UNIVERSITY OF OXFORD

Bioinformatics for Genomic Medicine

Lecture – part II

Silvia Salatino, PhD

High-throughput Bioinformatician
Wellcome Centre for Human Genetics, Oxford

Email: silvia@well.ox.ac.uk

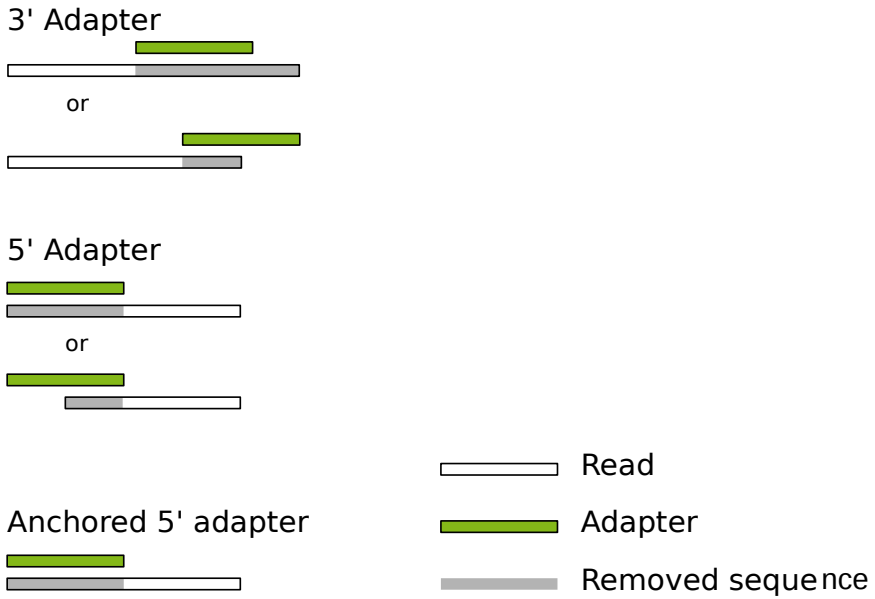
Module for the DPhil programme
Genomic Medicine and Statistics

22-26.10.2018

Variant analysis workflow scheme – adapter trimming

Adapter trimming, lane merging

There are lots of trimming tools doing either or both types of data processing: **Trimmomatic**, **Skewer**, **FASTX-Toolkit**, **Trim Galore**, **Cutadapt**, etc.

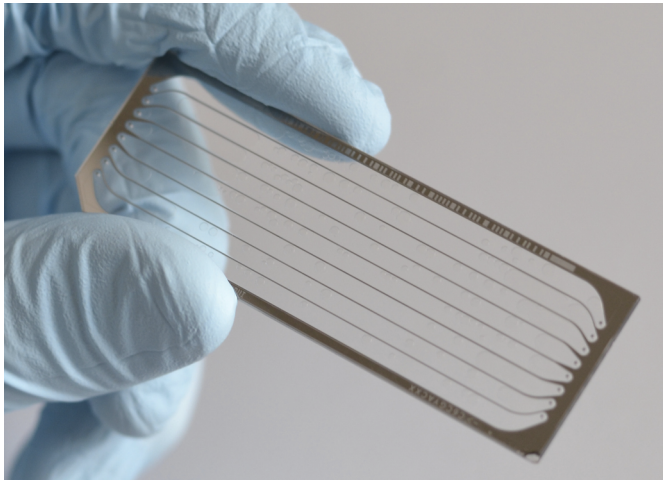


Variant analysis workflow scheme – lane merging

Adapter trimming, lane merging

Often a samples are pooled (thanks to specific barcodes) and sequenced on multiple lanes of the flowcell.

Therefore, after sequencing, this fragmented information needs to be merged into a unique file per sample.



Merging can be done on either FASTQ or BAM files, generally using **samtools** or **Picard tools**.

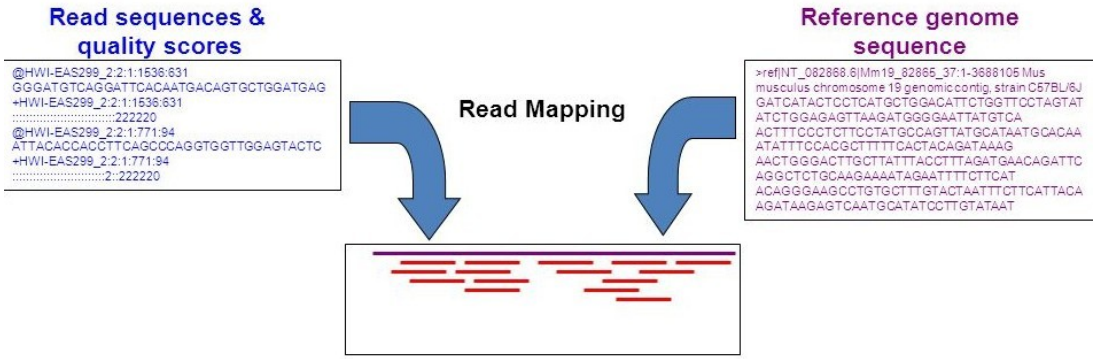
Variant analysis workflow scheme – mapping to the reference genome

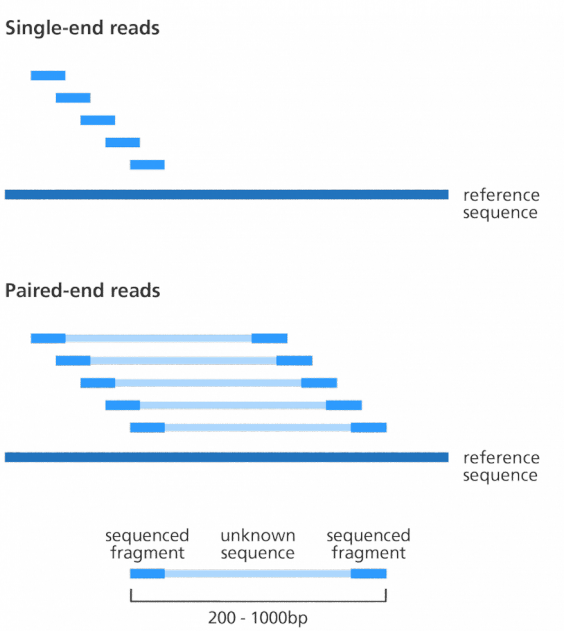
3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome





There is a continuously increasing number of mapping tools, each with different features and performances...

Short-read sequence alignment tools - 1

Short-read sequence alignment [edit]

Name	Description	paired-end option	Use FASTQ quality	Gapped	Multi-threaded	License	Link	Reference	Year
Arioc	Computes Smith-Waterman gapped alignments and mapping qualities on one or more GPUs. Supports BS-seq alignments. Processes 100,000 to 500,000 reads per second (varies with data, hardware, and configured sensitivity).	Yes	No	Yes	Yes	Free, BSD	github	[28]	2015
BarraCUDA	A GPGPU accelerated Burrows-Wheeler transform (FM-index) short read alignment program based on BWA, supports alignment of indels with gap openings and extensions.	Yes	No	Yes	Yes, POSIX Threads and CUDA	Free, GPL	link		
BBMap	Uses a short kmers to rapidly index genome; no size or scaffold count limit. Higher sensitivity and specificity than Burrows-Wheeler aligners, with similar or greater speed. Performs affine-transform-optimized global alignment, which is slower but more accurate than Smith-Waterman. Handles Illumina, 454, PacBio, Sanger, and Ion Torrent data. Splice-aware; capable of processing long indels and RNA-seq. Pure Java; runs on any platform. Used by the Joint Genome Institute.	Yes	Yes	Yes	Yes	Free, BSD	link		2010
BFAST	Explicit time and accuracy tradeoff with a prior accuracy estimation, supported by indexing the reference sequences. Optimally compresses indexes. Can handle billions of short reads. Can handle insertions, deletions, SNPs, and color errors (can map ABI SOLiD color space reads). Performs a full Smith Waterman alignment.				Yes, POSIX Threads	Free, GPL	link	[29]	2009
BigBWA	Runs the Burrows-Wheeler Aligner-BWA on a Hadoop cluster. It supports the algorithms BWA-MEM, BWA-ALN, and BWA-SW, working with paired and single reads. It implies an important reduction in the computational time when running in a Hadoop cluster, adding scalability and fault-tolerancy.	Yes	Low quality bases trimming	Yes	Yes	Free, GPL 3	link	[30]	2015
BLASTN	BLAST's nucleotide alignment program, slow and not accurate for short reads, and uses a sequence database (EST, sanger sequence) rather than a reference genome.						link		
BLAT	Made by Jim Kent. Can handle one mismatch in initial alignment step.				Yes, client-server	Proprietary, freeware for academic and noncommercial use	link	[31]	2002
Bowtie	Uses a Burrows-Wheeler transform to create a permanent, reusable index of the genome; 1.3 GB memory footprint for human genome. Aligns more than 25 million Illumina reads in 1 CPU hour. Supports Maq-like and SOAP-like alignment policies	Yes	Yes	No	Yes, POSIX Threads	Free, Artistic	link	[32]	2009
HIVE-hexagon	Uses a hash table and bloom matrix to create and filter potential positions on the genome. For higher efficiency uses cross-similarity between short reads and avoids realigning non unique redundant sequences. It is faster than bowtie and bwa and allows indels and divergent sensitive alignments on viruses, bacteria, and more conservative eukaryotic alignments.	Yes	Yes	Yes	Yes	Proprietary, freeware for academic and noncommercial users registered to HIVE deployment instance	link	[33]	2014
BWA	Uses a Burrows-Wheeler transform to create an index of the genome. It's a bit slower than bowtie but allows indels in alignment.	Yes	Low quality bases trimming	Yes	Yes	Free, GPL	link	[34]	2009
BWA-PSSM	A probabilistic short read aligner based on the use of position specific scoring matrices (PSSM). The aligner is adaptable in the sense that it can take into account the quality scores of the reads and models of data specific biases, such as those observed in Ancient DNA, PAR-CLIP data or genomes with biased nucleotide compositions.[35]	Yes	Yes	Yes	Yes	Free, GPL	link	[35]	2014
CASHX	Quantify and manage large quantities of short-read sequence data. CASHX pipeline contains a set of tools that can be used together, or separately as modules. This algorithm is very accurate for perfect hits to a reference genome.				No	Proprietary, freeware for academic and noncommercial use	link		
Cloudburst	Short-read mapping using Hadoop MapReduce				Yes, Hadoop MapReduce	Free, Artistic	link		
CUDA-EC	Short-read alignment error correction using GPUs.				Yes, GPU enabled		link		
CUSHAW	A CUDA compatible short read aligner to large genomes based on Burrows-Wheeler transform	Yes	Yes	No	Yes (GPU enabled)	Free, GPL	link	[36]	2012
CUSHAW2	Gapped short-read and long-read alignment based on maximal exact match seeds. This aligner supports both base-space (e.g. from Illumina, 454, Ion Torrent and PacBio sequencers) and ABI SOLiD color-space read alignments.	Yes	No	Yes	Yes	Free, GPL	link		2014
CUSHAW2-GPU	GPU-accelerated CUSHAW2 short-read aligner.	Yes	No	Yes	Yes	Free, GPL	link		
CUSHAW3	Sensitive and accurate base-space and color-space short-read alignment with hybrid seeding	Yes	No	Yes	Yes	Free, GPL	link	[37]	2012
drFAST	Read mapping alignment software that implements cache obliviousness to minimize main/cache memory transfers like mrFAST and mrsFAST, however designed for the SOLiD sequencing platform (color space reads). It also returns all possible map locations for improved structural variation discovery.	Yes	Yes, for structural variation	Yes	No	Free, BSD	link		
ELAND	Implemented by Illumina. Includes ungapped alignment with a finite read length.								

