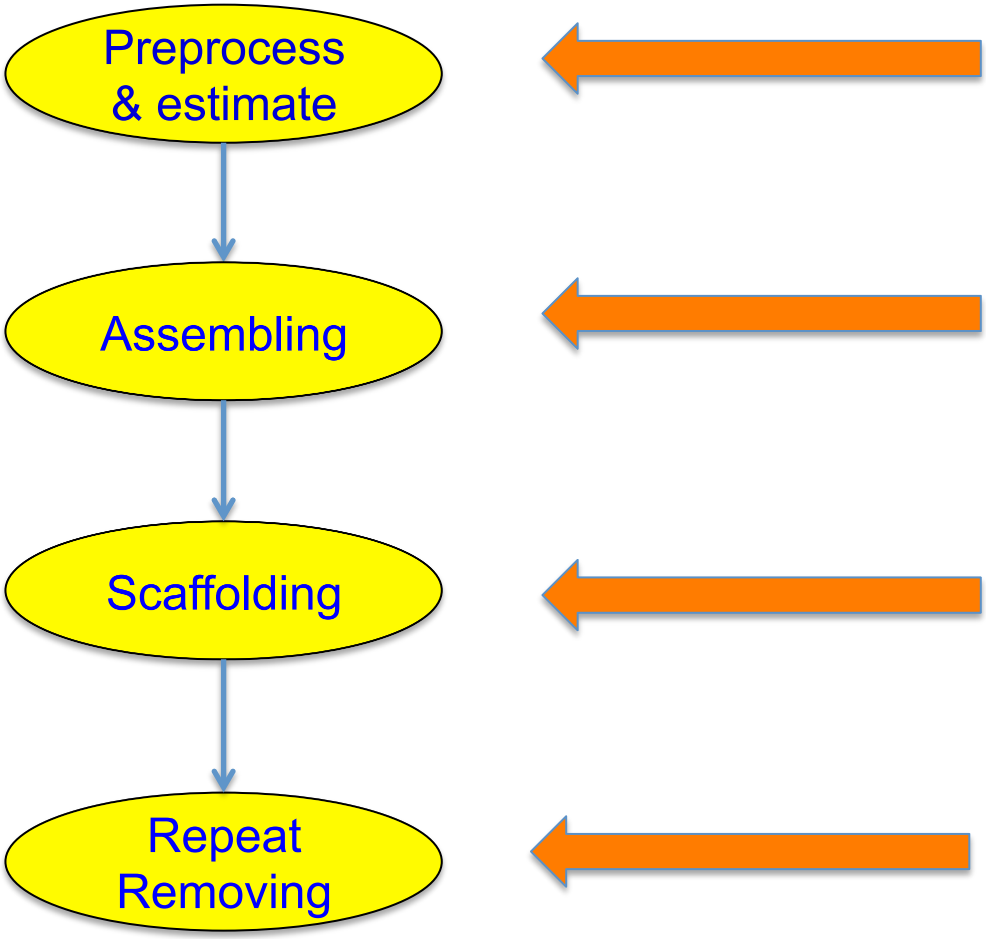
Next-generation sequencing

Lecture 7

De novo Assembly



Preprocess & estimate

Assembling

Scaffolding

Repeat Removing

30-50X coverage, paired-end sequencing, insert size 1k-100kbp, multiple librar

- **Velvet**: small genomes
- **ABySS**: large genome

Using mate pairs.
BAMBUS, SSPACE,
GRASS

Need to deal with repeat at any steps during assembly. Assemblers can do some work for repeat control

Assessing Assembly Quality

- Why do we need QC?
  - Misassembly correction is expensive
  - some assemblers have a simple quality-control method that does not capture larger errors

- Common measures of quality:
  - number and sizes of contigs (N50)
    - Assumption: few large contigs is better than many small contigs.
    - True because there are less gaps in the former, but, does not account for the possibility of misassemblies.
  - And more ..
  - Compare with a complete sequence

Whole genome sequencing

- *De Novo* sequencing
- Mapping assembly (Reference-guided assembly) (Resequencing)

"DNA resequencing is the task of sequencing a DNA region for an individual given that a reference sequence for this region is already available for the specific species. "

- ✓ Whole genome assembly
- ✓ Variant discovery and applications (GWAS)

  Discover or quantitate rare sequence variants

  HIV mutants within a single patient

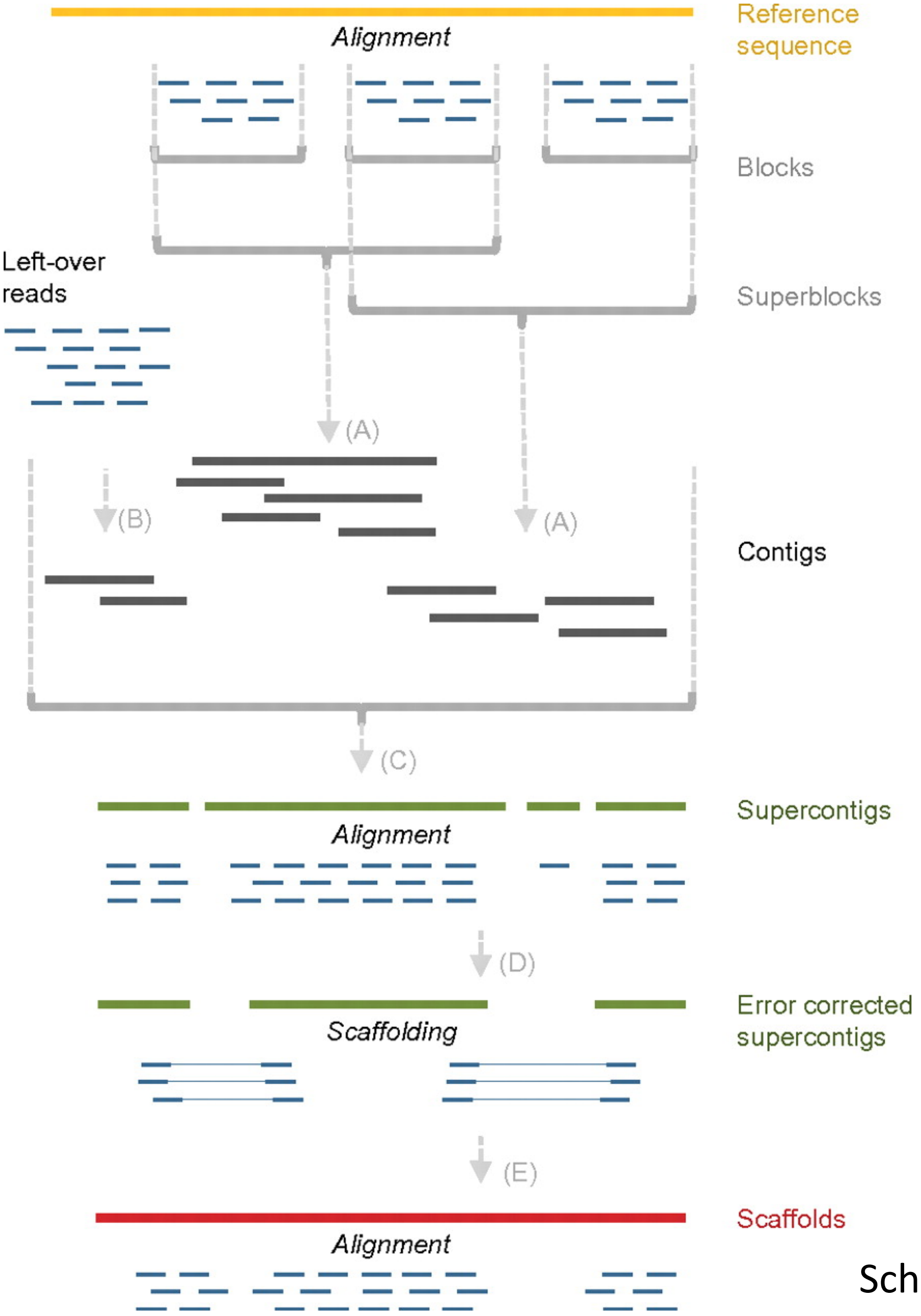
  Scan for mutations in tumor samples

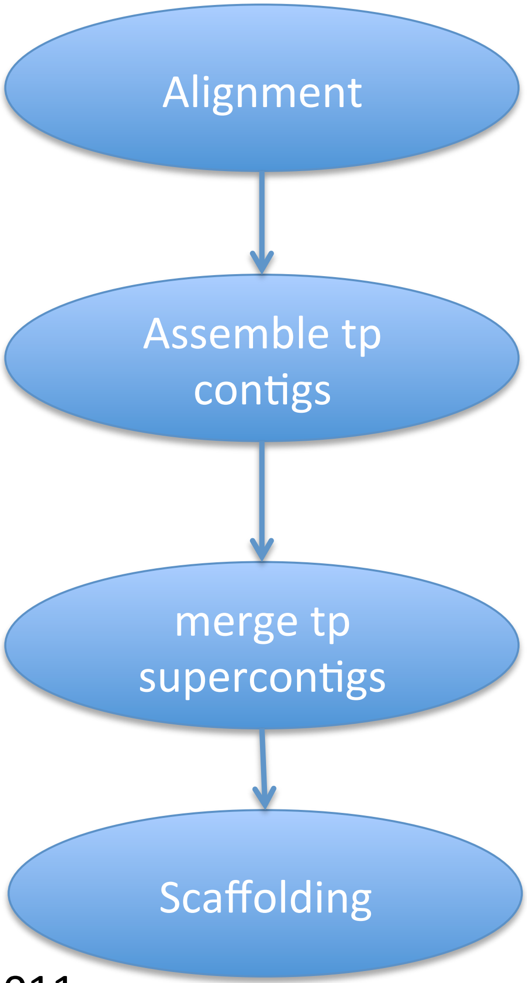
Resequencing

(mutation discovery/genotyping)

- A lot of current sequencing effort is spent on re-sequencing genomes of known species
  - Individual humans (1000 Genomes Project)
  - Experimental organisms – looking for genetic variation, copy number variation
- Challenge is to (quickly) align millions of sequence reads to a reference genome with some percent of mismatches
- Challenge to accurately call SNPs and indels
- Problems with repeated sequences – both tandem and dispersed repeats

