

# Practical Bioinformatics

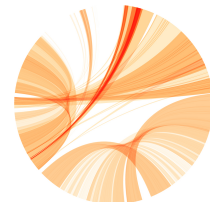
## Variant Calling

### Part 2

Stefan Wyder  
stefan.wyder@uzh.ch  
**URPP Evolution**  
[www.evolution.uzh.ch](http://www.evolution.uzh.ch)



**Universität  
Zürich**<sup>UZH</sup>



**URPP**

# Variant Calling Workflow



# Alignment Postprocessing

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

# INDEL Realignment

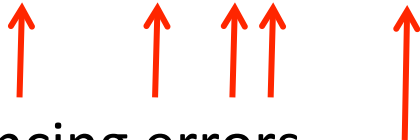
```
TAAATAATGAAATTTATTTCTCAGAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG++++AGGGT++++GCACTCTCTGCTTCATAAATGGGTCTCTTGCCGCAAAAAAATCTGTTTGCTCCTCCAGATTCATCAA
<-  TGGAATTTATTTCTCAGAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG*****AGG
<-  TGGAATTTATTTCTCAAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG*****AGG
<-  GGAAATTTATTTCTCAGAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG*****AGGG
->  GGAAATTTATTTCAACAGAGTAATGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGCTTCTAAGTCTGCTG*****AGGG
->                                     CAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG*****AGGGTAGGGCGCACTCTCTGCTTCATAAATGGGTCTCTTGC
->  ATTTCTCAGAGTACTGGAAGCTGGGACTCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG*****AGGGT*****AGGGTGC
<-  GTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGT*****GCACTCTCTGCT
<-  AATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGT*****GCACTCTCTGCTTCATAAATGGGTCTC
->  ATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGT*****GCACTCTCTGCTTCATAAATGGGTCTCTTGCCGCA
<-  GTCTGGTGAGGGTAGGGT*****GCACTCTCTGCTTCATAAATGGGTCTCTTGCCGCAAAAAAATCTGTTTGCTCCTCCAG
```

```
TAAATAATGAAATTTATTTCTCAGAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTG++++AGGGTGCACTCTCTGCTTCATAAATGGGTCTCTTGCCGCAAAAAAATCTGTTTGCTCCTCCAGATTCATCAA
<-  TGGAATTTATTTCTCAGAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGG
<-  TGGAATTTATTTCTCAAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGG
<-  GGAAATTTATTTCTCAGAGTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGG
->  GGAAATTTATTTCAACAGAGTAATGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGCTTCTAAGTCTGCTGAGGG
->                                     CAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGCGCACTCTCTGCTTCATAAATGGGTCTCTTGC
->  ATTTCTCAGAGTACTGGAAGCTGGGACTCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGTGCACTCTCTGCT
<-  GTACTGGAAGCTGGGAATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGTGCACTCTCTGCT
<-  AATCCAAGATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGTGCACTCTCTGCTTCATAAATGGGTCTC
->  ATCAAAATGCCAGCAGATTCTAAGTCTGGTGAGGGTAGGGTGCACTCTCTGCTTCATAAATGGGTCTCTTGCCGCA
<-  GTCTGGTGAGGGTAGGGTGCACTCTCTGCTTCATAAATGGGTCTCTTGCCGCAAAAAAATCTGTTTGCTCCTCCAG
```

# SNP Discovery: Goal

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

GTTACTGTCGTTGTAATACTCCACGATGTC  
GTTACTGTCGTTGTAATACTCCACGATGTC  
GTTACTGTCGTTGTAATACTCCACGATGTC  
GTTACTGTCGTTGTAATACTCCACAATGTC  
GTTACTGTCGTTGTAATgCTCCACGATGTC  
GTTACTGTCGTTGTAATACTCCACAATGTC  
GTTACTGTCGTTGTAATACTCCACGATGTC  
GTTACTGTCGTGGTAATACTCCACaATGTC  
GTTACTGTCGTTGTAATACTCCACaATGTC  
GTTAaTGTCGTTGTAATACTCCACGATGTC  
GTTACTGTCGTTGTAcTACTCCACGATGTC  
GTTACTGTCGTTGTAATACTCCACaATGTC