Short-read sequence alignment tools - 2

ERNE	Extended Randomized Numerical alignEr for accurate alignment of NGS reads. It can map bisulfite-treated reads.	Yes	Low quality bases trimming	Yes	Multithreading and MPI-enabled	Free, GPL 3	link		
GASSST	Finds global alignments of short DNA sequences against large DNA banks				Multithreading	CeCILL version 2 License.	link	[38]	2011
GEM	High-quality alignment engine (exhaustive mapping with substitutions and indels). More accurate and several times faster than BWA or Bowtie 1/2. Many standalone biological applications (mapper, split mapper, mappability, and other) provided.	Yes	Yes	Yes	Yes	Dual, freeware for noncommercial use; GEM source is currently unavailable	link	[39]	2012
Genalice MAP	Ultra fast and comprehensive NGS read aligner with high precision and small storage footprint.	Yes	Low quality bases trimming	Yes	Yes	Proprietary, commercial	link		
Geneious Assembler	Fast, accurate overlap assembler with the ability to handle any combination of sequencing technology, read length, any pairing orientations, with any spacer size for the pairing, with or without a reference genome.				Yes	Proprietary, commercial	link		
GensearchNGS	Complete framework with user-friendly GUI to analyse NGS data. It integrates a proprietary high quality alignment algorithm and plug-in ability to integrate various public aligner into a framework allowing to import short reads, align them, detect variants, and generate reports. It is made for resequencing projects, namely in a diagnostic setting.	Yes	No	Yes	Yes	Proprietary, commercial	link		
GMAP and GSNAP	Robust, fast short-read alignment. GMAP: longer reads, with multiple indels and splices (see entry above under Genomics analysis); GSNAP: shorter reads, with one indel or up to two splices per read. Useful for digital gene expression, SNP and indel genotyping. Developed by Thomas Wu at Genentech. Used by the National Center for Genome Resources (NCGR) in Alpheus.	Yes	Yes	Yes	Yes	Proprietary, freeware for academic and noncommercial use	link		
GNUMAP	Accurately performs gapped alignment of sequence data obtained from next-generation sequencing machines (specifically of Solexa-Illumina) back to a genome of any size. Includes adaptor trimming, SNP calling and Bisulfite sequence analysis.		Yes, also supports Illumina *_int.txt and *_prb.txt files with all 4 quality scores for each base		Multithreading and MPI-enabled		link	[40]	2009
iSAAC	Fully uses all the computing power available on one server node; thus, it scales well over a broad range of hardware architectures, and alignment performance improves with hardware abilities	Yes	Yes	Yes	Yes	Free, BSD	github paper		
LAST	Uses adaptative seeds and copes more efficiently with repeat-rich sequences (e.g. genomes). For example: it can align reads to genomes without repeat-masking, without becoming overwhelmed by repetitive hits.	Yes	Yes	Yes	No	Free, GPL	link	[41]	2011
MAQ	Ungapped alignment that takes into account quality scores for each base.					Free, GPL	link		
mrFAST, mrsFAST	Gapped (mrFAST) and ungapped (mrsFAST) alignment software that implements cache obliviousness to minimize main/cache memory transfers. They are designed for the Illumina sequencing platform and they can return all possible map locations for improved structural variation discovery.	Yes	Yes, for structural variation	Yes	No	Free, BSD	mrFAST mrsFAST		
MOM	MOM or maximum oligonucleotide mapping is a query matching tool that captures a maximal length match within the short read.				Yes		link		
MOSAIK	Fast gapped aligner and reference-guided assembler. Aligns reads using a banded Smith-Waterman algorithm seeded by results from a k-mer hashing scheme. Supports reads ranging in size from very short to very long.				Yes		link		
MPscan	Fast aligner based on a filtration strategy (no indexing, use q-grams and Backward Nondeterministic DAWG Matching)						link	[42]	2009
Novoalign & NovoalignCS	Gapped alignment of single end and paired end Illumina GA I & II, ABI Colour space & ION Torrent reads. High sensitivity and specificity, using base qualities at all steps in the alignment. Includes adapter trimming, base quality calibration, Bi-Seq alignment, and option to report multiple alignments per read.	Yes	Yes	Yes	Multi-threading and MPI versions available with paid license	Proprietary, freeware single threaded version for academic and noncommercial use	Novocraft		
NextGENe	Developed for use by biologists performing analysis of next generation sequencing data from Roche Genome Sequencer FLX, Illumina GA/HiSeq, Life Technologies Applied BioSystems' SOLiD System, PacBio and Ion Torrent platforms.	Yes	Yes	Yes	Yes	Proprietary, commercial	Softgenetics		
NextGenMap	Flexible and fast read mapping program (twice as fast as BWA), achieves a mapping sensitivity comparable to Stampy. Internally uses a memory efficient index structure (hash table) to store positions of all 13-mers present in the reference genome. Mapping regions where pairwise alignments are required are dynamically determined for each read. Uses fast SIMD instructions (SSE) to accelerate alignment calculations on CPU. If available, alignments are computed on GPU (using OpenCL/CUDA) further reducing runtime 20-50%.	Yes	No	Yes	Yes, POSIX Threads, OpenCL/CUDA, SSE	Free	Official GitHub Page	[43]	2013
Omixon Variant Toolkit	Includes highly sensitive and highly accurate tools for detecting SNPs and indels. It offers a solution to map NGS short reads with a moderate distance (up to 30% sequence divergence) from reference genomes. It poses no restrictions on the size of the reference, which, combined with its high sensitivity, makes the Variant Toolkit well-suited for targeted sequencing projects and diagnostics.	Yes	Yes	Yes	Yes	Proprietary, commercial	www.omixon.com		
PALMapper	Efficiently computes both spliced and unspliced alignments at high accuracy. Relying on a machine learning strategy combined with a fast mapping based on a banded Smith-Waterman-like algorithm, it aligns around 7 million reads per hour on one CPU. It refines the originally proposed QPALMA approach.				Yes	Free, GPL	link		

Short-read sequence alignment tools - 3

Partek Flow	For use by biologists and bioinformaticians. It supports ungapped, gapped and splice-junction alignment from single and paired-end reads from Illumina, Life technologies Solid TM, Roche 454 and Ion Torrent raw data (with or without quality information). It integrates powerful quality control on FASTQ/Qual level and on aligned data. Additional functionality include trimming and filtering of raw reads, SNP and InDel detection, mRNA and microRNA quantification and fusion gene detection.	Yes	Yes	Yes	Multiprocessor-core, client-server installation possible	Proprietary, commercial, free trial version	[1]		
PASS	Indexes the genome, then extends seeds using pre-computed alignments of words. Works with base space, color space (SOLID), and can align genomic and spliced RNA-seq reads.	Yes	Yes	Yes	Yes	Proprietary, freeware for academic and noncommercial use	PASS_HOME		
PerM	Indexes the genome with periodic seeds to quickly find alignments with full sensitivity up to four mismatches. It can map Illumina and SOLiD reads. Unlike most mapping programs, speed increases for longer read lengths.				Yes	Free, GPL	link	[44]	
PRIMEX	Indexes the genome with a k-mer lookup table with full sensitivity up to an adjustable number of mismatches. It is best for mapping 15-60 bp sequences to a genome.	No	No	Yes	No, multiple processes per search		link	[2]	2003
QPalma	Can use quality scores, intron lengths, and computation splice site predictions to perform and performs an unbiased alignment. Can be trained to the specifics of a RNA-seq experiment and genome. Useful for splice site/intron discovery and for gene model building. (See PALMapper for a faster version).				Yes, client-server	Free, GPL 2	link		
RazerS	No read length limit. Hamming or edit distance mapping with configurable error rates. Configurable and predictable sensitivity (runtime/sensitivity tradeoff). Supports paired-end read mapping.					Free, LGPL	link		
REAL, cREAL	REAL is an efficient, accurate, and sensitive tool for aligning short reads obtained from next-generation sequencing. The programme can handle an enormous amount of single-end reads generated by the next-generation Illumina/Solexa Genome Analyzer. cREAL is a simple extension of REAL for aligning short reads obtained from next-generation sequencing to a genome with circular structure.		Yes		Yes	Free, GPL	link		
RMAP	Can map reads with or without error probability information (quality scores) and supports paired-end reads or bisulfite-treated read mapping. There are no limitations on read length or number of mismatches.	Yes	Yes	Yes		Free, GPL 3	link		
rNA	A randomized Numerical Aligner for Accurate alignment of NGS reads	Yes	Low quality bases trimming	Yes	Multithreading and MPI-enabled	Free, GPL 3	link		
RTG Investigator	Extremely fast, tolerant to high indel and substitution counts. Includes full read alignment. Product includes comprehensive pipelines for variant detection and metagenomic analysis with any combination of Illumina, Complete Genomics and Roche 454 data.	Yes	Yes, for variant calling	Yes	Yes	Proprietary, freeware for individual investigator use	link		
Segemehl	Can handle insertions, deletions, mismatches; uses enhanced suffix arrays	Yes	No	Yes	Yes	Proprietary, freeware for noncommercial use	link	[45]	2009
SeqMap	Up to 5 mixed substitutions and insertions-deletions; various tuning options and input-output formats					Proprietary, freeware for academic and noncommercial use	link		
Shrec	Short read error correction with a suffix tree data structure				Yes, Java		link		
SHRiMP	Indexes the reference genome as of version 2. Uses masks to generate possible keys. Can map ABI SOLiD color space reads.	Yes	Yes	Yes	Yes, OpenMP	Free, [[BSD licenses| style="background: #9FF; color: black; vertical-align: middle; text-align: center; " class="free table-free"|Free, BSD]] derivative	link	[46] [47]	2009-2011
SLIDER	Slider is an application for the Illumina Sequence Analyzer output that uses the "probability" files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences.						link		
SOAP, SOAP2, SOAP3, SOAP3-dp	SOAP: robust with a small (1-3) number of gaps and mismatches. Speed improvement over BLAT, uses a 12 letter hash table. SOAP2: using bidirectional BWT to build the index of reference, and it is much faster than the first version. SOAP3: GPU-accelerated version that could find all 4-mismatch alignments in tens of seconds per one million reads. SOAP3-dp, also GPU accelerated, supports arbitrary number of mismatches and gaps according to affine gap penalty scores.	Yes	No	Yes, SOAP3-dp	Yes, POSIX Threads; SOAP3, SOAP3-dp need GPU with CUDA support	Free, GPL	link	[48][49]	
SOCS	For ABI SOLiD technologies. Significant increase in time to map reads with mismatches (or color errors). Uses an iterative version of the Rabin-Karp string search algorithm.				Yes	Free, GPL	link		
SparkBWA	Integrates the Burrows-Wheeler Aligner—BWA on a Apache Spark framework running atop Hadoop. Version 0.2 of October 2016, supports the algorithms BWA-MEM, BWA-backtrack, and BWA-ALN. All of them work with single-reads and paired-end reads.	Yes	Low quality bases trimming	Yes	Yes	Free, GPL 3	link	[50]	2016
SSAHA, SSAHA2	Fast for a small number of variants					Proprietary, freeware for academic and noncommercial use	link		
Stampy	For Illumina reads. High specificity, and sensitive for reads with indels, structural variants, or many SNPs. Slow, but speed increased dramatically by using BWA for first alignment pass.	Yes	Yes	Yes	No	Proprietary, freeware for academic and noncommercial use	link	[51]	2010
SToRM	For Illumina or ABI SOLiD reads, with SAM native output. Highly sensitive for reads with many errors, indels (full from 0 to 15, extended support otherwise). Uses spaced seeds (single hit) and a very fast SSE-SSE2-AVX2-AVX-512 banded alignment filter. For fixed-length reads only, authors recommend SHRiMP2 otherwise.	No	Yes	Yes	Yes, OpenMP	Free	link	[52]	2010

