic calling**
- Structural variation

Cancer somatic calling

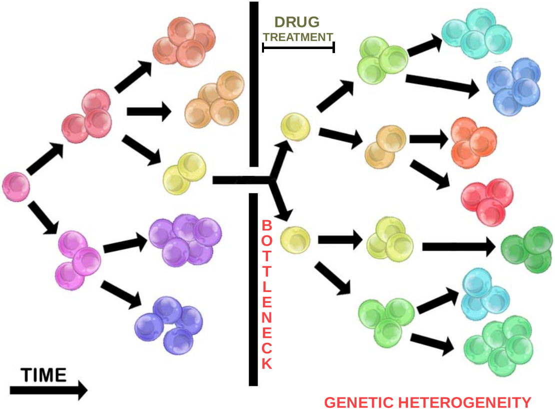


Chr.	Position	Mutation	Amino acid change
1	10379	C > T	Ser > Asn
4	26081	G > A	Arg > Lys
2	16205	T > C	Glu > Asp
7	36252	G > T	Gly > Trp
...	...	...	...

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



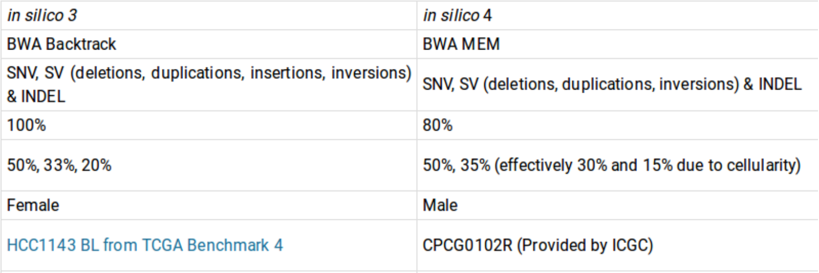
http://en.wikipedia.org/wiki/Tumour_heterogeneity

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

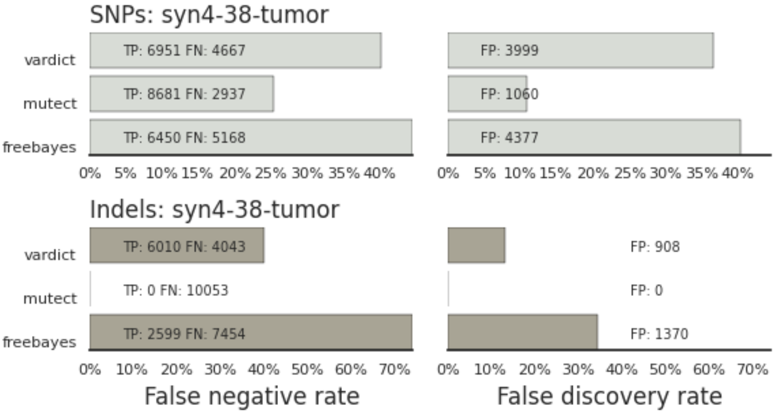
DREAM synthetic dataset 4



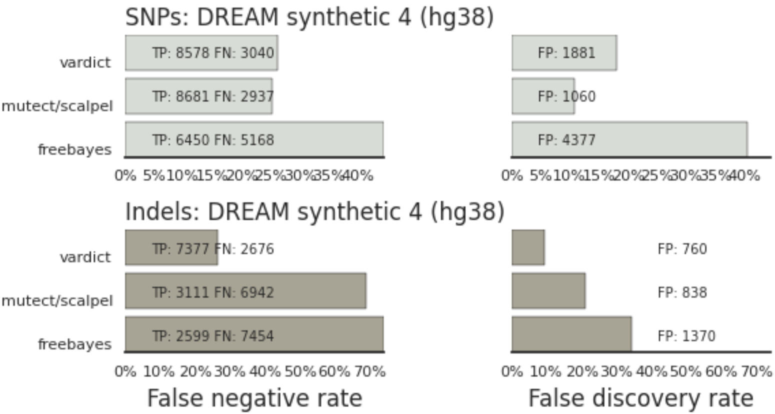
in silico 3	*in silico* 4
BWA Backtrack	BWA MEM
SNV, SV (deletions, duplications, insertions, inversions) & INDEL	SNV, SV (deletions, duplications, inversions) & INDEL
100%	80%
50%, 33%, 20%	50%, 35% (effectively 30% and 15% due to cellularity)
Female	Male
HCC1143 BL from TCGA Benchmark 4	CPCG0102R (Provided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

