Structure, Measurement & Analysis of Genetic Variation

Sven Cichon, PhD

Professor of Medical Genetics, Director,
Division of Medcial Genetics, University of Basel

Institute of Neuroscience and Medicine (INM-1),
Research Center Jülich

Institute of Human Genetics,
University of Bonn

JÜLICH
FORSCHUNGSZENTRUM

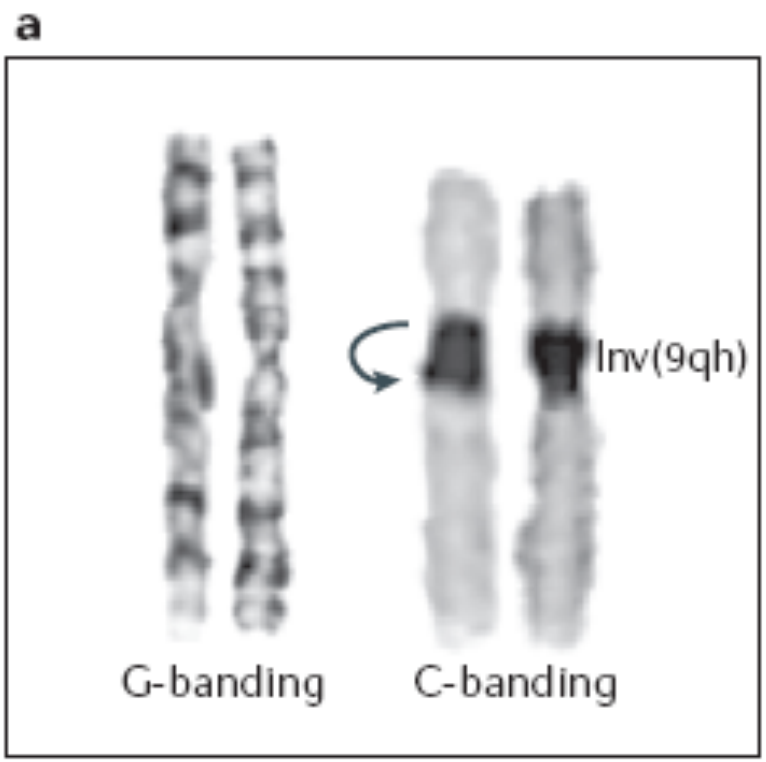
Universitätsspital
Basel

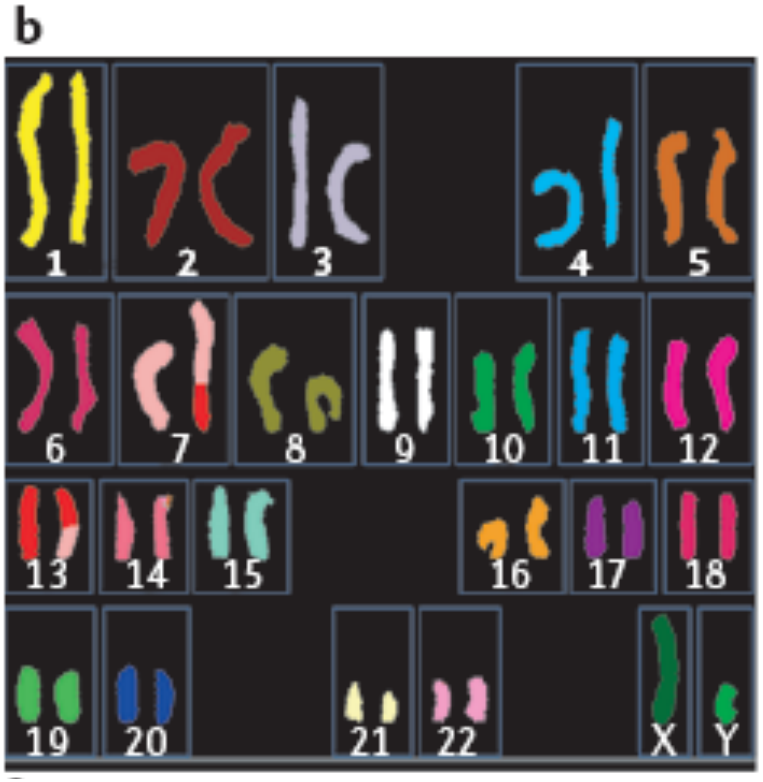
Learning objectives

- Get a feel for the variability in the human genome
- What are the most important types of genetic variation?
- Possible functional consequences of genetic variation
- Molecular genetic techniques to individually measure genetic variation

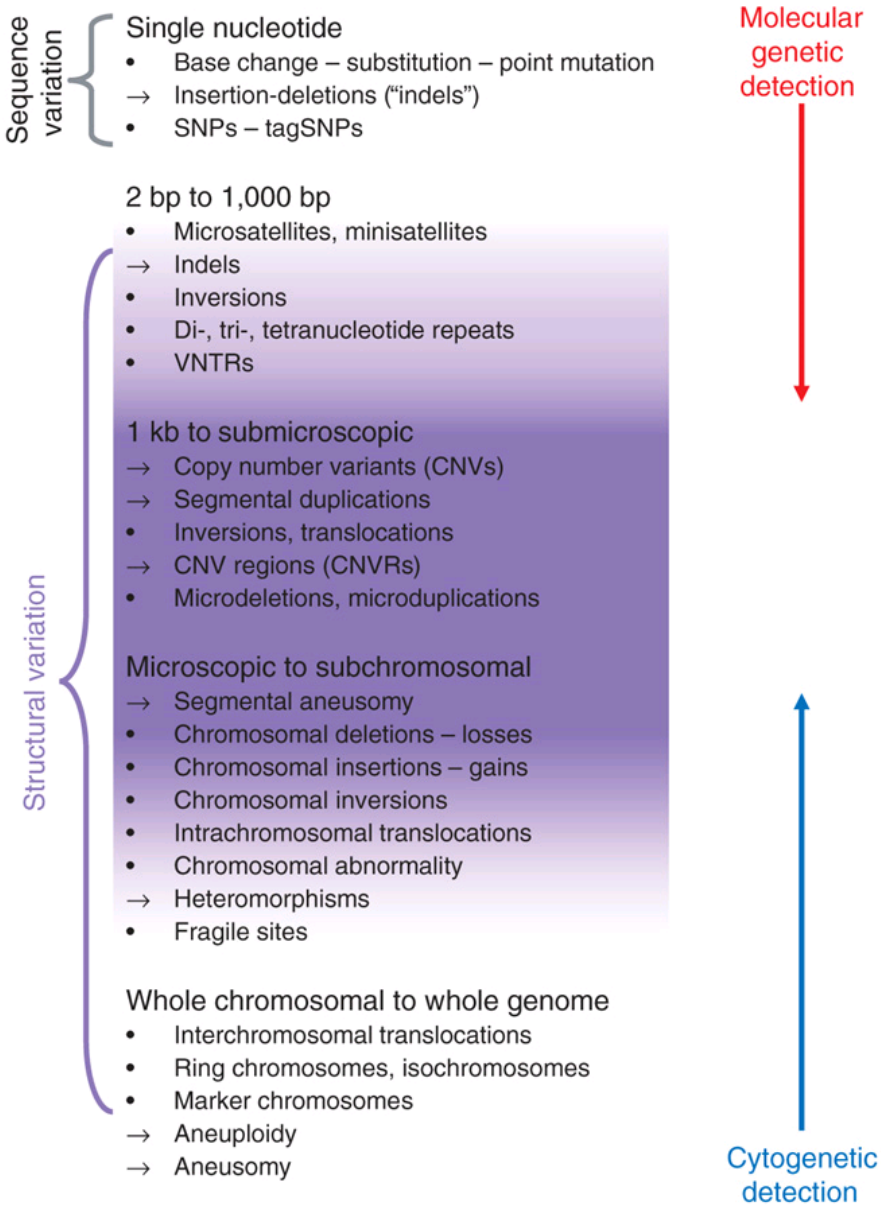
  (as a prerequisite to investigate its influence on brain imaging traits)

Genome variation is visible under the microscope already....