Reference-guided assembly





Alignment

Assemble tp contigs

merge tp supercontigs

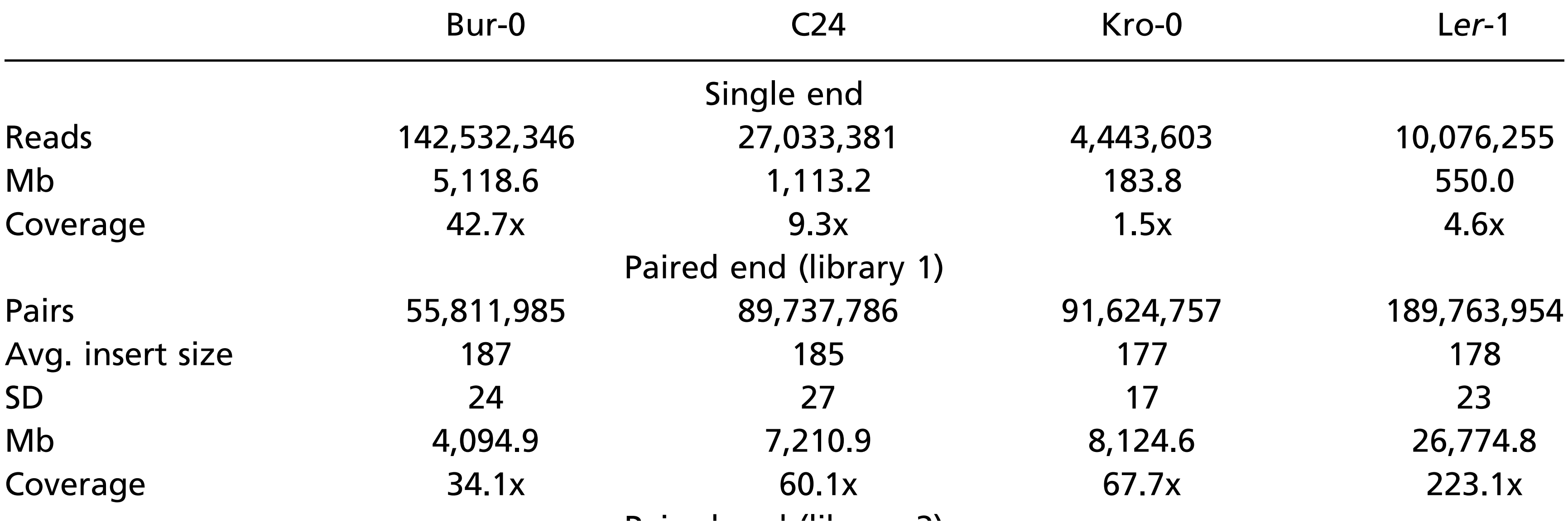
Scaffolding

Schneeberger, 2011

An example
Four Arabidopsis thaliana genomes

- Landsberg erecta (Ler-1), C24, Bur-0, Jro-0 strains

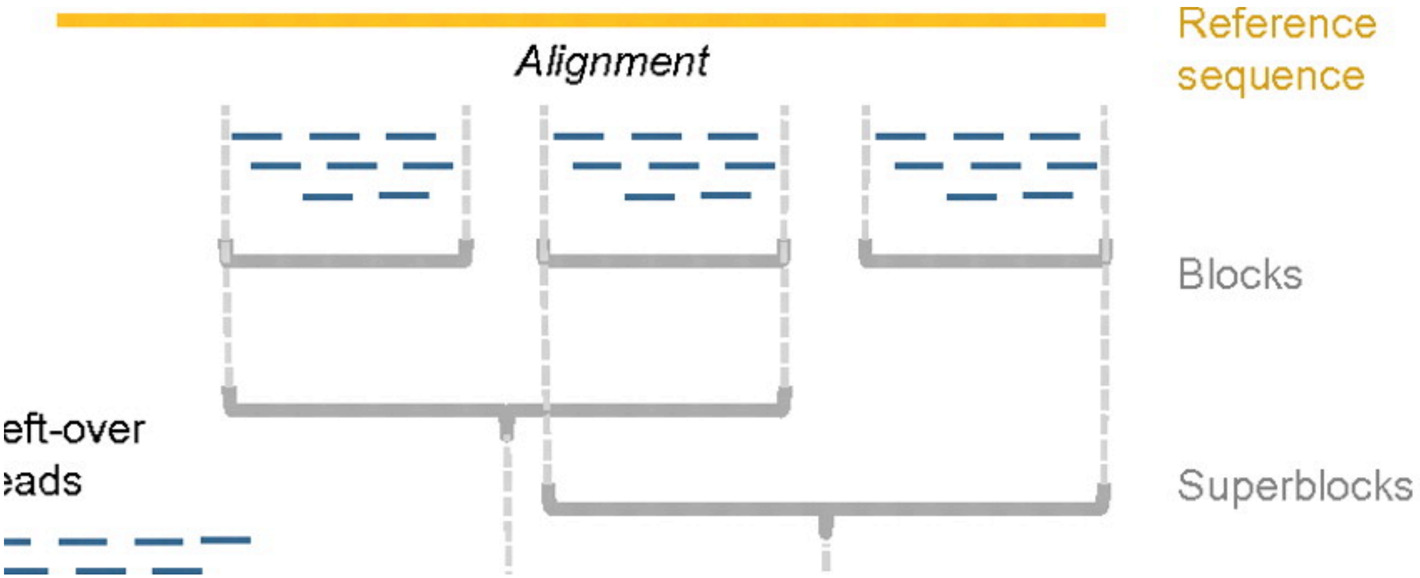
Read statistics



	Bur-0	C24	Kro-0	L*er*-1
Single end				
Reads	142,532,346	27,033,381	4,443,603	10,076,255
Mb	5,118.6	1,113.2	183.8	550.0
Coverage	42.7x	9.3x	1.5x	4.6x
Paired end (library 1)				
Pairs	55,811,985	89,737,786	91,624,757	189,763,954
Avg. insert size	187	185	177	178
SD	24	27	17	23
Mb	4,094.9	7,210.9	8,124.6	26,774.8
Coverage	34.1x	60.1x	67.7x	223.1x
Paired end (library 2)				

- 2 libraries (one single end and one paired end)
- Insert size 180 bp
- Read length 36-80 bp
- 30x – 200x coverage

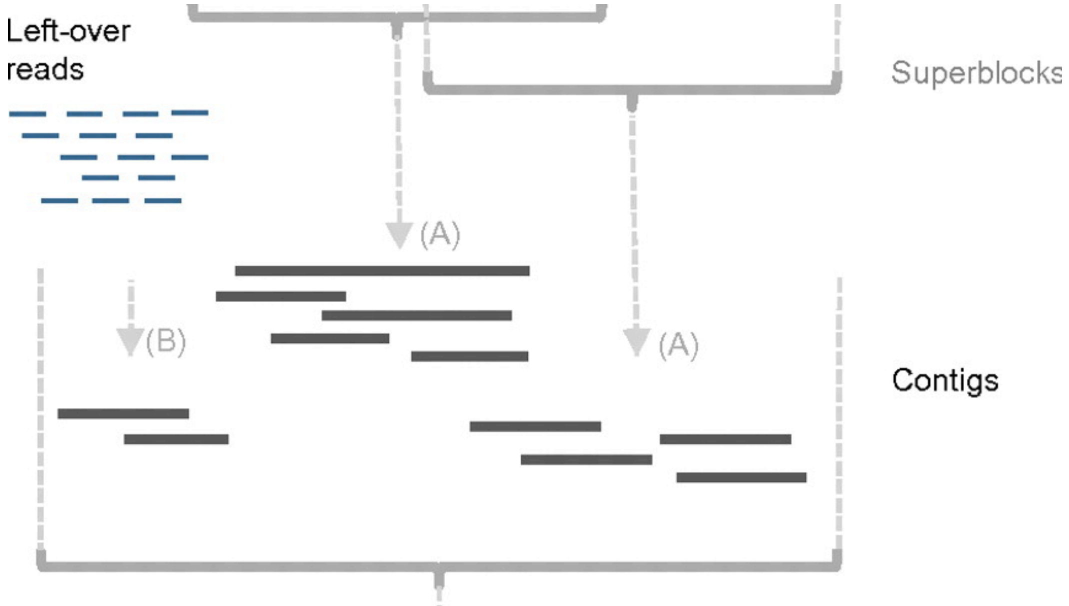
Reference-guided assembly step 1: Alignment



- Align the short reads against the reference sequence with GenomeMapper.
- Adjacent blocks were combined into superblocks, with neighboring superblocks sharing at least one block.

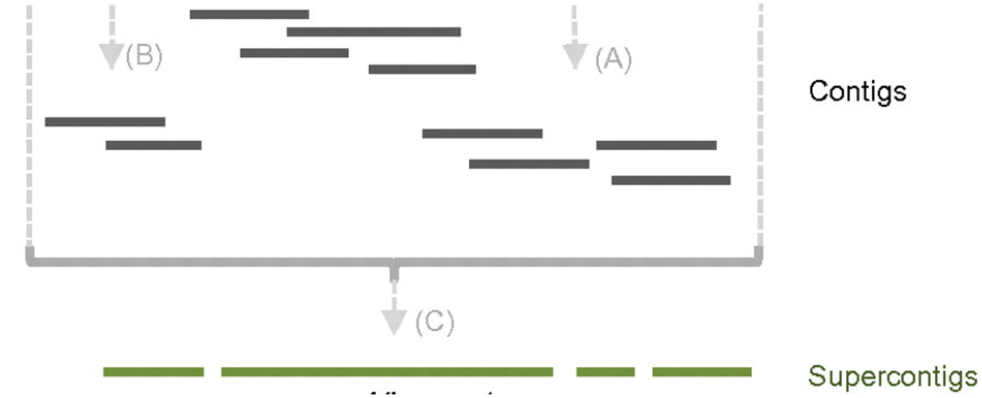
Blocks = regions with constant coverage or adjacent regions connected by aligned mate pairs.

Reference-guided assembly step 2: Assemble to contigs



- Reads corresponding to each superblock were assembled separately using the de Bruijn graph-based assemblers. (Both ABySS and Velvet with eight different kmer sizes).
- All leftover reads (unaligned) are assembled using VELVET, to get nonreference sequences.

Reference-guided assembly step 3: to supercontigs

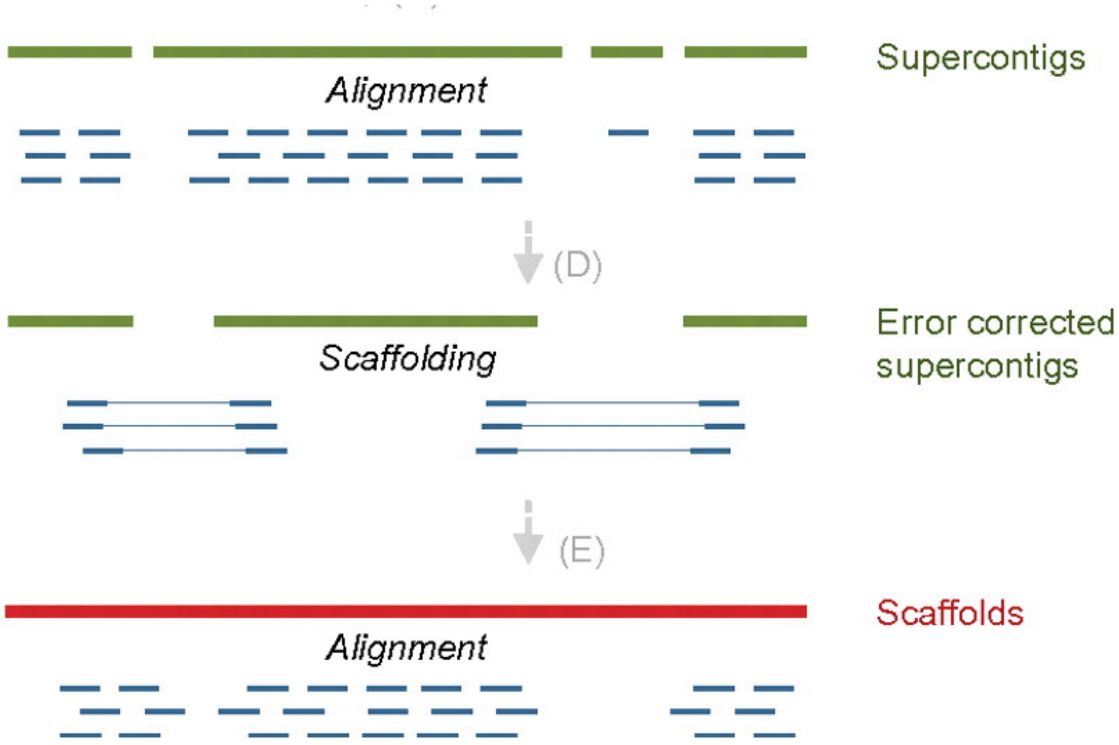


Due to different assemblies, redundancy is introduced into the contigs.

