Next-generation sequencing

- Variation Discovery

Lecture 8

Outline

- Definition and motivation
- SNP distribution and characteristics
- SV detection strategy, tools and data summary
- Applications in other projects

Different types of variations

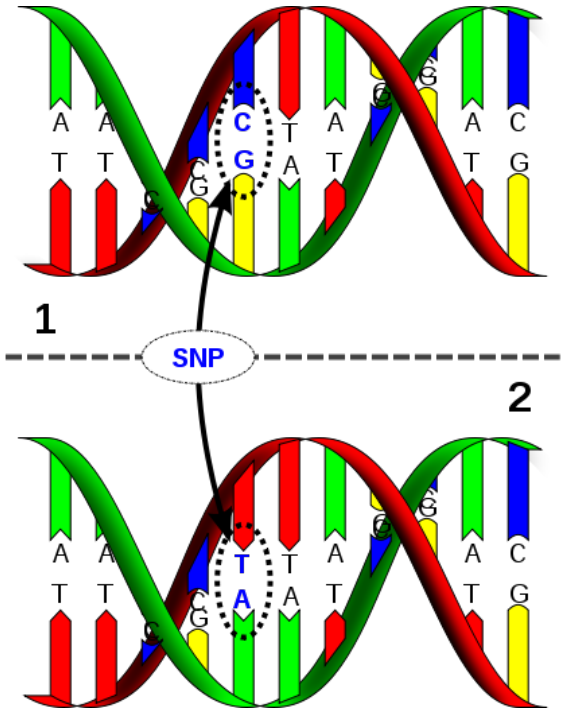
- SNP: Single Nucleotide Polymorphism
- Structural variations

  CNV: Copy Number Variation

  InDel: Insertion/Deletion

Polymorphism

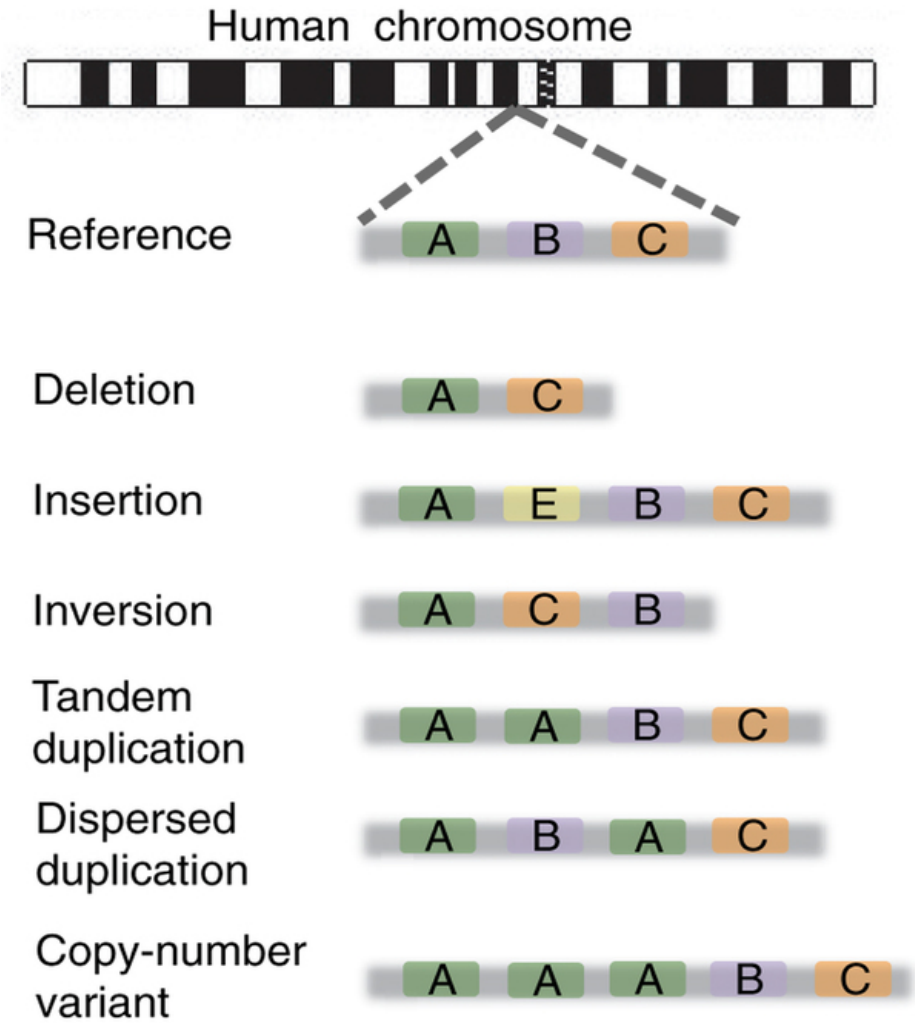
- Single Nucleotide Polymorphism (SNP) — sites/ genes with “common” variation, less common allele frequency ≥1%, otherwise called rare variant and not polymorphic



SNP Distribution in human

- Most common, 1 SNP / 100-300 bp
- Most mutations lost within a few generations
- 2/3 are CT differences
- In non-coding regions, often less SNPs at more conserved regions
- In coding regions, often more synonymous (amino acid change) than non-synonymous SNPs

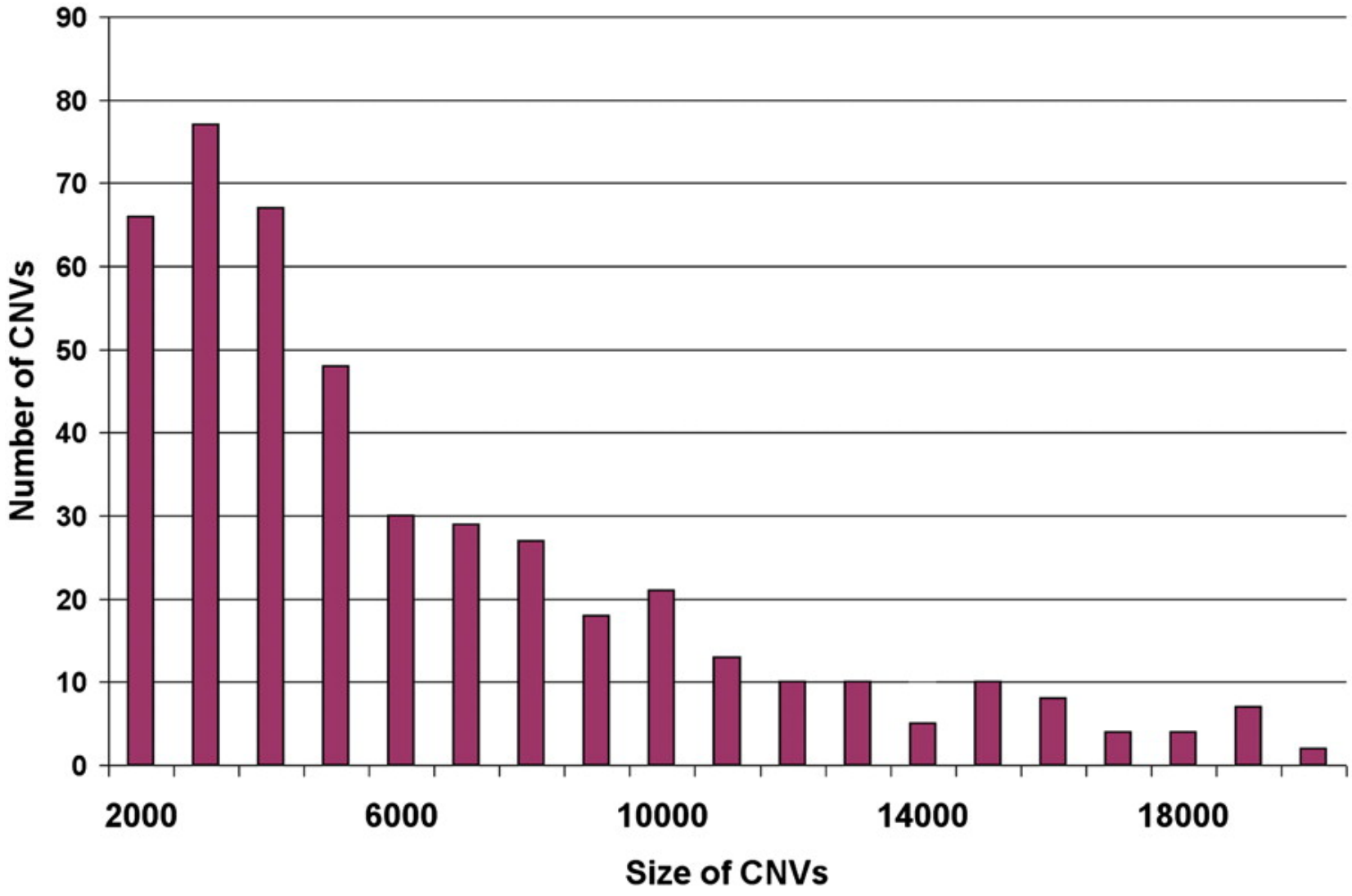
Other types of genomic variations



- Structural variation occurs in all forms and sizes.
- Often involves repetitive regions of the genome and complex rearrangements
- Genome structural variation encompasses polymorphic rearrangements 50 base pairs to hundreds of kilobases in size.
- And affects about 0.5% of the genome of a given individual.

Nature Methods 9, 133 -137 (2012) doi:10.1038/nmeth.1858

Size Distribution of CNV in Human Genome

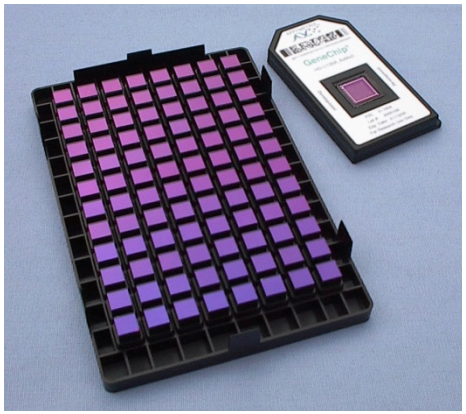


Why study structural variations?

- SVs can serve as genetic markers to identify genomic regions associated with disease
- Disease-associated SVs, regardless of function, have potential for clinical applications, including prediction of disease risk, treatment response and prognosis
- May be responsible for aberrant gene expression and protein function that drive disease processes or play a role in drug response
- Most genetic variations in the human genome are silent variations. i.e. have no phenotypic effect.