Short-read sequence alignment tools - 4

Subread, Subjunc	Superfast and accurate read aligners. Subread can be used to map both gDNA-seq and RNA-seq reads. Subjunc detects exon-exon junctions and maps RNA-seq reads. They employ a novel mapping paradigm named *seed-and-vote*.	Yes	Yes	Yes	Yes	Free, GPL 3	link link		
Taipan	De-novo assembler for Illumina reads					Proprietary, freeware for academic and noncommercial use	link		
UGENE	Visual interface both for Bowtie and BWA, and an embedded aligner	Yes	Yes	Yes	Yes	Free, GPL	link		
VelociMapper	FPGA-accelerated reference sequence alignment mapping tool from TimeLogic. Faster than Burrows-Wheeler transform-based algorithms like BWA and Bowtie. Supports up to 7 mismatches and/or indels with no performance penalty. Produces sensitive Smith-Waterman gapped alignments.	Yes	Yes	Yes	Yes	Proprietary, commercial	TimeLogic		
XpressAlign	FPGA based sliding window short read aligner which exploits the embarrassingly parallel property of short read alignment. Performance scales linearly with number of transistors on a chip (i.e. performance guaranteed to double with each iteration of Moore's Law without modification to algorithm). Low power consumption is useful for datacentre equipment. Predictable runtime. Better price/performance than software sliding window aligners on current hardware, but not better than software BWT-based aligners currently. Can manage large numbers (>2) of mismatches. Will find all hit positions for all seeds. Single-FPGA experimental version, needs work to develop it into a multi-FPGA production version.					Proprietary, freeware for academic and noncommercial use	link		
ZOOM	100% sensitivity for a reads between 15-240 bp with practical mismatches. Very fast. Support insertions and deletions. Works with Illumina & SOLiD instruments, not 454.				Yes (GUI), no (CLI)	Proprietary, commercial	link	[53]	

Source: List of sequence alignment software (2017, September 3), Wikipedia, The Free Encyclopedia

Variant analysis workflow scheme – SAM / BAM file formats

3. Bioinf. Analysis II

Adapter trimming, lane merging

V

Read mapping to ref. genome

The most commonly used output format of alignment programs is **SAM** and its binary version **BAM** (preferably used to save disk space).

Mandatory fields (followed by any number of optional fields):

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Example of 3 reads in a SAM file:

Headers

Alignments

```
1 @HD VN:1.0  SO:unsorted
2 @SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
3 @PG ID:bowtie2  PN:bowtie2  VN:2.1.0
4 gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
  008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
  ATCACCAACCATCTGGTAGCGATGAT  2222222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
  YT:Z:UU
5 gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
  008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
  CCCCCGCCAAAATCACCAACCATCTG  2222222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
  :UU
6 gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
  008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
  GACGCGTACAGGAAACACCGAAAAAA  2222222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
      YT:Z:UU
```

Variant analysis workflow scheme – deduplication

3. Bioinf. Analysis II

Adapter trimming, lane merging

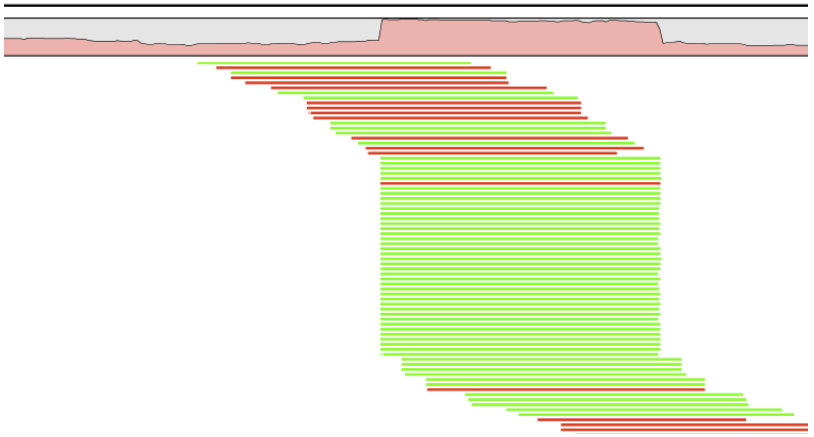
V

Read mapping to ref. genome

V

Read deduplication

The main purpose of read de-duplication (= keep 1 read per position) is to reduce the effects of a possible (and common) **PCR amplification bias** that could have happened during library construction and might produce inaccurate variant calling results.



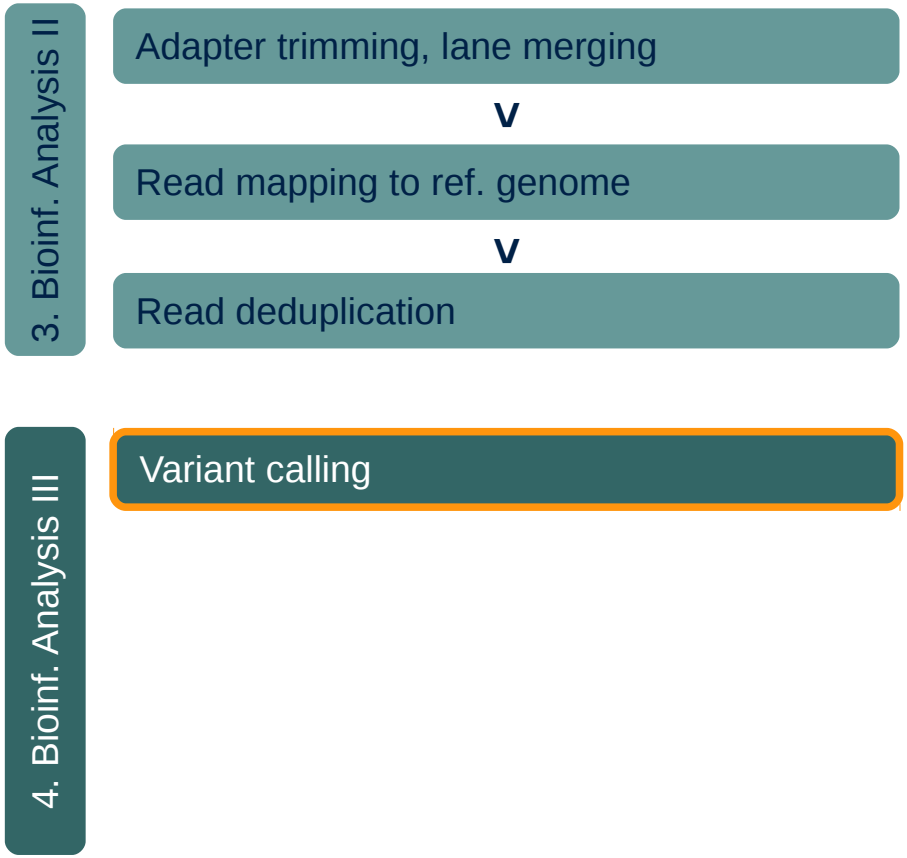
Source: Quiagen (http://resources.qiagenbioinformatics.com/manuals/)

This can be done, for example, by using **Picard**'s MarkDuplicate tool (or, alternatively, samtools).

Gain the computational benefit of reducing the reads to process downstream

Risk of setting a hard cap to the dynamic range of measurements and losing potentially true signal

Variant analysis workflow scheme – variant calling overview



Variant calling is required to **identify variant positions**, **genotype likelihood** and **allele frequencies** from an input BAM file, returning as output a VCF file.

Challenges of finding variants from NGS data:

- base calling errors
- mapping errors
- low coverage sequencing

Variant calling tools can be grouped in 4 categories:

- **germline callers** *(e.g. GATK, FreeBayes, Platypus, VarScan2)*
- **somatic callers** *(e.g. GATK, MuTect, VarScan2, SAMtools)*
- **CNV identification software** *(e.g. ExomeDepth, CNVseq)*
- **SV identification software** *(e.g. BreakDancer, FusionMap)*

For a comprehensive list of the available software, look at:
"Pabinger et al., Brief Bioinform. 2014 Mar;15(2):256-78."

Variant analysis workflow scheme – somatic variant calling

3. Bioinf. Analysis II

Adapter trimming, lane merging

V

Read mapping to ref. genome

V

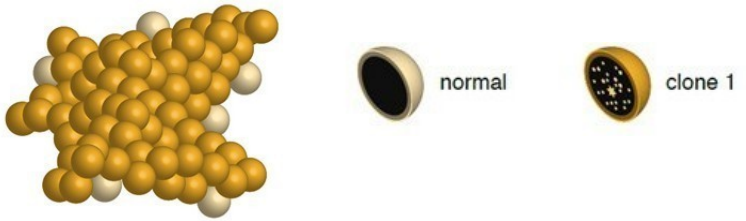
Read deduplication

4. Bioinf. Analysis III

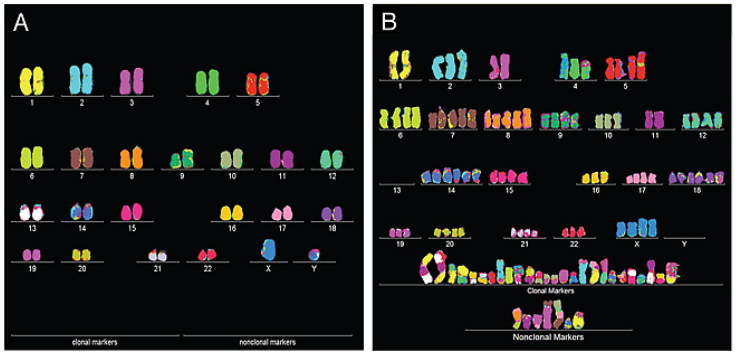
Variant calling

Somatic variant calling has additional challenges:

- **low tumour purity** (% of normal cells in the sample)



- **tumour ploidy**



(source: University of Berkeley)

- **tumour heterogeneity** (multiple clones)

(source: Govindan et al., Cell 2012)

Variant analysis workflow scheme – variant calling for SNPs and InDels

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

v

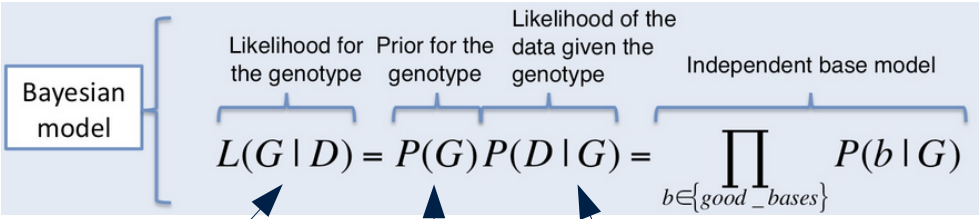
Read deduplication

4. Bioinf. Analysis III

Variant calling

Early SNP callers used a **naïve approach** of counting reads for each allele that have passed a mapping threshold → not good enough when coverage of a given variant is low!