The homology guided Sanger assembler AMOScmp merge all contigs of each chromosome arm into nonredundant supercontigs

Reference-guided assembly

step 4 and 5: Error correction and Scaffolds



1. All original reads are aligned against supercontigs.
2. Differences between supercontigs and reads indicates misassemblies.
3. Correct or remov supercontigs.

- Read pairs with ends that aligned to different supercontigs were used for scaffolding with BAMBUS.

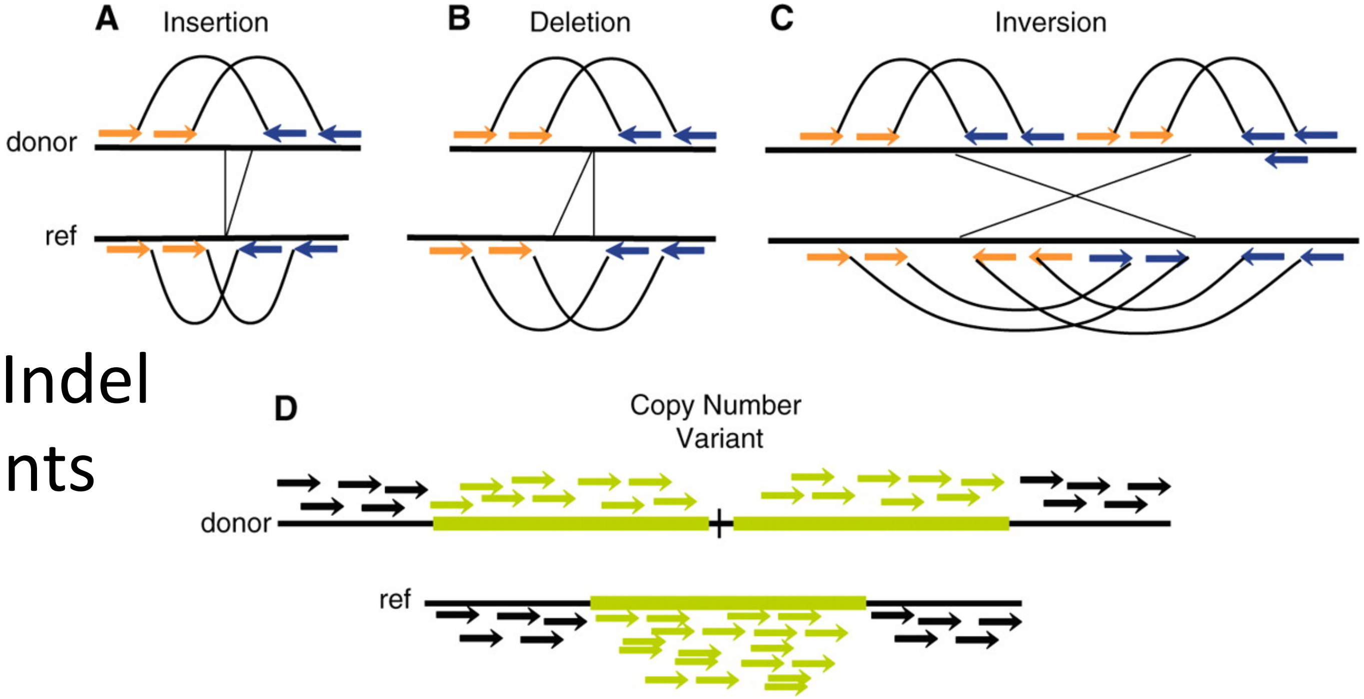
Assembly statistics

	Bur-0	C24	Kro-0	Ler-1
Coverage	83.2x	75.0x	72.7x	322.4x
Libraries	2	2	2	2
N50 (kbp)	193	109	161	297
Scaffolds	2526	2052	2670	1528
Total Length (Mbp)	101	101.3	99.9	100.8
Longest Scaffold (Mbp)	4	3.6	5.1	1.3

Ref genome
105.2Mbp

Variant discovery

Recent advances in sequencing technology make it possible to comprehensively catalog genetic variation in population samples, creating a foundation for understanding human disease, ancestry and evolution.



1. SNP and Micro-Indel
2. Structural Variants

Dalca and Brudno, 2010

Find variant with genome comparison

- MUMmer
- MUMmer is a system for rapidly aligning entire genomes, whether in complete or draft form.
- http://mummer.sourceforge.net/

Find variant with genome comparison

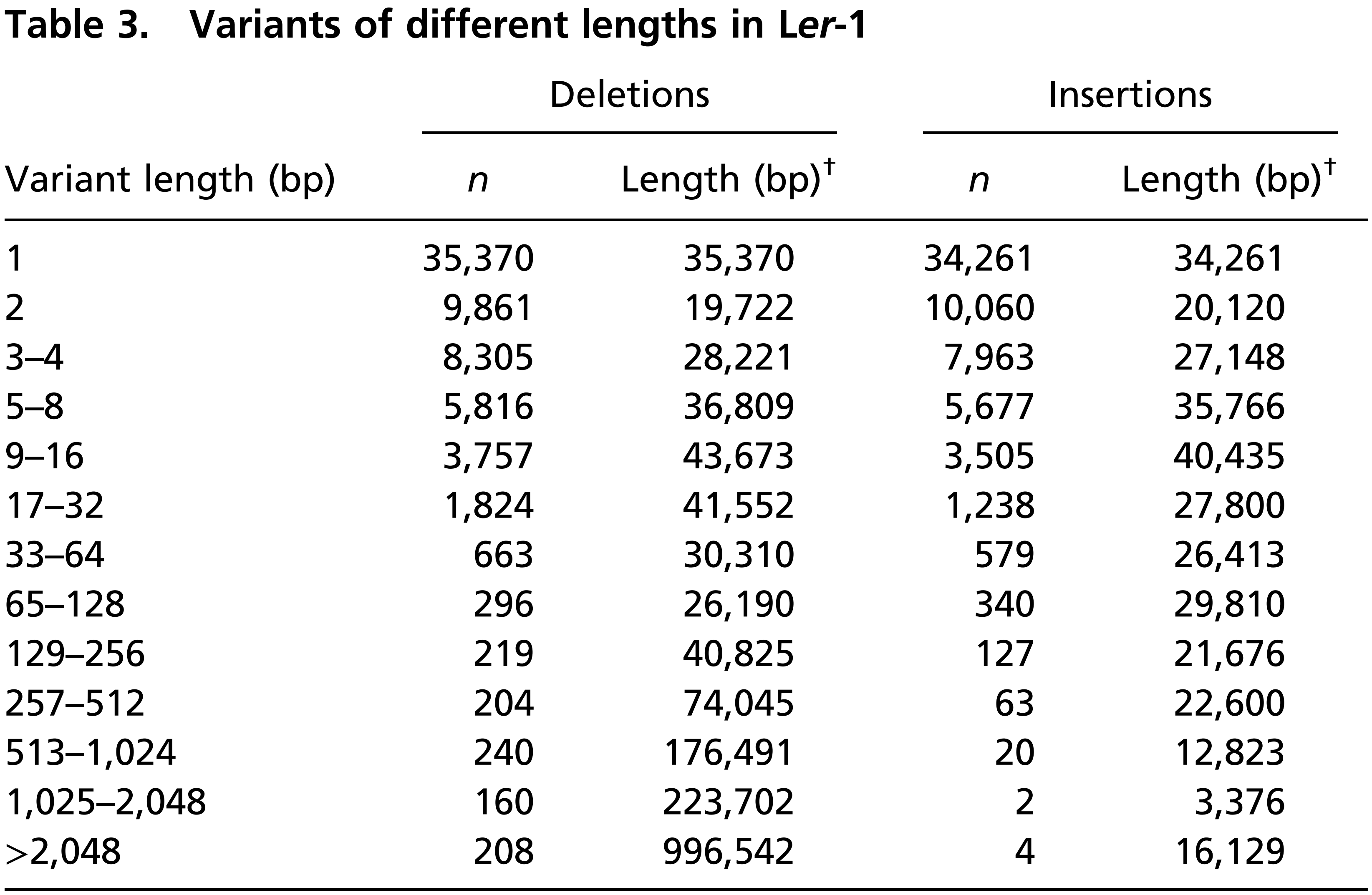
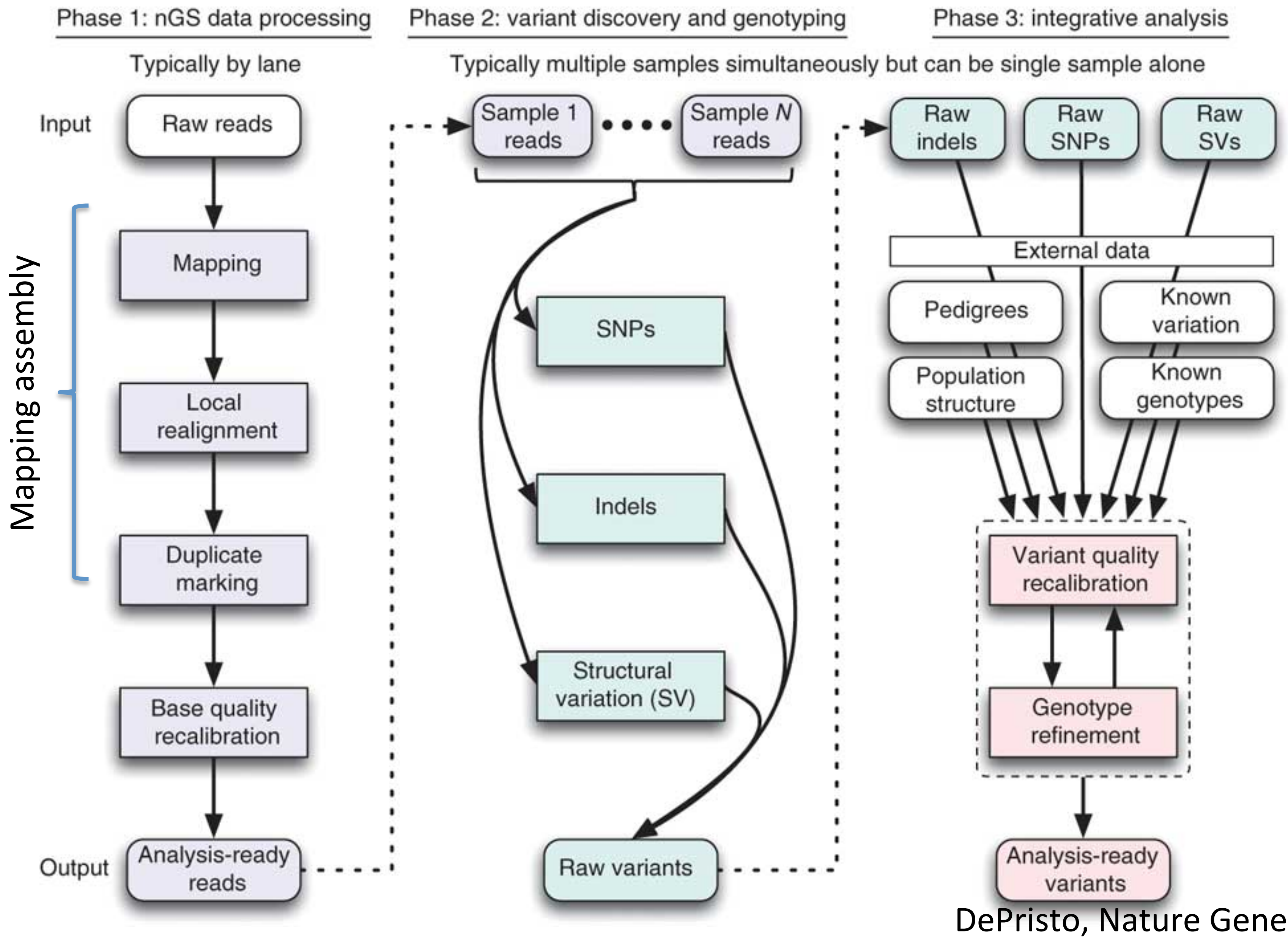