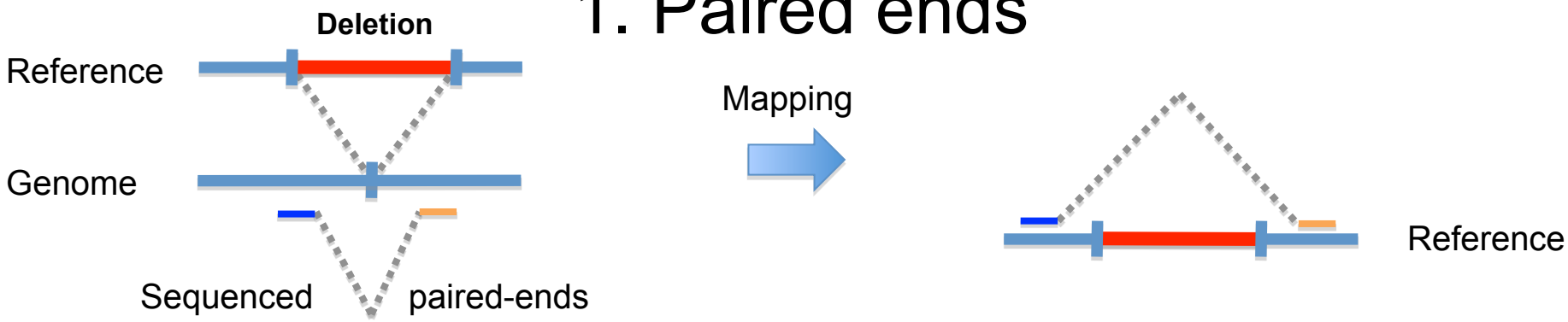
NGS is excellent for variation detection

- NGS technology enables the unprecedented sequencing coverage and high-throughput.
- NGS will be central in genomic and medical genetic studies. Genotype and SNP calling would be essential foundations.

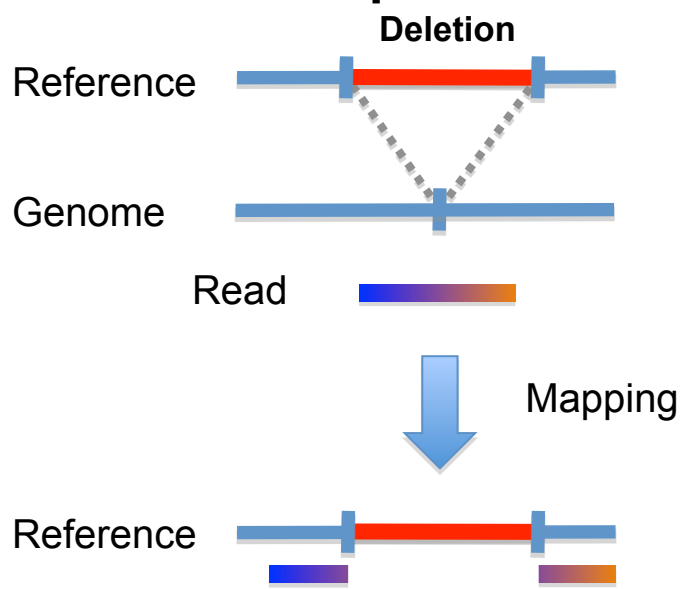


High Throughput DNA Sequencing based Methods to detect SVs

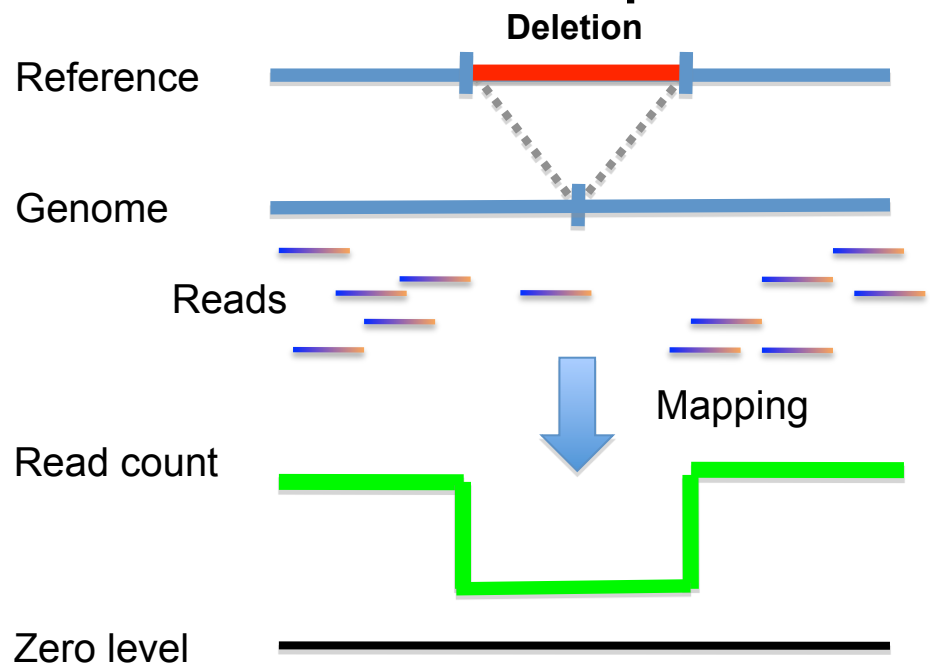
1. Paired ends



3. Split read



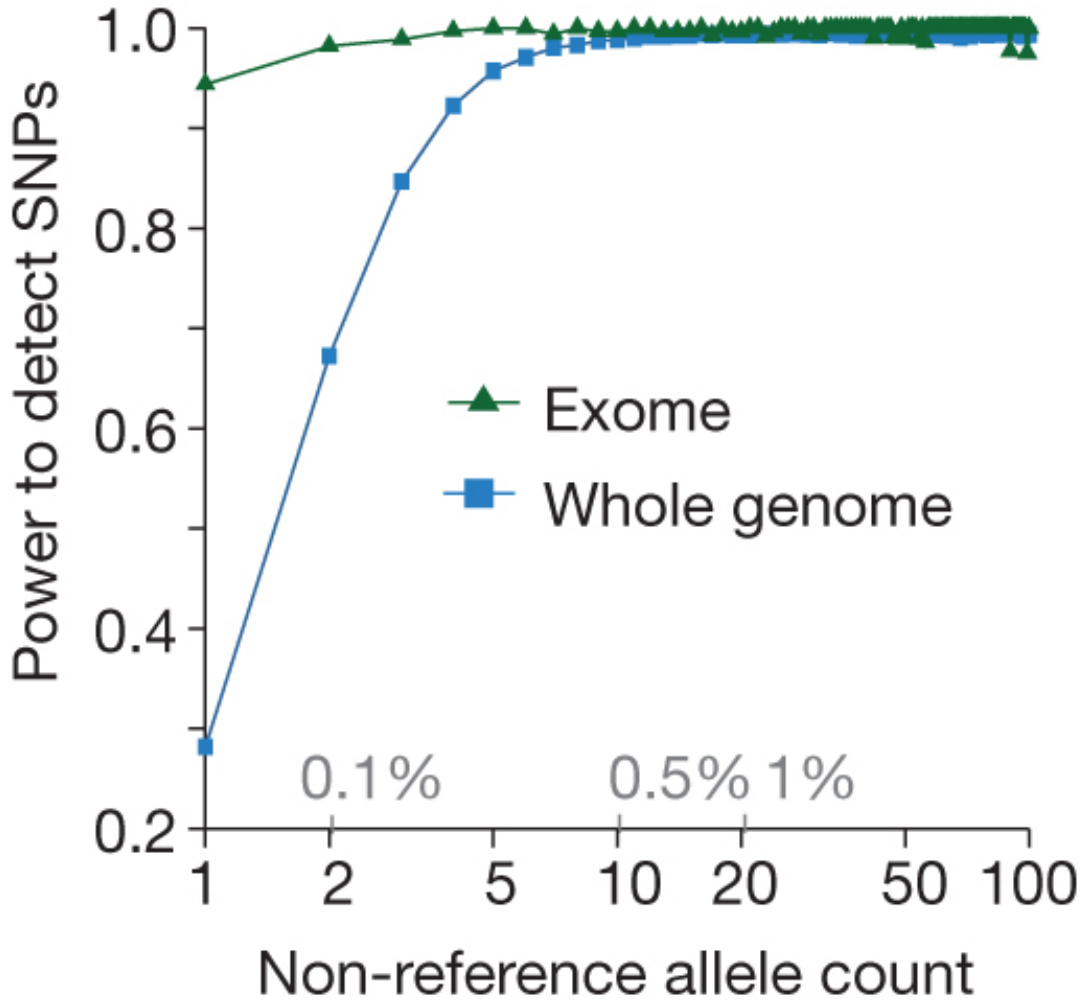
2. Read depth



Challenges of Variation detection using NGS

- NGS data can suffer from high error rates due to multiple factors, including base-calling and alignment errors.
- Many NGS studies rely on low-coverage sequencing, for which there is high probability that only one or two chromosomes of a diploid individual has been sampled at a specified site.
- Such uncertainty influences downstream analyses based on the inferred SNPs and genotypes, e.g., identification of rare mutations, estimation of allele frequency and association mapping.

Power and accuracy



The 1000 Genomes Project Consortium Nature **491**, 56-65 (2012) doi:10.1038/nature11632

How to call a variant?