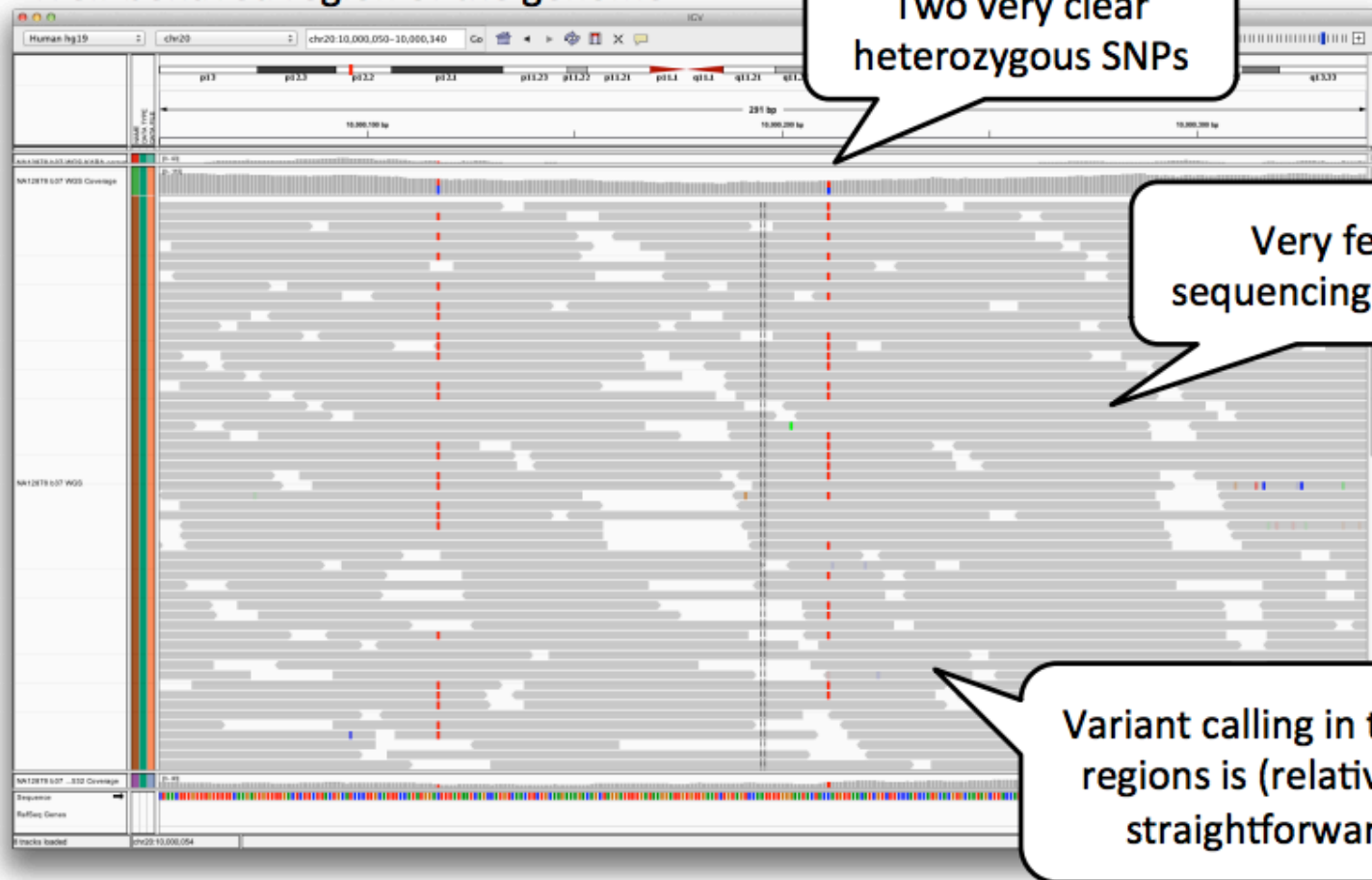


sequencing errors

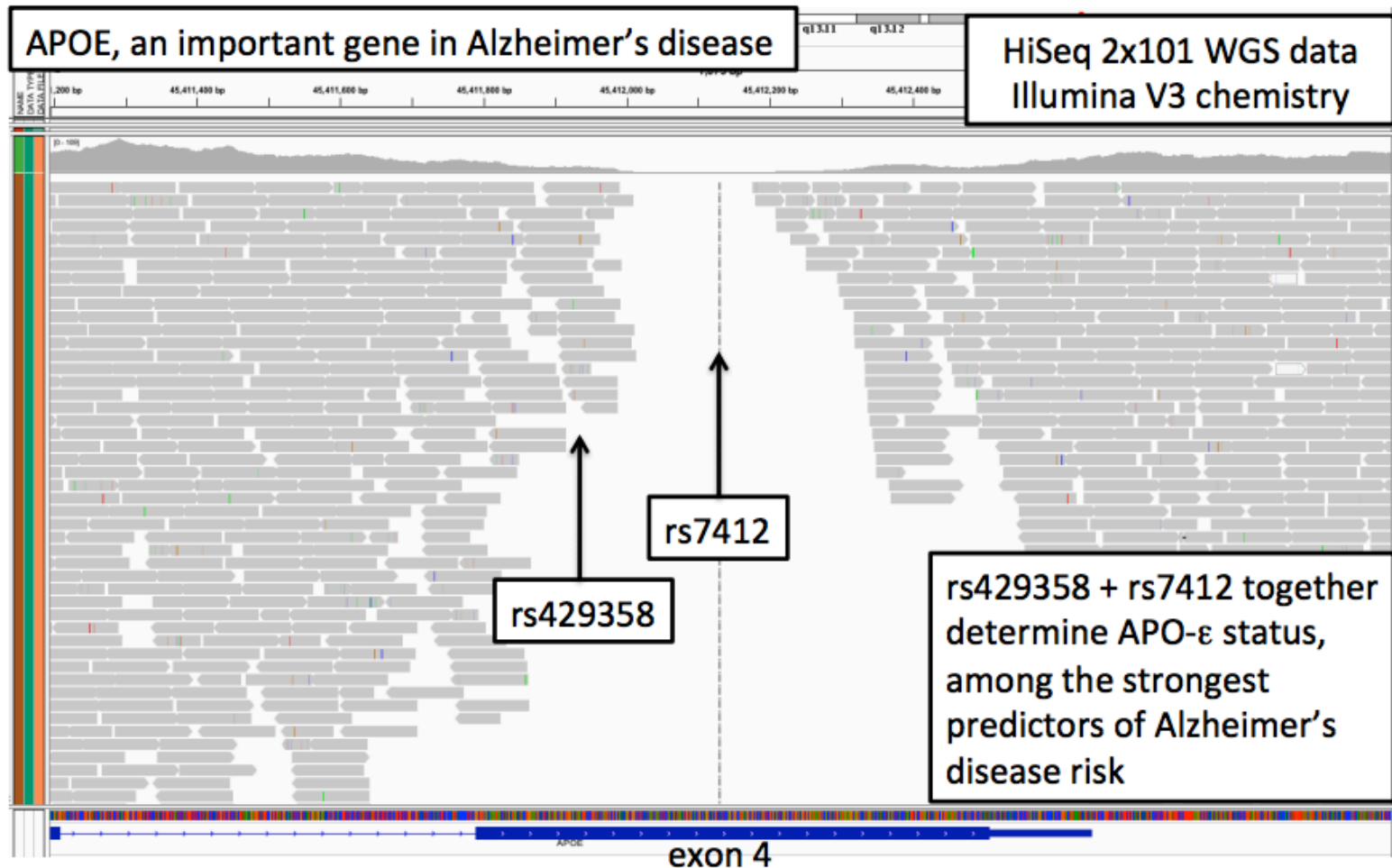
SNP

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

## Well-behaved region of the genome

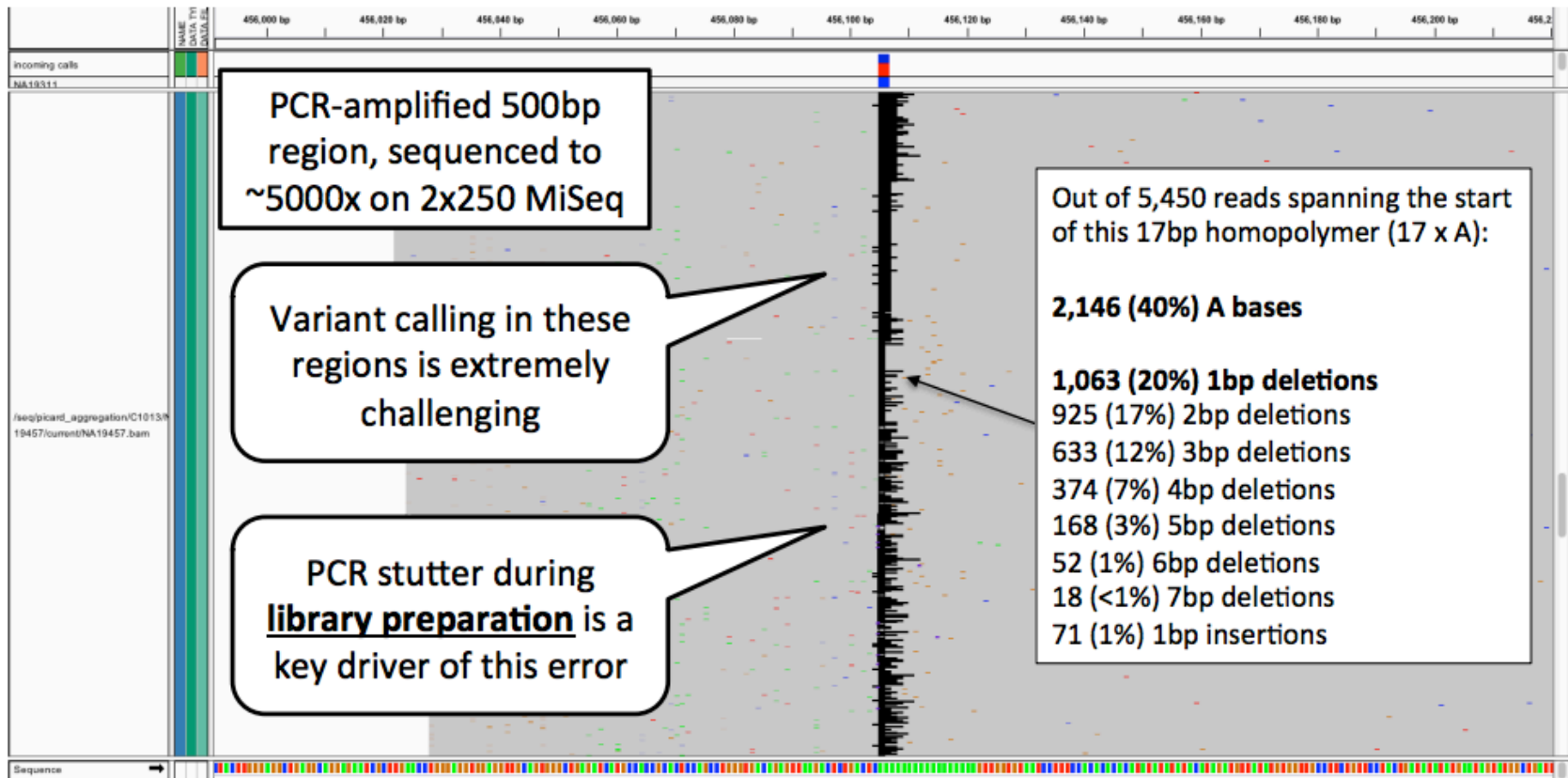


... but lack of coverage blinds us to many genomic regions



# ...it can get even worse

## Poorly-behaved region of the genome





# 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)

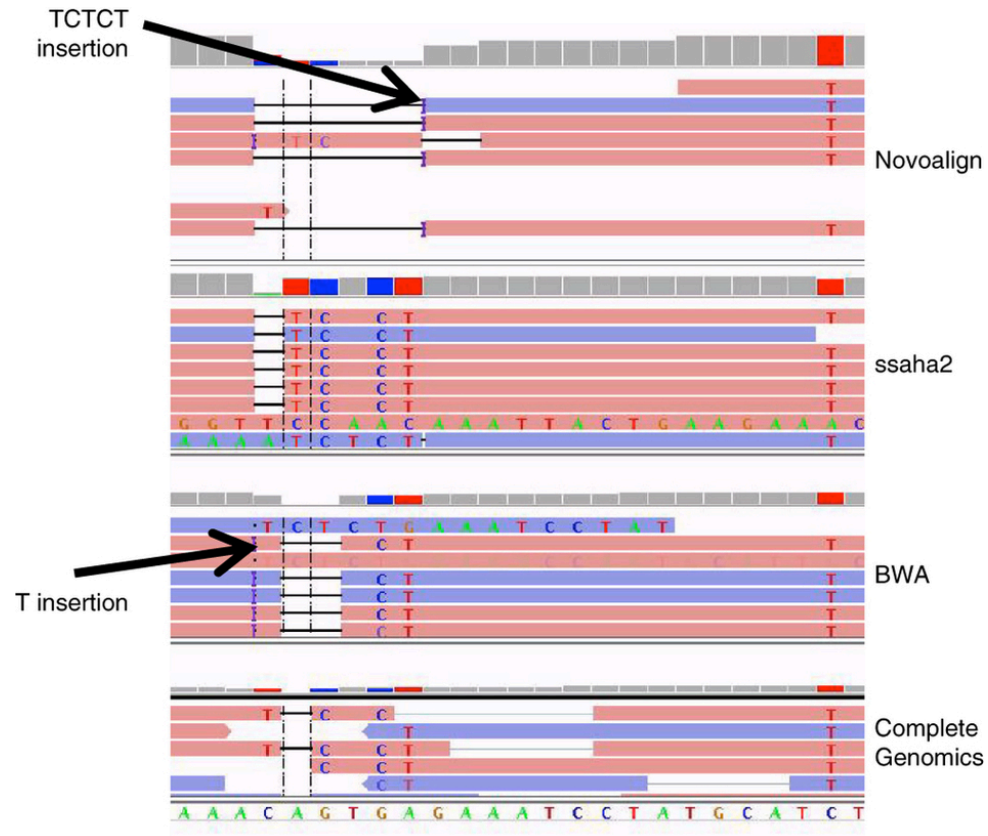
# Possible problems

- 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



# VCF format

**VCF header**

```
##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=alt)">
##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">
```

**Mandatory header lines**

**Optional header lines** (meta-data about the annotations in the VCF body)

**Body**

#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

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

**Large SV**

**Insertion**

**Other event**

**Phased data** (G and C above are on the same chromosome)

# 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

# VCF info field

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

22	49582364	.	A	G	198.96	0
AB=0.67; AC=3; AF=0.50; AN=6; DP=87; Dels=0.00; HRun=1; MQ=71.31; MQ0=22; QD=2.29; SB=-31.76 GT:DP:GQ		<div>0</div> <div>1</div>				
		AC	No. chromosomes carrying alt allele		AB	Allele balance of ref/alt in hets
		AN	Total no. of chromosomes		Hrun	Length of longest contiguous homopolymer
		AF	Allele frequency		MQ	RMS MAPQ of all reads
		DP	Depth of coverage		MQ0	No. of MAPQ 0 reads at locus
		QD	QUAL score over depth		SB	Estimated SB score
		0/1:12:99.00		0/1:11:89.43		0/1:28:37.78

