Variant Calling

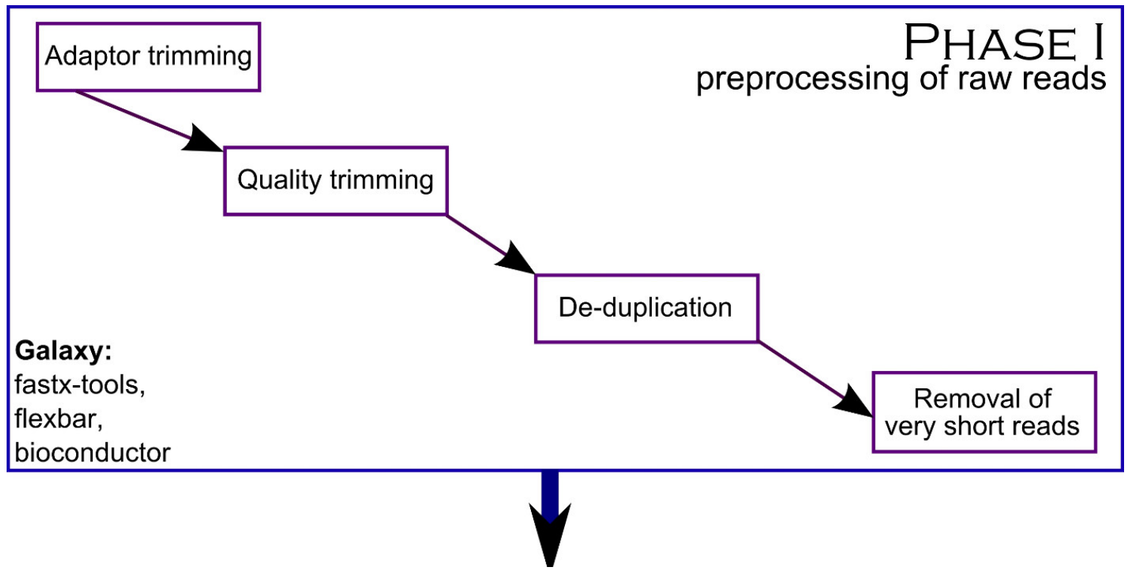
MEDT32/33

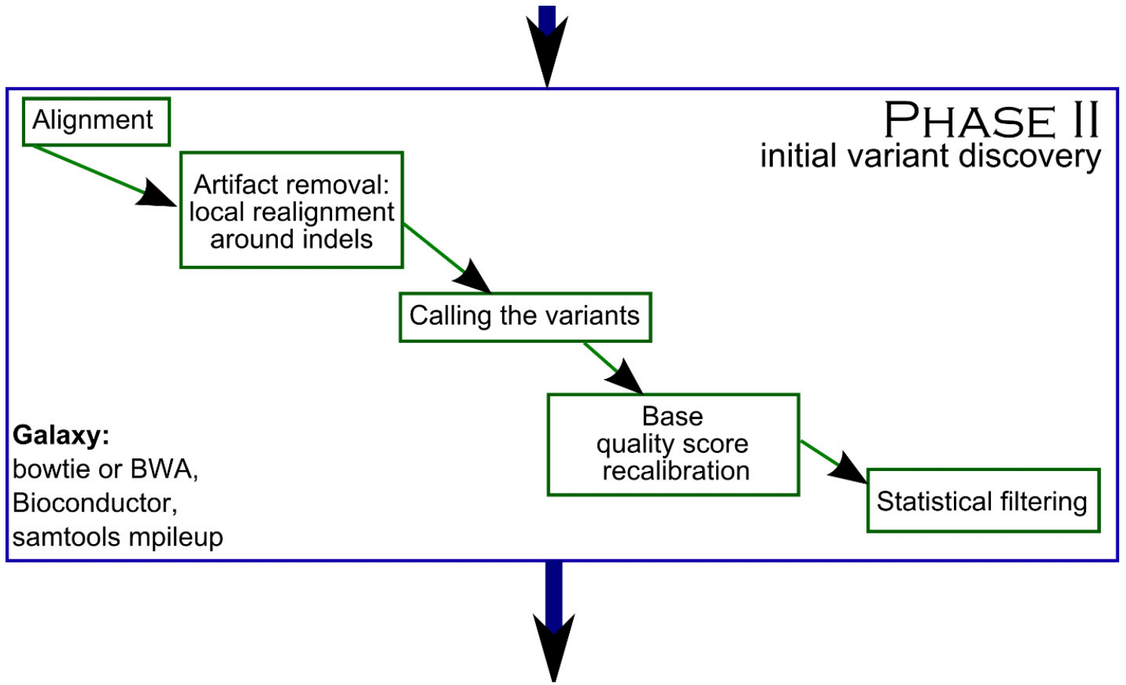
Variant calling

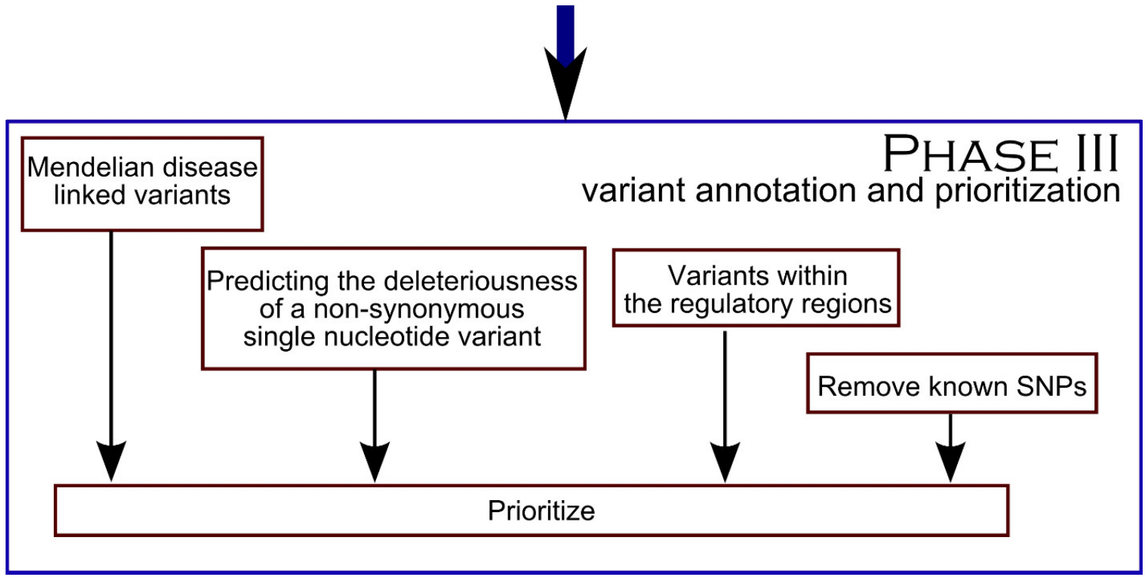
- challenging due to rapidly changing algorithms
- relatively low concordance between methods
- comparing approaches:
  - http://bcbio.wordpress.com/2013/05/06/framework-for-evaluating-variant-detection-methods-comparison-of-aligners-and-callers/

http://genomemedicine.com/content/5/3/28/abstract

overview


Adaptor trimming
Quality trimming
De-duplication
Removal of very short reads
PHASE I
preprocessing of raw reads
Galaxy:
fastx-tools,
flexbar,
bioconductor


Alignment
Artifact removal:
local realignment
around indels
Calling the variants
Base
quality score
recalibration
Statistical filtering
PHASE II
initial variant discovery
Galaxy:
bowtie or BWA,
Bioconductor,
samtools mpileup


PHASE III
variant annotation and prioritization
Mendelian disease linked variants
Predicting the deleteriousness of a non-synonymous single nucleotide variant
Variants within the regulatory regions
Remove known SNPs
Prioritize

- Pabinger et al. (2013) provides a good discussion of the common tools and approaches for variant calling.
- Also see the older Nielsen et al. (2011).