Variant calling brief pipeline

Library construction and sequencing

Sequencing quality control
FASTQ files

↓

Read mapping
Sort, index, remove duplicates

Mapping quality control
BAM files

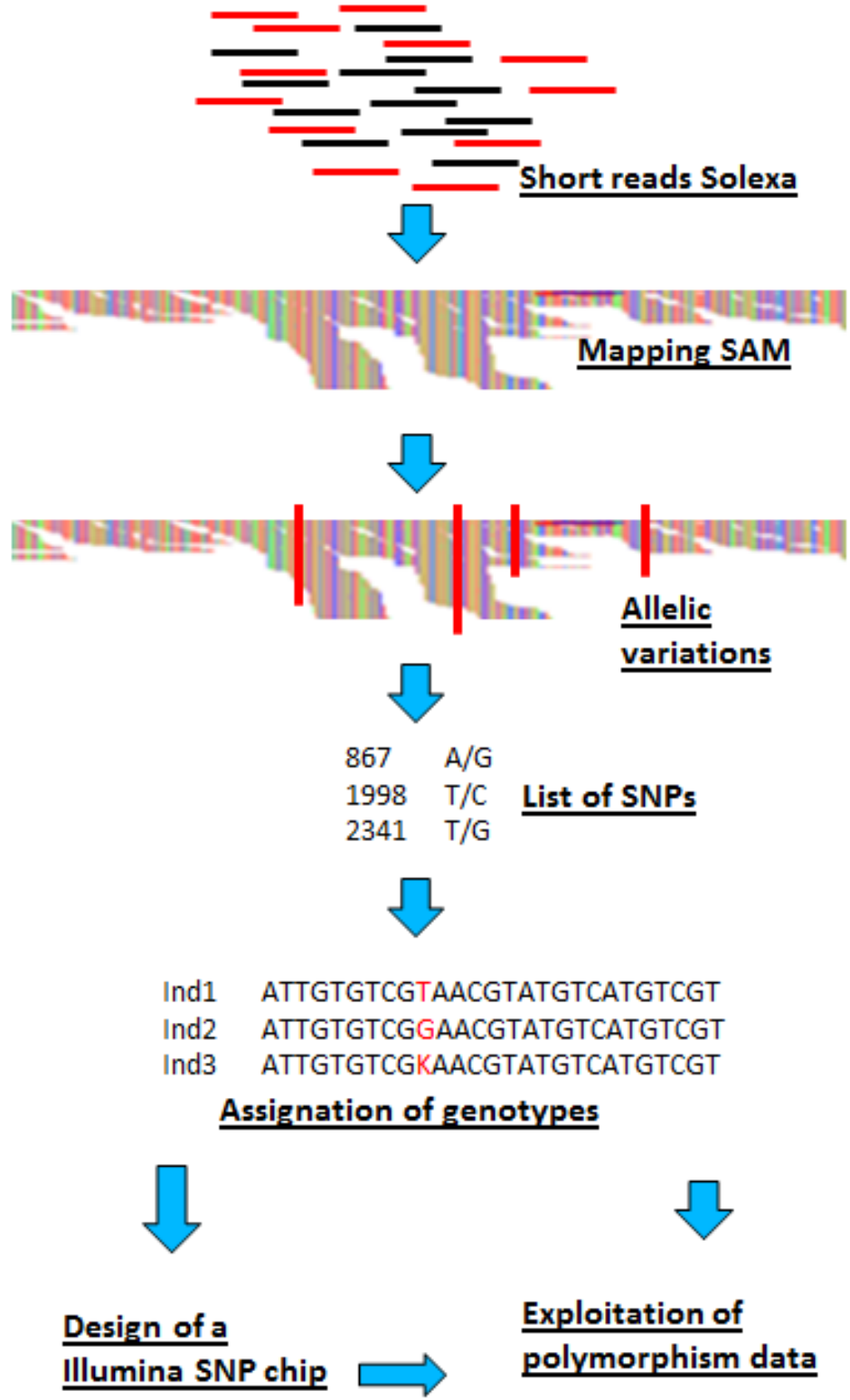
↓

Variant Calling

Variant calling quality control
VCF files

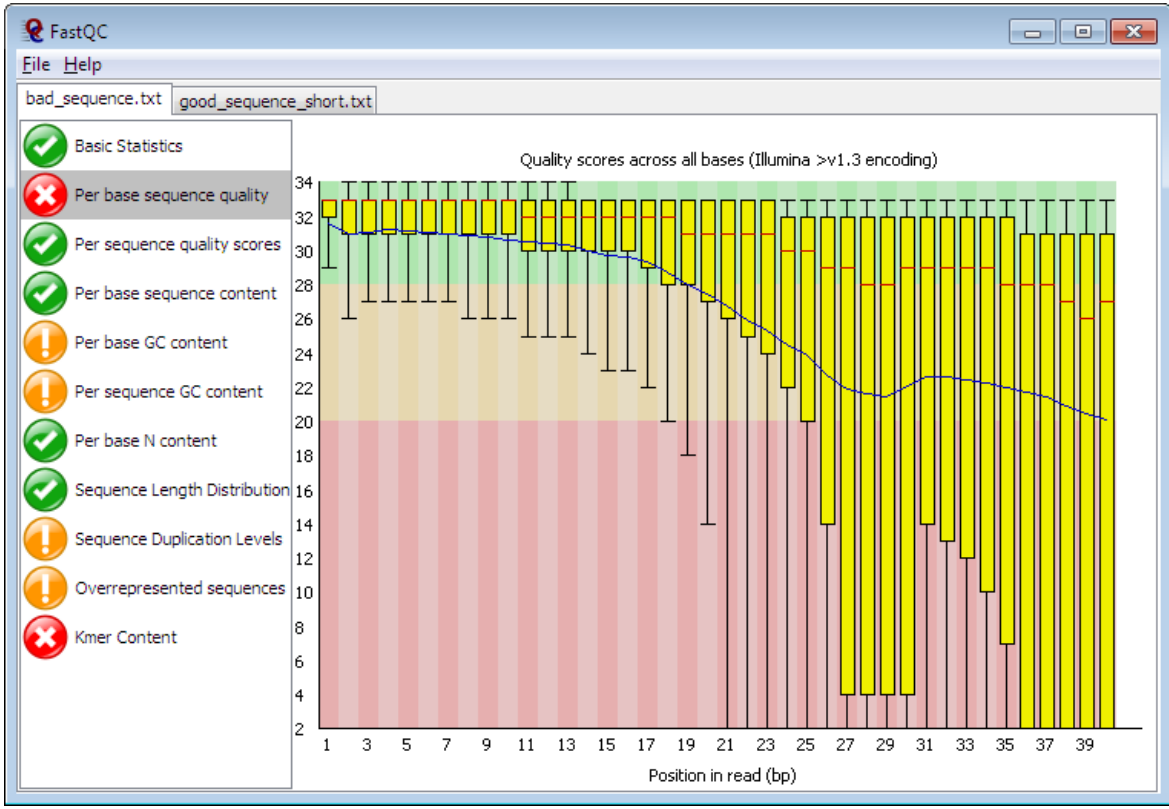
↓

Identification of potentially causal variants
Individualized care and counseling

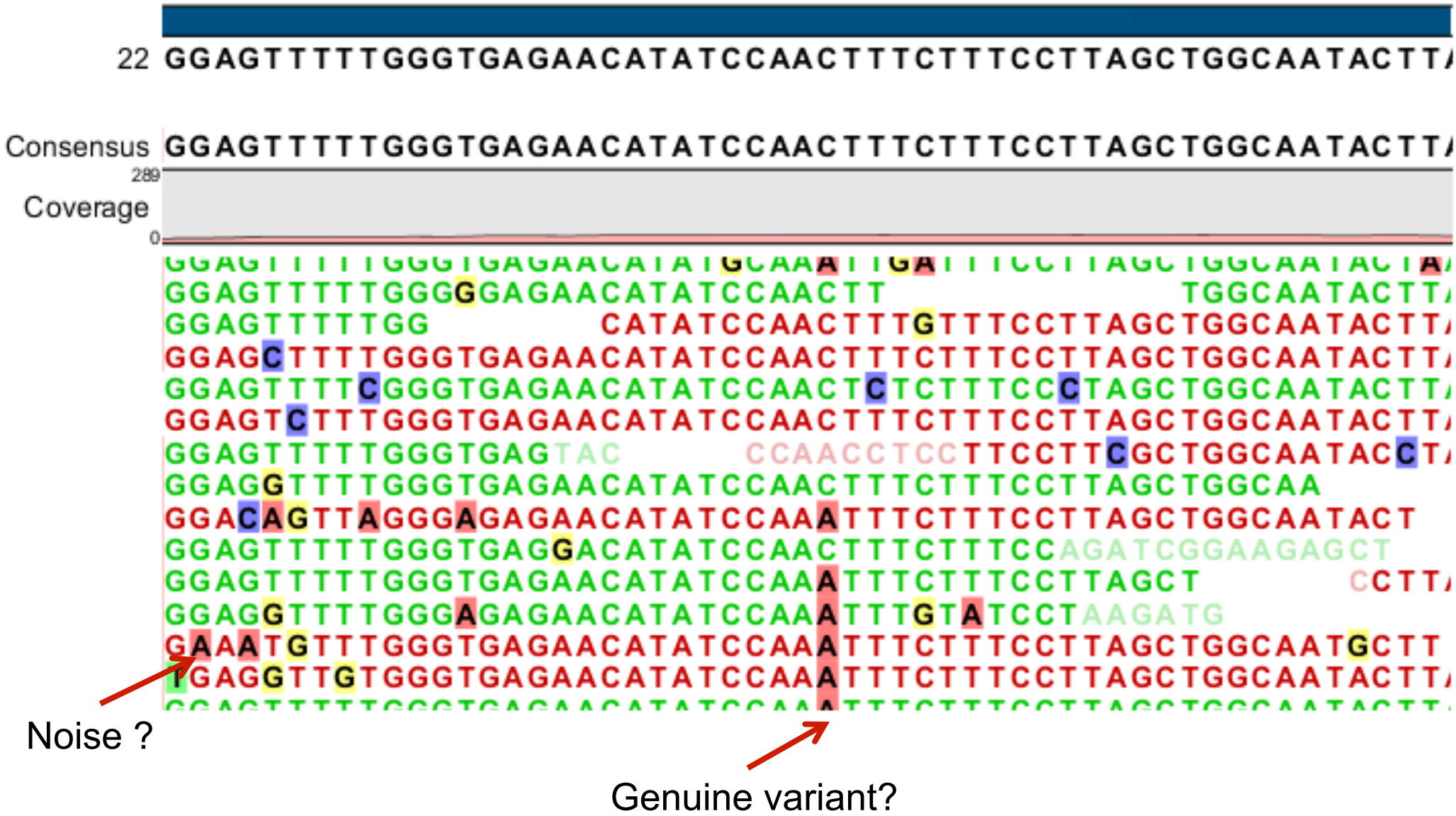


Pre-processing for Variation Calling

- Typically, analyses would first involve a filtering step in which only high-confidence bases would be kept.
- The least stringent cutoff used would be a Phred-type quality score of > 20, which corresponds to 1% error rate in base calling.



How to determine a genuine variant?



Probabilistic Methods (I)

- For moderate or low sequencing depths, genotype calling based on fixed cutoffs will typically lead to under-calling of heterozygous genotypes;
- The use of a simple filtering based on quality score leads to a loss of information regarding individual read qualities;
- The early methods for genotype calling typically does not provide measures of uncertainty in the genotype inference.
- Therefore, probabilistic methods have been developed that use the quality score to provide a posterior probability for each genotype.

Prior probability of genotypes

- Example: Assuming
  - heterozygous SNP rate 0.001
  - homozygous SNP rate 0.0005
  - Transition/transversion ratio 2