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

VarDict sensivitity/precision before



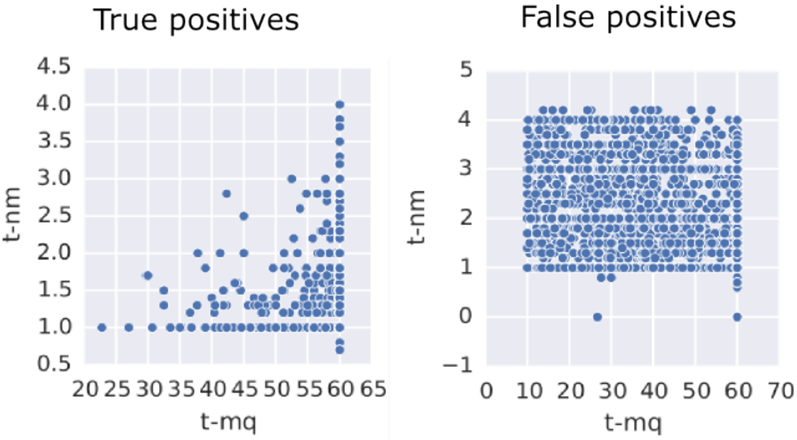
VarDict sensivitity/precision after



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



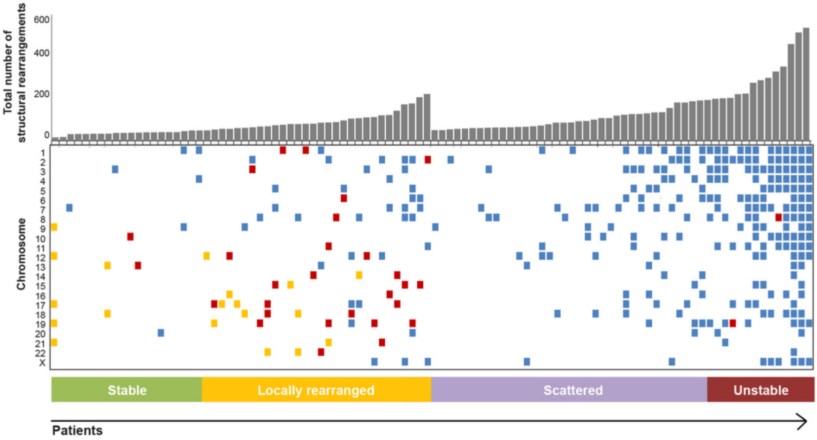
True positives

False positives

Outline: science

- Human build 38 + HLA
- Low frequency somatic calling
- **Structural variation**