More advanced SNP callers use a **Bayesian approach** to calculate likelihoods of each possible genotype, accounting for low coverage datasets and indels:



Source: GATK slides

prob. of the genotype G given the observed sequencing data D

prob. of the genotype G (how likely do we expect to see it based on previous observations, studies of the population, etc.)

conditional prob. of observing a particular sequencing dataset (reads) from a given genotype; can be estimated, for example, using a HMM to produce the likelihoods of each read (through its per base quality scores) against each haplotype

Variant analysis workflow scheme – variant calling for SVs and CNVs (1)

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

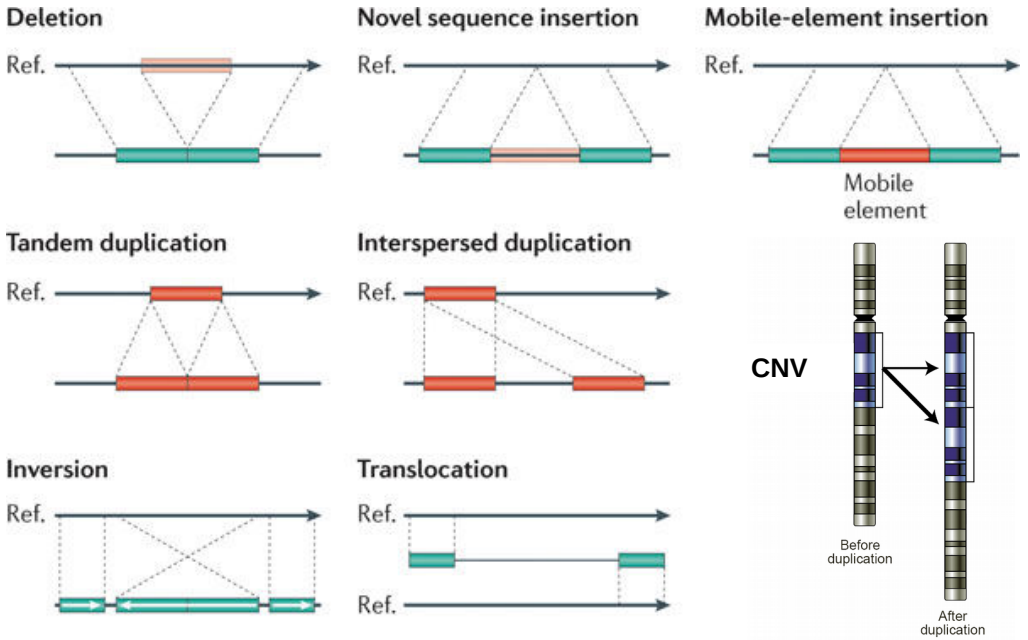
v

Read deduplication

4. Bioinf. Analysis III

Variant calling

There are different types of SVs and CNVs:



Adapted from: Alkan et al., Nat Rev Genet. 2011 May;12(5):363-76.

SV callers from short read sequence data can be classified in 3 categories, based on the approach they adopt:

- **split reads**
- **paired-end reads**
- **read depth**

Read depth analysis is particularly effective for exome data as it does not rely on sequencing into or near the CNV breakpoints.

Variant analysis workflow scheme – variant calling for SVs and CNVs (2)

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

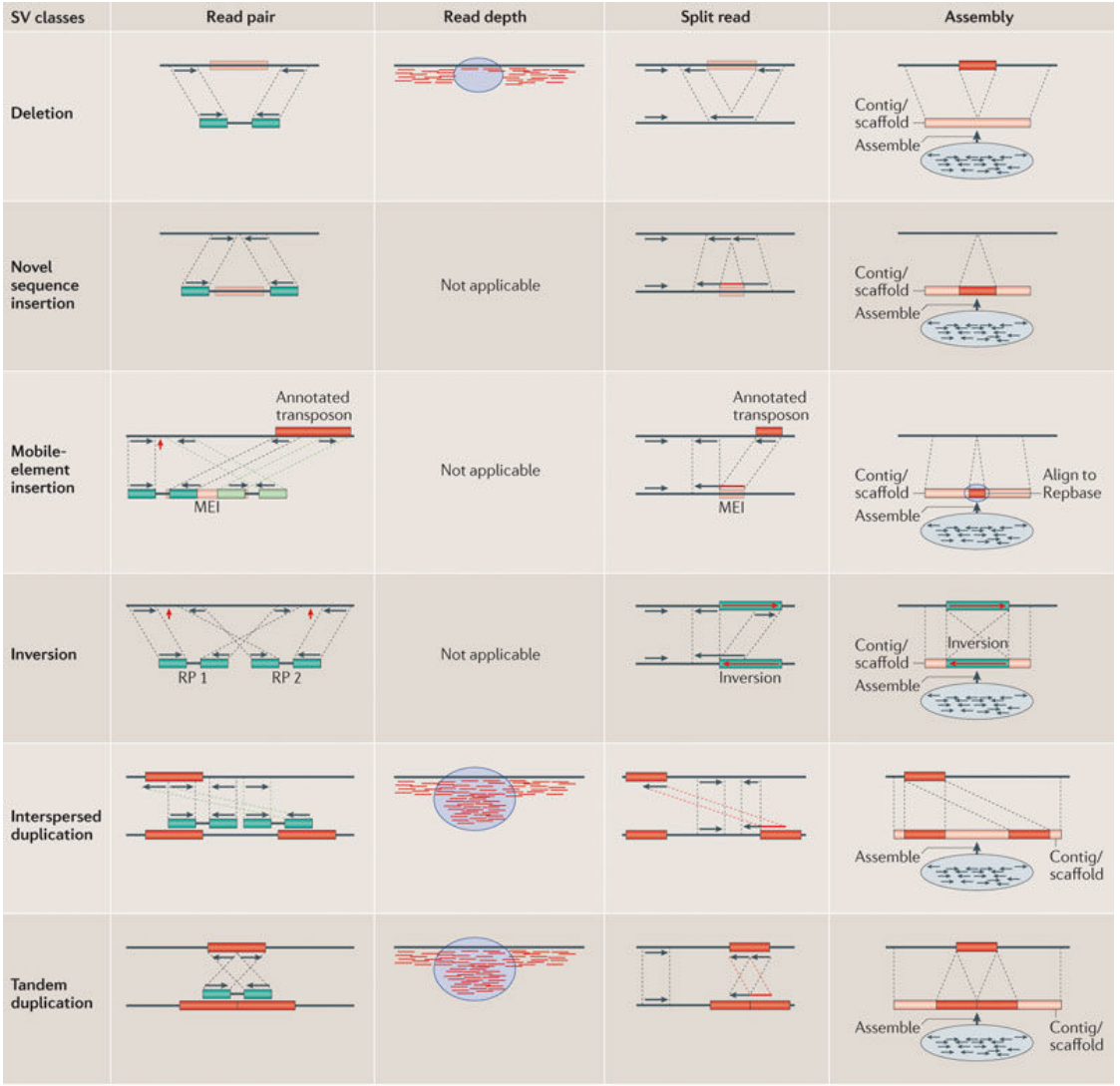
Read mapping to ref. genome

v

Read deduplication

4. Bioinf. Analysis III

Variant calling



Alkan et al., Nat Rev Genet. 2011 May;12(5):363-76.

Variant analysis workflow scheme – VCF file format

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

v

Read deduplication

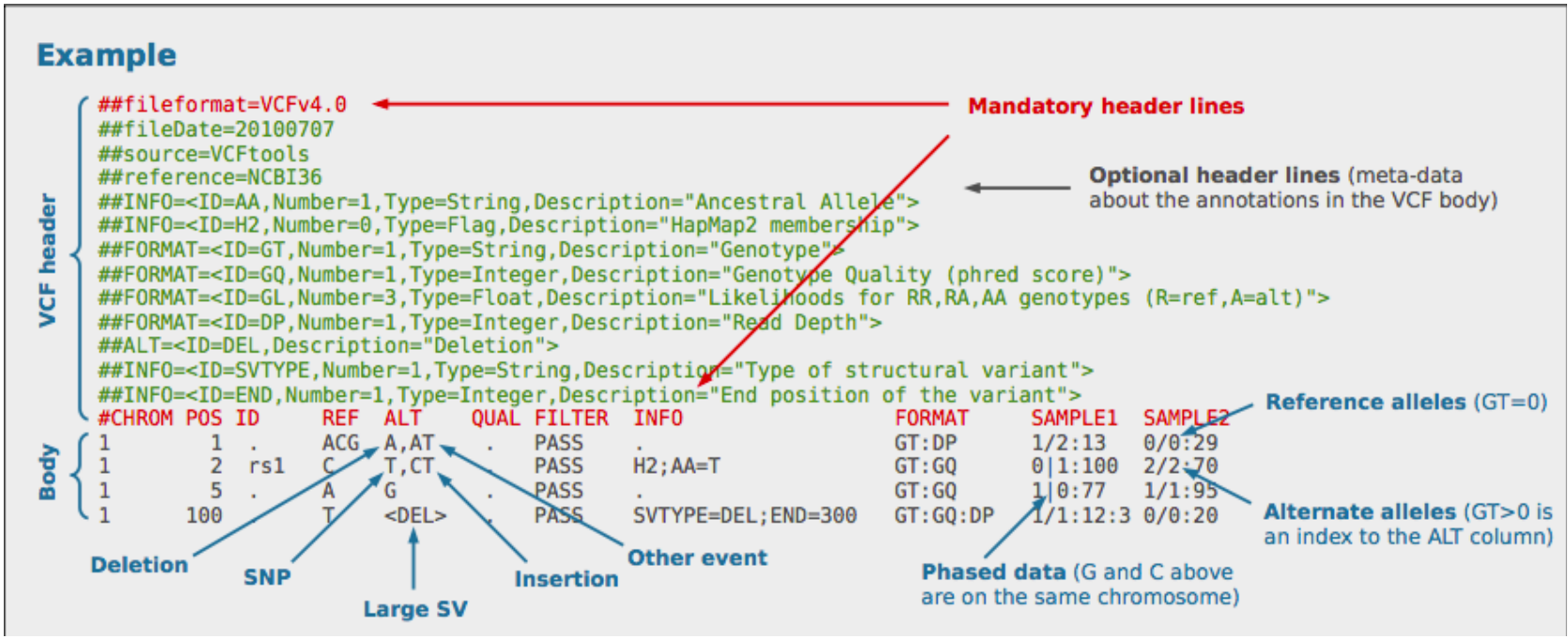
4. Bioinf. Analysis III

Variant calling

The typical output of variant calling is a **VCF file**, invented and used within the 1000 Genome Project to report all variant positions (as well as any possible information about them) and to promote the use of a standardised format.

A VCF file contains:
- **meta-information lines**,
- **a header line**,
- **multiple data lines**.

Each data line represents one variant (SNP, insertion or deletion), as well as quality scores, allele frequencies, database annotations and genotypes for one or more samples.



Variant analysis workflow scheme – Cohort variant calling approaches (1)

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

v

Read deduplication

4. Bioinf. Analysis III

Variant calling

When dealing with cohorts, there are three common approaches:

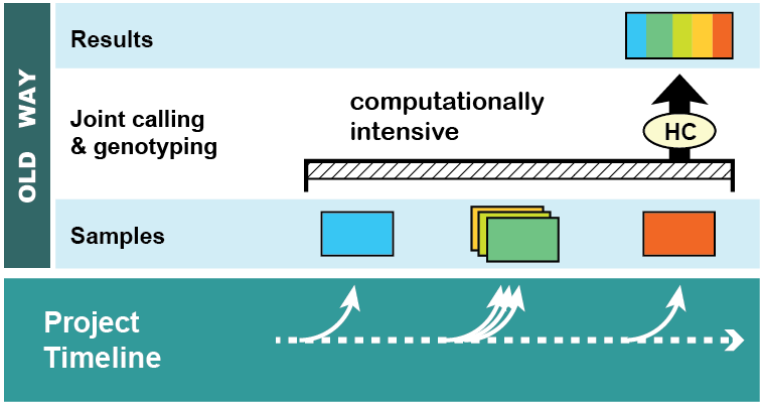
- **single sample calling**: sample BAMs are analyzed individually, and individual call sets are combined more downstream;
- **batch calling**: sample BAMs are analyzed in separate batches, and batch call sets are merged more downstream;
- **joint calling**: variants are called simultaneously across all sample BAMs, generating a single call set for the entire cohort.