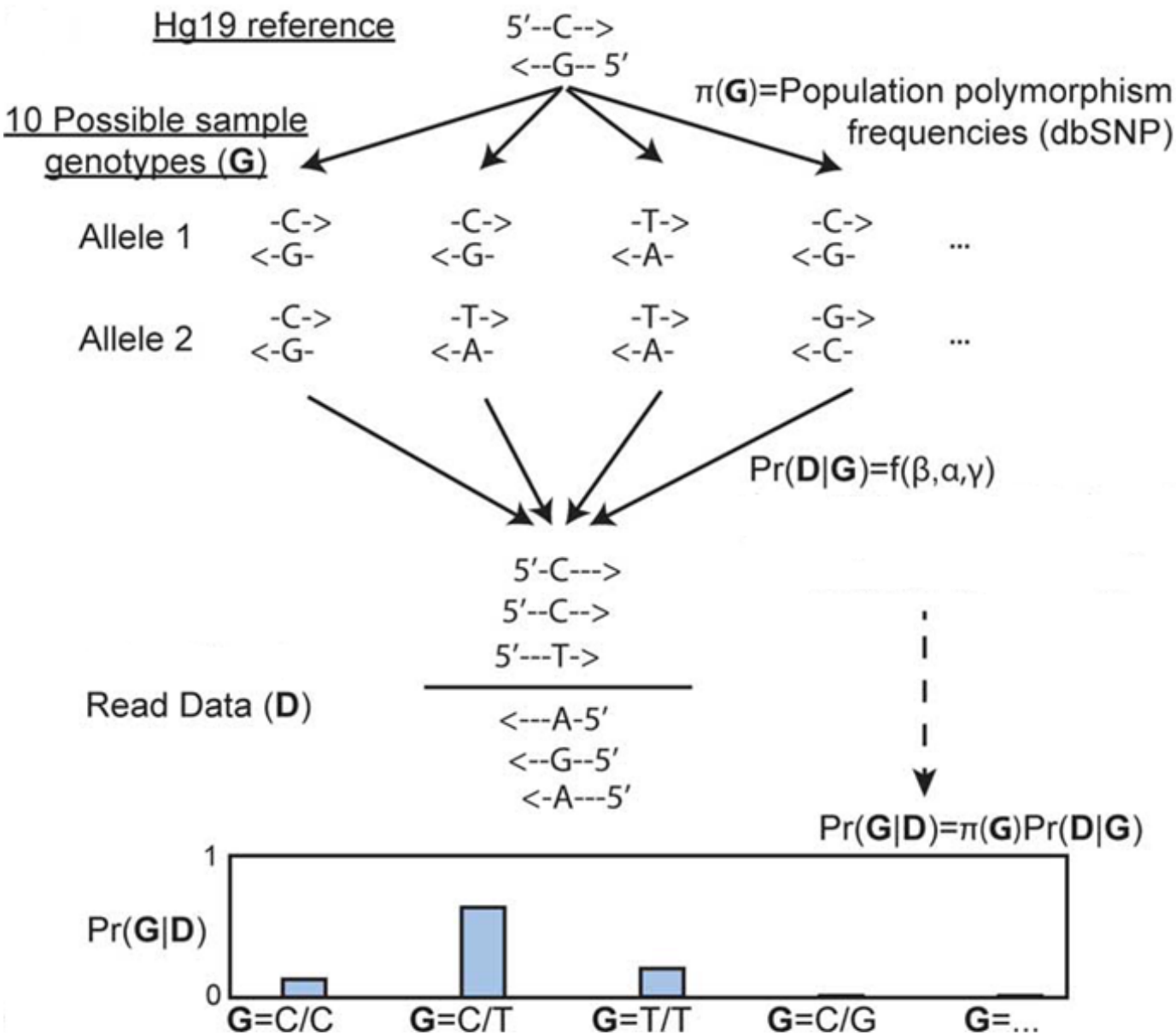
Table: Ts/Tv ratio in human genome

	A	C	G	T
A	3.33×10^{-4}	1.11×10^{-7}	6.67×10^{-4}	1.11×10^{-7}
C		8.33×10^{-5}	1.67×10^{-4}	2.78×10^{-8}
G			0.9985	1.67×10^{-4}
T				8.33×10^{-5}

Probabilistic Methods (II)

- In brief, it is assumed that one can compute a genotype likelihood, $p(D|G)$, for a genotype G.
- The symbol D represents all of the read data for a particular individual at a particular site.
- In conjunction with a genotype prior, $p(G)$, Bayes' formula is used to calculate $p(G|D)$, which is the posterior probability of genotype G.
- Finally, the genotype with the highest posterior probability is generally chosen, and this probability, or perhaps the ratio between the highest and the second highest probabilities, is used as a measure of confidence.

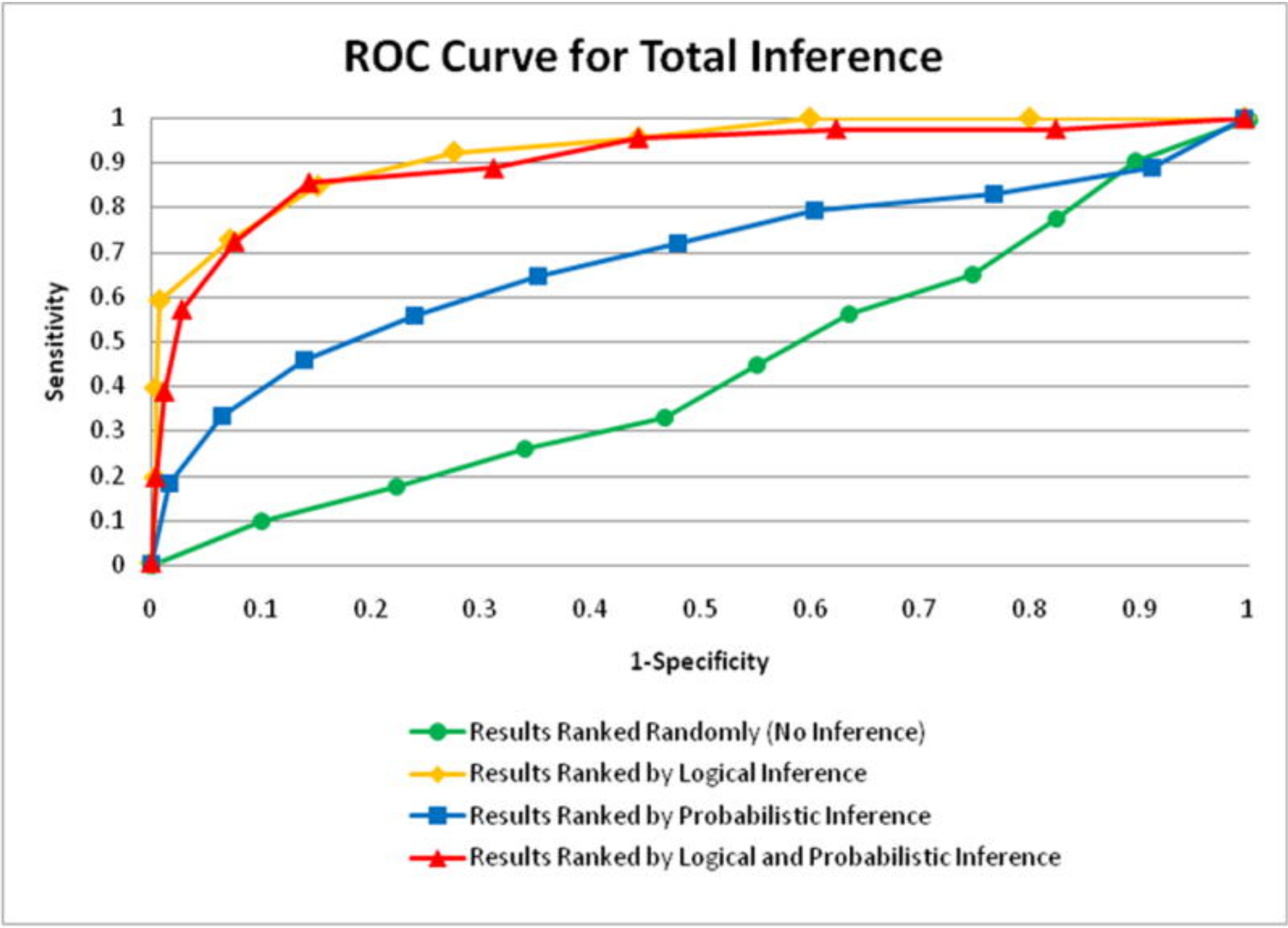
The SNP calling step



Liu et al. Genome Biology 2012, 13:R61 http://genomebiology.com/2012/13/7/R61

Advantages of Probabilistic Methods

- Incorporate errors introduced in base calling, alignment and assembly.
- Coupled with prior information, such as allele frequencies and patterns of linkage disequilibrium.
- Reduce and quantify the uncertainty associated with SNP and genotype calling.
- Provide measures of statistical uncertainty when calling genotypes.
- Lead to higher accuracy of genotype calling.



Shen et al. 2012. doi:10.1016/j.jbi.2009.12.002

Variant calling tools

- GATK
- VarScan
- Samtools

GATK (Genome Analysis ToolKit)

- Package for analysis of NGS data.
- Developed for the analysis of Human medical resequencing projects(1000 Genomes, The Cancer Genome Atlas).
- Includes tools for depth analysis, quality score recalibration, SNP/InDel discovery.

PREPROCESS:

*** Index genome (Picard)**
*** Convert Illumina reads to Fastq format**
*** Convert Illumina 1.6 read quality scores to standard Sanger scores**

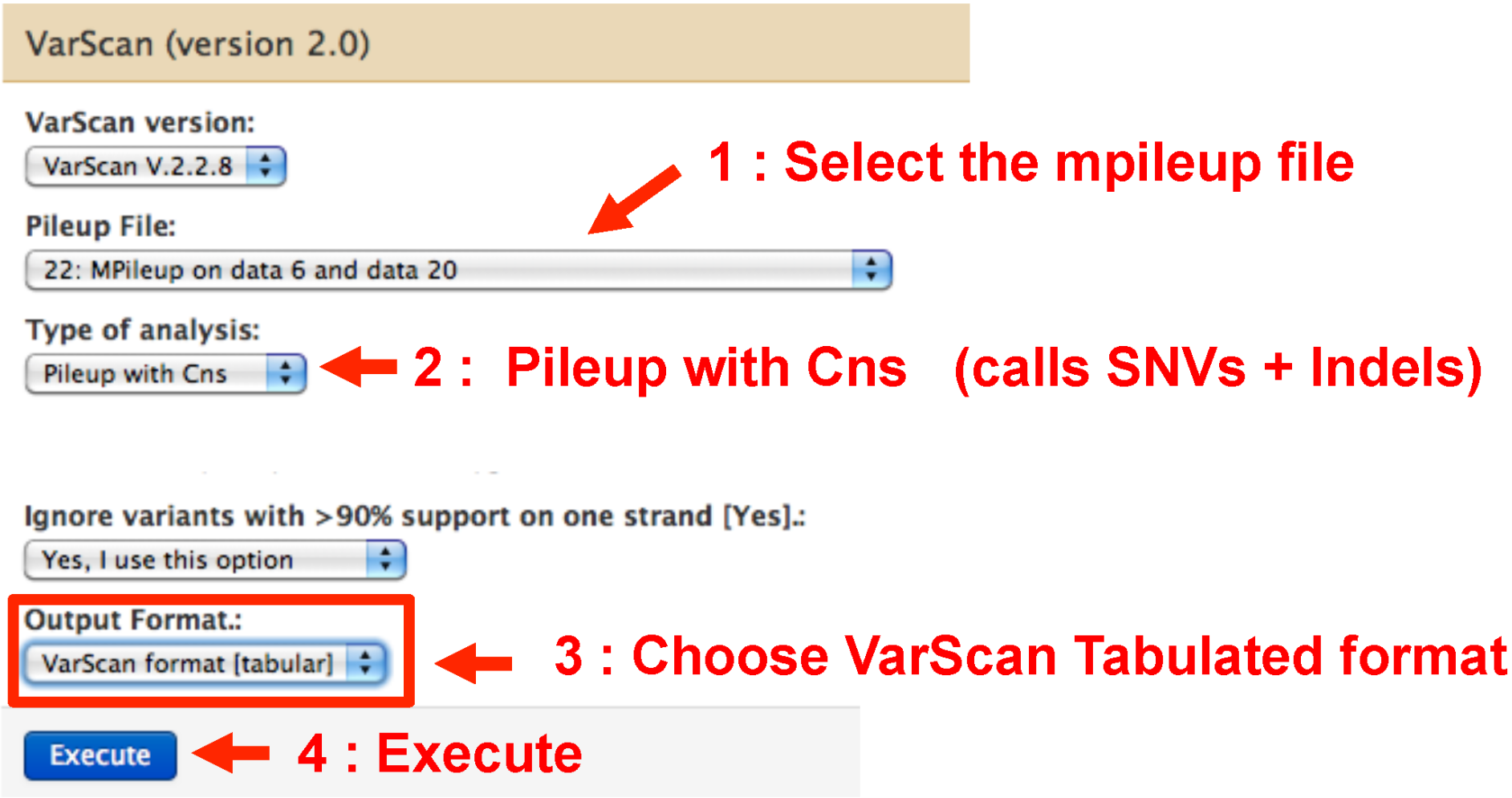
FOR EACH SAMPLE:

1. **Align samples to genome (BWA), generates SAI files.**
2. **Convert SAI to SAM (BWA)**
3. **Convert SAM to BAM binary format (SAM Tools)**
4. **Sort BAM (SAM Tools)**
5. **Index BAM (SAM Tools)**
6. **Identify target regions for realignment**
7. **Realign BAM to get better Indel calling**
8. **Reindex the realigned BAM (SAM Tools)**
9. **Call Indels (Genome Analysis Toolkit)**
10. **Call SNPs (Genome Analysis Toolkit)**
11. **View aligned reads in BAM/BAI (Integrated Genome Viewer)**

VarScan

- Mutation caller written in **Java** (no installation required) working with Pileup files of **Targeted, Exome, and Whole-Genome** sequencing data
- Multi-platforms: Illumina, SOLiD, Life/PGM, Roche/454
- Detection of different kinds of variants (SNVs/Indels) :
  - Germline variants in individual samples
  - Multi-sample variants shared or private in multi-sample datasets
- VarScan specificity is to be able to work with Tumor/ Normal pairs:
  - Somatic and germline mutation, LOH events in tumor-normal pairs
  - Somatic copy number alterations (CNAs) in tumor-normal exome data

VarScan


VarScan (version 2.0)
VarScan version:
VarScan V.2.2.8
1 : Select the mpileup file
Pileup File:
22: MPileup on data 6 and data 20
Type of analysis:
Pileup with Cns
2 : Pileup with Cns (calls SNVs + Indels)
Ignore variants with >90% support on one strand [Yes].:
Yes, I use this option
Output Format.:
VarScan format [tabular]
3 : Choose VarScan Tabulated format
Execute
4 : Execute

Variants can be hard to find

- DNPs — double nucleotide polymorphisms
- TNPs — triple nucleotide polymorphisms
- Small indels next to SNPs
- 30+ bp indels
- Homopolymer indels
- Homopolymer indel and SNP together
- Indels in palindromes
- Dense regions of variants

A comparison of genotype-caller accuracies

Table 1 | **A list of available non-commercial NGS genotype-calling software**

Software	Available from	Calling method	Prerequisites	Comments	Refs
SOAP2	http://soap.genomics.org.cn/index.html	Single-sample	High-quality variant database (for example, dbSNP)	Package for NGS data analysis, which includes a single individual genotype caller (SOAPsnp)	15
realSFS	http://128.32.118.212/thorfinn/realSFS/	Single-sample	Aligned reads	Software for SNP and genotype calling using single individuals and allele frequencies. Site frequency spectrum (SFS) estimation	-
Samtools	http://samtools.sourceforge.net/	Multi-sample	Aligned reads	Package for manipulation of NGS alignments, which includes a computation of genotype likelihoods (samtools) and SNP and genotype calling (bcftools)	53
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Multi-sample	Aligned reads	Package for aligned NGS data analysis, which includes a SNP and genotype caller (Unifed Genotyper), SNP filtering (Variant Filtration) and SNP quality recalibration (Variant Recalibrator)	32,33
Beagle	http://faculty.washington.edu/browning/beagle/beagle.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation, phasing and association that includes a mode for genotype calling	42
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation and phasing, including a mode for genotype calling. Requires fine-scale linkage map	44
QCall	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	Multi-sample LD	'Feasible' genealogies at a dense set of loci, genotype likelihoods	Software for SNP and genotype calling, including a method for generating candidate SNPs without LD information (NLDA) and a method for incorporating LD information (LDA). The 'feasible' genealogies can be generated using Margarita (http://www.sanger.ac.uk/resources/software/margarita)	54
MaCH	http://genome.sph.umich.edu/wiki/Thunder	Multi-sample LD	Genotype likelihoods	Software for SNP and genotype calling, including a method (GPT_Freq) for generating candidate SNPs without LD information and a method (thunder_glf_freq) for incorporating LD information	-

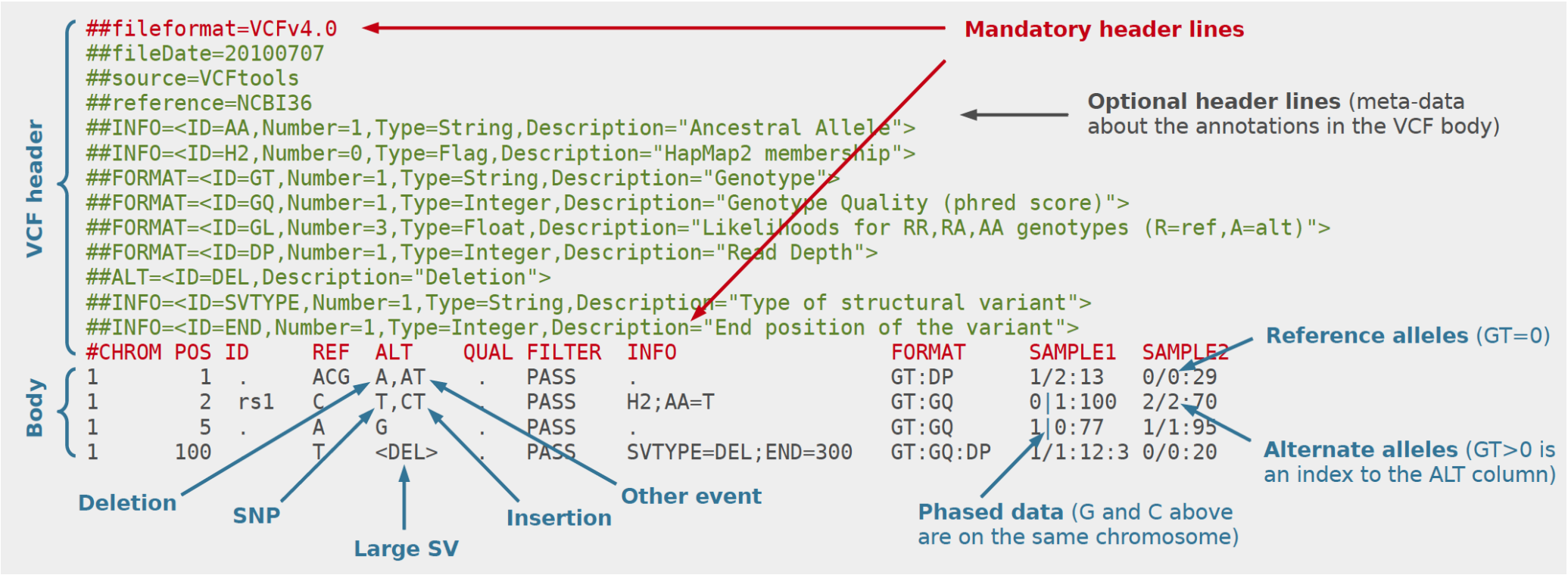
Nature Reviews Genetics 12, 443-451 (June 2011) doi:10.1038/nrg2986

Output format of variation calling tools

- All variation calling tools adopted the same or similar file format, which is called Variant Call Format (VCF).

VCF format (Variant Call Format)

- Variant Call Format
  - First developed in the 1000 genome project
  - Standardized format for storing the most prevalent types of sequence variation, including SNPs, indels and larger structural variants, together with rich annotations.
  - Usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome.



VCF format (Variant Call Format)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

- Type 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 format (Variant Call Format)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

- QUAL: phred p-value of the variant call quality
  - If ALT <> '.', QUAL = -log10[p-value(no variant)]
  - If ALT = '.', QUAL = -log10[p-value(variant)]
  - Higher QUAL value -> less mistake

- Filter:
  - PASS – if this position passed all the filters in the header files
  - q10;s50 – list of filters that are not met

In headers:

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

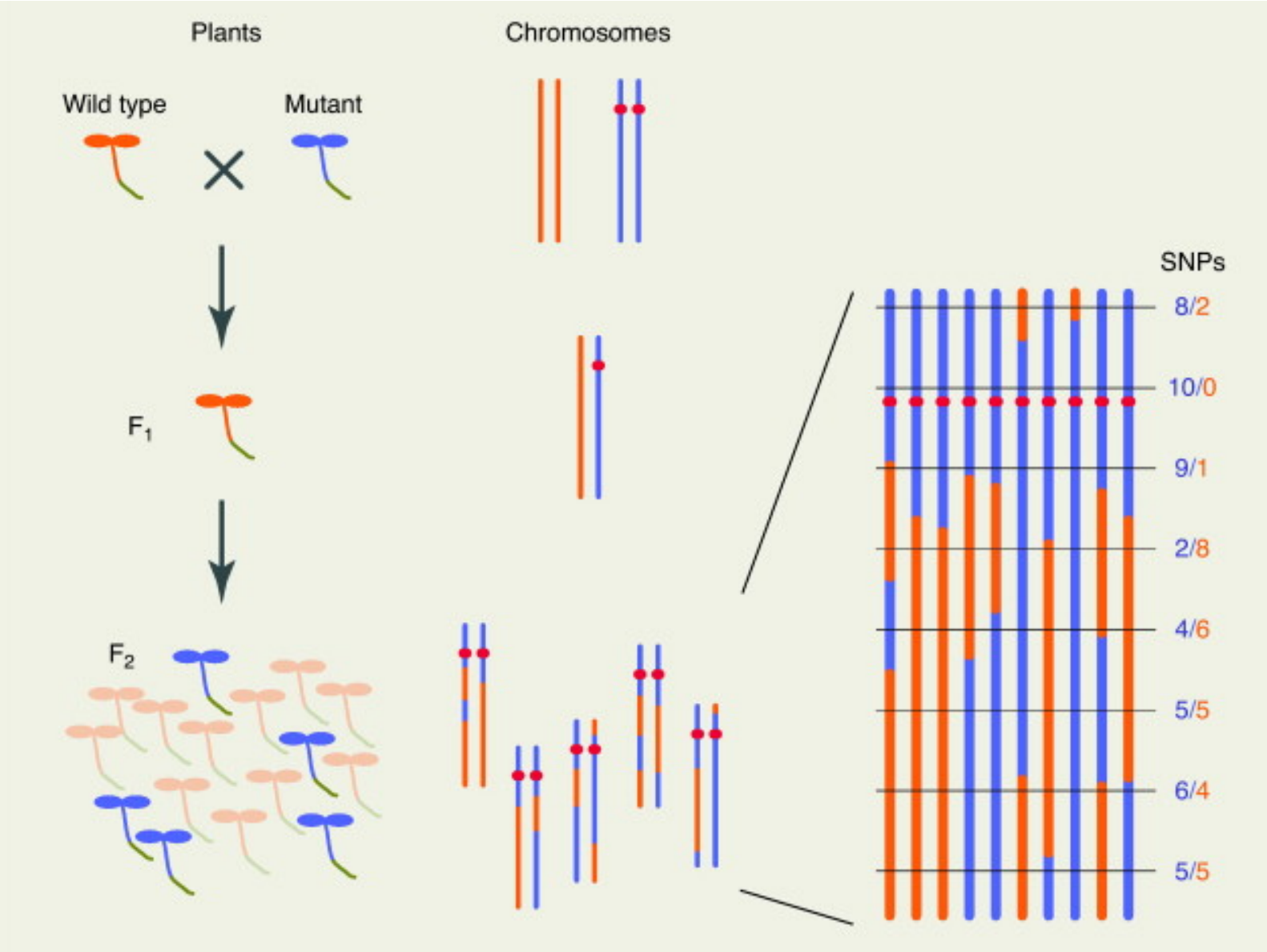
Applications of SV detection using NGS

- **BSR-seq**: Bulked Segregant RNA-Seq
- **GWAS**: Genome-wide association studies
- **eQTL**: expression quantitative trait loci

