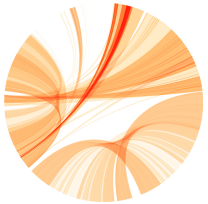
Practical Bioinformatics

Variant Calling

Part 2

Stefan Wyder
stefan.wyder@uzh.ch
URPP Evolution
www.evolution.uzh.ch

Universität Zürich[UZH]



URPP

Variant Calling Workflow

Alignment Postprocessing

- Filtering reads based on flags
  (use samtools or Picard)
- Duplicate Removal (PCR)
- Indel Realignment
- Base Recalibration

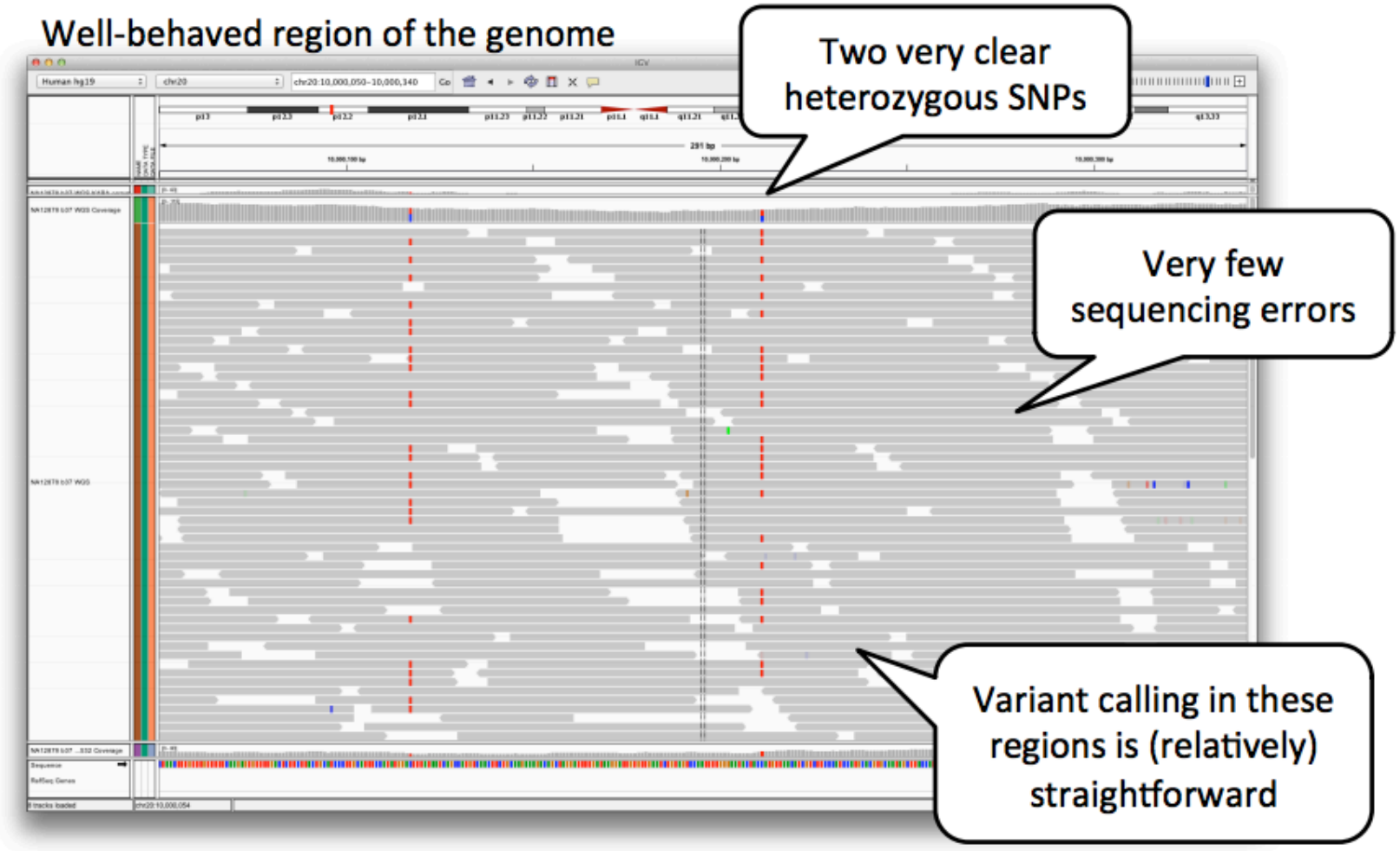
INDEL Realignment

SNP Discovery: Goal

Distinction of system noise (instrument errors, PCR errors, ..) from real variation