Advantages of joint variant calling include:

1. clearer distinction between homozygous reference sites and sites with missing data
2. greater sensitivity for low-frequency variants
3. greater ability to filter out false positives

Drawbacks of joint variant calling are:

1. scaling & infrastructure (you need a lot of compute available)
2. the “N+1 problem” (joint calling doesn’t allow for incremental analysis; every time you get even one new sample sequence, you have to re-call all samples from scratch)



Source: GATK forums

Variant analysis workflow scheme – Cohort variant calling approaches (2)

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

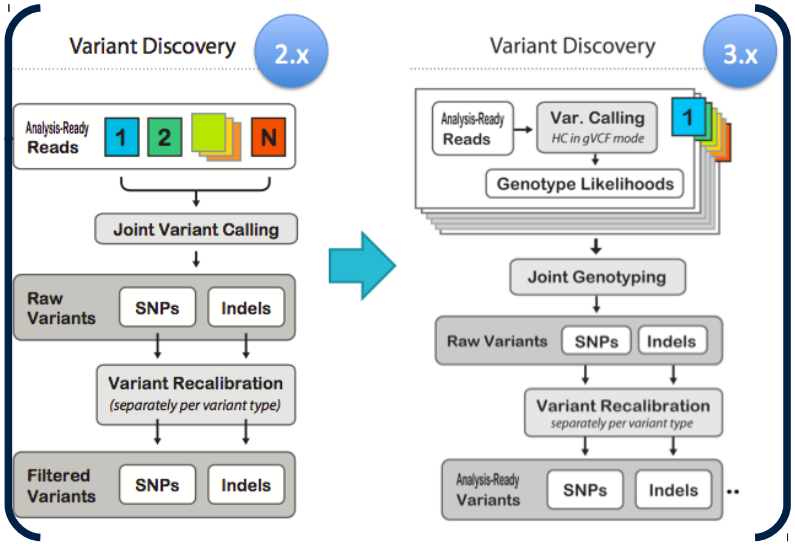
Read mapping to ref. genome

v

Read deduplication

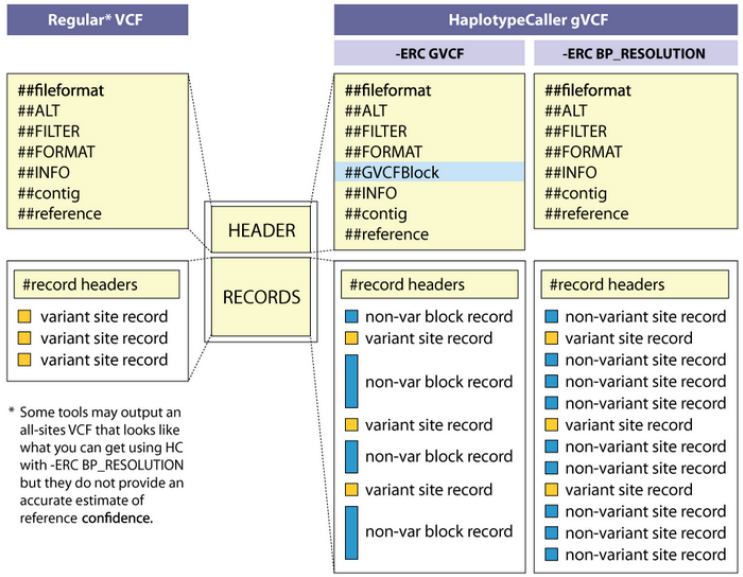
4. Bioinf. Analysis III

Variant calling

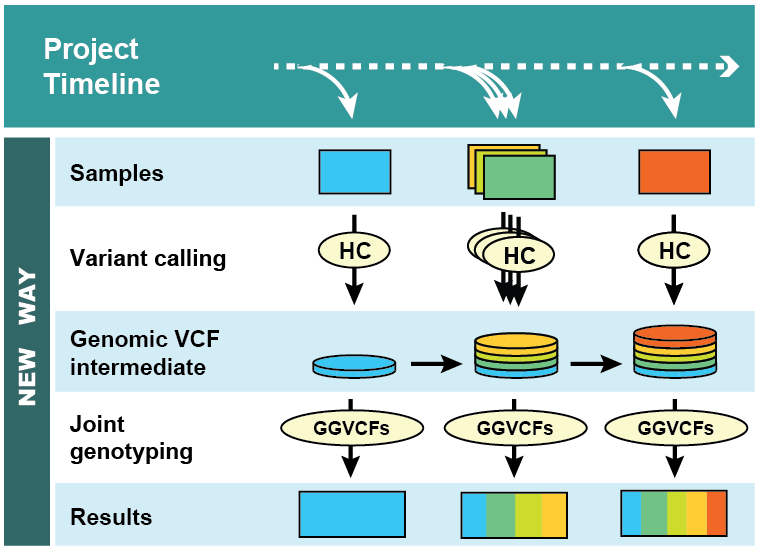


GATK solution: **joint variant genotyping**

1) call variants individually on each sample to get a comprehensive record of genotype likelihoods for each site in the genome (or exome), as a ***gVCF file*** →



2) perform a joint genotyping analysis of the gVCFs produced for all samples in a cohort. This way, you obtain the same accurate genotyping results, without the computational burden of exponential runtimes as the available cohort grows.



Source: GATK forums

Variant analysis workflow scheme – variant annotation frameworks

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

v

Read deduplication

4. Bioinf. Analysis III

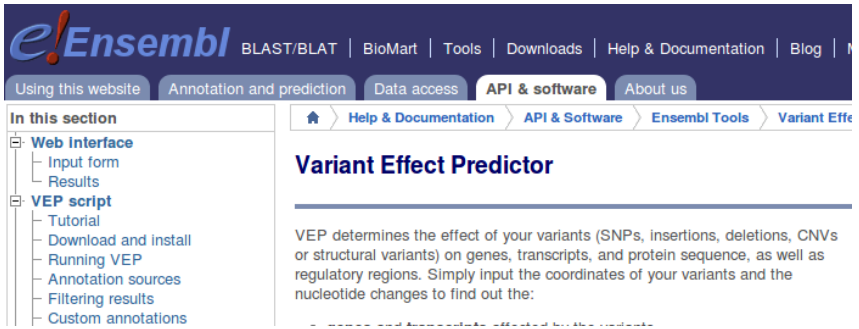
Variant calling

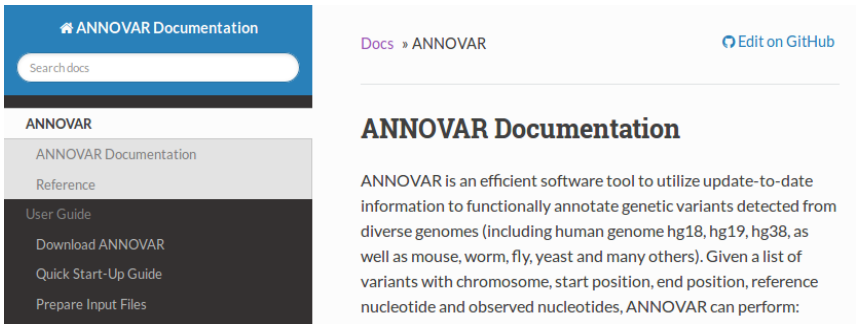
v

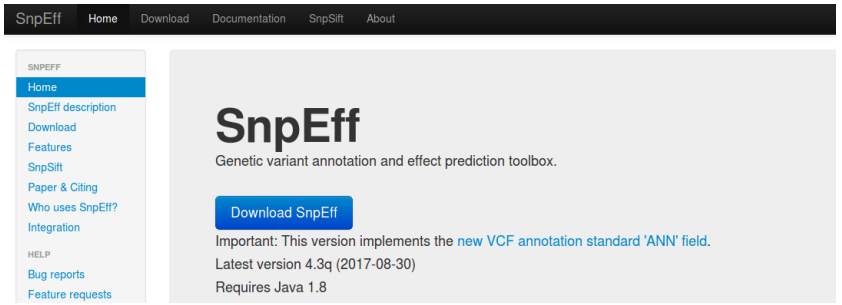
Variant annotation

Variant annotation is crucial for a more informative variant prioritisation downstream in the analysis workflow.

The most commonly used frameworks to annotate VCF files are:

- 

- 

- 

Variant analysis workflow scheme – variant annotation tools

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

v

Read deduplication

4. Bioinf. Analysis III

Variant calling

v

Variant annotation

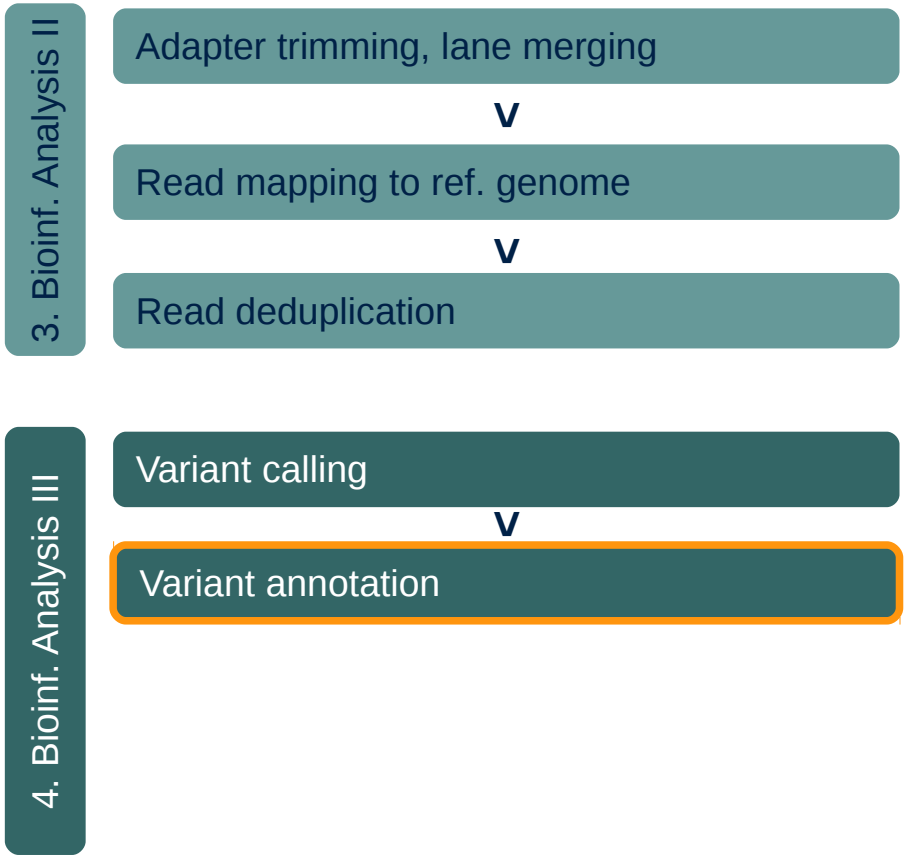
There are also a number of tools that allow you to manipulate VCF and BCF (binary version) files, as well as to add custom annotations. An example is the **bcftools suite**:

LIST OF COMMANDS

For a full list of available commands, run **bcftools** without arguments. For a full list of available options, run **bcftools** *COMMAND* without arguments.