S Pabinger, A Dander, M Fischer, R Snajder, M Sperk, M Efremova, B Krabichler, MR. Speicher, J Zschocke, Z Trajanoski (2013) A survey of tools for variant analysis of nextgeneration genome sequencing data. Briefings in Bioinformatics (21 January 2013).
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from nextgeneration sequencing data. Nature Reviews 12:443451.

- (2
  - Most aligners do a good job of hard and soft clipping reads now so it's not as crucial. Best practice is to throw things right in and only trim if absolutely necessary since alignment errors and variants will get filtered downstream.
  -
  -
    - /)
- Selection of the tool to use depends on the amount of adaptor sequence leftover in the data.

Quality trimming

- inspect reads in bulk
- quality tends to drop off toward one end of the read.

Schmieder R and Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863864.

Removal of very short reads

- short reads are likely to align to multiple (wrong) locations on reference

pre-alignment processing does help reduce number of problem reads but it also doesn't scale well to multiple samples.

All of the fastq manipulation is disk intensive, since - reading and writing big files,

so gets slow trying to do a large number in parallel.

blog post by Eric Vallabh Minikel
2012

Initiation Variants Alignment

GATK

- Reads that align on the edges of indels often get mapped with mismatching bases that might look like evidence for SNPs. We look for the most consistent placement of the reads with respect to the indel in order to clean up these artifacts.

Alignment

- bwa for shorter read (< 75bp)
- bwa mem for longer reads
- outperform bowtie2 in most alignment comparisons
- novoalign (http://www.novocraft.com/main/index.php) requires a license. It is slower than bwa but more comprehensive since it does full smith-waterman. Newer version (3.0+) handle both short and long reads.

Base quality score recalibration

- The per base estimate of error (base quality score)
- estimates provided by the sequencing machines are often inaccurate
- empirically accurate error model through recalibration

GATK: QSR

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. (2011) A framework for variation discovery and genotyping using nextgeneration DNA sequencing data. Nat Genet. 43(5):4918.

Calling

- Only those variants should be kept that have a high confidence score:
- minimum Q30 for deep coverage data (>10x per sample) and
- minimum Q4 (if <=100 samples) or Q10 (if >100 samples) for shallower coverage.
- The variant calls are usually produced in form of VCF files (1020 M of hard disk space per file),

http://www.1000genomes.org/wiki/Analysis/vcf4.0

- quality of variant call sets is expected to increase substantially if multiple variant callers are used

- GATK HaplotypeCaller
- - FreeBayes (https://github.com/ekg/freebayes, http://gkno.me/)
- - GATK UnifiedGenotyper
- - samtools (performs poorly on indels)

Statistical filtering

- raw VCF files frequently have many sites that are not really genetic variants
- separate out the false positive machine artifacts from the true positive genetic variants
- variant quality score recalibration

http://gatkforums.broadinstitute.org/discussion/39/variantqualityscorerecalibrationvqsr

GEMINI - integrative exploration of genetic variation and genome annotations.

- Annotate Functional Impact of Each Variant
- integrated genetic variation (from VCF files) with a wealth of genome annotations into a unified database framework
- 1,000,000 variants times 1,000 samples yields one billion genotypes