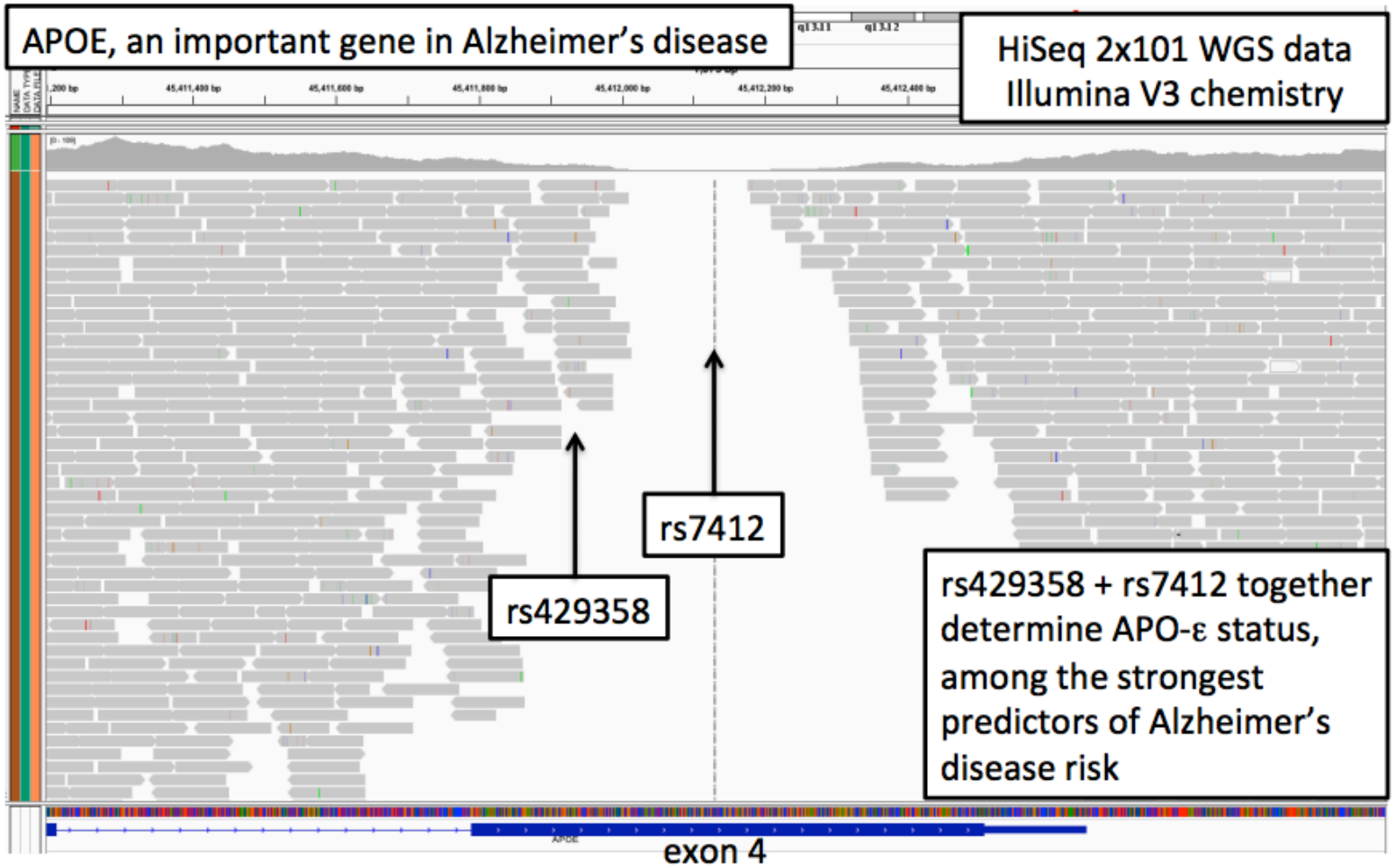
bioinformatics.ca

Analysis of SNPs in well-behaved regions of the genome is pretty simple



http://www.broadinstitute.org/gatk//events/2247/agbt_2013_depristo.pdf

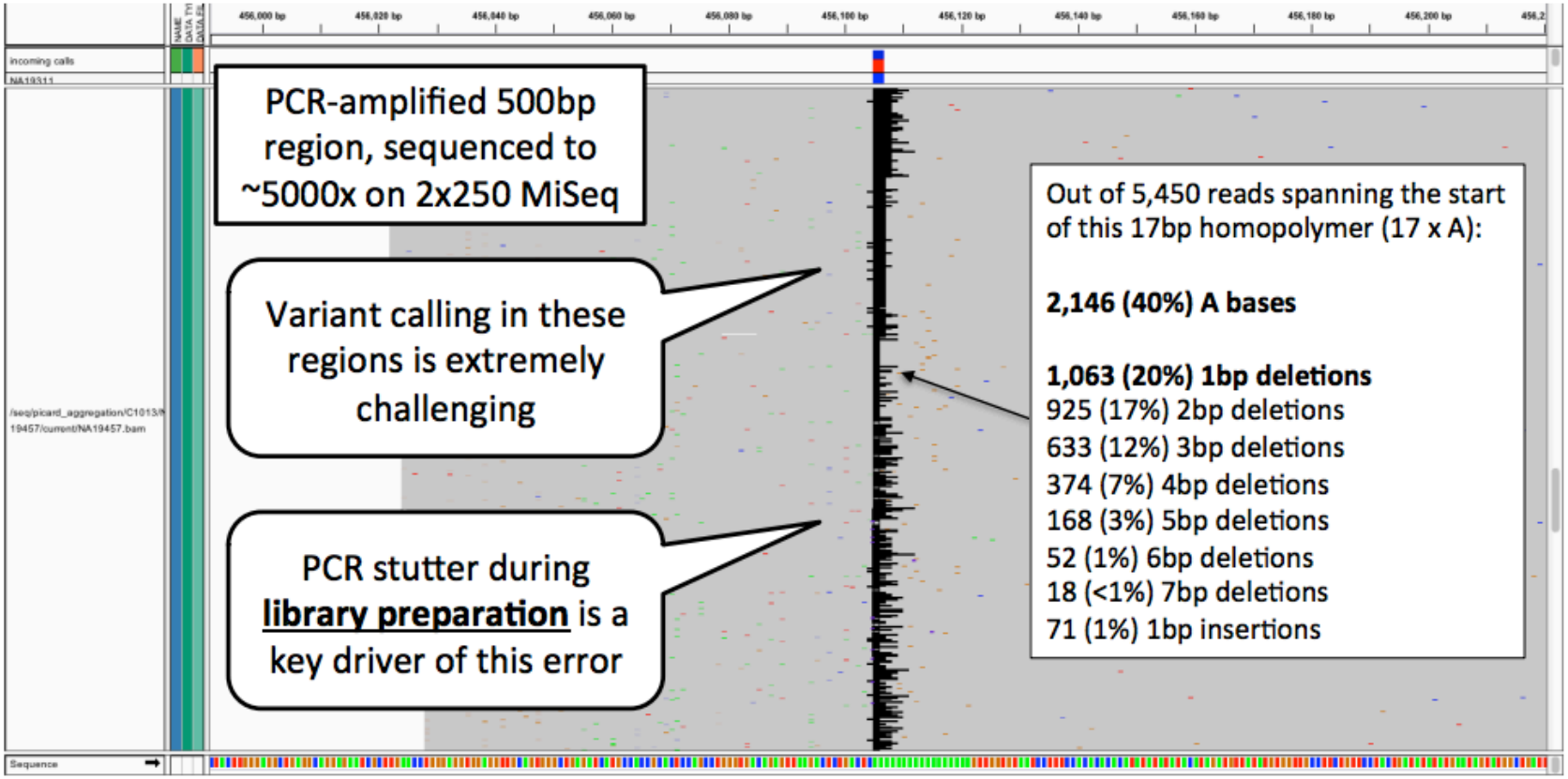
… but lack of coverage blinds us to many genomic regions



http://www.broadinstitute.org/gatk//events/2247/agbt_2013_depristo.pdf

…it can get even worse

Poorly-behaved region of the genome



http://www.broadinstitute.org/gatk//events/2247/agbt_2013_depristo.pdf

SNP Calling

- modelling various error types
- expected distribution of calls
  (homozygous AA, homozygous variant BB, heterozygous AB)