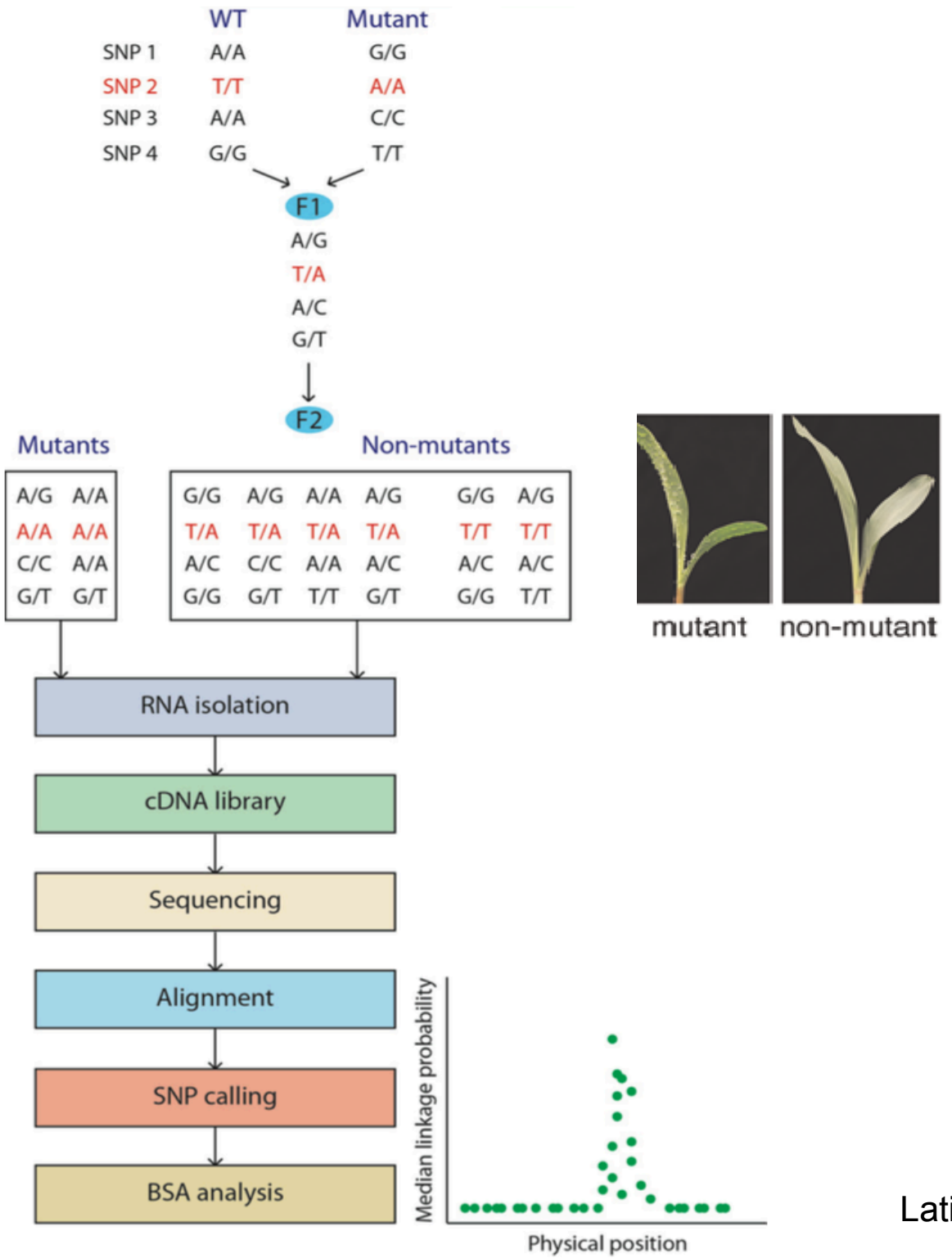
Bulked Segregant RNA-Seq (BSR-Seq)

- Bulked segregant analysis (BSA) is an efficient method to rapidly and efficiently map genes responsible for mutant phenotypes.
- BSA requires quantitative genetic markers that are polymorphic in the mapping population.

Diagram of BSR-Seq



doi:10.1016/j.tplants.2011.02.006 Trends in Plant Science, May 2011, Vol. 16, No. 5



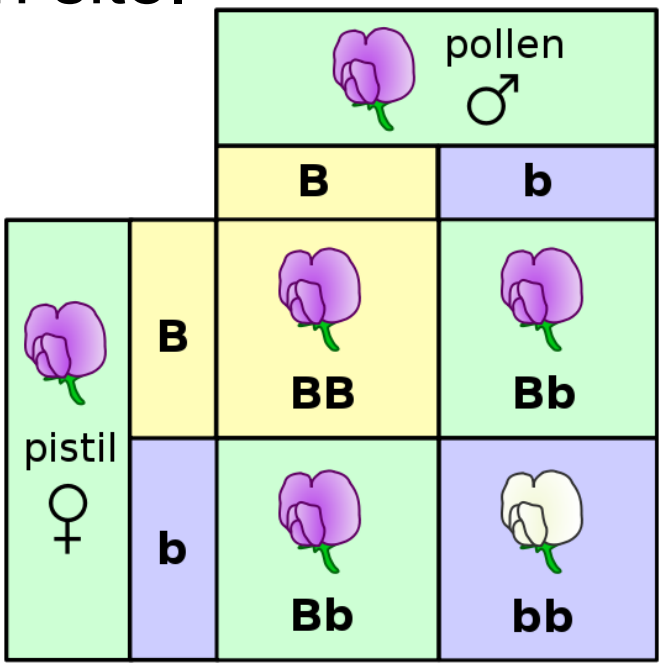
Latitia et al. 2013

Applications of SV detection using NGS

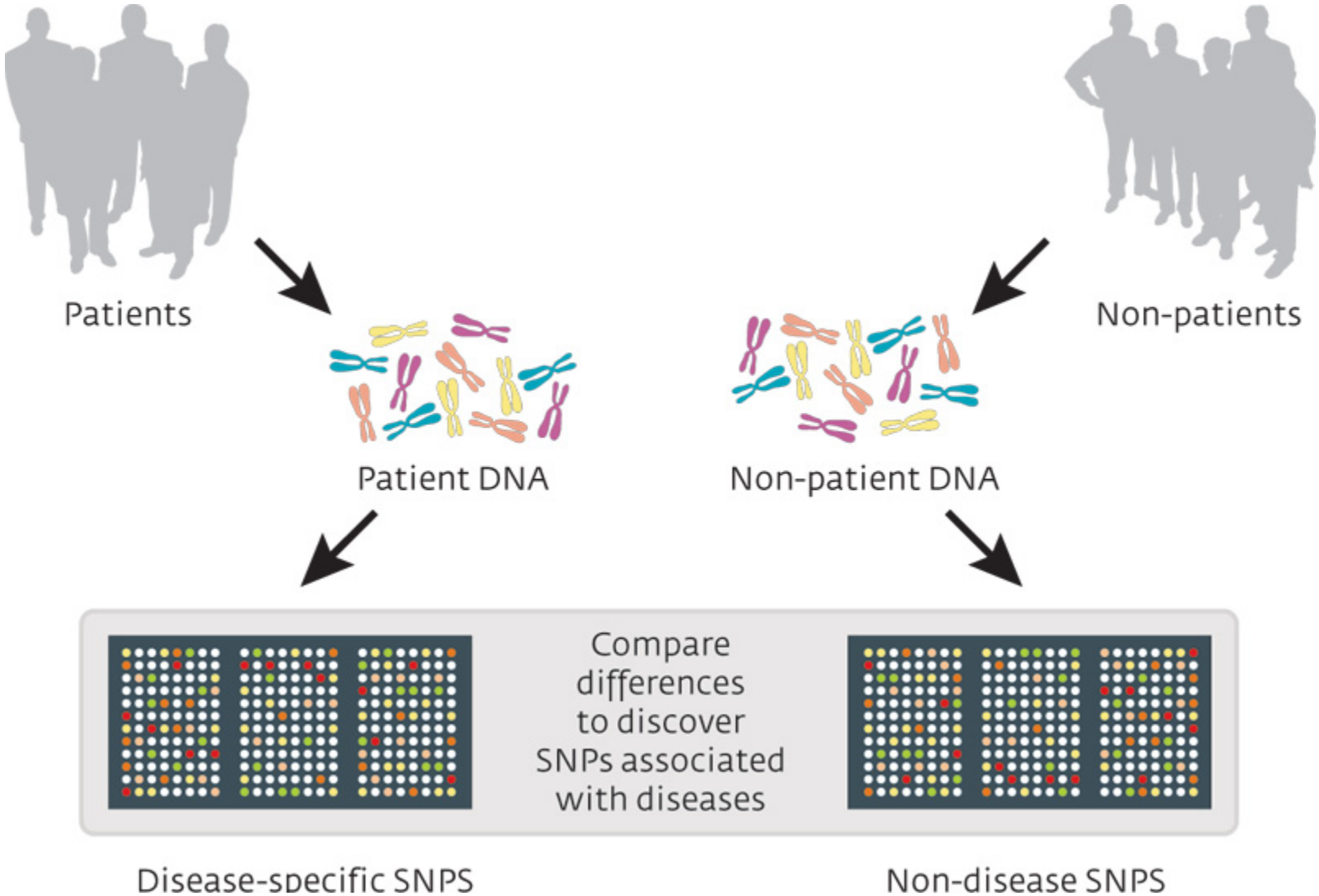
- **BSR-seq**: Bulked Segregant RNA-Seq
- **GWAS**: Genome-wide association studies
- **eQTL**: expression quantitative trait loci

Genotype

- The genetic makeup of a cell, an organism, or an individual (i.e. the specific allele makeup of the individual) usually with reference to a specific character.
- Genotype calling: Determines the genotype for each individual at each site.