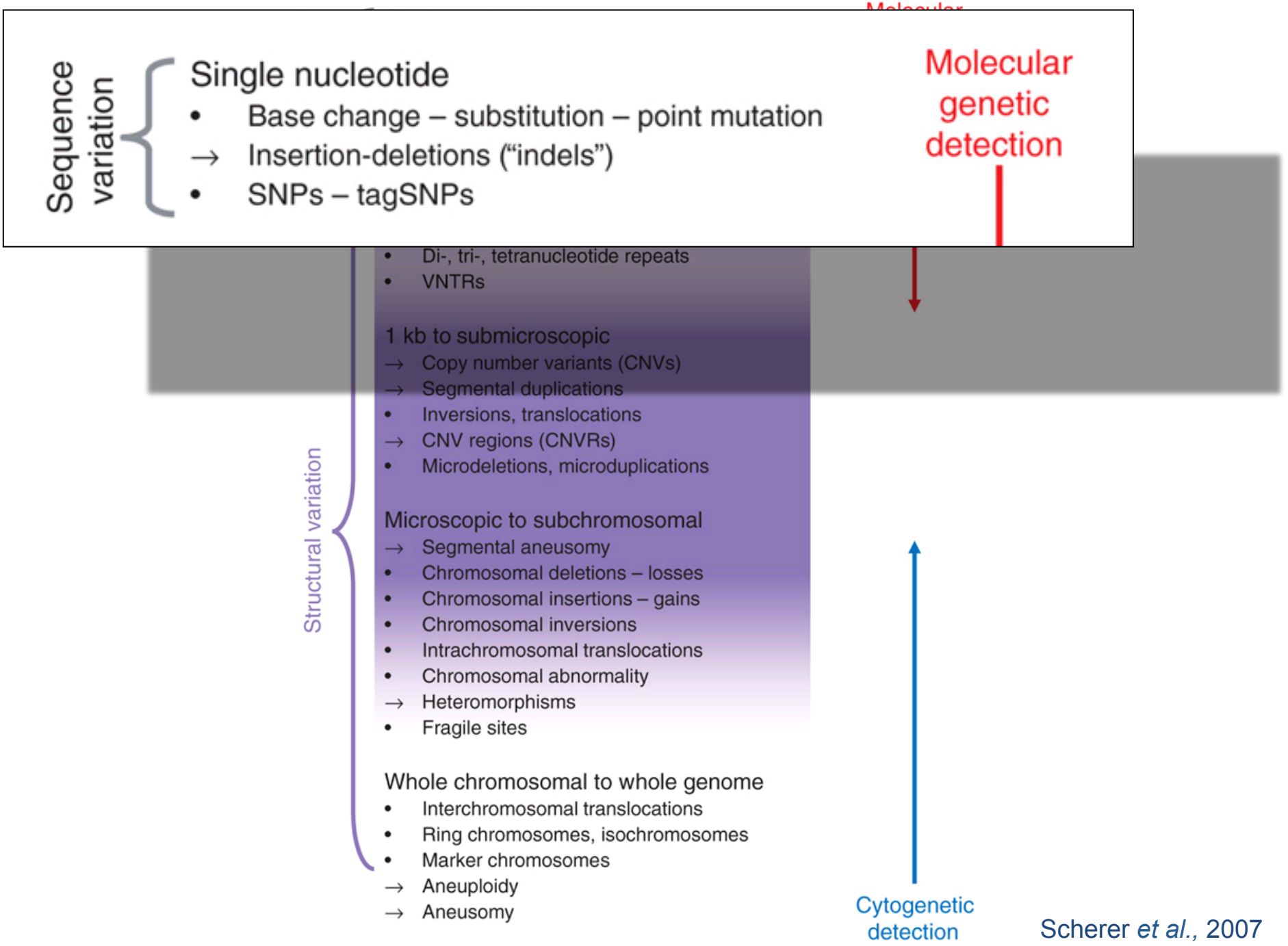




....but it gets enormous at the submicroscopic level

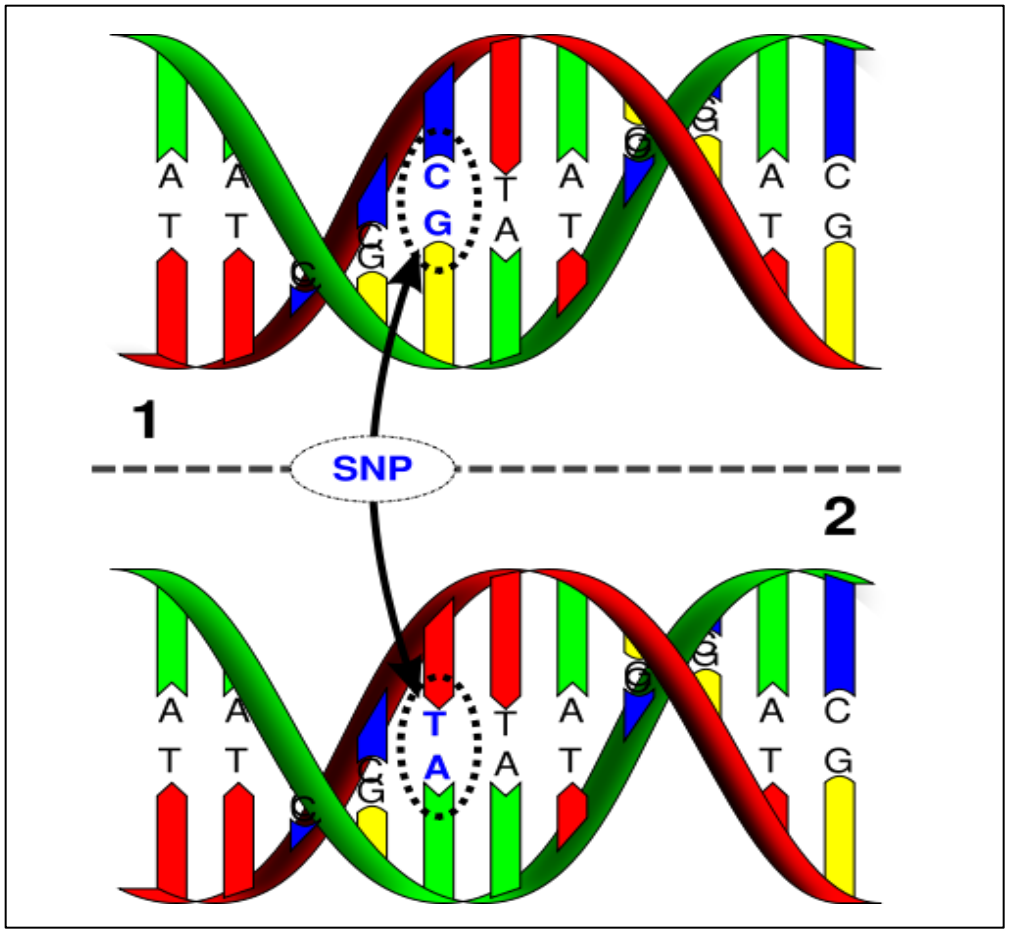


Scherer *et al.*, 2007



Scherer *et al.*, 2007

Single nucleotide polymorphism (SNP)