- Correct genotyping depends on
  - sequence quality values
  - read depth
  - correct alignment

SNP Calling 2

- To gain sensitivity some SNP callers allow multi-sample variant calling (multiple individuals/samples from the same or closely related species)

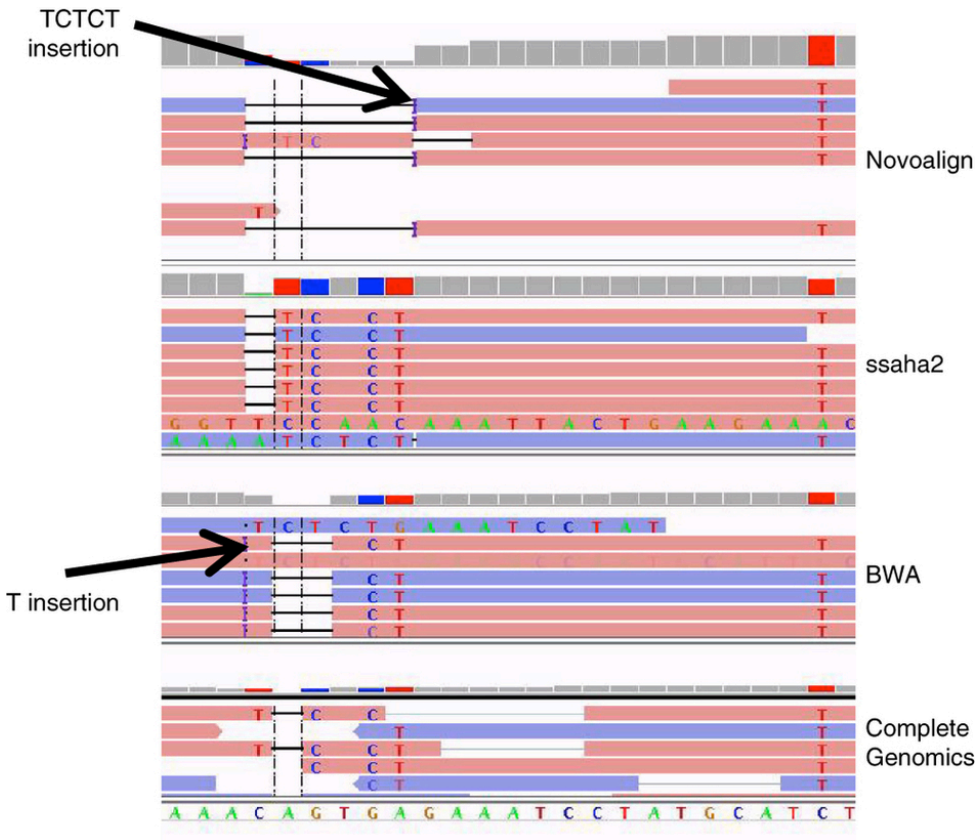
Possible problems

- inadequate read coverage of a region
- mismapped reads / errors in the alignment
  - segmental duplication
  - processed pseudogenes
  - close paralogues
  - repetitive sequences
  - small but complex indels
- incomplete/missassembled reference genome