- **annotate** .. edit VCF files, add or remove annotations
- **call** .. SNP/indel calling (former "view")
- **cnv** .. Copy Number Variation caller
- **concat** .. concatenate VCF/BCF files from the same set of samples
- **consensus** .. create consensus sequence by applying VCF variants
- **convert** .. convert VCF/BCF to other formats and back
- **csq** .. haplotype aware consequence caller
- **filter** .. filter VCF/BCF files using fixed thresholds
- **gtcheck** .. check sample concordance, detect sample swaps and contamination
- **index** .. index VCF/BCF
- **isec** .. intersections of VCF/BCF files
- **merge** .. merge VCF/BCF files files from non-overlapping sample sets
- **mpileup** .. multi-way pileup producing genotype likelihoods
- **norm** .. normalize indels
- **plugin** .. run user-defined plugin
- **polysomy** .. detect contaminations and whole-chromosome aberrations
- **query** .. transform VCF/BCF into user-defined formats
- **reheader** .. modify VCF/BCF header, change sample names
- **roh** .. identify runs of homo/auto-zygosity
- **sort** .. sort VCF/BCF files
- **stats** .. produce VCF/BCF stats (former vcfcheck)
- **view** .. subset, filter and convert VCF and BCF files

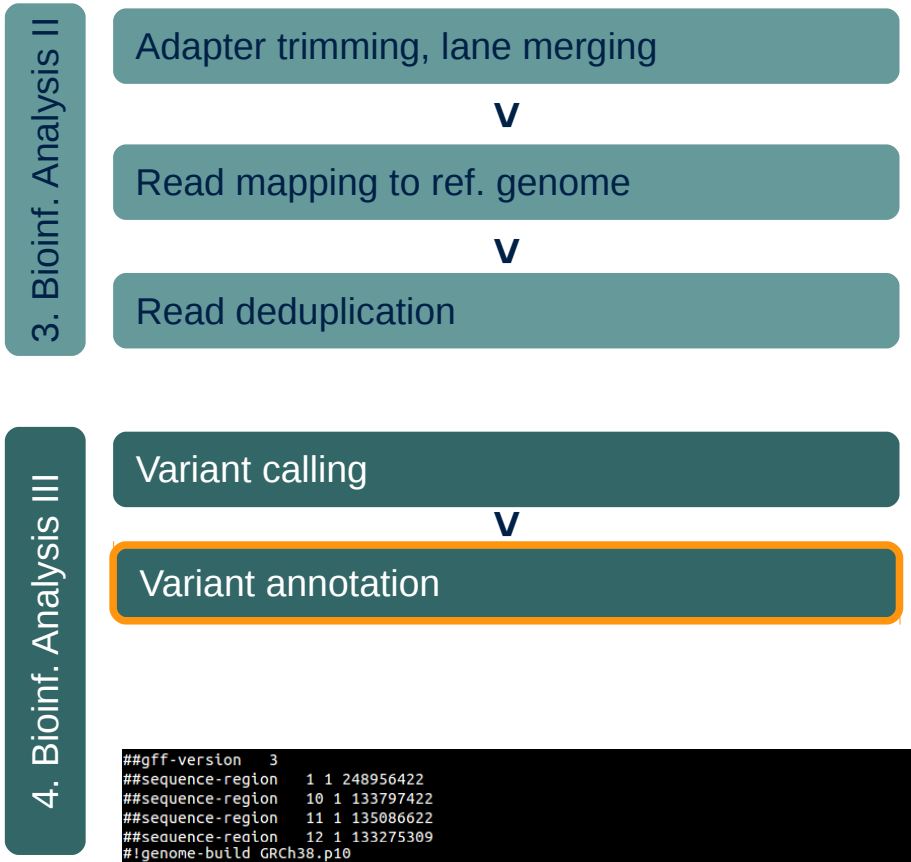
Variant analysis workflow scheme – variant annotation databases



The sources most commonly used to annotate variants include:

- **large genomic projects** (e.g. 1000 Genomes Project, UK10K, ExAC, ESP6500, WGS500, HapMap, etc.)
- **databases of known variantion** (e.g. dbSNP, dbVar, etc.)
- **consequence on the protein sequence** (e.g. stop lost, frameshift variant, missense variant, etc.)
- **location of the variant** (e.g. upstream of a transcript, in a coding sequence, in non-coding RNA, in regulatory regions)
- **overlapping (or nearest) gene** symbol, HGNC identifier, etc.
- **overlapping (or nearest) protein** product, Uniprot ID, AA pos.
- **conservation scores** (e.g. PhastCons, PhyloP, GERP++, etc.)
- **clinically relevant variants** (e.g. ClinVar, HGMD, etc.)
- **somatic mutations in cancer** (e.g. COSMIC, TCGA, etc.)
- **overlap with segmental duplications or repeated regions**
- **scores for deleteriousness** based on sequence homology and physical properties of the AA (e.g. SIFT, PolyPhen, etc.)
- **motifs** (i.e. alterations to transcription factor binding sites)

Variant analysis workflow scheme – GFF/GTF file format



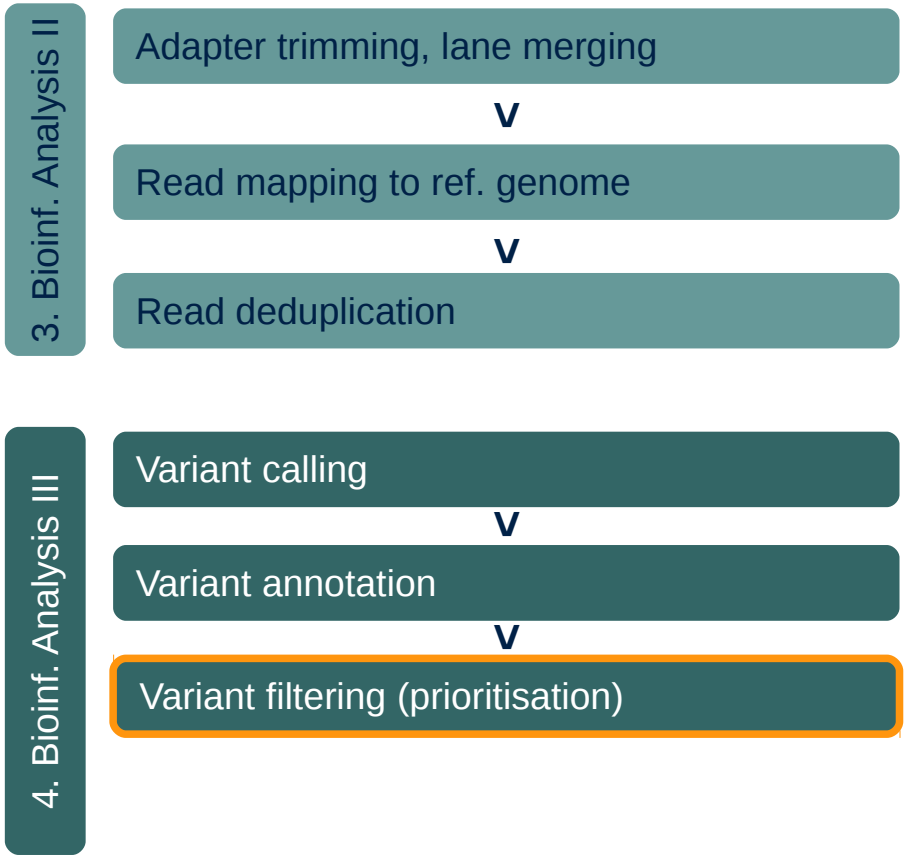
The annotation databases are often stored as GFF/GTF files.

They contain a header section (starting with '#') and 9 tab-separated fields:

1. **seqname**: chromosome or scaffold name
2. **source**: program that generated this feature, or DB
3. **feature**: feature type (e.g. gene, exon, ...)
4. **start**: start position of the feature
5. **end**: end position of the feature
6. **score**: floating point value
7. **strand**: either + or -
8. **frame**: one of '0', '1', or '2' to indicate the base nº in the codon
9. **attribute**: semicolon-separated list of tag-value pairs

```
##gff-version   3
##sequence-region   1 1 248956422
##sequence-region   10 1 133797422
##sequence-region   11 1 135086622
##sequence-region   12 1 133275309
#!genome-build GRCh38.p10
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.25
#!genebuild-last-updated 2017-06
1       havana  gene    11869   14409   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
1       havana  transcript      11869   14409   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name "DDX11L1"; gene_sourc
1       havana  exon    11869   12227   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "1"; gene_name "DDX11L1"; g
1       havana  exon    12613   12721   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "2"; gene_name "DDX11L1"; g
1       havana  exon    13221   14409   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "3"; gene_name "DDX11L1"; g
1       havana  transcript      12010   13670   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; gene_name "DDX11L1"; gene_sourc
1       havana  exon    12010   12057   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "1"; gene_name "DDX11L1"; g
1       havana  exon    12179   12227   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "2"; gene_name "DDX11L1"; g
1       havana  exon    12613   12697   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "3"; gene_name "DDX11L1"; g
1       havana  exon    12975   13052   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "4"; gene_name "DDX11L1"; g
1       havana  exon    13221   13374   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "5"; gene_name "DDX11L1"; g
1       havana  exon    13453   13670   .       +       .       gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "6"; gene_name "DDX11L1"; g
1       havana  gene    14404   29570   .       -       .       gene_id "ENSG00000227232"; gene_version "5"; gene_name "WASH7P"; gene_source "havana"; gene_biotype "unprocessed_pseudogene";
1       havana  transcript      14404   29570   .       -       .       gene_id "ENSG00000227232"; gene_version "5"; transcript_id "ENST00000488147"; transcript_version "1"; gene_name "WASH7P"; gene_source
1       havana  exon    29534   29570   .       -       .       gene_id "ENSG00000227232"; gene_version "5"; transcript_id "ENST00000488147"; transcript_version "1"; exon_number "1"; gene_name "WASH7P"; ge
1       havana  exon    24738   24891   .       -       .       gene_id "ENSG00000227232"; gene_version "5"; transcript_id "ENST00000488147"; transcript_version "1"; exon_number "2"; gene_name "WASH7P"; ge
```

Variant analysis workflow scheme – filtering tools



Why do we need to filter variants?
To narrow down the list of candidate variants and to identify pathogenic/disease-associated mutations.