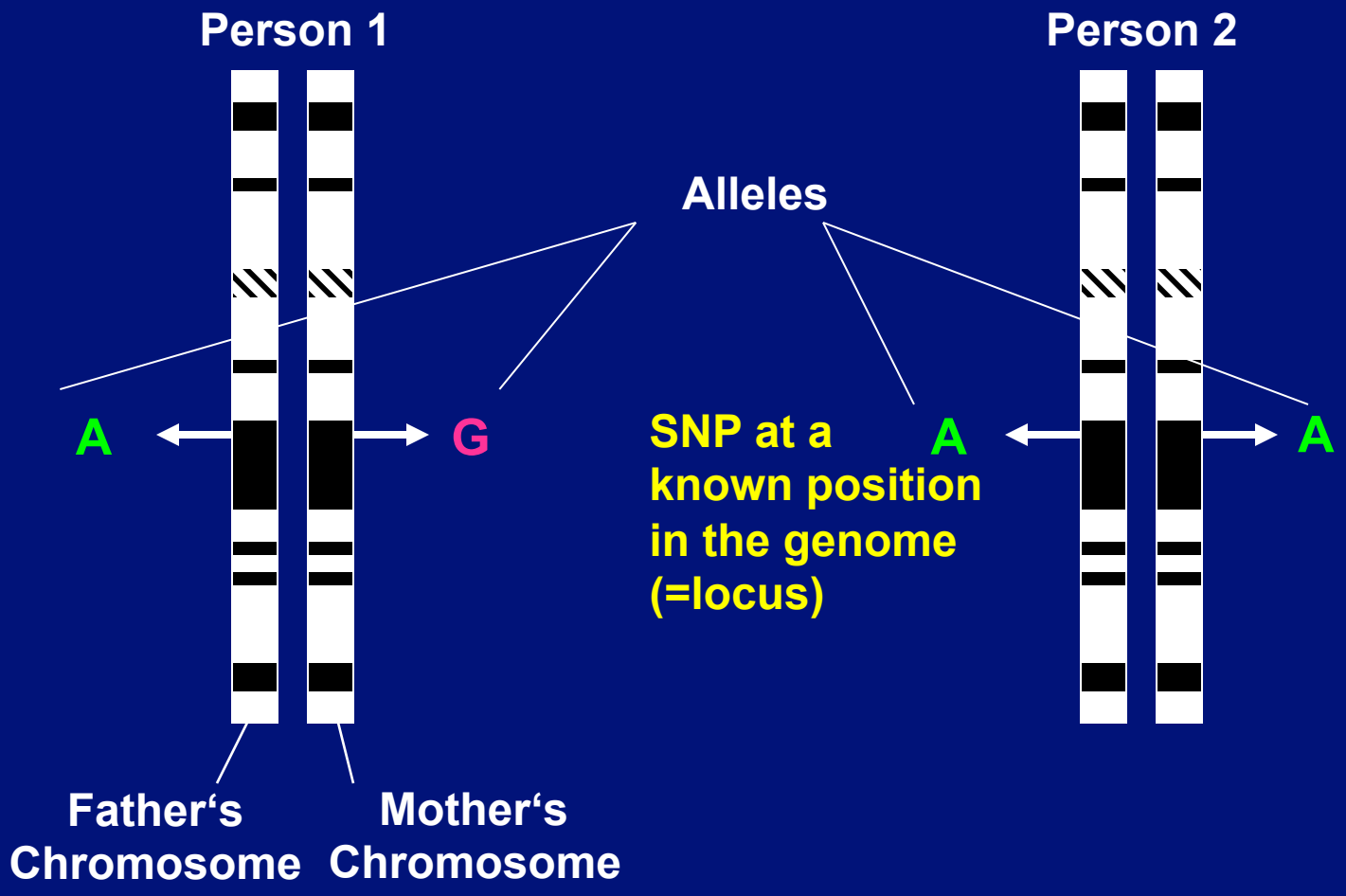
C-allele: 70% frequency

C = major allele

T-allele: 30% frequency

T = minor allele

Terminology: alleles, genotypes, SNPs,.....



Heterozygous genotype

Homozygous genotype

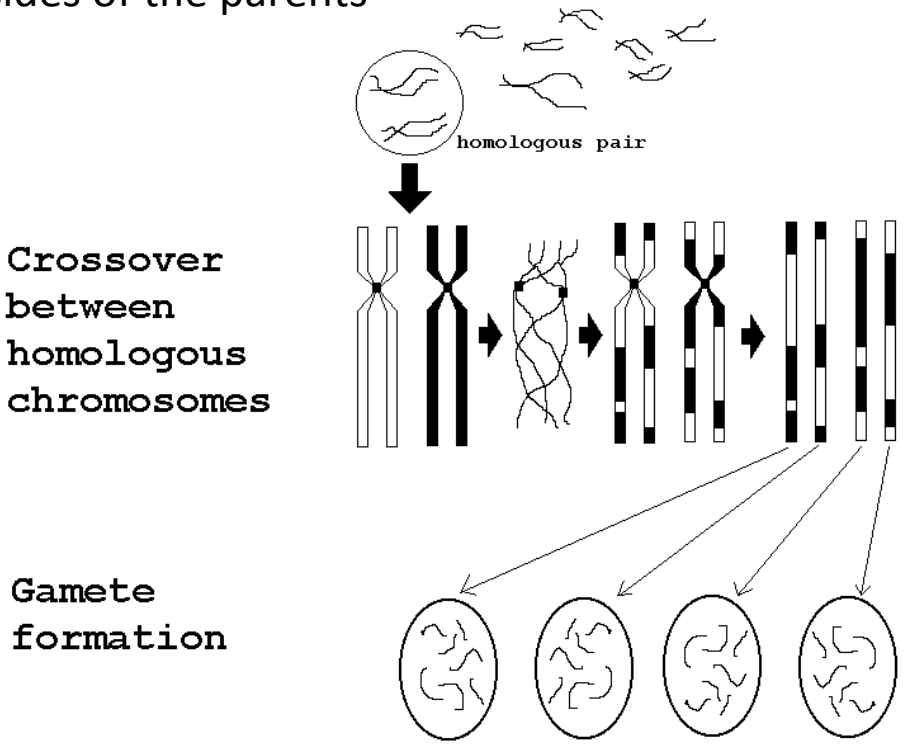
How many SNPs in the human genome ?

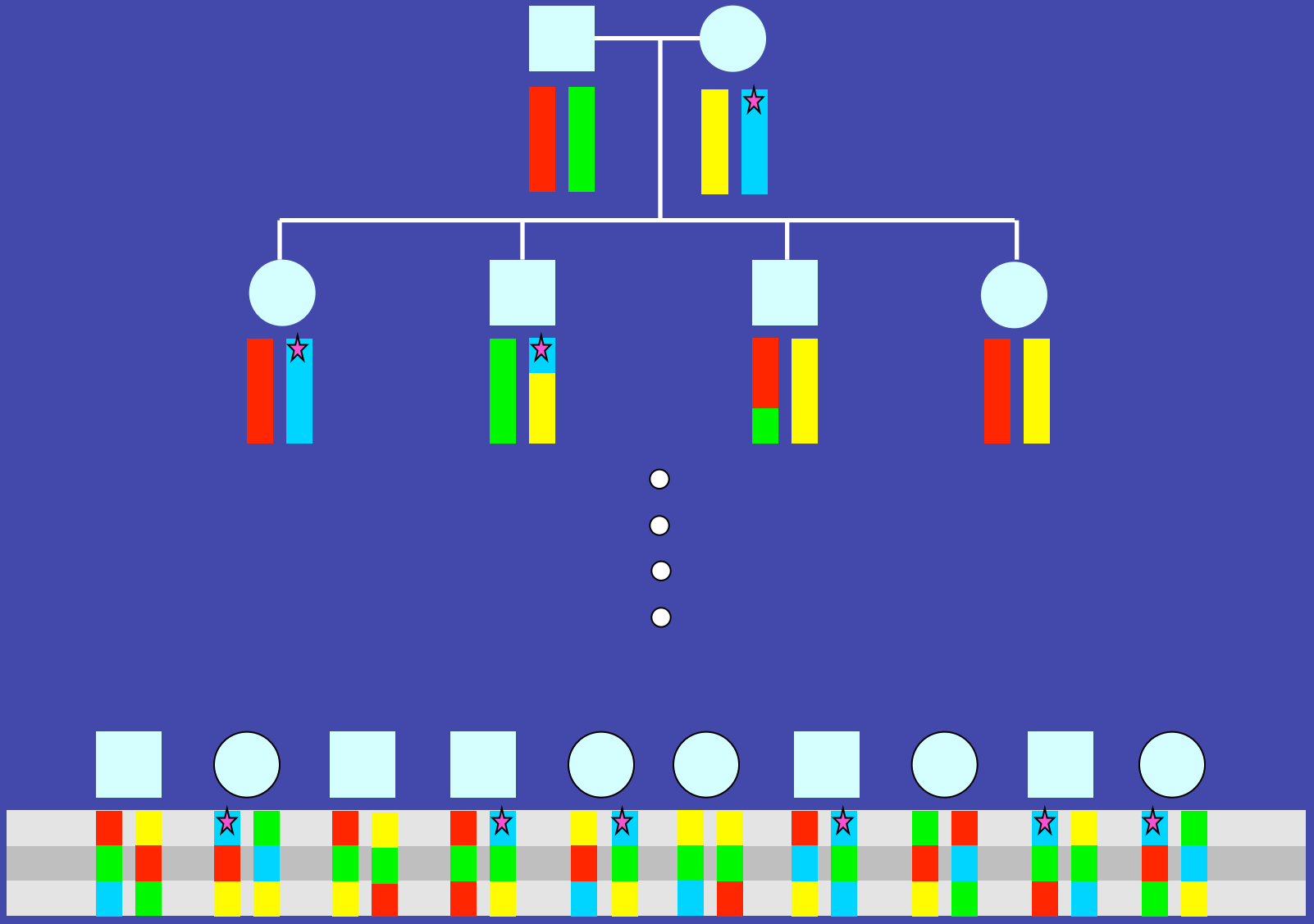
minor allele frequency	No. of SNPs (Mio.)	SNP density (1 SNP/bp)
1	12.0	290
5	7.1	450
10	5.3	600
20	3.3	960
30	2.0	1,570
40	0.97	3,280

Based on mutation rate and population size it can be assumed that every base pair of the human genome exists in a mutated form in at least several individuals.