http://gemini.readthedocs.org/en/latest/

Challenges

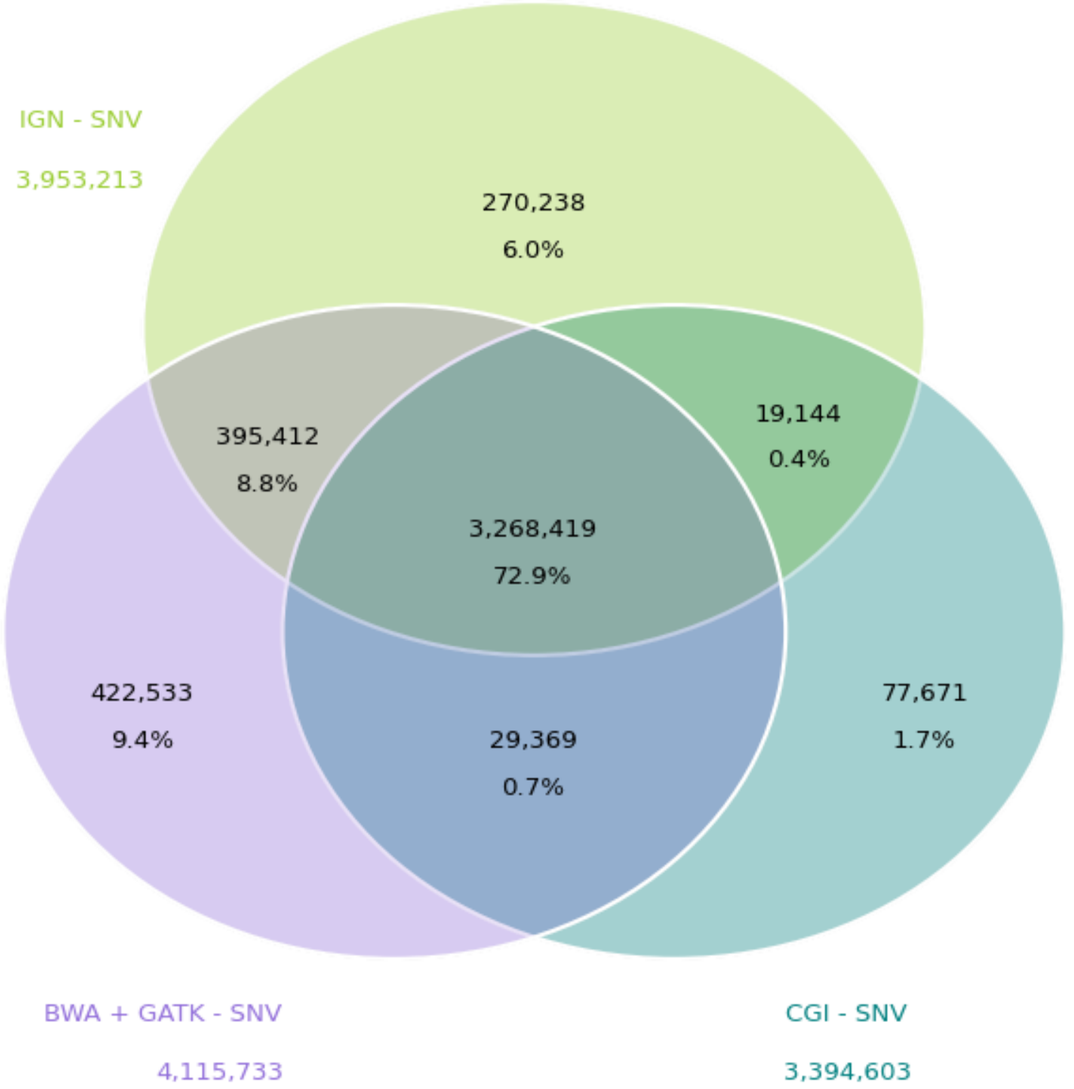
- Variants being called individually and not in a haplotype aware manner
- GC rich regions or regions that produce PCR or sequencing artifacts
- Regions of the genome with low mappability or high homology to other regions

Poor similarity between the sample's genome and the reference.

- Structural variants : inversions or translocations
- tandem repeats or mobile element insertions
- reference sequence is unable to capture the allelic diversity of the population and the sample

Heterozygote call

- 30x genome : heterozygote on average to be covered by 15 reads to support the mutant allele
- loci where a heterozygote legitimately exists but it gets sampled only once or twice
  - 1.74 times when calling 4 million variants



State ofVariant calling : don;t panic!

http://blog.goldenhelix.com/?p=1725

TOOLS

- **comp_hets**: Identifying potential compound heterozygotes
- **de_novo**: Identifying potential de novo mutations.
- **autosomal_recessive**: Find variants meeting an autosomal recessive model.
- **autosomal_dominant**: Find variants meeting an autosomal dominant model.
- **pathways**: Map genes and variants to KEGG pathways.
- **interactions**: Find genes among variants that are interacting partners.
- **lof_sieve**: Filter LoF variants by transcript position and type
- **annotate**: adding your own custom annotations
- **region**: Extracting variants from specific regions or genes
- **windower**: Conducting analyses on genome "windows".
- **stats**: Compute useful variant statistics.