Complex variants have multiple representations

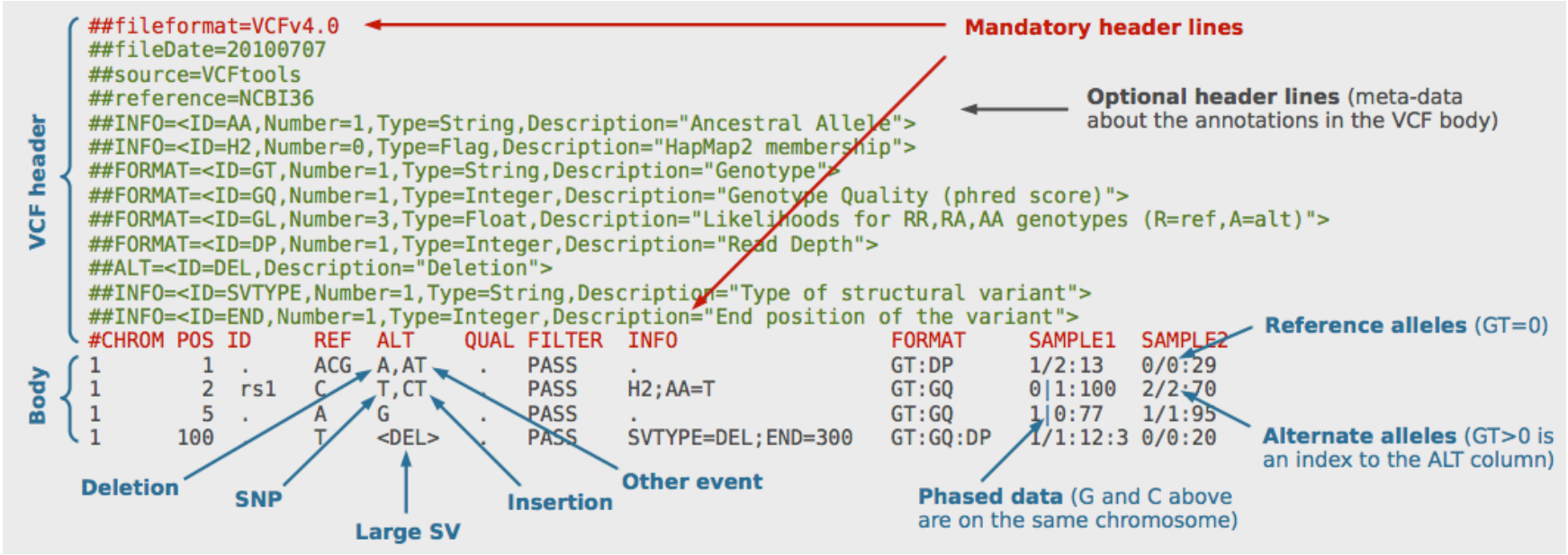
Different data - and mappers

6 bases CAGTGA are replaced by the 5 bases TCTCT
1: 114841792–114841797



Zook et al. Nature Biotechnology (2014)