Structural variants critical in cancer

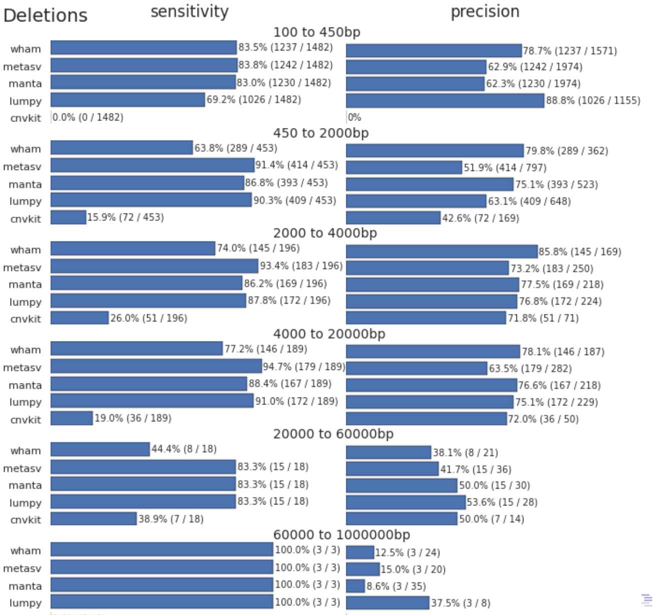


http://www.nature.com/nature/journal/v518/n7540/full/nature14169.html

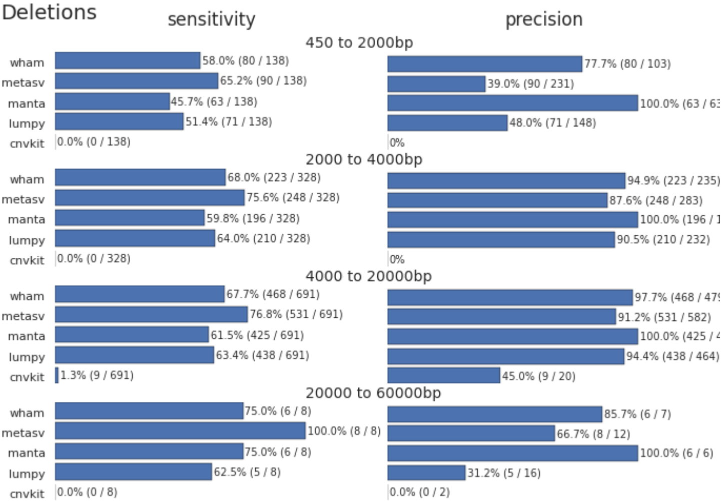
Improvements in speed, sensitivity and precision

- 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

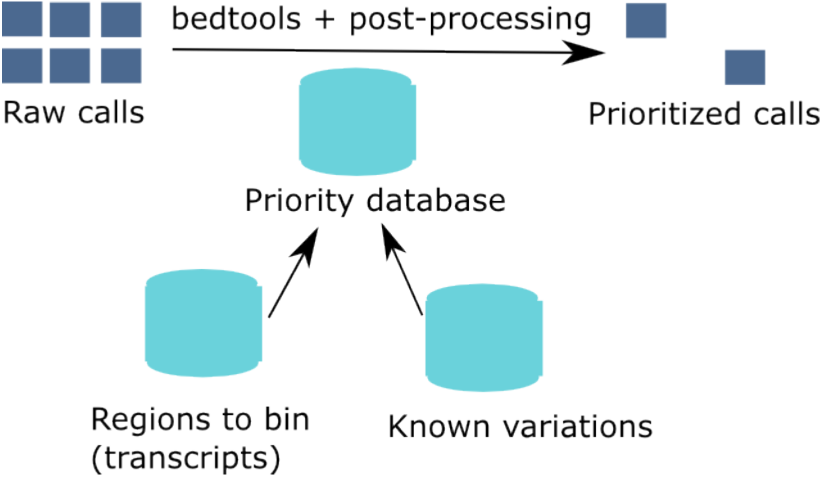
Results: Germline deletions



Results: Somatic deletions

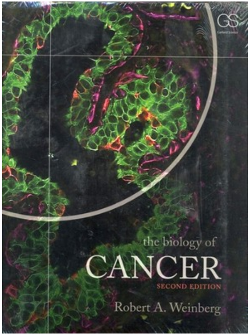


Prioritize in previously known regions



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

```
---
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
- Motivate for using open source community resources
- bcbio validated variant calling
- Science
  - Human build 38 + HLA
  - Cancer calling of low frequency variants
  - Structural variation
- Practical calling example

http://bcb.io