Table 3. Variants of different lengths in L*er*-1

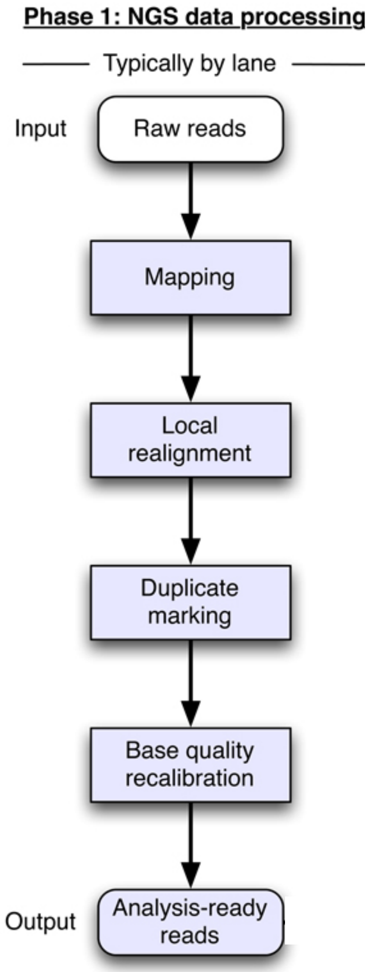
	Deletions		Insertions	
Variant length (bp)	*n*	Length (bp)[†]	*n*	Length (bp)[†]
1	35,370	35,370	34,261	34,261
2	9,861	19,722	10,060	20,120
3–4	8,305	28,221	7,963	27,148
5–8	5,816	36,809	5,677	35,766
9–16	3,757	43,673	3,505	40,435
17–32	1,824	41,552	1,238	27,800
33–64	663	30,310	579	26,413
65–128	296	26,190	340	29,810
129–256	219	40,825	127	21,676
257–512	204	74,045	63	22,600
513–1,024	240	176,491	20	12,823
1,025–2,048	160	223,702	2	3,376
>2,048	208	996,542	4	16,129

A framework for Variant discovery



DePristo, Nature Genetics, 2011

A framework for variation discovery



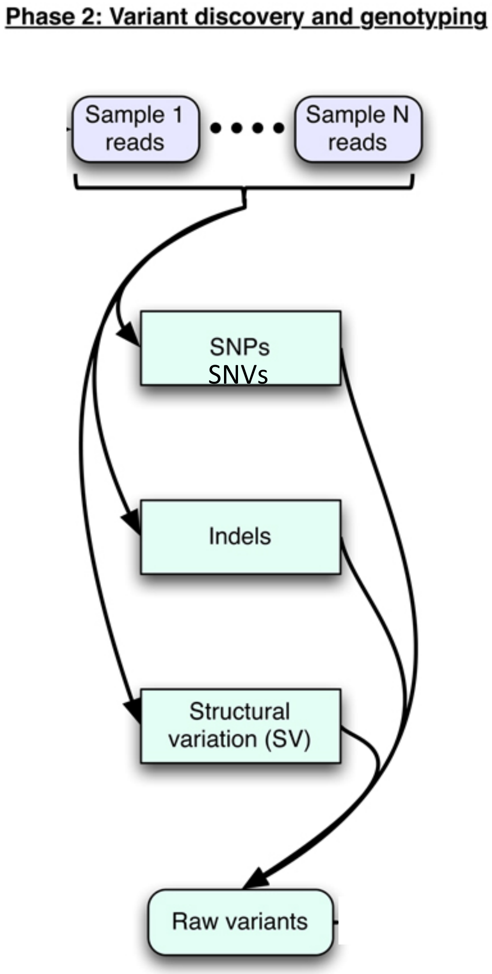
Phase 1: Mapping

- Place reads with an initial alignment on the reference genome using mapping algorithms
- Refine initial alignments
  - local realignment around indels
  - molecular duplicates are eliminated
- Generate the technology-independent SAM/BAM alignment map format

Accurate mapping crucial for variation discovery

DePristo, Nature Genetics, 2011

A framework for variation discovery

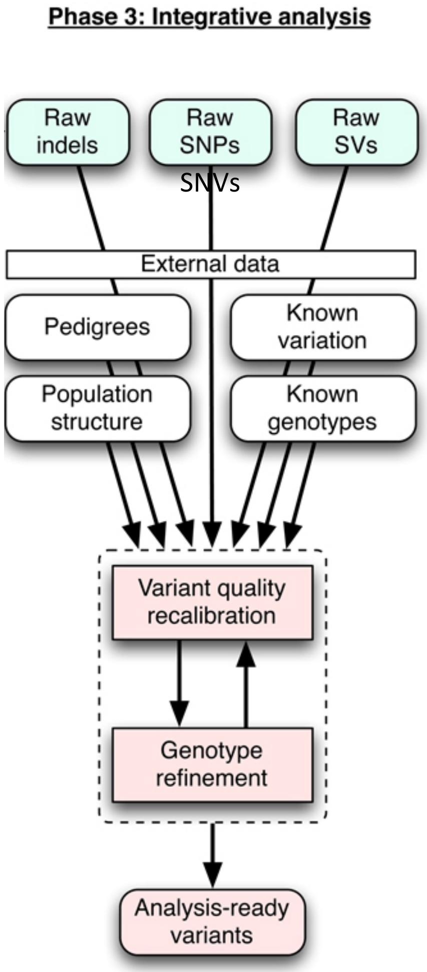

Phase 2: Variant discovery and genotyping
Sample 1 reads
Sample N reads
SNPs
SNVs
Indels
Structural variation (SV)
Raw variants

Phase 2: Discovery of raw variants

- Analysis-ready SAM/BAM files are analyzed to discover all sites with statistical evidence for an alternate allele present among the samples
- SNPs, SNVs, short indels, and SVs

DePristo, Nature Genetics, 2011

A framework for variation discovery

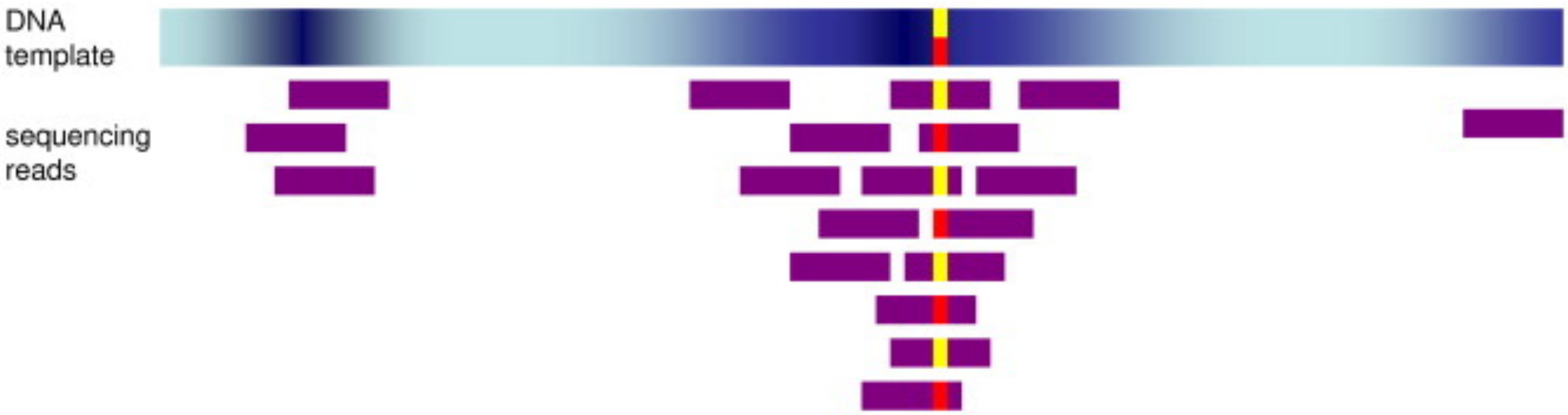


Phase 3: Discovery of analysis-ready variants

- technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium, and family and population structure are integrated with the raw variant calls from Phase 2 to separate true polymorphic sites from machine artifacts
- at these sites high-quality genotypes are determined for all samples