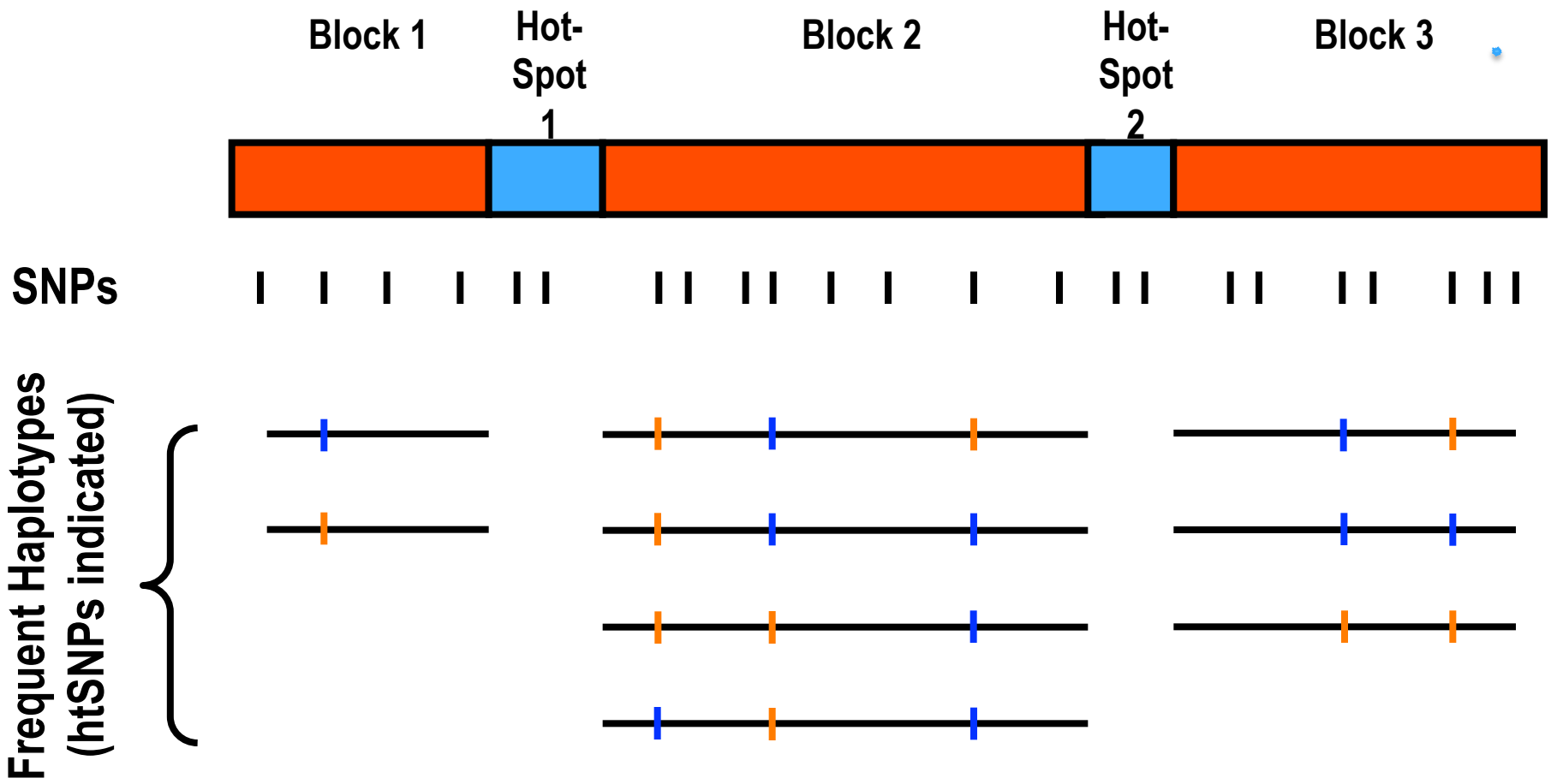
Crossing-over and Recombination

- During meiosis, two homologous chromosomes, one from mom and one from dad, twist around each other
- Large segments of DNA are exchanged and recombined
- Gamete (egg cell or sperm cell) is formed, each carries one set of chromosomes from both sides of the parents



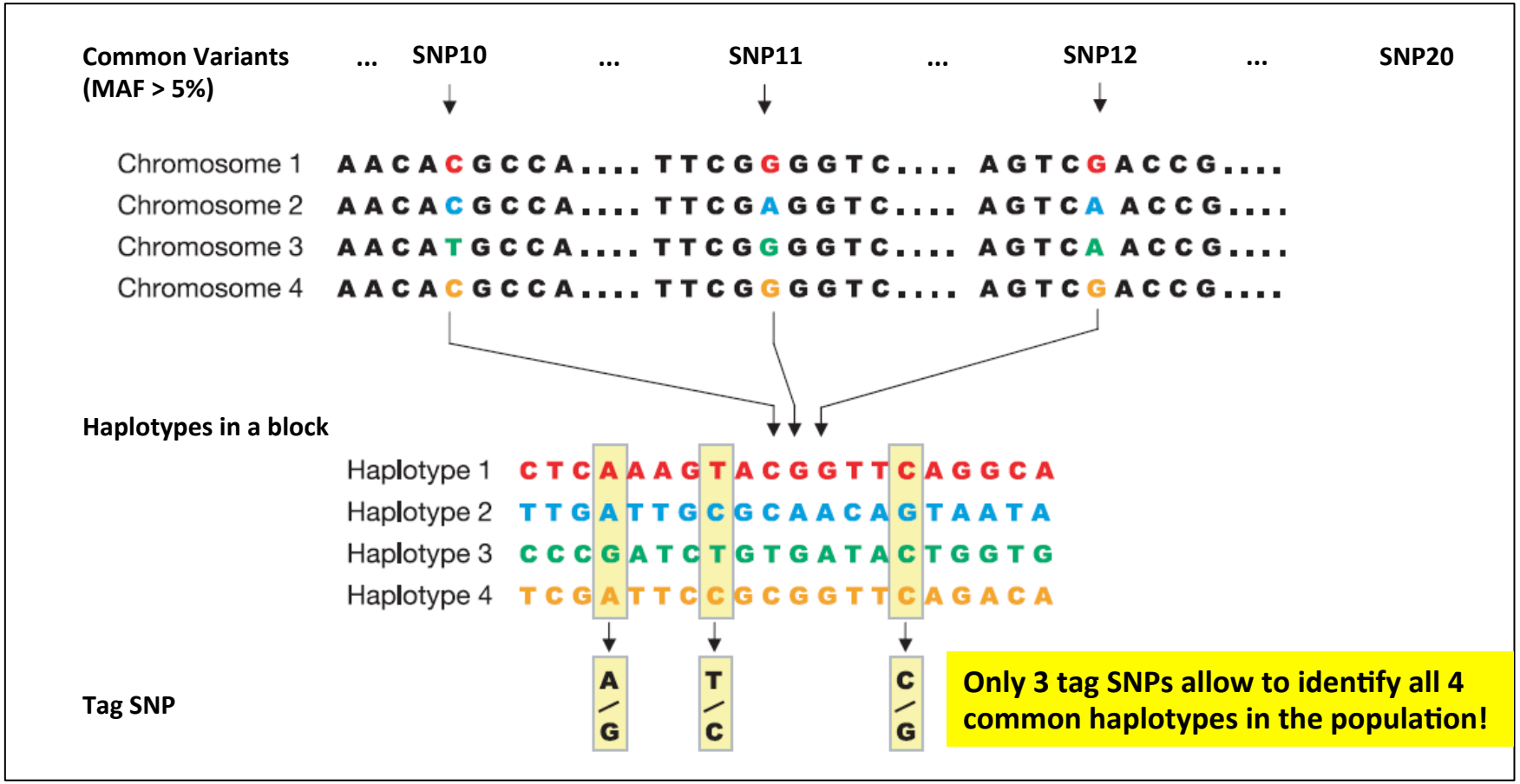


Recombination does not happen everywhere– The block structure of the human genome