INFO field

Heterozygous genotype A/G in all three individuals

# Variant Filtering

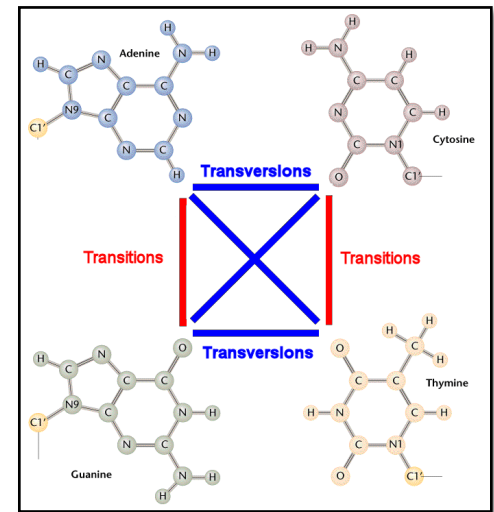
- The optimal threshold for filtering has to be determined empirically
- trade-off sensitivity <-> specificity
- which metric of variant call confidence?

## 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

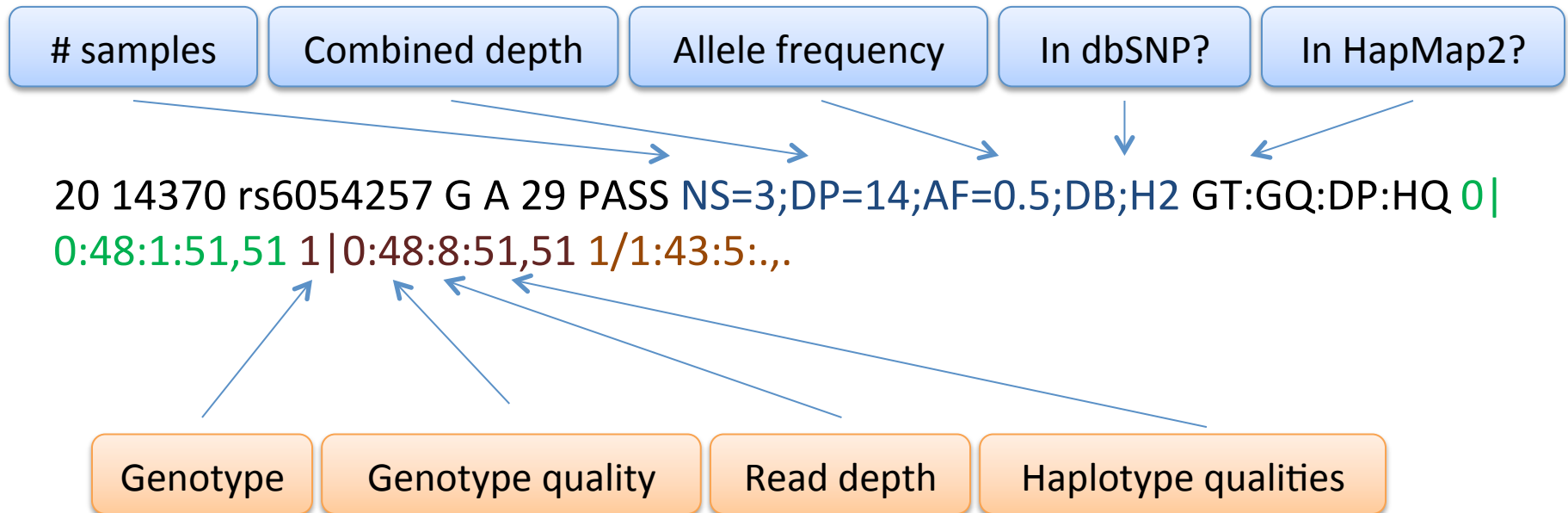
## Illumina Sequencing

- 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

## Data Pre-processing