DePristo, Nature Genetics, 2011

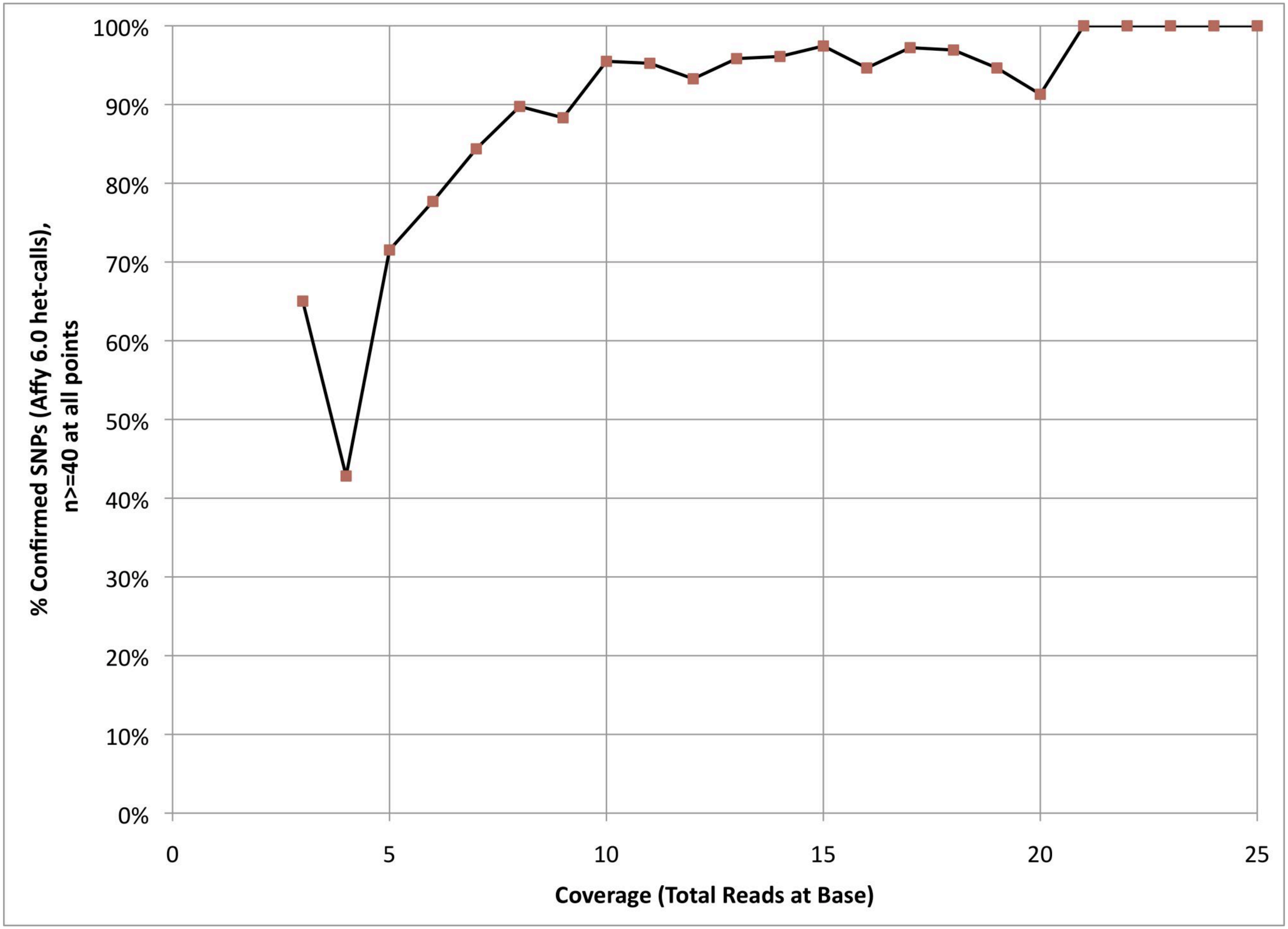
Variant discovery



1. SNP and Micro-Indel

Morozova and Marra, Genomics, 2008

8-10X coverage sufficient for high-quality SNP calls



Evaluating SNP call quality

Did I get the right number of calls?

- The number of SNP calls should be close to the average human heterozygosity of 1 variant per 1000 bases
- Only detects gross under/over calling

Concordance with hapmap chip results?

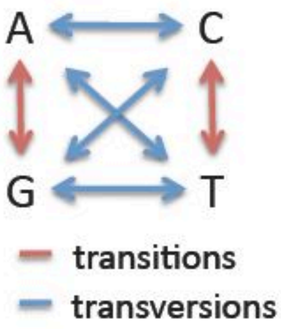
- Often we have genotype chip data that indicates the hom-ref, het, hom-var status at millions of sites
- Good SNP calls should be >99.5% consistent these chip results, and >99% of the variable sites should be found
- The chip sites are in the better parts of the genome, and so are not representative of the difficulties at novel sites

What fraction of my calls are already known?

- dbSNP catalogs most common variation, so most of the true variants found will be in dbSNP
- For single sample calls, ~90 of variants should be in dbSNP
- Need to adjust expectation when considering calls across samples

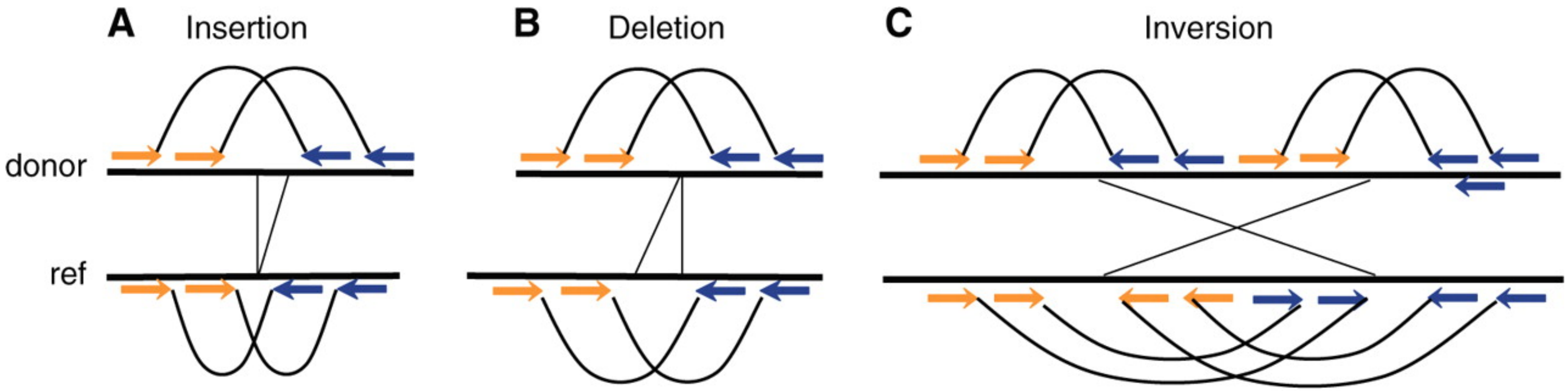
Reasonable transition to transversion ratio (Ti/Tv)?

- Transitions are twice as frequent as transversions (see *Ebersberger, 2002*)
  - Validated human SNP data suggests that the Ti/Tv should be ~2.1 genome-wide and ~2.8 in exons
- FP SNPs should has Ti/Tv around 0.5
- Ti/Tv is a good metric for assessing SNP call quality



1Kg

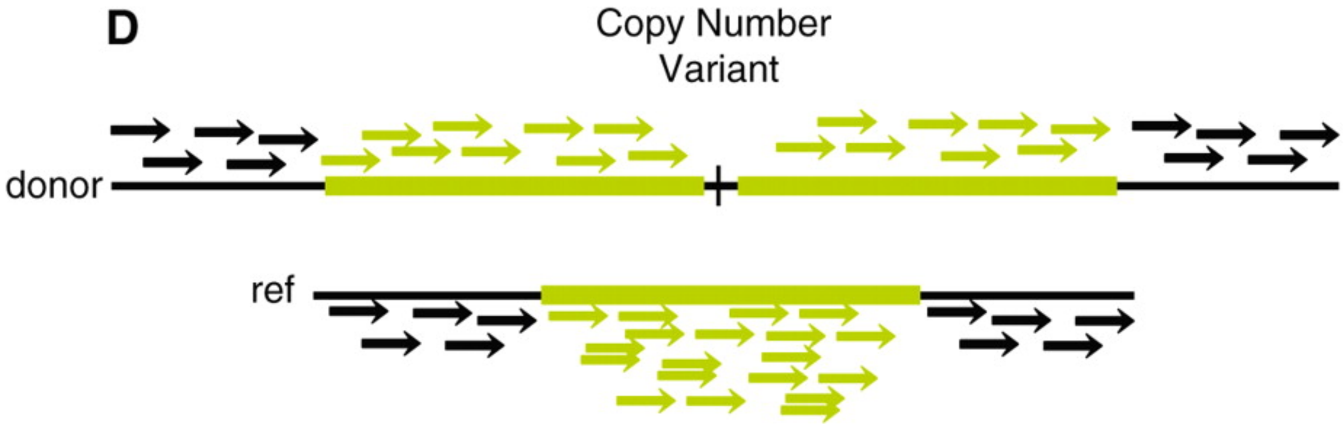
Variant discovery



Indels

Dalca and Brudno, 2010

Variant discovery



Dalca and Brudno, 2010

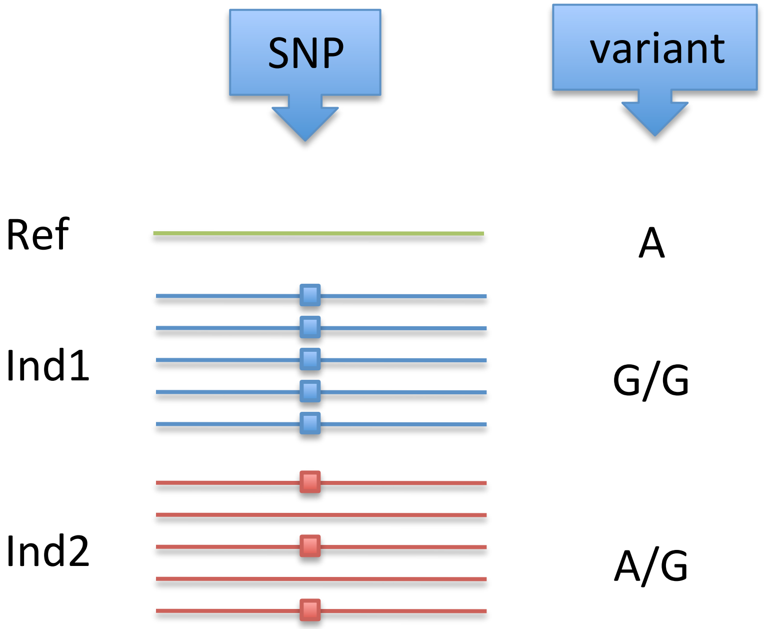
Open Factors of a Variant’s Fidelity

```
How do we know the quality is good?
```

(N) Number of reads supporting that site,
(P_v) Probability of that platform-specific variant change,
(QVD) The average deviation of the quality values,
(T) The set of alignments with unique start sites,
(D) PCR Duplicates,
(S) Strand representation (half on one, half on the other),
(Z) Zygosity change (CNV regions)
(C) Cellular heterogeneity

Variant calling methods

- > 15 different algorithms
- Three categories
  - Allele counting
  - Probabilistic methods, e.g. Bayesian model
    - to quantify statistical uncertainty
    - Assign priors based on observed allele frequency of multiple samples
  - Heuristic approach
    - Based on thresholds for read depth, base quality, variant allele frequency, statistical significance



http://seqanswers.com/wiki/Software/list

Nielsen R, Paul JS, Albrechtsen A, Song YS.. Nat Rev Genet. 12(6):443-51.