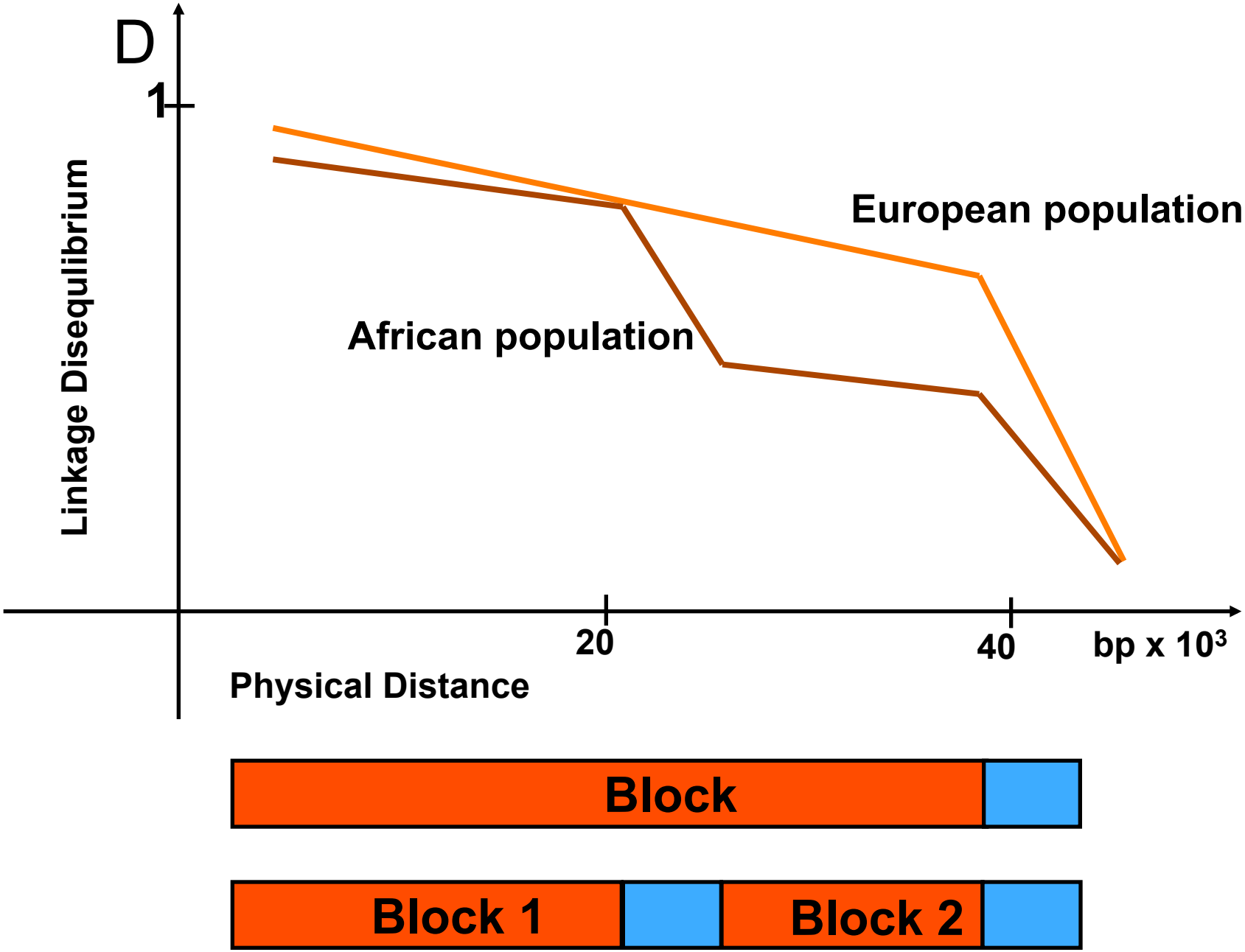


Many SNPs within a block are in “linkage disequilibrium” (LD)

Tagging (or “tag) SNPs in Haplotype Blocks

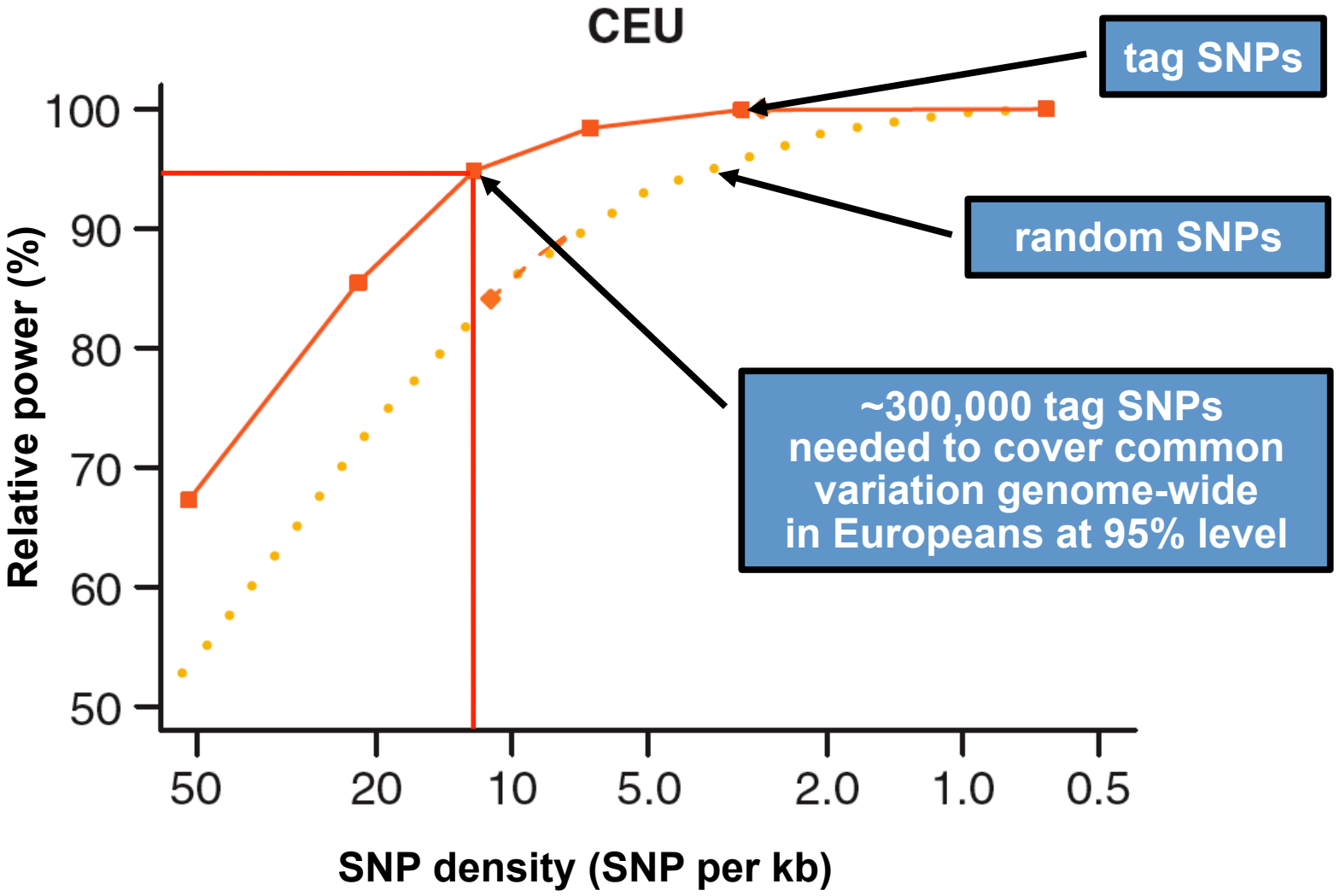


Adapted from Balding DJ (2006) *Nature Reviews Genetics*.

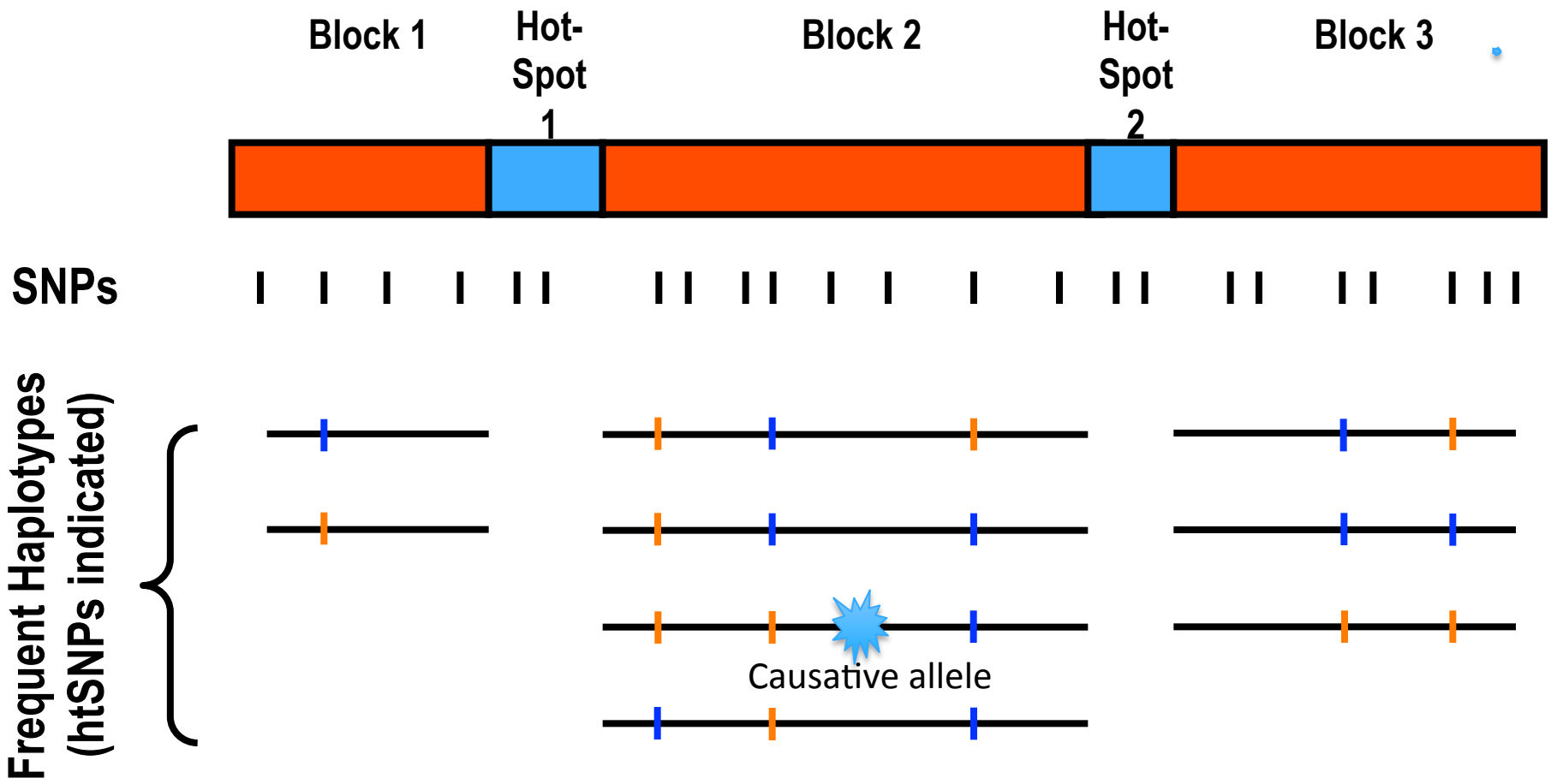

D
1
Linkage Disequilibrium
European population
African population
20
40
bp x 10³
Physical Distance
Block
Block 1
Block 2