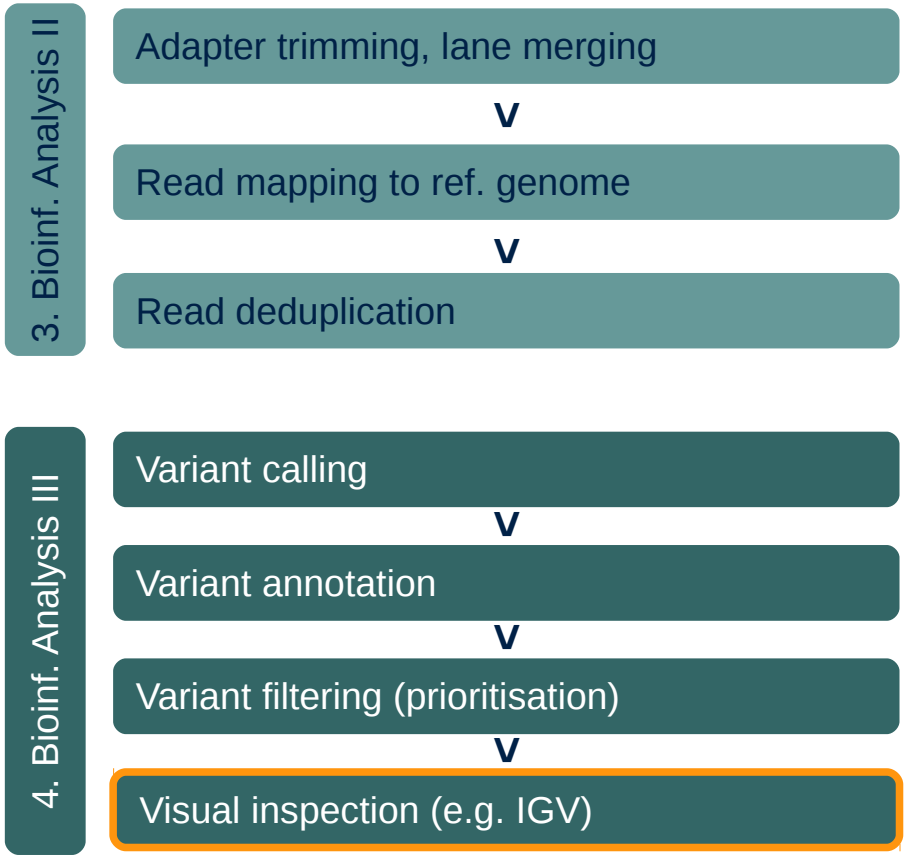
How can we filter variants?
Depending on the experimental context, researchers tune the thresholds of a number of potential filters. The most common steps include:

- quality scores: *remove poor quality variants*
- allele frequency: *discard common polymorphisms*
- presence in public databases
- effect on protein translation / splicing / binding sites: *prioritise variants with high functional impact on transcription*
- inheritance model (X-linked, autosomal recessive, etc.)
- read coverage of a given variant
- etc...

Variant filtering tools can be classified in:

- **command-line** *(e.g. VCFtools, GEMINI)*
- **GUI-based tools** *(e.g. Ingenuity, VariantStudio, VCF-Miner)*
- **both options available** *(e.g. BrowseVCF)*

Variant analysis workflow scheme – visualisation tools (1)



Visual representation of data can provide a crucial help for the interpretation of obtained results. Most visualisation tools allow the user to:

- **interpret** sequence data of *de novo* or re-sequencing experiments
- **browse** mapped experimental data (e.g. SNPs or SVs) in combination with different types of annotation
- **compare** sequences from multiple organisms or individuals.

The two main types of visualisation tools are:

- **web-based applications** running on dedicated web servers

  offer a variety of genomic annotations
  no need to install packages and dependencies
  data upload poses privacy and security issues

- **stand-alone tools**, most of which with a GUI

  can be used on any platform (Windows, Mac, Linux)
  can be significantly quicker for large datasets (as it's local)
  need to download annotations and keep them up-to-date

Most commonly used software include: Integrative Genome Viewer (IGV), UCSC Genome Browser, Vista and Savant.

Variant analysis workflow scheme – visualisation tools (2)

3. Bioinf. Analysis II

Adapter trimming, lane merging

v

Read mapping to ref. genome

v

Read deduplication

4. Bioinf. Analysis III

Variant calling

v

Variant annotation

v

Variant filtering (prioritisation)

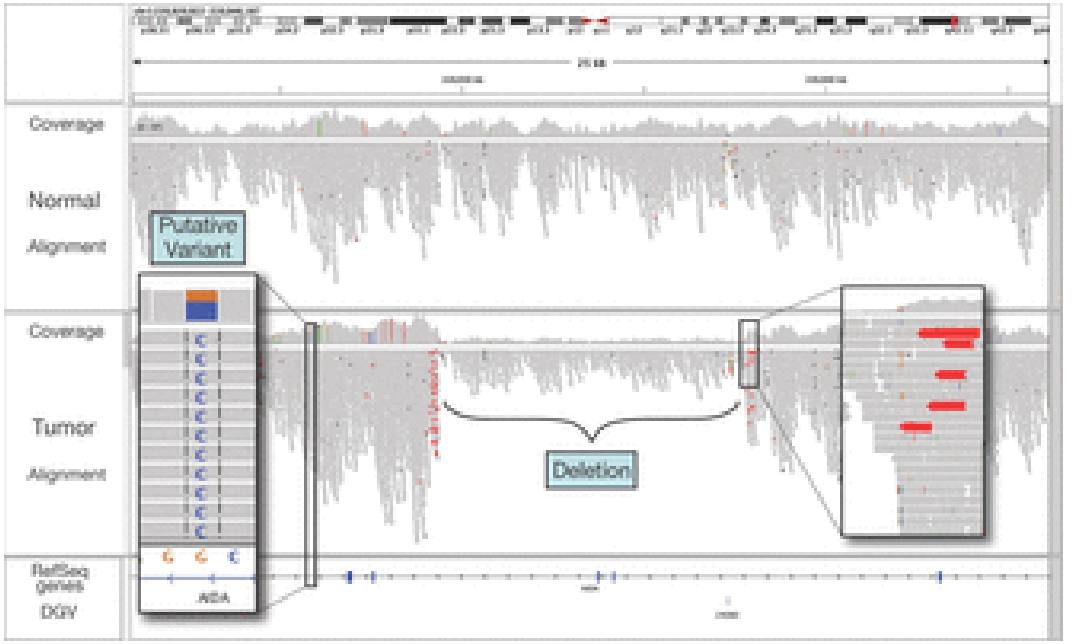
v

Visual inspection (e.g. IGV)

Considerations to keep in mind when interpreting aligned sequences on a browser:

- reads mapped with many mismatches should not be trusted
- mutations only backed by a small fraction of reads should be discarded
- reads should only be trusted for further processing if they align at a unique starting position

(Source: Pabinger et al., Brief Bioinform 2014)



Variant analysis workflow scheme – validation by Sanger sequencing

3. Bioinf. Analysis II

- Adapter trimming, lane merging
- Read mapping to ref. genome
- Read deduplication

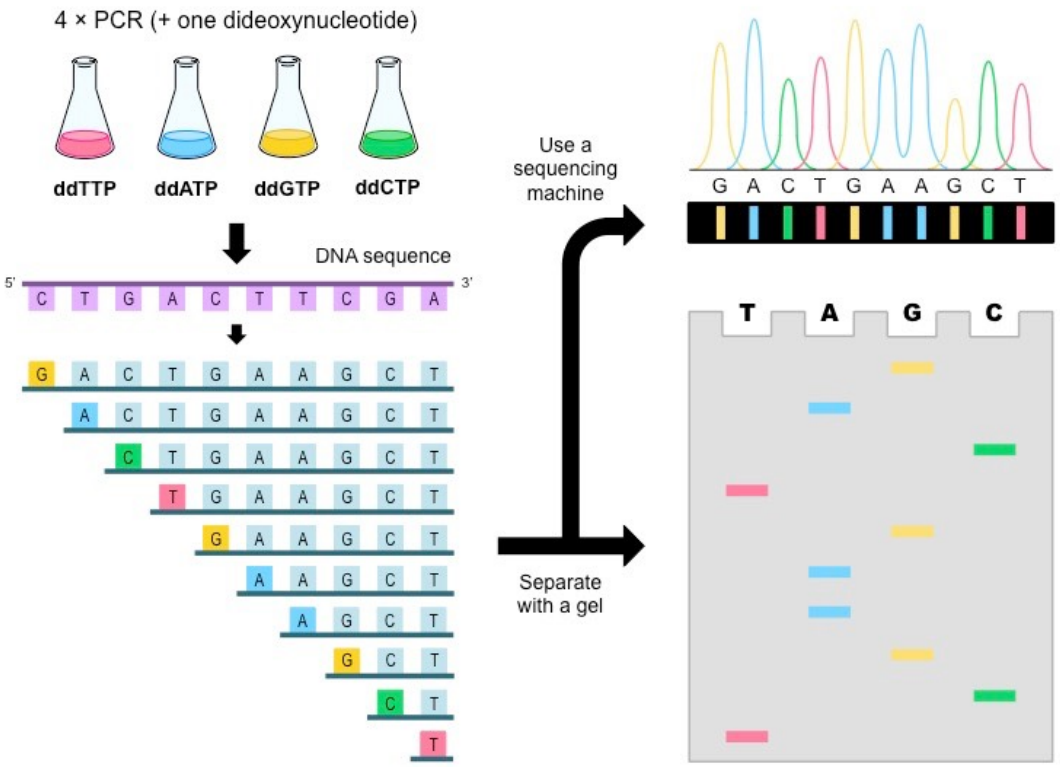
4. Bioinf. Analysis III

- Variant calling
- Variant annotation
- Variant filtering (prioritisation)
- Visual inspection (e.g. IGV)

5. Wet lab

Confirmation by:

- Sanger sequencing
- MLPA (Multiplex Ligation-dependent Probe Amplification, a multiplex PCR method detecting copy numbers at 1bp resolution)



Variant analysis workflow scheme – validation by MLPA

3. Bioinf. Analysis II

- Adapter trimming, lane merging
- Read mapping to ref. genome
- Read deduplication

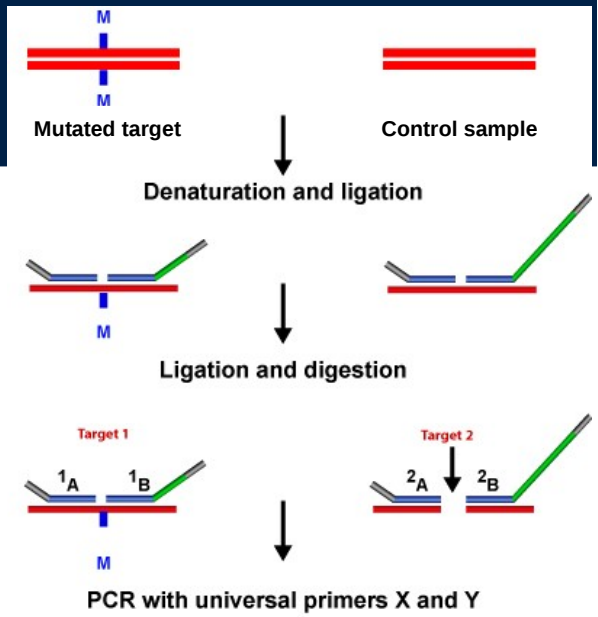
4. Bioinf. Analysis III

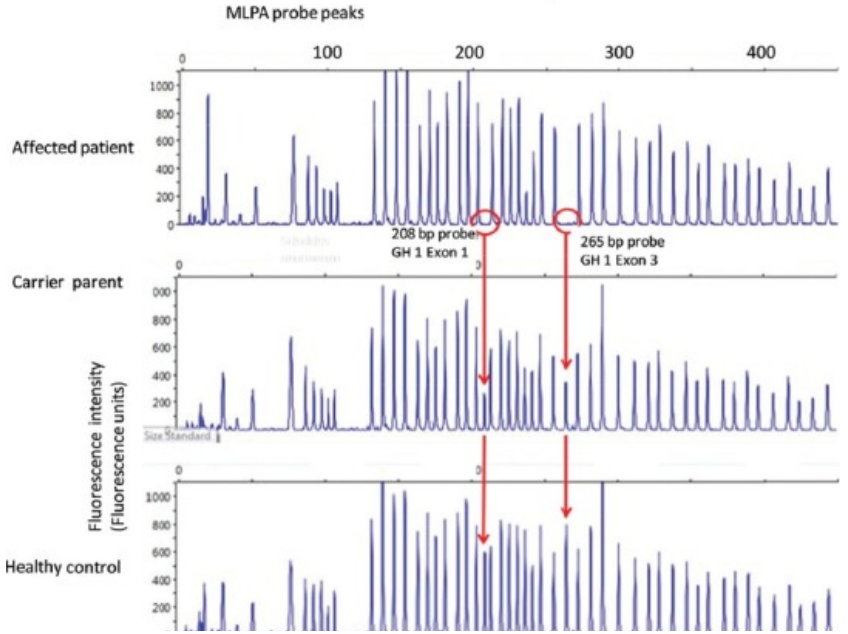
- Variant calling
- Variant annotation
- Variant filtering (prioritisation)
- Visual inspection (e.g. IGV)

5. Wet lab

Confirmation by:

- Sanger sequencing
- MLPA (Multiplex Ligation-dependent Probe Amplification, a multiplex PCR method detecting copy numbers at 1bp resolution)





Extra steps – phenotype-based analysis

Many genetic conditions show overlapping features (→ “disease families”).