Variant callers

Name	Category	Tumor/Normal Pairs	Metric	Reference
Bambino	Allele Counting	Yes	SNP Score	Edmonson, M.N. et al. (2011)
JointSNVMix (Fisher)	Allele Counting	Yes	Somatic probability	Roth, A. et al. (2012)
Somatic Sniper	Heuristic	Yes	Somatic Score	Larson, D.E. et al. (2012)
VarScan 2	Heuristic	Yes	Somatic p-value	Koboldt, D. et al. (2012)
Genome Analysis ToolKit (GATK)	Bayesian	No	Phred QUAL	DePristo, M.A. et al. (2011)

the SAM/BAM format. Bioinformatics 27 (6): 865-866 (2011).

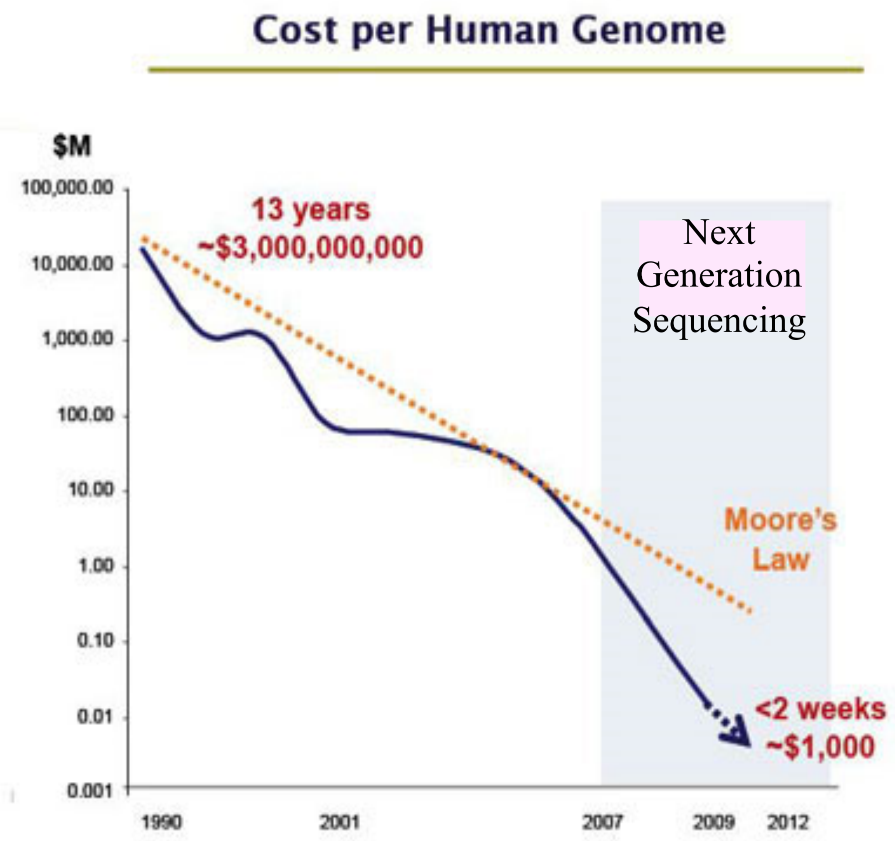
Roth, A. et al. JointSNVMix : A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/ Tumour Paired Next Generation Sequencing Data. Bioinformatics (2012).

Larson, D.E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 28(3):311-7 (2012).

Koboldt, D. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research DOI: 10.1101/gr.129684.111 (2012).

DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

Accelerating Technology & Plummeting Cost



Commodore PET 2001 with 4KB memory is $795 in 1977 (=$2,800 in current $)

iPad has 16GB memory and is $499.

Personal genome sequencing Applications

Human genetic variation

- Single Nucleotide Polymorphisms (SNPs)
- Small insertion/deletions (Indels)
- Structural Variation (SV)

Linking genetic variants to disease

- Functional categorization of SNPs
- Genome-wide association studies(GWAS)

Why study genetic variation?

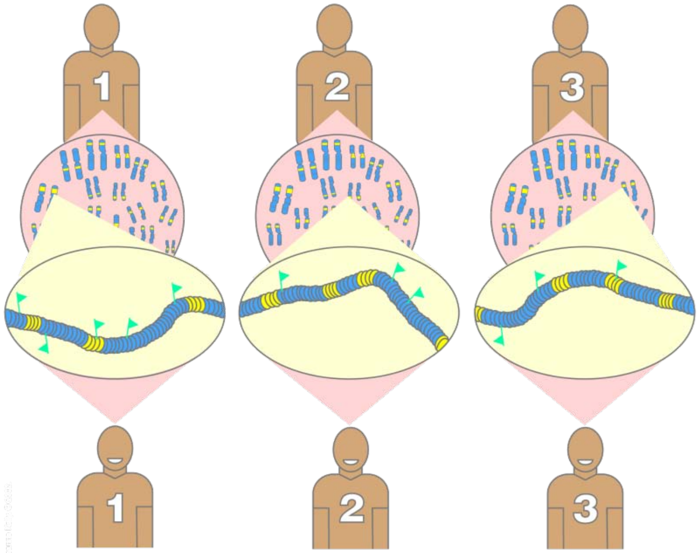
- SNPs can serve as genetic markers to identify genomic regions associated with disease.
- Disease-associated SNPs, regardless of function, have potential for clinical applications, including prediction of disease risk, treatment response, and prognosis.
- Maybe responsible for aberrant gene expression and protein function that drive disease processes or play a role in drug response.

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

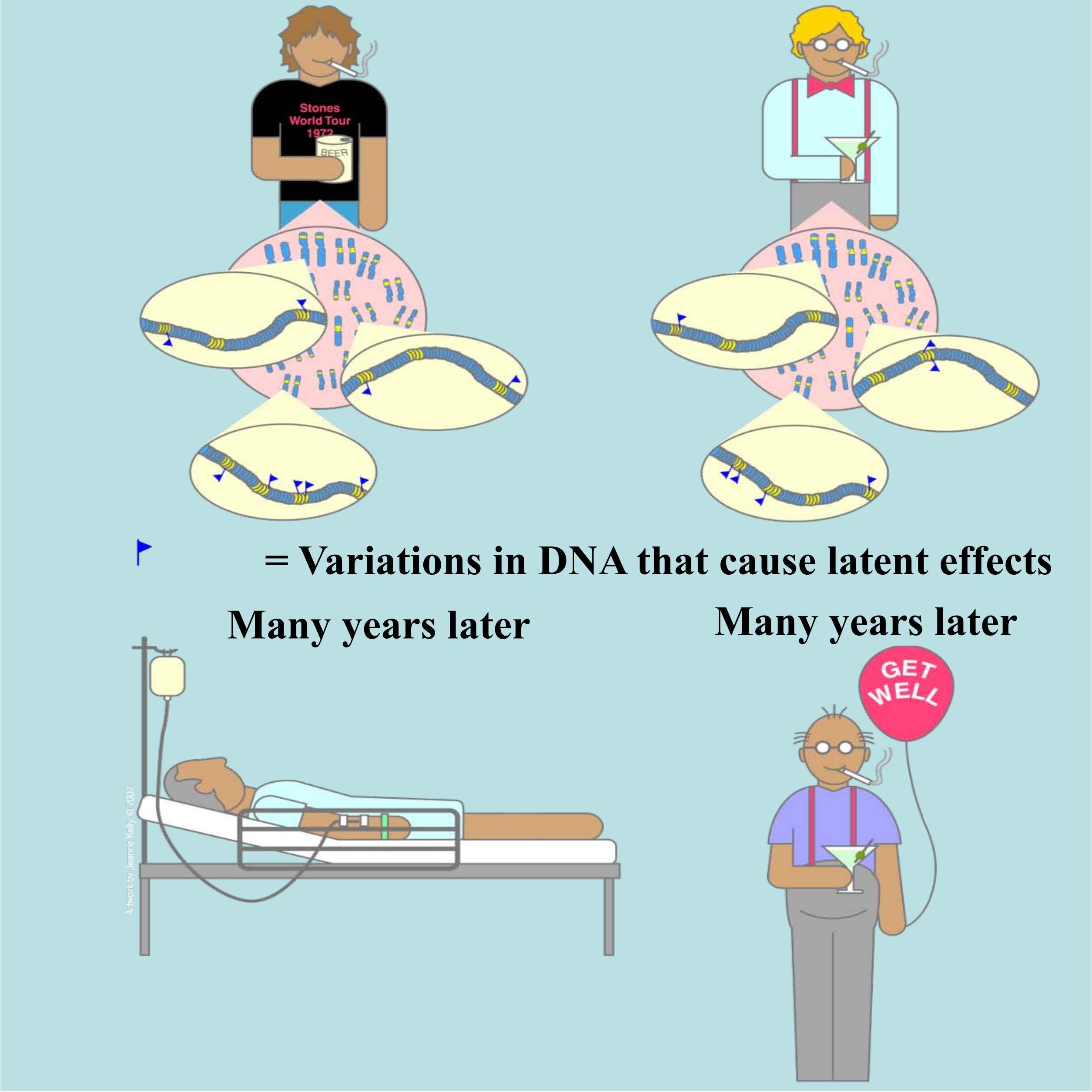
Most genetic variations have no effect

- Most genetic variations in the human genome are silent variations, i.e. have no phenotypic effect
- Do not occur in coding or regulatory regions of genesor are within these regions but have no effect



⚑ = Variations in DNA that cause no changes

Some genetic variations do have effect. We need to identify them.