Genome-wide association studies (GWAS)



From:www.mpg.de


MAY 27, 2013
TIME
THE
ANGELINA
EFFECT
Angelina Jolie's double mastectomy puts genetic testing in the spotlight. What her choice reveals about calculating risk, cost and peace of mind
BY JEFFREY KLUGER & ALICE PARK
time.com
Time

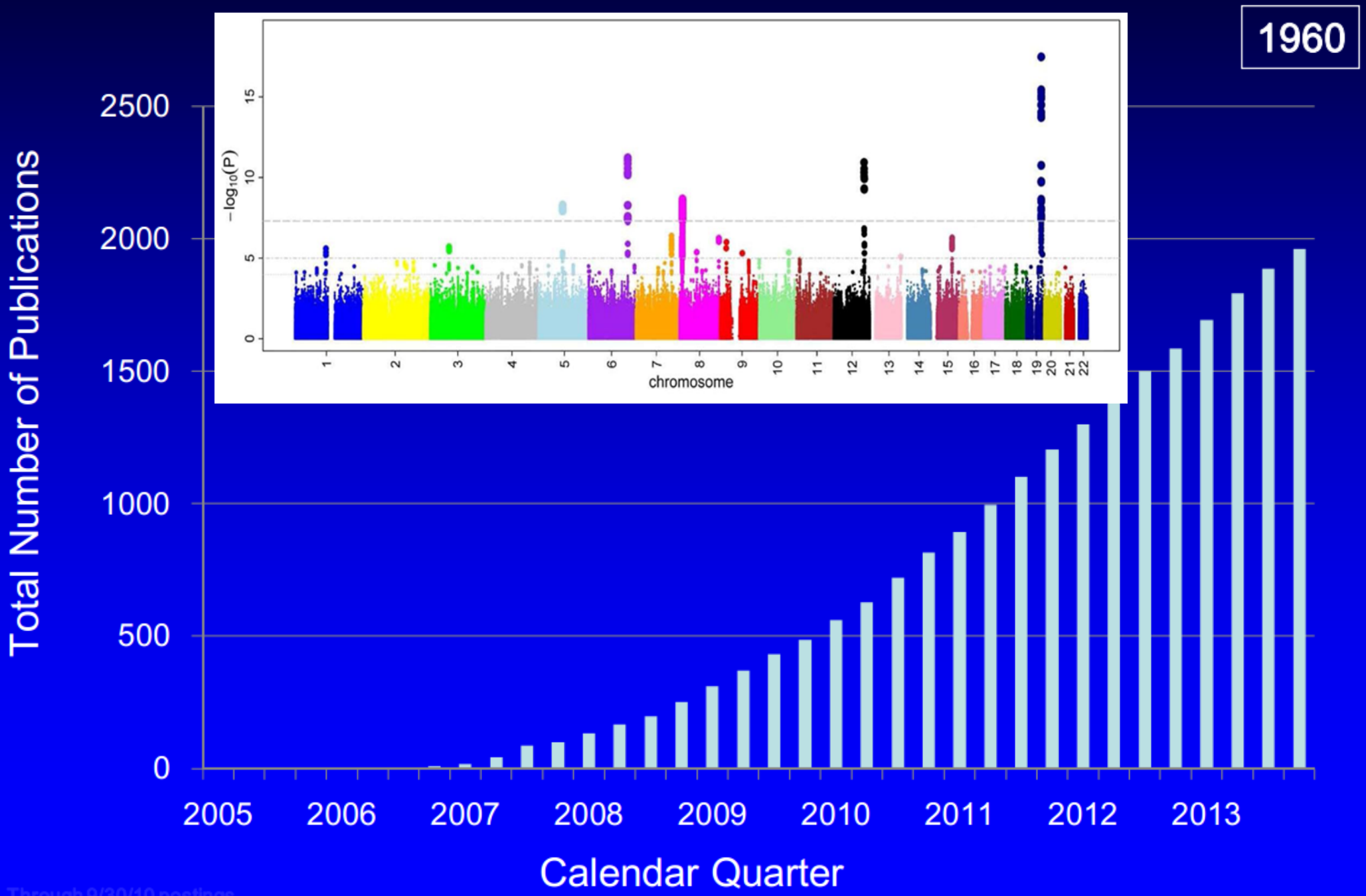
Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Published GWA Reports, 2005 – 2013



Through 9/30/10 postings

SNP Discovery: dbSNP database

dbSNP
NCBI SNP database

dbSNP Home Page

http://www.ncbi.nlm.nih.gov/SNP/

Variation Discovery ▾ Science Sites ▾ UCSC Bookmarks ▾ ComparativeGen ▾ SNPWorkshop ▾ News ▾

dbSNP Home Page

NCBI

Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search SNP for Go Clear

Limits Preview/Index History Clipboard Details

dbSNP BUILD 123

GENERAL
Contact Us
dbSNP Homepage
SNP Science Primer
Announcements
dbSNP Summary
FTP Download
Server
Getting Started
Build History
Handle Request

DOCUMENTATION
FAQ
dbSNP Handbook
Overview
How to Submit
RefSNP Summary Info
Schema
Database
PDF
Changes NEW
Genotype
Data Formats
Heterozygosity
Computation

dbSNP Search Options

Entrez SNP | ID Numbers | Submission Info | Batch | Locus Info | Free Form | Easy Form | Between Markers

ANNOUNCEMENT

- NEW! Search SNP in Mouse.
- NEW! dbSNP genotype data are now available on the web and on our FTP site (more info).
- ALERT! xml brief and submission format reports are dropped from ftp dump starting build 116. Please contact snp-admin with concerns.

Search by IDs

Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

Reference cluster ID(rs#)

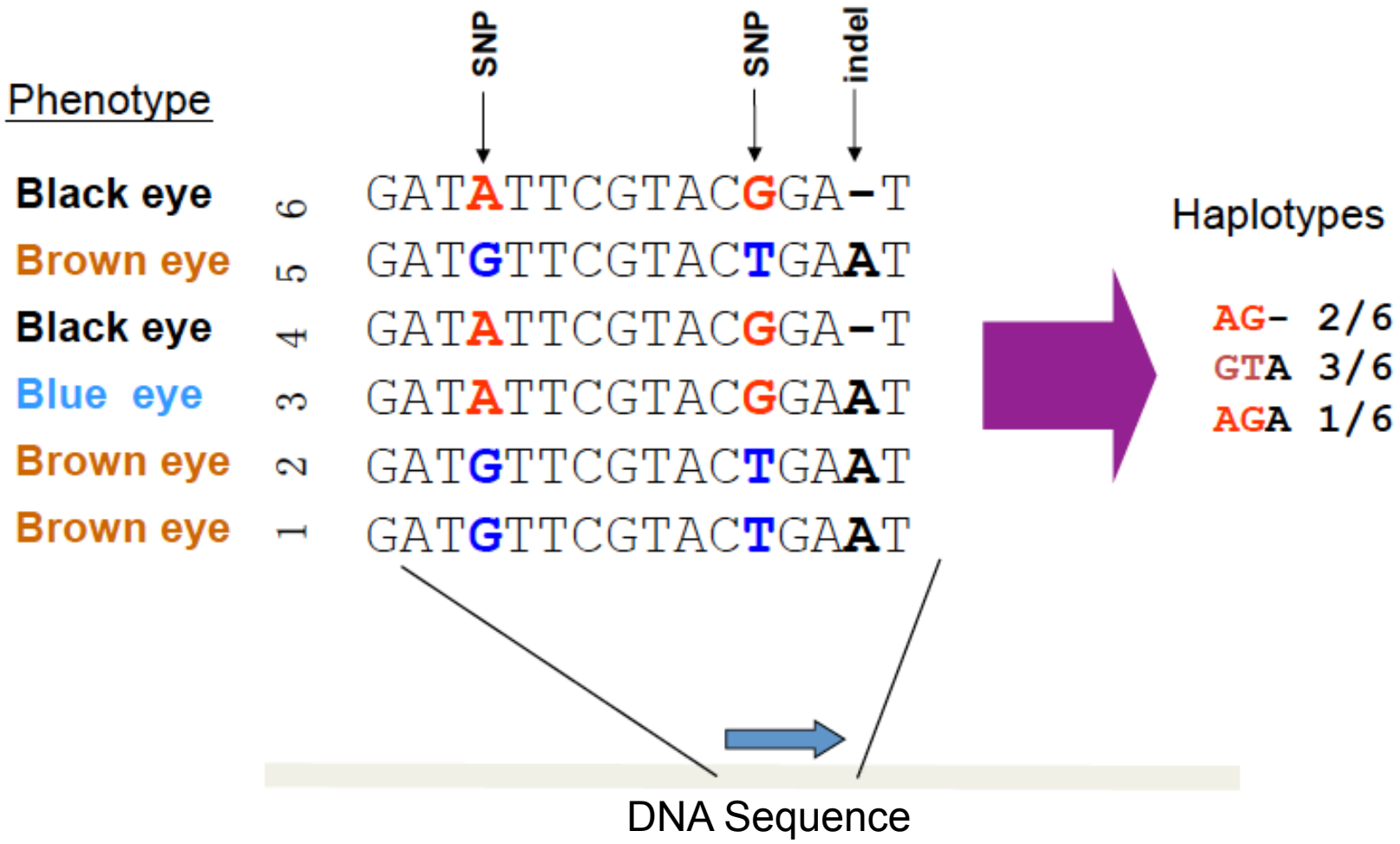
Search Reset

Advanced ID Search

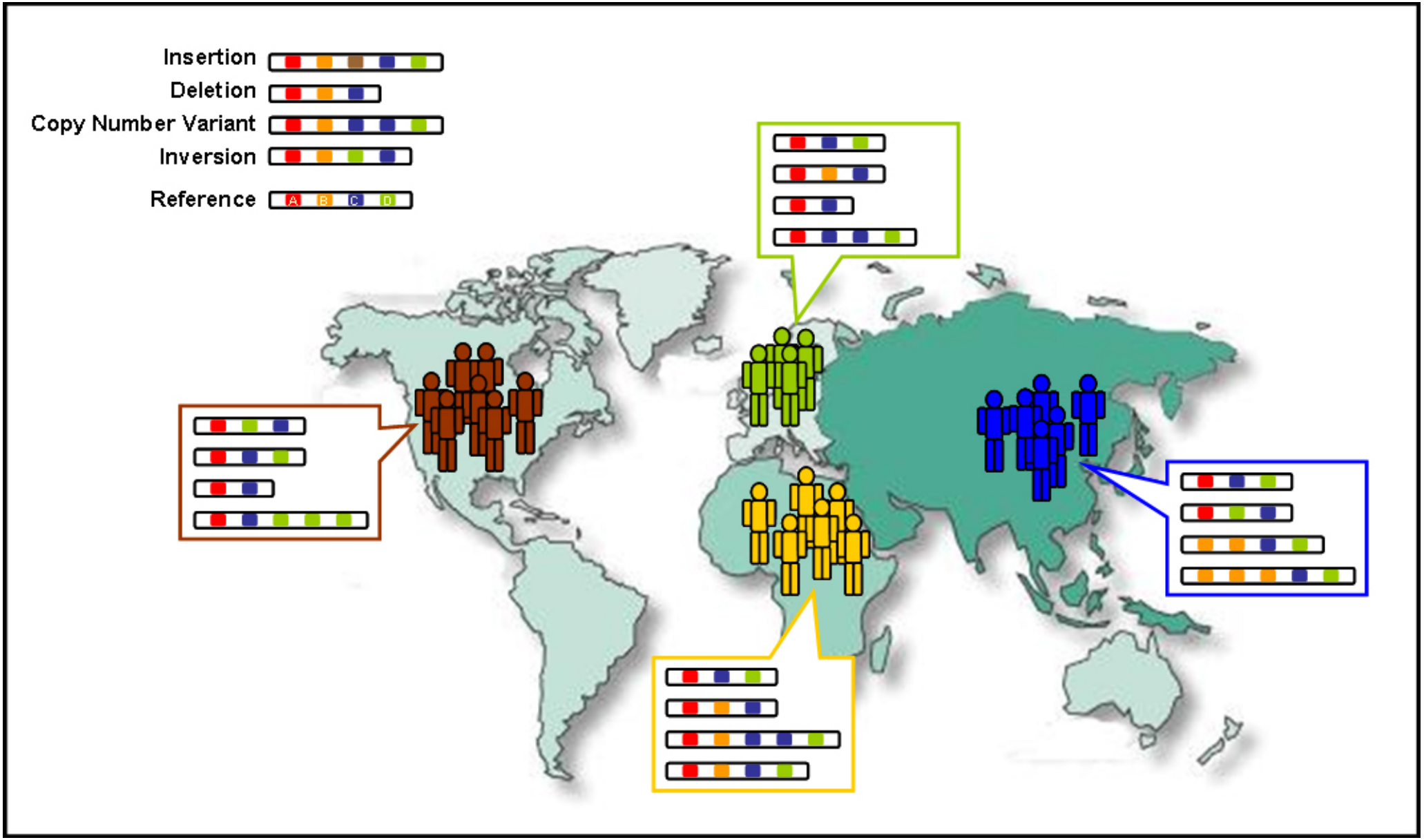
Submission Information

- By Submitter
- New Batches
- Method
- Population
  - Detail (Description, Handle, and ID)
  - Class (Based on geographic location)
- Publication
- Chromosome Report