The HapMap Project and 1000Genomes Project define patterns of genetic variation across human genomes

How many SNPs needed to cover most of the common variation?

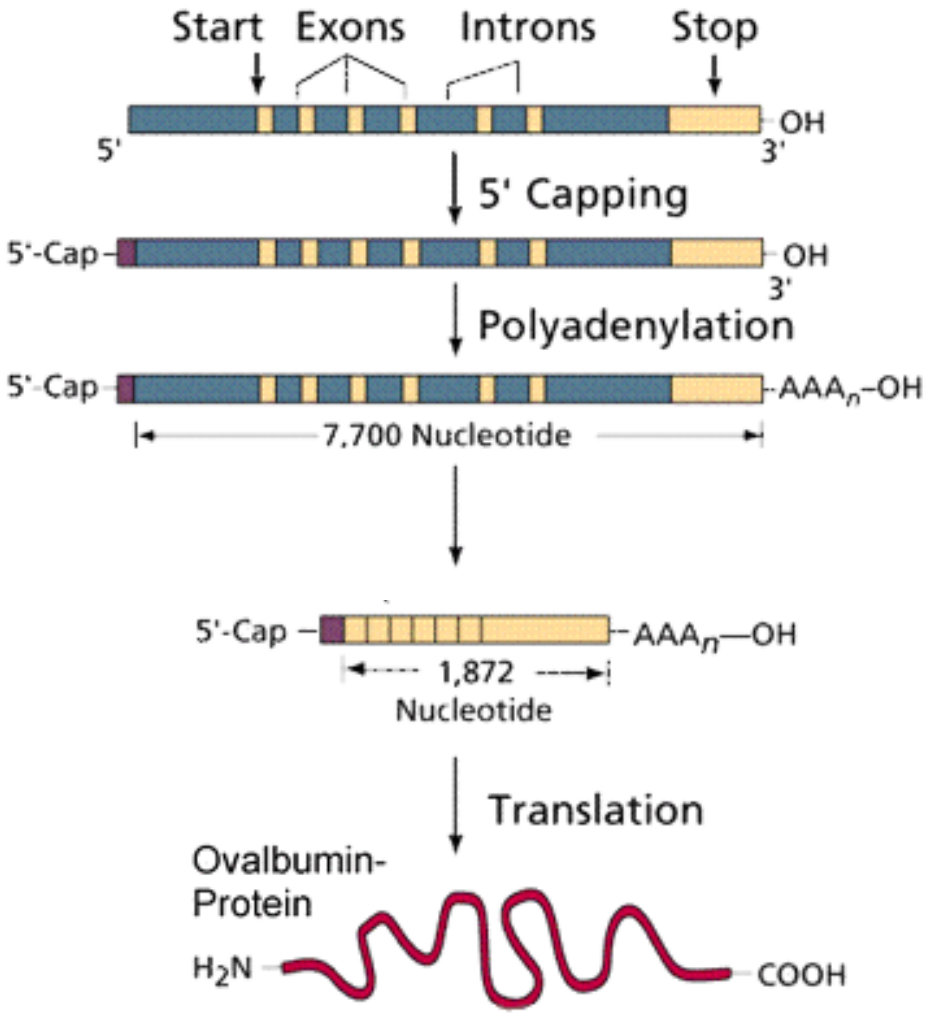


Implications of block structure and linkage disequilibrium for identifying causal alleles for a (imaging) trait



Location of the causative allele (within a gene? where in the gene?) has impact on its functional relevance

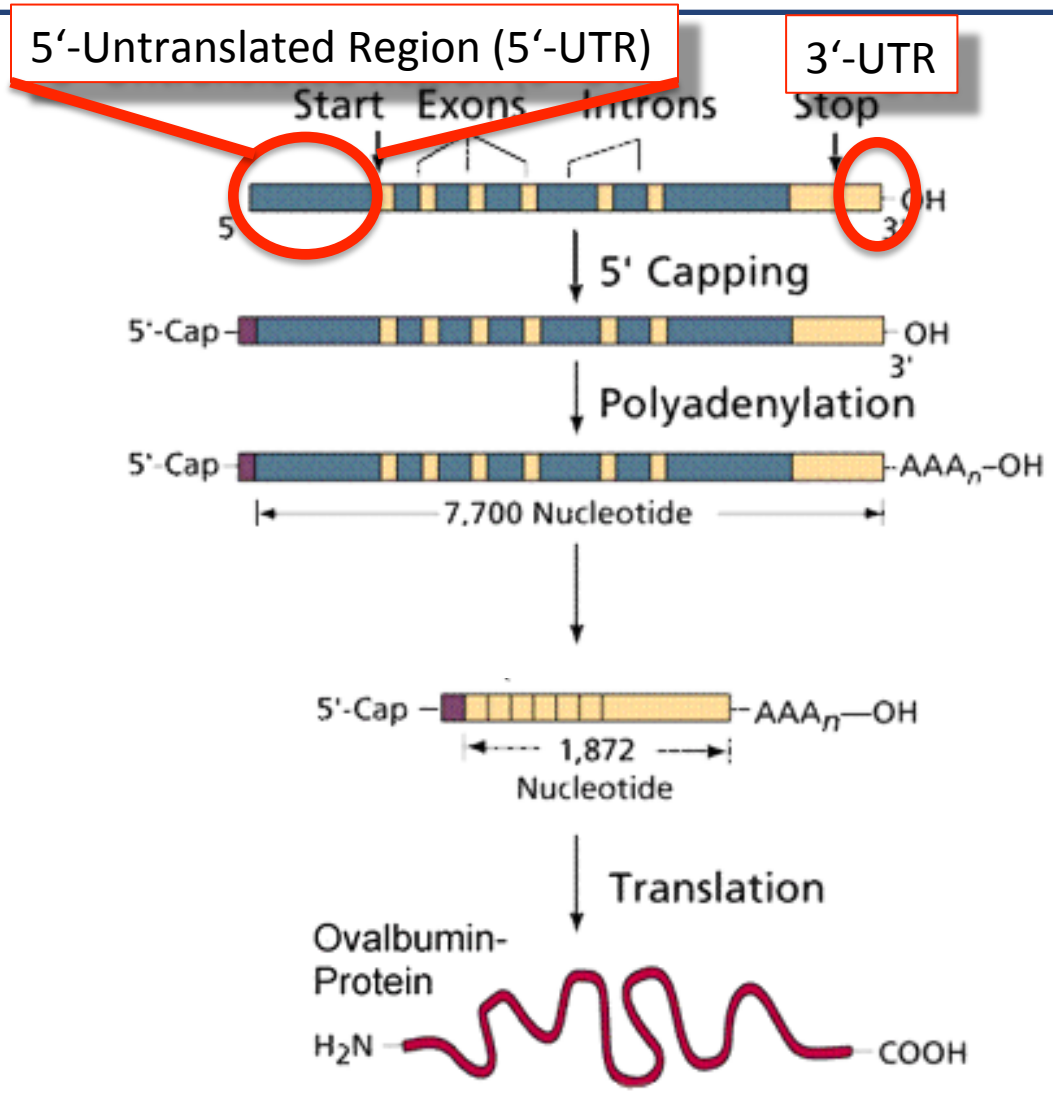
Splicing and translation of mRNA are controlled by regulatory regions