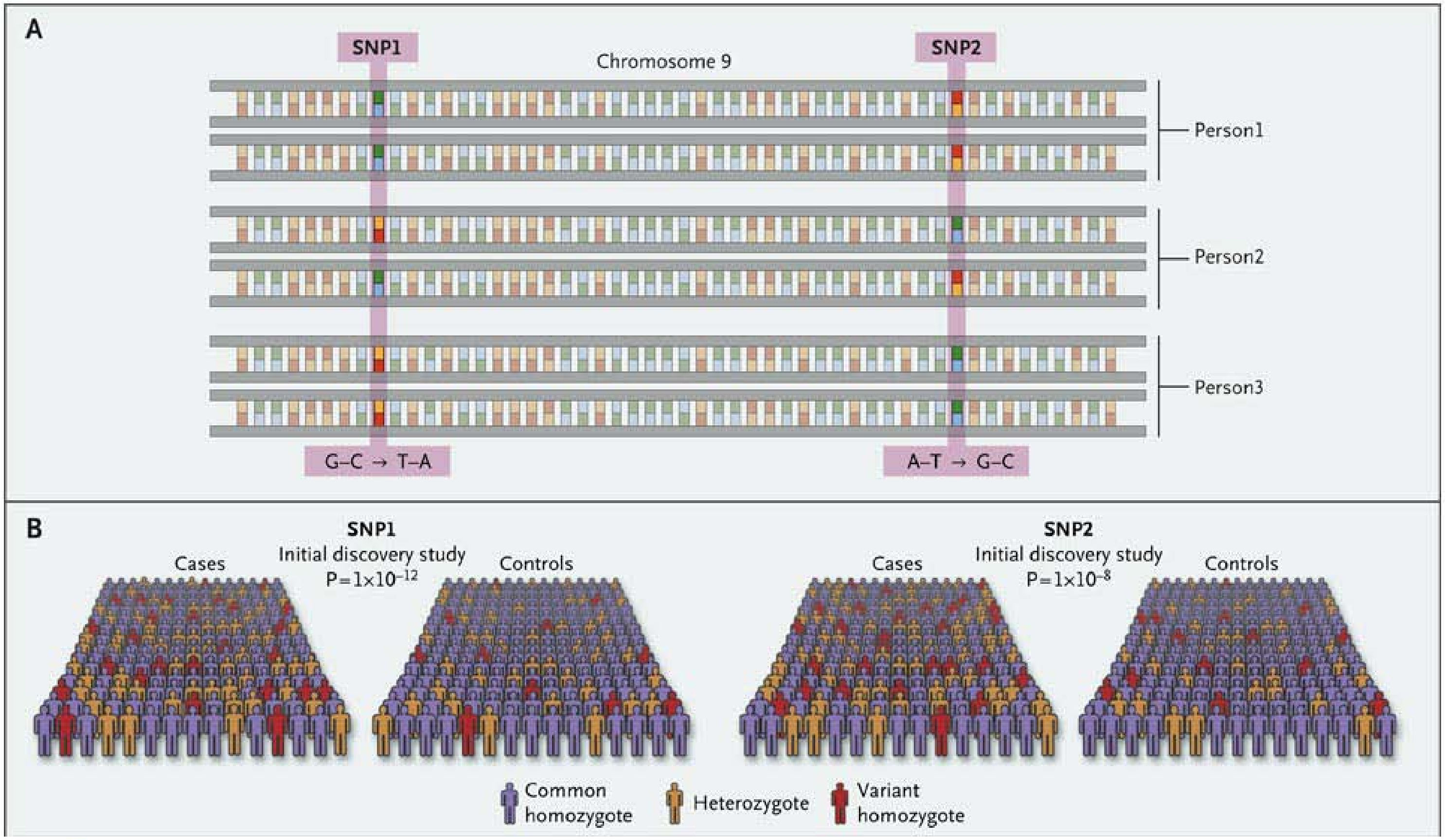


⚑ = Variations in DNA that cause latent effects

Many years later

Many years later

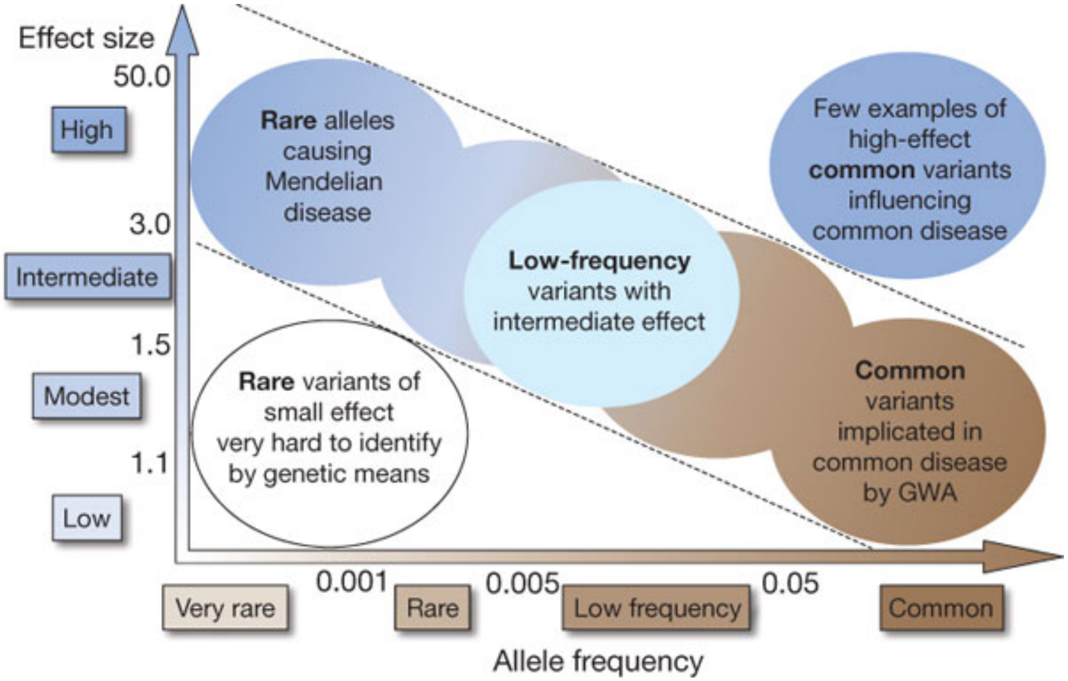
(GWAS) are commonly used to link genetic variations (mostly SNPs) with disease or health-related traits



Manolio T. N Engl J Med 2010;363:166-176.

Key underlying principles for GWAS

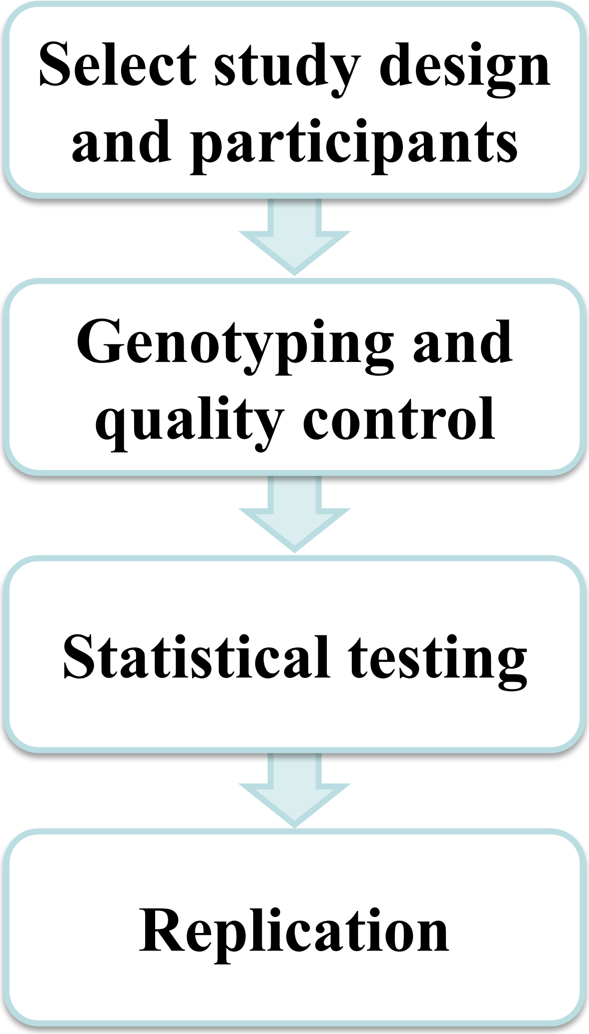
- **'Common disease, common variant' hypothesis** posits that common variants present in more than 1–5% of the population contribute to common diseases
- GWAS generally do not capture rare variants



Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Manolio et al. Nature 461, 747-753

Typical GWAS approach



Select study design and participants

Selection of a large number of individuals with the phenotype (disease or trait) of interest and an appropriate comparison group

Genotyping and quality control

DNA isolation, genotyping, and application of quality control measures

Statistical testing

Statistical analysis to test for associations between the genetic variants (SNPs) passing quality thresholds and the phenotype

Replication

Replication of genotyping in independent samples using a subset of SNPs found to be significant in the initial study or experimental investigation of functional implications

Quality Control

- Poor study design and errors in genotype calling can introduce systematic bias in association studies.
  - Increase in false positive error rate and decrease in power.
- Assess data quality to remove sub-standard genotypes, samples and SNPs from subsequent association analysis.

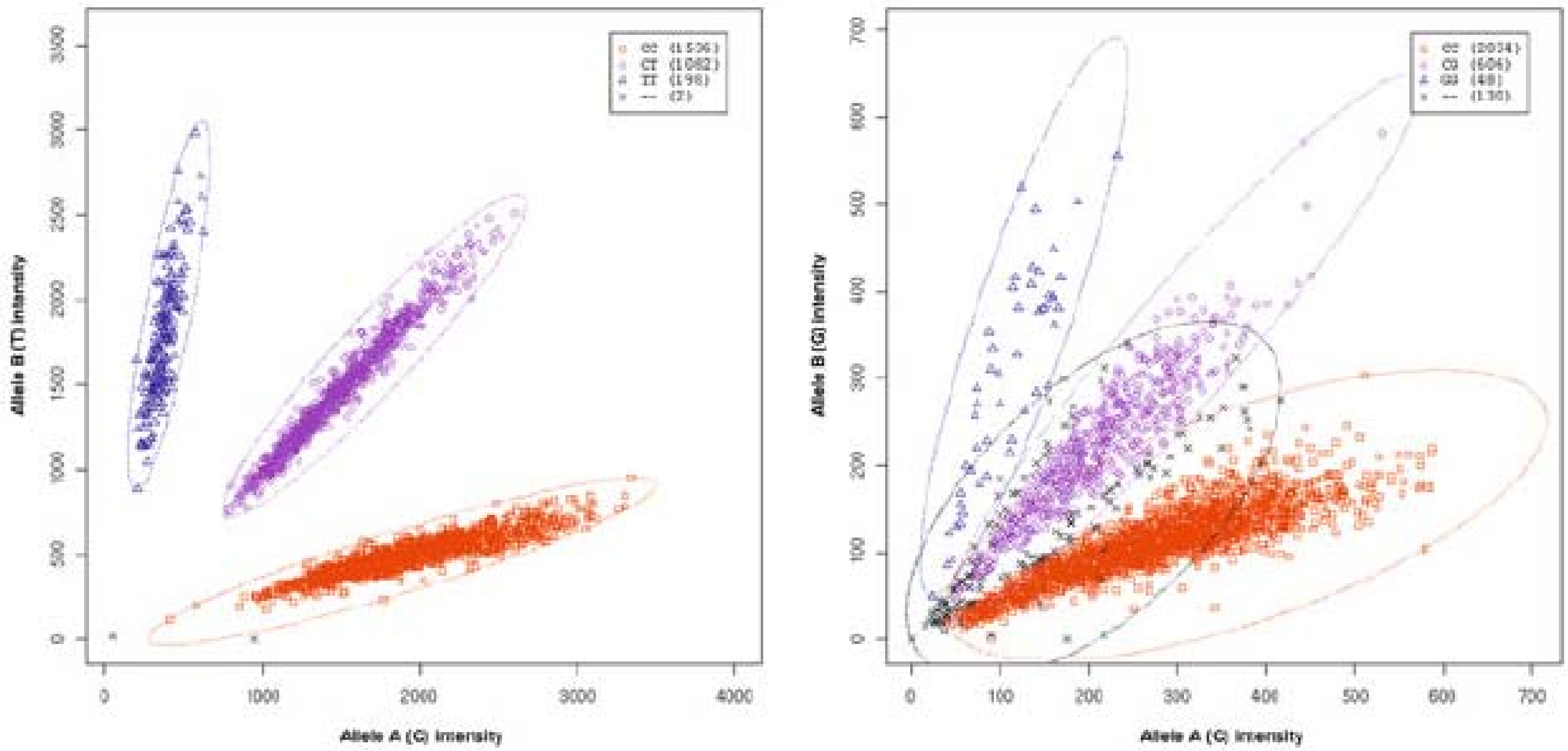
Sample quality control

- Remove samples on the basis of:
  - Low call rate (poor DNA quality).
  - Outlying heterozygosity across autosomes (DNA sample contamination or inbreeding).
  - Duplication or relatedness based on identity-by-state (samples should be independent).
  - Mismatches with external information (sample mix-up).
  - Outlying population ancestry (confounding due to population structure).