From SNP to Haplotype



1000 Genome Project



Catalogs of human genetic variation

The 1000 Genomes Project
http://www.1000genomes.org/
SNPs and structural variants genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using NGS technologies

HapMap
http://hapmap.ncbi.nlm.nih.gov/
Identify and catalog genetic similarities and differences

dbSNP
http://www.ncbi.nlm.nih.gov/snp/
Database of SNPs and multiple small-scale variations that include indels, microsatellites, and non-polymorphic variants

COSMIC
http://www.sanger.ac.uk/genetics/CGP/cosmic/
Catalog of Somatic Mutations in Cancer

Applications of SV detection using NGS

- Useful websites and tools
- **BSR-seq**: Bulked Segregant RNA-Seq
- **GWAS**: Genome-wide association studies
- **eQTL**: expression quantitative trait loci

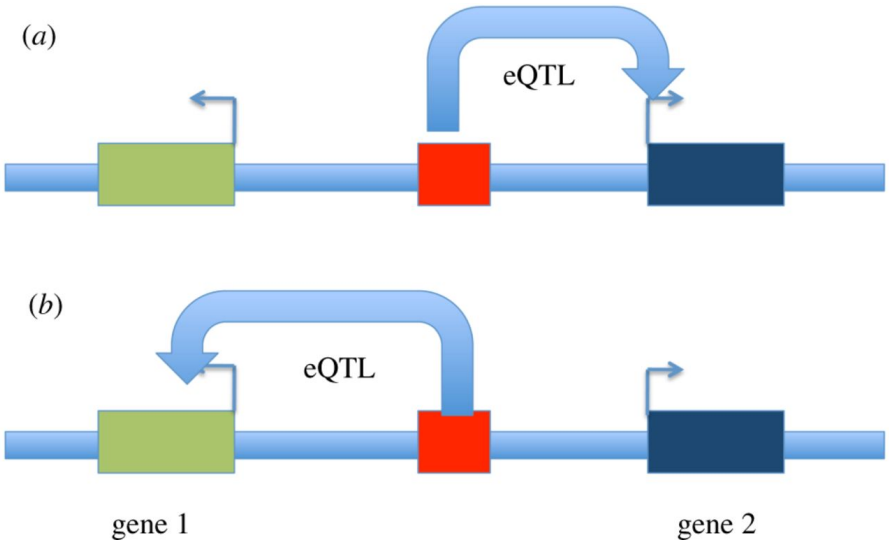
QTL (Quantitative Trait Locus)

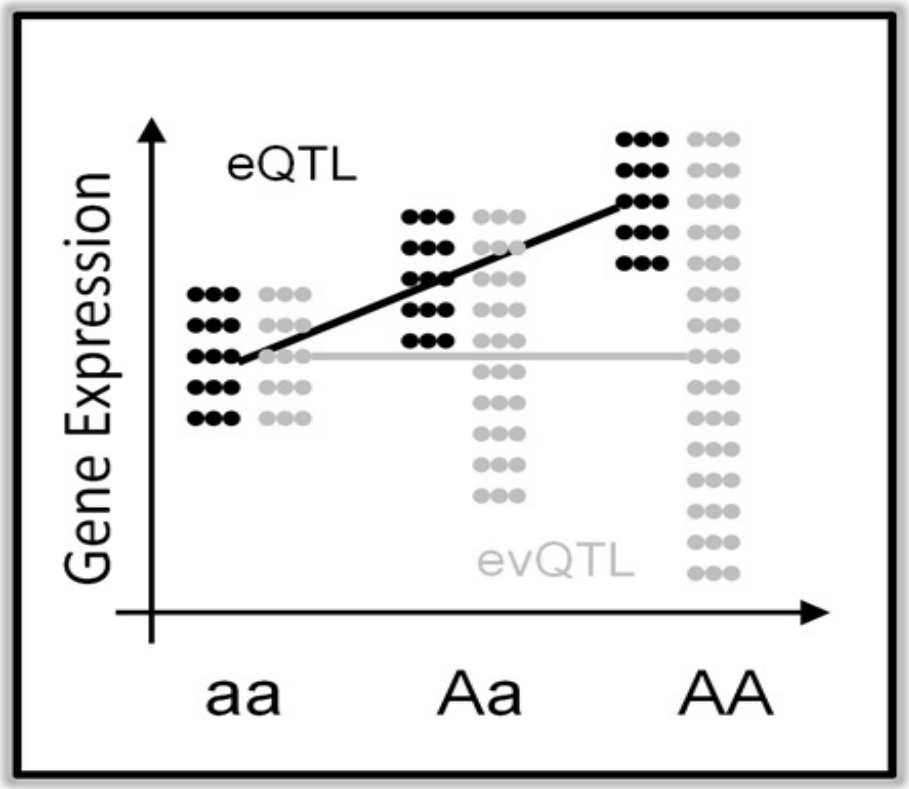
Genetic locus (QTL; L), Disease (D)

- More than 1000 monogenic Mendelian diseases controlling genes have been identified.
- Multiple genes, environmental factors, and interactions have limited the successes in human complex traits (such as cancer, diabetes, asthma).

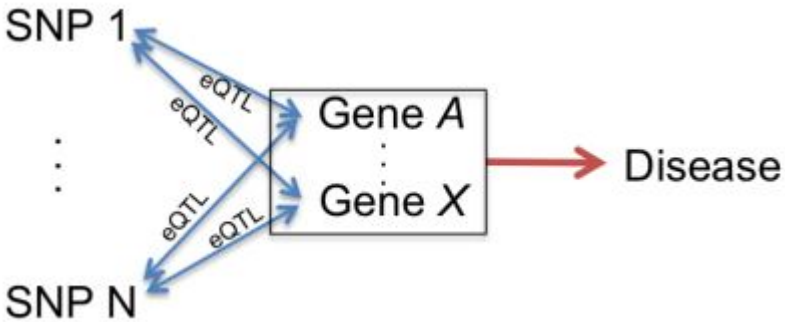
expression quantitative trait loci (eQTL)

An example of a cis-eQTL:





eQTL: identify SNPs that may influence the expression levels of a particular gene, from both gene expression and SNP-disease association results.

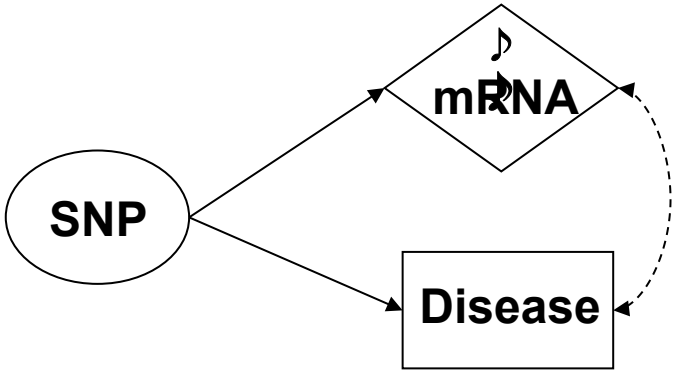


A integrative approach

- Models for causality
  - Causal Model

  - Reactive Model

  - Independent Model



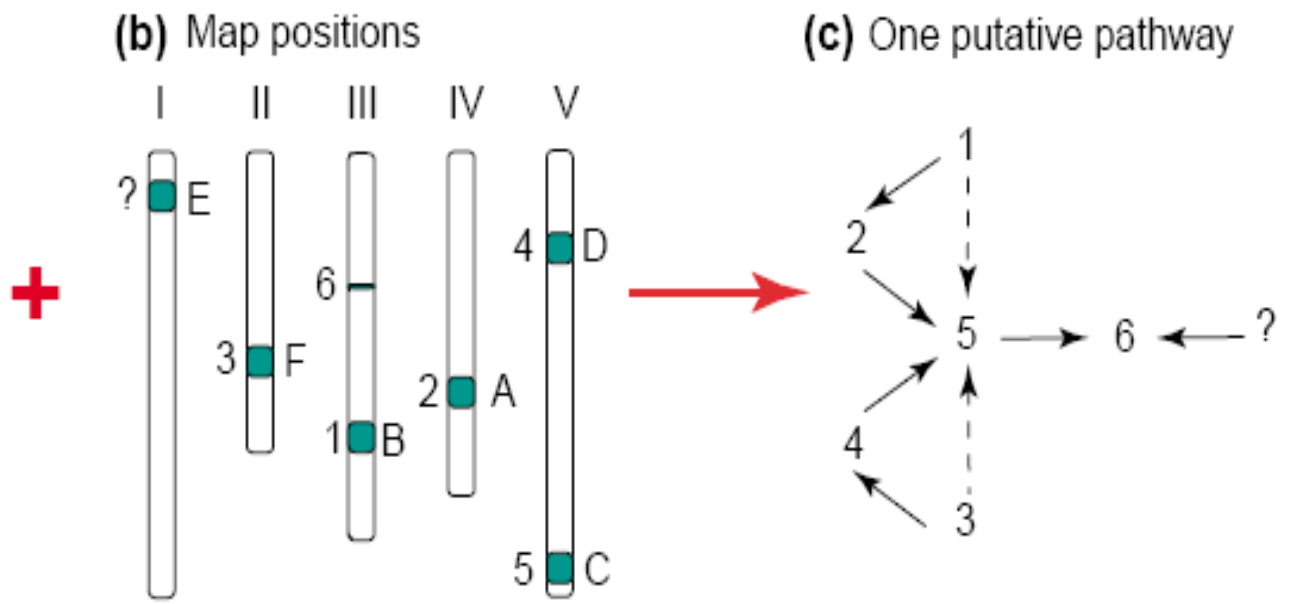
How to identify the eQTL using NGS data?

Constructing regulatory networks for eQTL

Genetic locus

Expression

	A	B	C	D	E	F	…	all
1		*						
2	*	*						
3								
4				*		*		
5	*	*		*		*		
6	*	*	*	*	*	*		
…								
all								



TRENDS in Genetics

Jansen, R.C. & Nap, J.P. (2001)
Trends Genet, **2001***, 17*, 388-391