Primary transcript

Spliced transcript
= mature mRNA,
Ready for translation

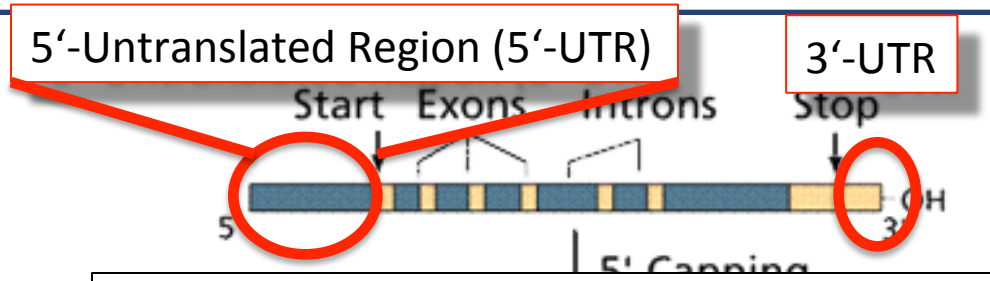
Splicing and translation of mRNA are controlled by regulatory regions



Primary transcript

Spliced transcript
= mature mRNA,
Ready for translation

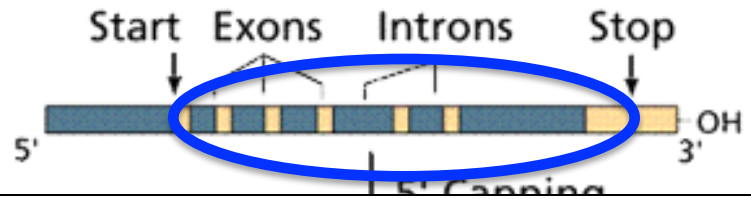
Splicing and translation of mRNA are controlled by regulatory regions



Primary transcript

- 5’-UTR and 3’-UTR important for protein translational control (influence stability of mRNA, subcellular localization, translational control)
- Not part of the protein
- SNPs or mutations in these regions may lead to reduction or acceleration of translation

Splicing and translation of mRNA are controlled by regulatory regions

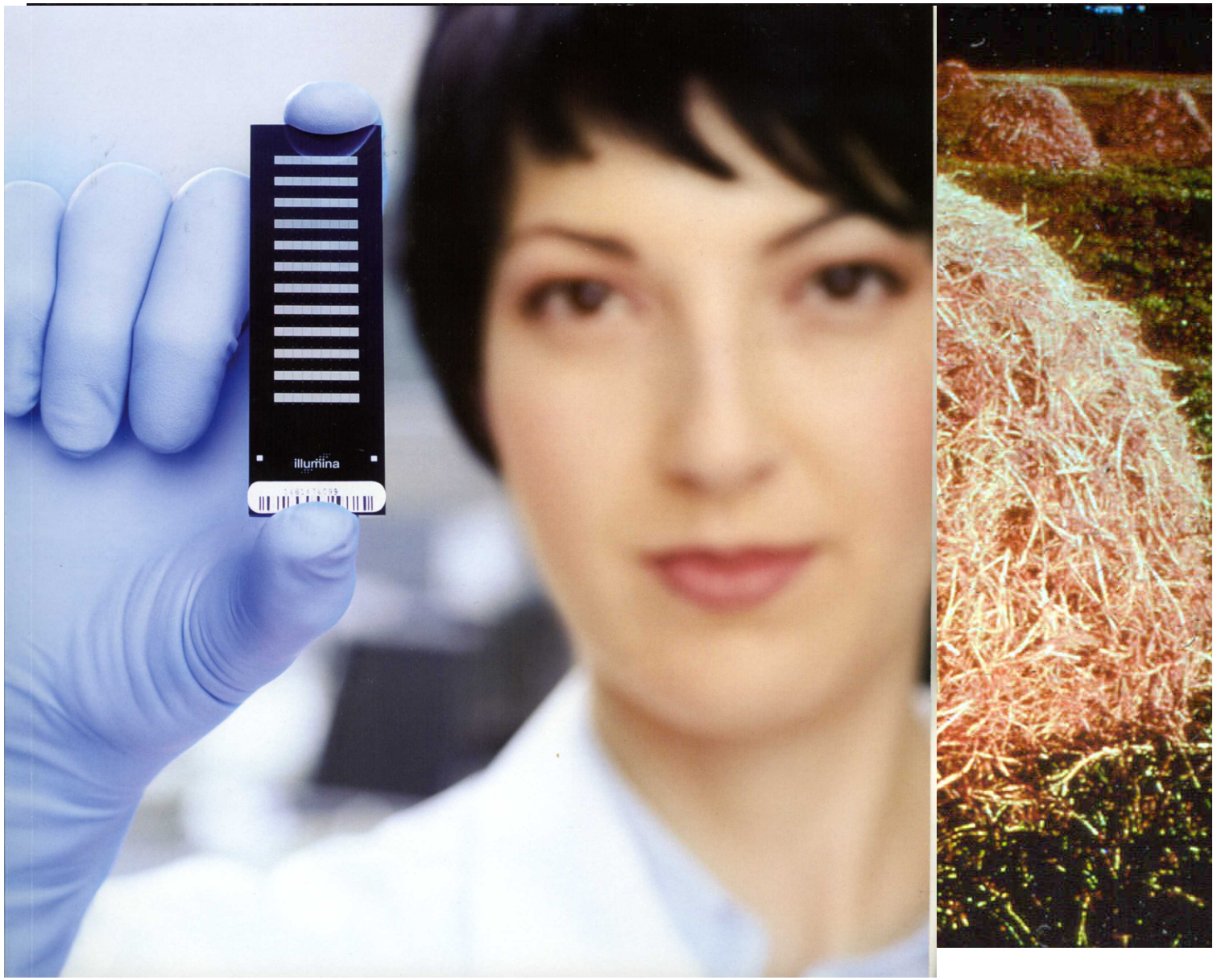


Primary transcript

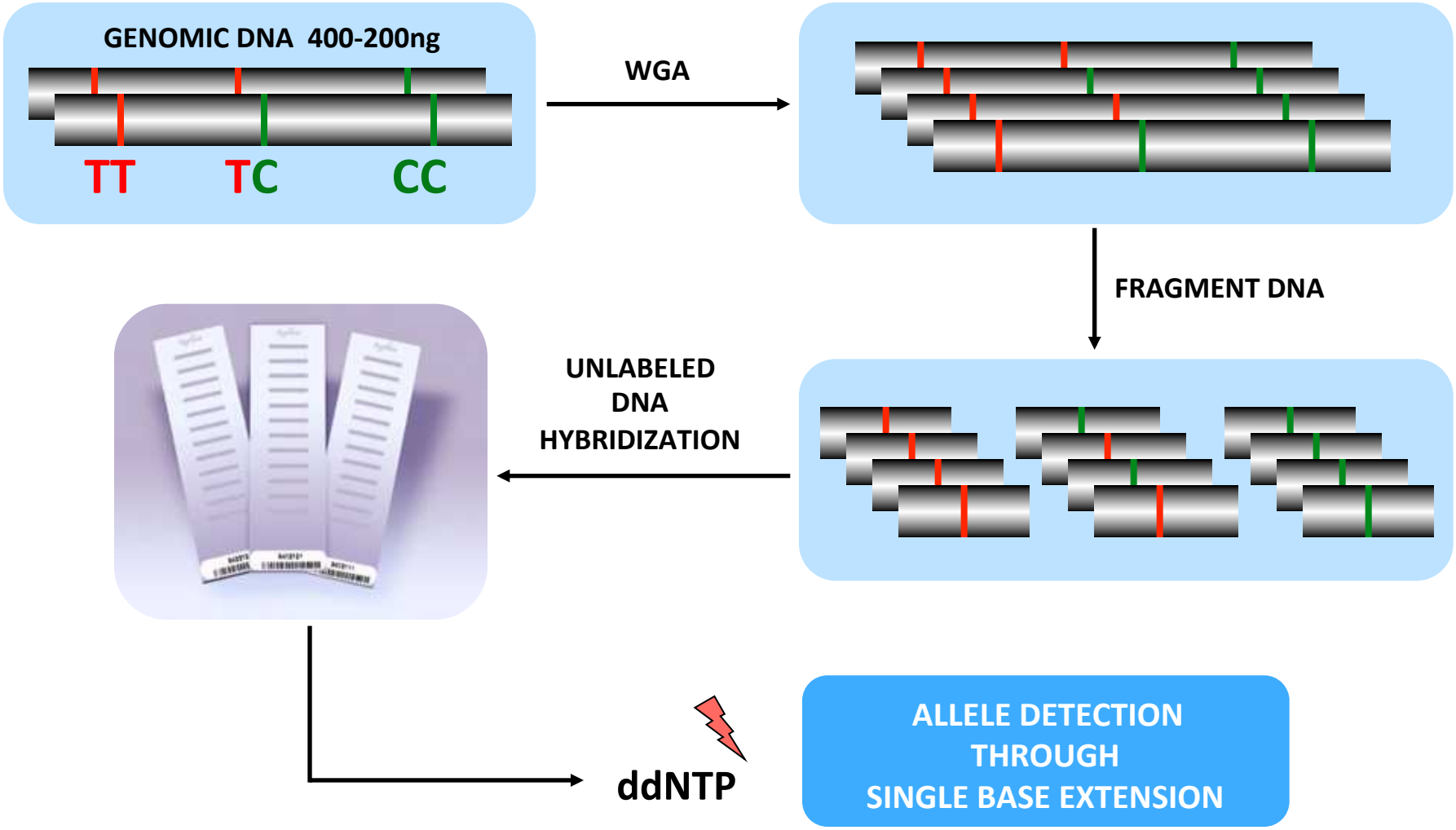
- 5'-UTR and 3'-UTR important for protein translational control (influence stability of mRNA, subcellular localization, translational control)
- Not part of the protein
- SNPs or mutations in these regions may lead to reduction or acceleration of translation
- SNPs or mutations in introns may influence gene expression or splicing
- SNPs or mutations in exons may influence protein structure/function (missense, nonsense, frameshift)