VCF format



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

VCF examples

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

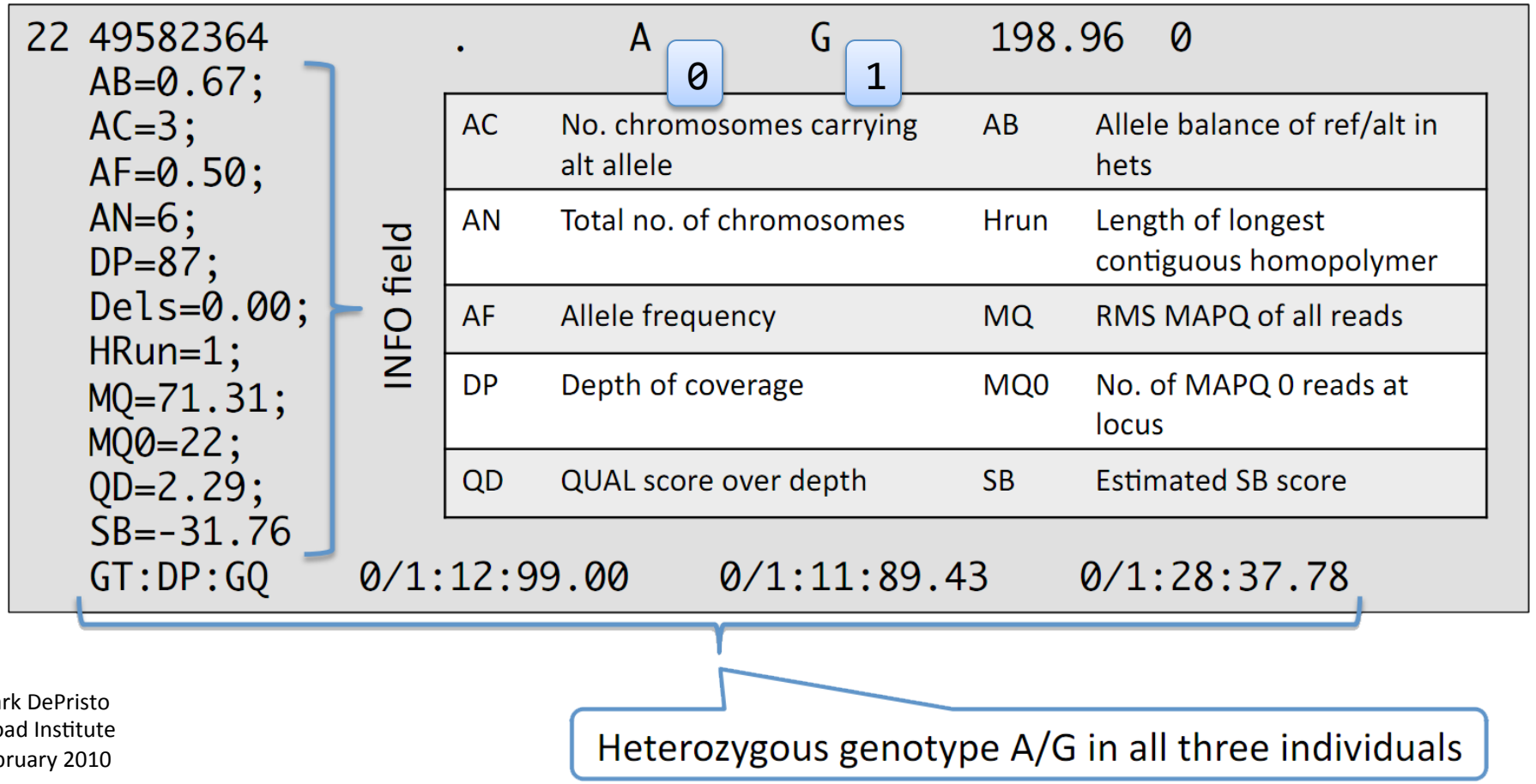
VCF representation

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

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

VCF info field

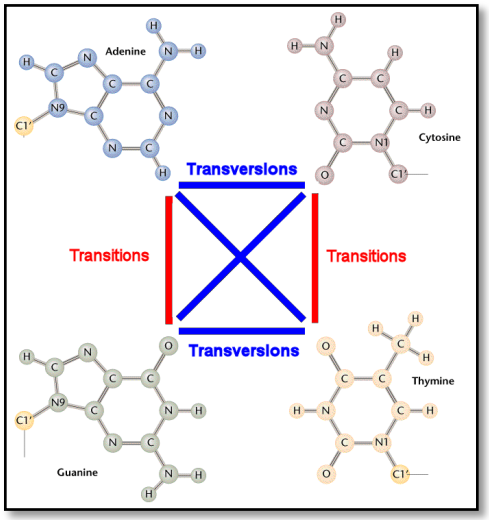
VCF record for an A/G SNP at 22:49582364



Mark DePristo
Broad Institute
February 2010

Variant Filtering

- The optimal threshold for filtering has to be determined empirically
- trade-off sensitivity <-> specificity
- which metric of variant call confidence?



Intrinsic

- Transitions:transversions ratio (Ti/Tv)
  (e.g. nuclear genes in humans close to 2)

Experimental Validation

- Small-scale validation (Sanger seq, qPCR, pyrosequencing, ...)
- Orthogonal data (e.g. microarrays, different seq platform)
- Concordance among Trios

Sources & Links

Article Collections

- Review Articles from Nature Reviews Genetics
- PLoS Computational Biology: Education

Material

- GATK http://www.broadinstitute.org/gatk/
- SEQanswers NGS forum http://seqanswers.com/
- Biostar http://biostars.org/
- List of Applications http://seqanswers.com/wiki/Special:BrowseData/