E.g.: Marfan syndrome

(FBN1 gene)

Congenital contractural arachnodactyly

(FBN2 gene)

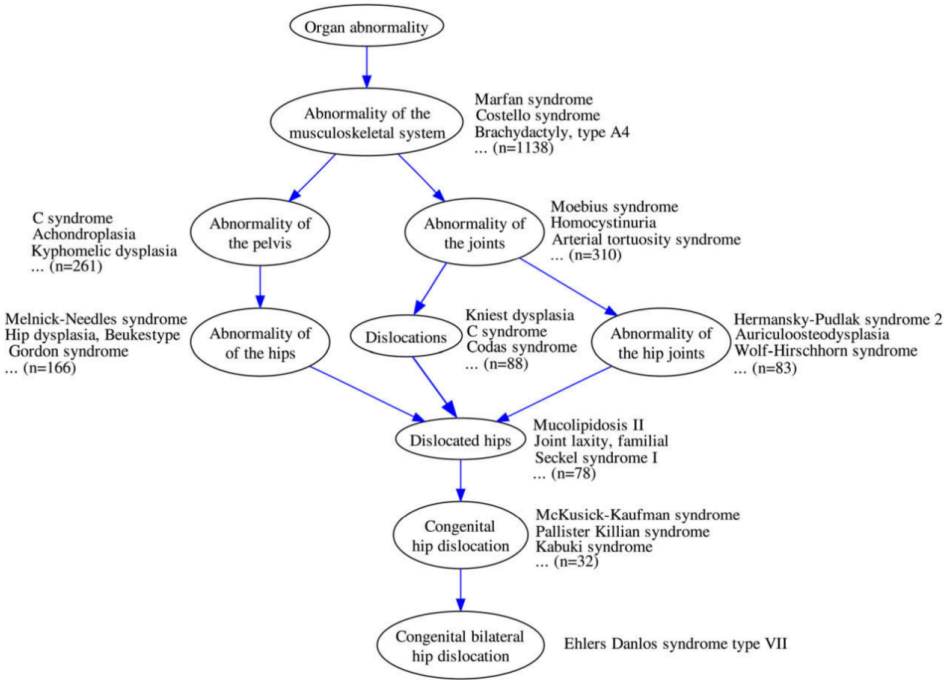
Phenotypic similarities within disease families may be related to dysfunction of a regulatory network or signalling pathway.

The **Human Phenotype Ontology** (**HPO**) database stores all phenotypic abnormalities commonly found in human monogenic (or “mendelian”) diseases.

When the diagnosis is not known in advance, or if a novel disease gene is being sought, **computational phenotype analysis** can measure the similarity between a set of query terms representing the patient's clinical manifestations and those representing each of the diseases in a database.

Exome analysis **tools**:
eXtasy, Phevor, Phen-Gen, Exomiser, PhenIX, etc.

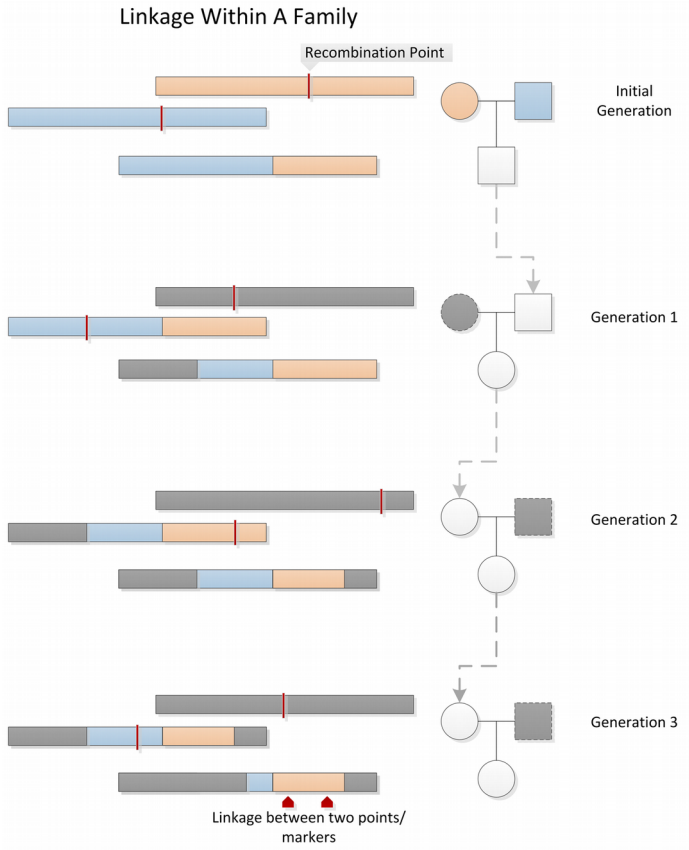
Source: Smedley et al. Genome Medicine, 2015

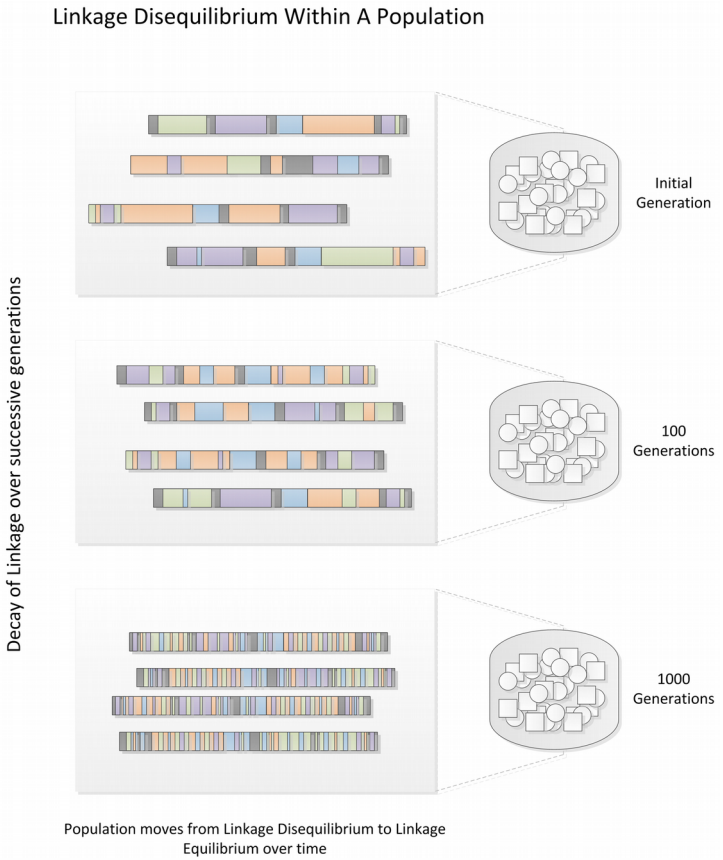


Source: Smedley et al. Am J Hum Genet. 2008

Extra steps – Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles at different loci in a given population. LD is influenced by many factors (selection, mutation rate, genetic drift, population structure, etc.).





Source: Bush et al. Plos Comp Bio 2012

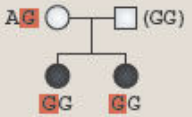
| = two genetic markers that remain linked on a chromosome rather than being broken apart by recombination events during meiosis.

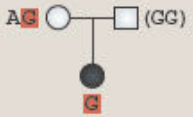
In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events, moving from LD to LE, as recombination eventually occur between every possible point on the chromosome.

Extra steps – Linkage analysis and GWAS

Exome and whole-genome sequencing is often preceded by genetic **Linkage analysis**, which allows variants outside of linkage peaks to be excluded. A measure for the likelihood of linkage is the **LOD score**, i.e. the logarithm of the odds that the loci are linked divided by the odds that the loci are unlinked.

Genome-Wide Association Studies (GWAS) examine SNPs throughout the genome in thousands of individuals (in disease and control groups) to identify alleles associated with disease. Comparison of allele frequencies between the two groups reveal genotypes that are overrepresented in the disease group and are therefore associated with disease risk.

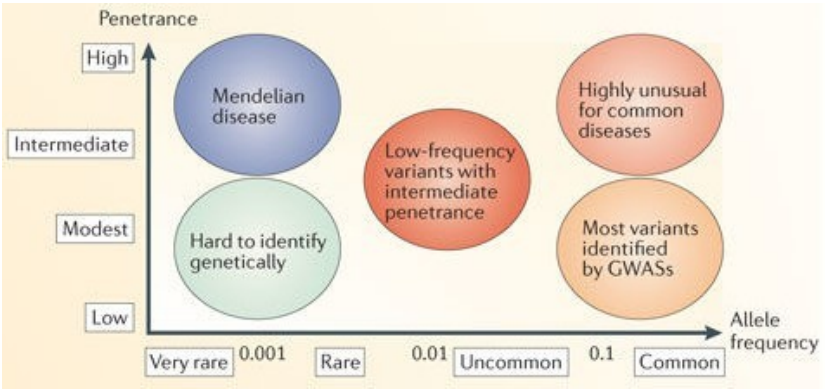




Property of mapping approach	Linkage analysis	Association analysis
Data type studied	Relatives	Unrelated or related individuals
Relevant parameter	Recombination fraction	Association statistic
Range of effect detected (linkage or association)	Long (≤5 Mb)	Short (≤100 kb)
Number of markers required for genome-wide coverage	Moderate (500–1,000)	Large (>100,000)
Statistics used	Cumbersome (requires tailor-made likelihood methods)	Elegant; can use the range of classical statistical tools
Dealing with correlated markers	Pose problems in presence of ungenotyped individuals	Can be handled efficiently
Biological basis of approach	Observe (or infer) recombination in pedigree data	Exploit unobserved recombination events in past generations
Dealing with allelic heterogeneity	Not a problem	Reduces power
Detecting genotyping errors	Potentially detected as Mendelian inconsistencies	Potentially detected only in family data, but not in case–control data
Most suitable application	Rare, dominant traits	Common traits

Nature Reviews | Genetics

Source: Ott et al. Nature Rev. Genet. 2011



Source: Sullivan et al. Nature Rev. Genet. 2012

Further reading...

- *"**Choice of transcripts and software** has a large **effect on variant annotation**"*
  https://www.ncbi.nlm.nih.gov/pubmed/24944579
- *"From FastQ data to high confidence variant calls: the **Genome Analysis Toolkit** best practices pipeline"*
  https://www.ncbi.nlm.nih.gov/pubmed/25431634
- *"**Phenotype-driven** strategies for exome prioritization of human Mendelian disease genes"*
  https://www.ncbi.nlm.nih.gov/pubmed/26229552
- *"Genotype and **SNP calling** from next-generation sequencing data"*
  https://www.ncbi.nlm.nih.gov/pubmed/21587300
- *"Computational methods and resources for the interpretation of genomic **variants in cancer**"*
  https://www.ncbi.nlm.nih.gov/pubmed/26111056

Questions?

silvia@well.ox.ac.uk