SNP quality control

- Remove SNPs on the basis of:
  - Low call rate, poor quality SNP.
  - Extreme deviation from Hardy-Weinberg equilibrium in cases, controls or both (genotyping error).
  - Extreme differential call rates between cases and controls (calling bias).
  - Study specific SNP QC filters (such as differences in allele frequencies between multiple control cohorts).
  - Low frequency SNPs (more prone to bias due to genotyping error and low power to detect association).
  - Visual inspection of cluster plots.

Intensity plots



(a) SNP with good genotyping quality

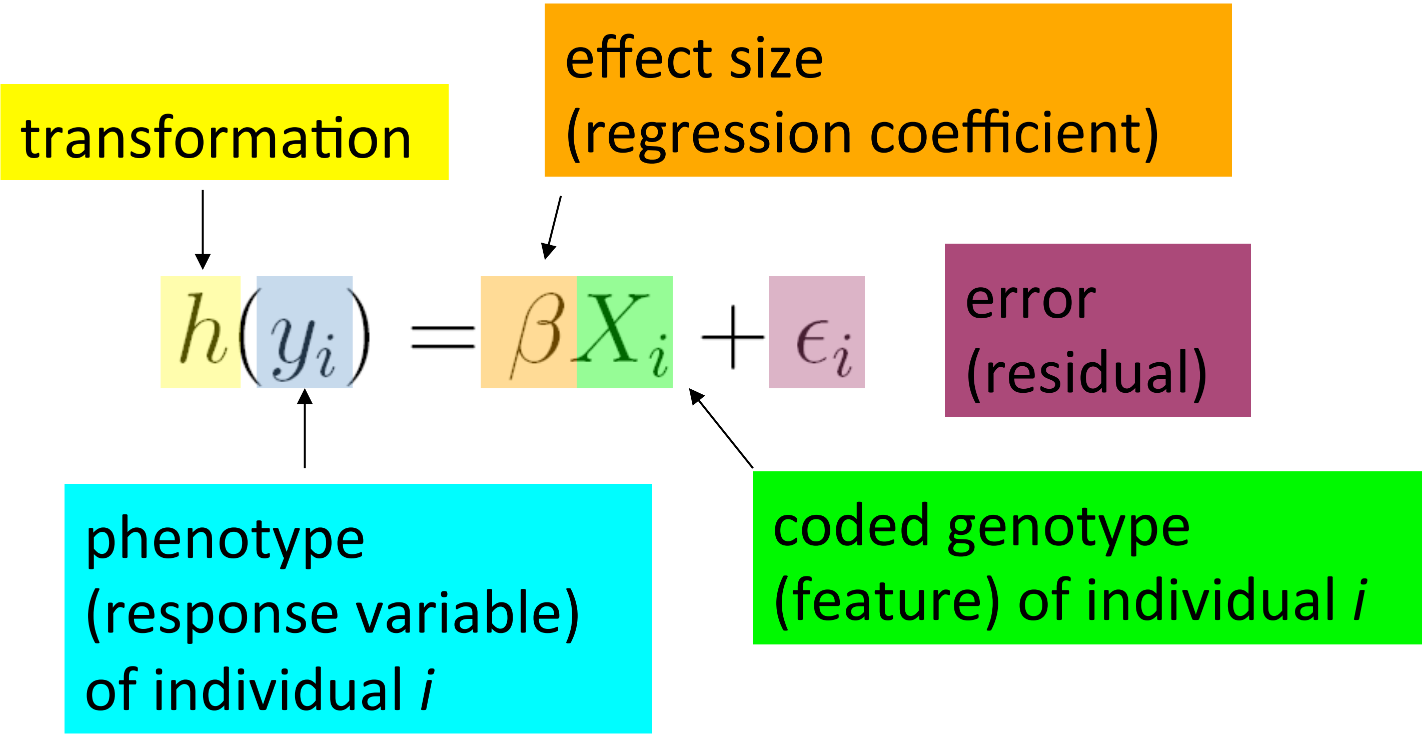
(b) SNP with poor genotyping quality

Laird and Lange (2011)

Statistical Testing

- Association analyses focus on the identification of SNPs that differ in allele (genotype) frequency between cases and controls.
- Basic analysis utilizes standard statistical epidemiological tools:
    - contingency table analysis;
    - logistic regression modelling.

Regression formalism



transformation

effect size
(regression coefficient)

$$h(y_i) = \beta X_i + \epsilon_i$$

error
(residual)

phenotype
(response variable)
of individual *i*

coded genotype
(feature) of individual *i*

Goal: Find effect size that explains best all (potentially transformed) phenotypes as a linear function of the genotypes

Statistical testing tools

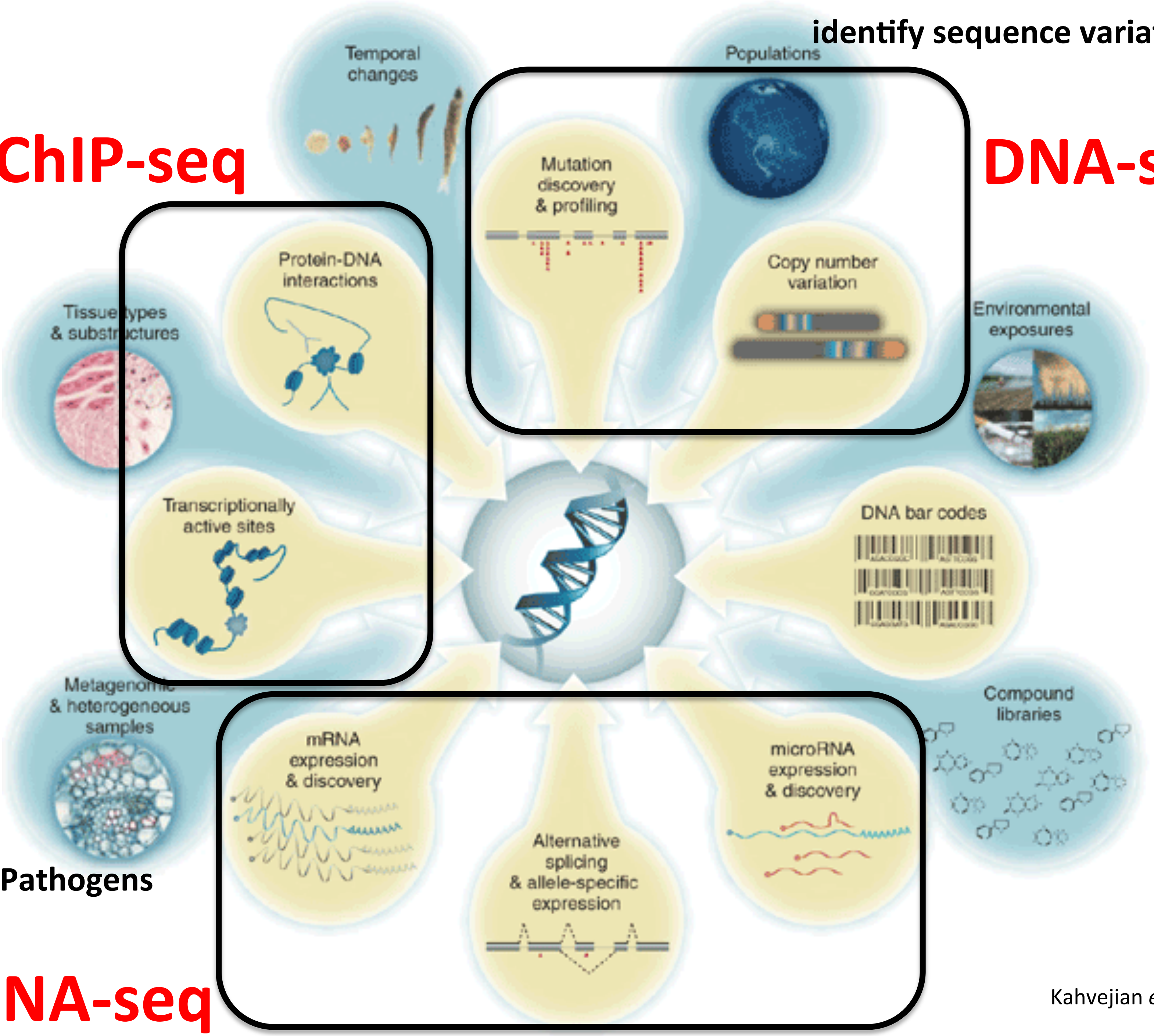
- Generalised linear modelling can be performed using standard statistical software, or some statistical software packages include specific libraries of routines to perform genetic analyses, such as R.
  - Define indicator variables for specific genetic models from the observed SNP genotype data.
- Specialised genetic analysis software:
  - ***PLINK***. Whole genome association analysis toolset designed to perform a range of basic, large-scale analyses. Allows for data management and basic QC analyses. Performs simple case-control tests of association.
  - ***SNPTEST***. Designed for analysis of whole genome association studies. Allows for flexible single-locus analysis of genotype data allowing for covariates.

Replication

- As in any association study, the most important step after the discovery of a novel association between a SNP and a trait is to validate or replicate the association in an independent studies
- To confirm positive association signals from an initial study, it is essential to replicate the result in independent samples from the same and/or different populations.
- Replication of positive association signals has not proved to be easy: will depend on power of both initial and replication studies.
- The results of the replication studies are likely to vary, so they are often combined in a meta-analysis to reach an overall conclusion.

GWAS Limitations

- Lack of functional information
- Many associated variants are not causal
- Statistical power issues. Statistical analysis entails an enormous number of association tests resulting in high potential for false-positive results
- “Missing heritability”


Temporal changes
Populations
Mutation discovery & profiling
Copy number variation
Protein-DNA interactions
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression

identify sequence variations

ChIP-seq

DNA-seq

Identify Pathogens

RNA-seq

Kahvejian *et al*, 2008