Sequencing Errors

- Error rate and error profile are technology-specific

Illumina Sequencing

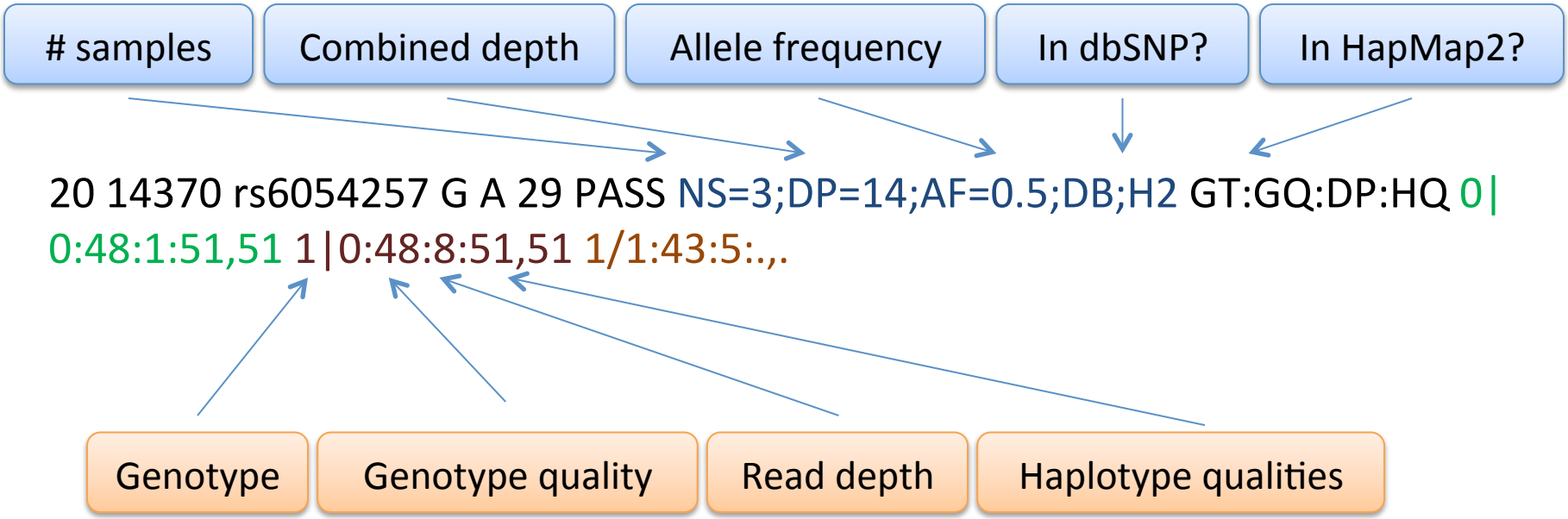
- Error Rate: > 0.1% (i.e. > 1 in 1000)
- mainly substitutions errors
- errors mostly at the end of reads

- PCR amplification bias

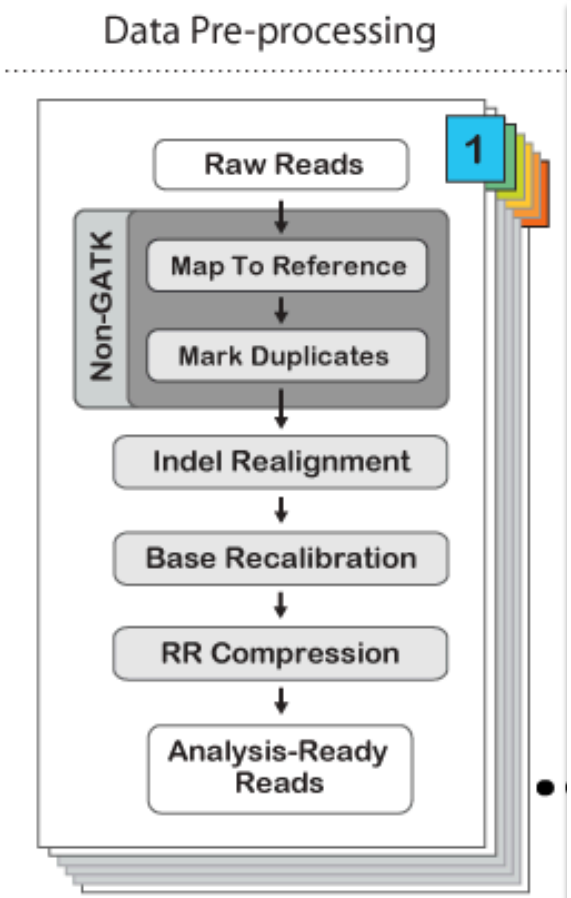
VCF

##fileformat=VCFv4.0

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003



GATK workflow



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