How to find the SNPs that influence your brain imaging phenotype? Screen the whole genome (GWAS)

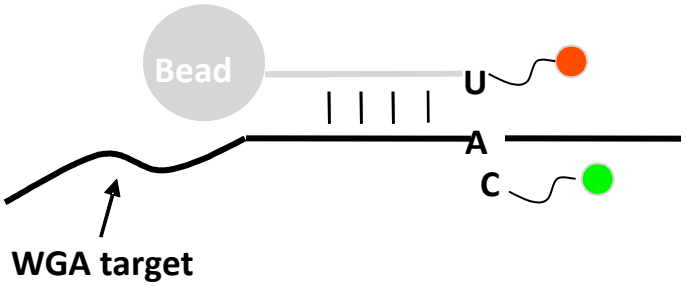
How to find the SNPs that influence your brain imaging phenotype? Screen the whole genome (GWAS) using SNP arrays

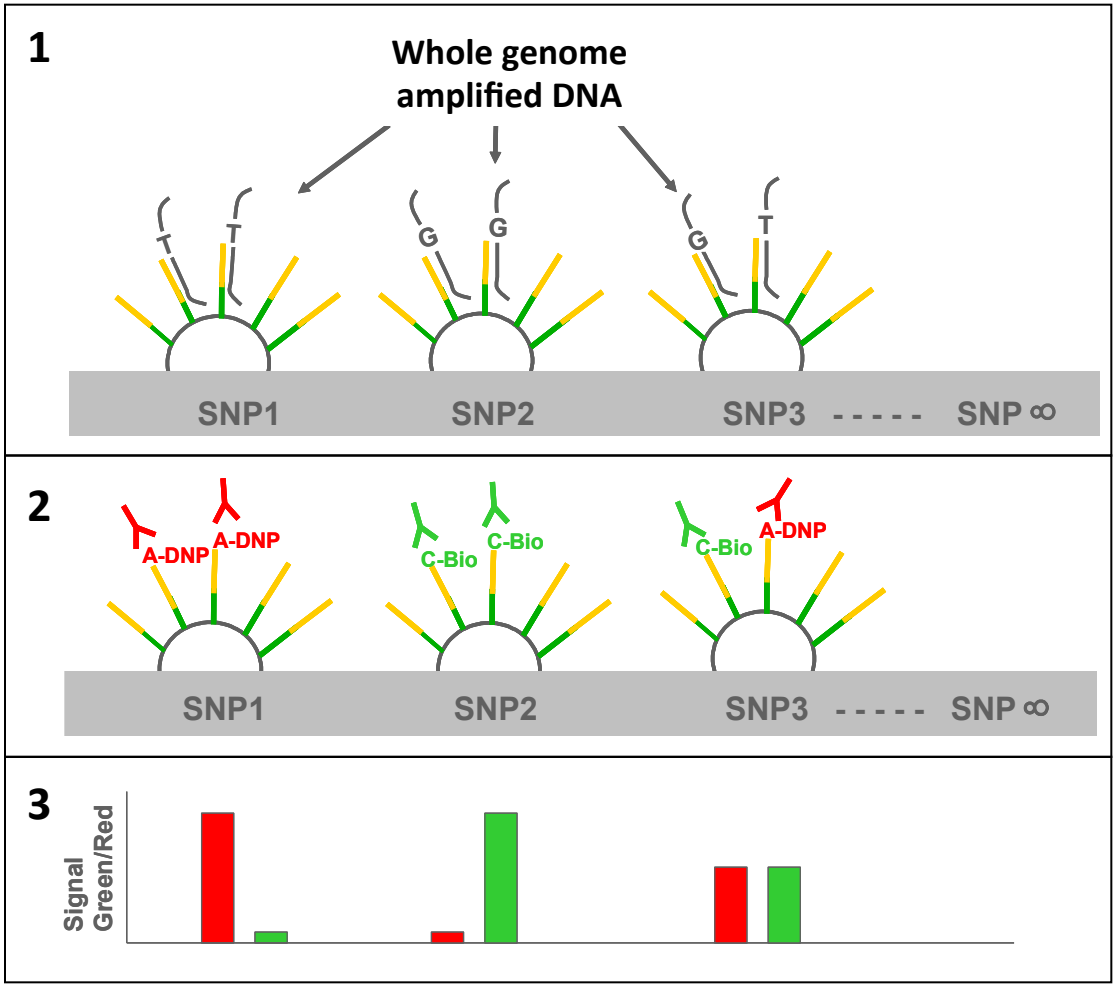


BeadArray Technology (Illumina)

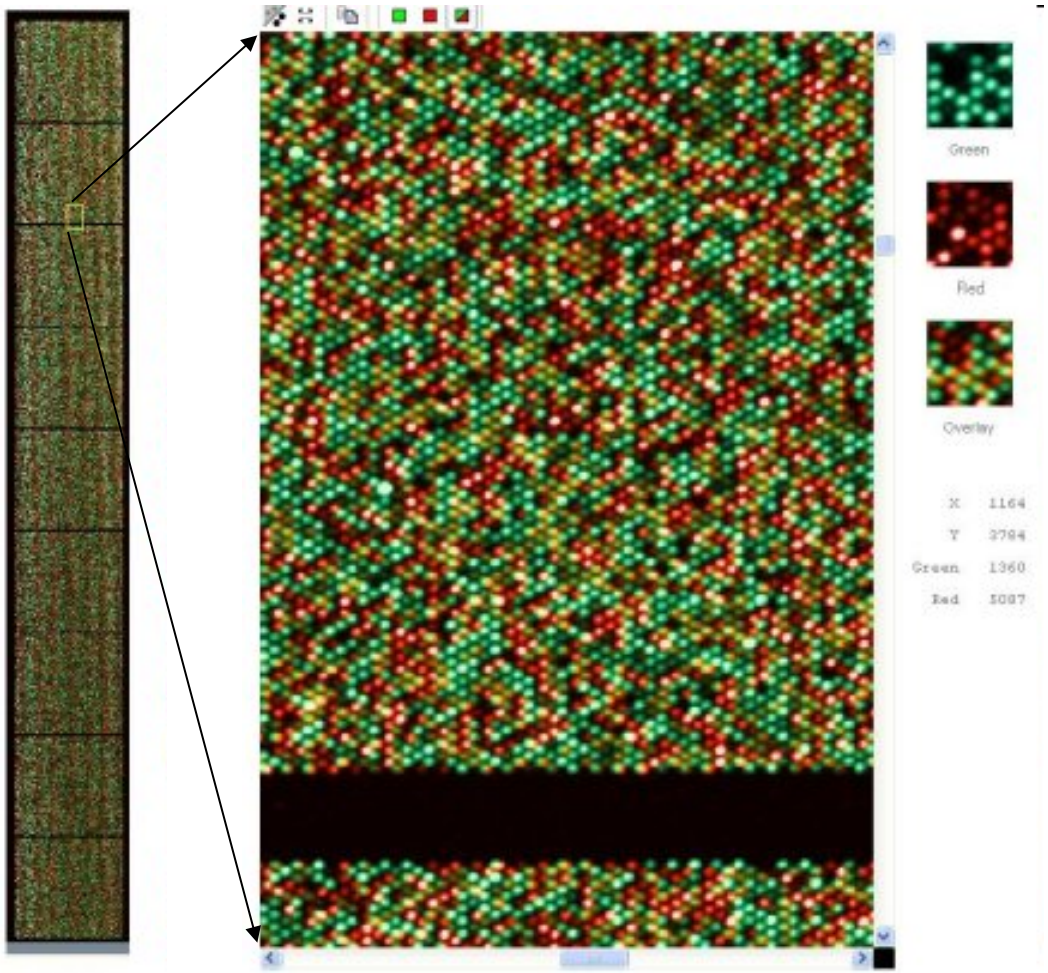


BeadArray Technology (Illumina)

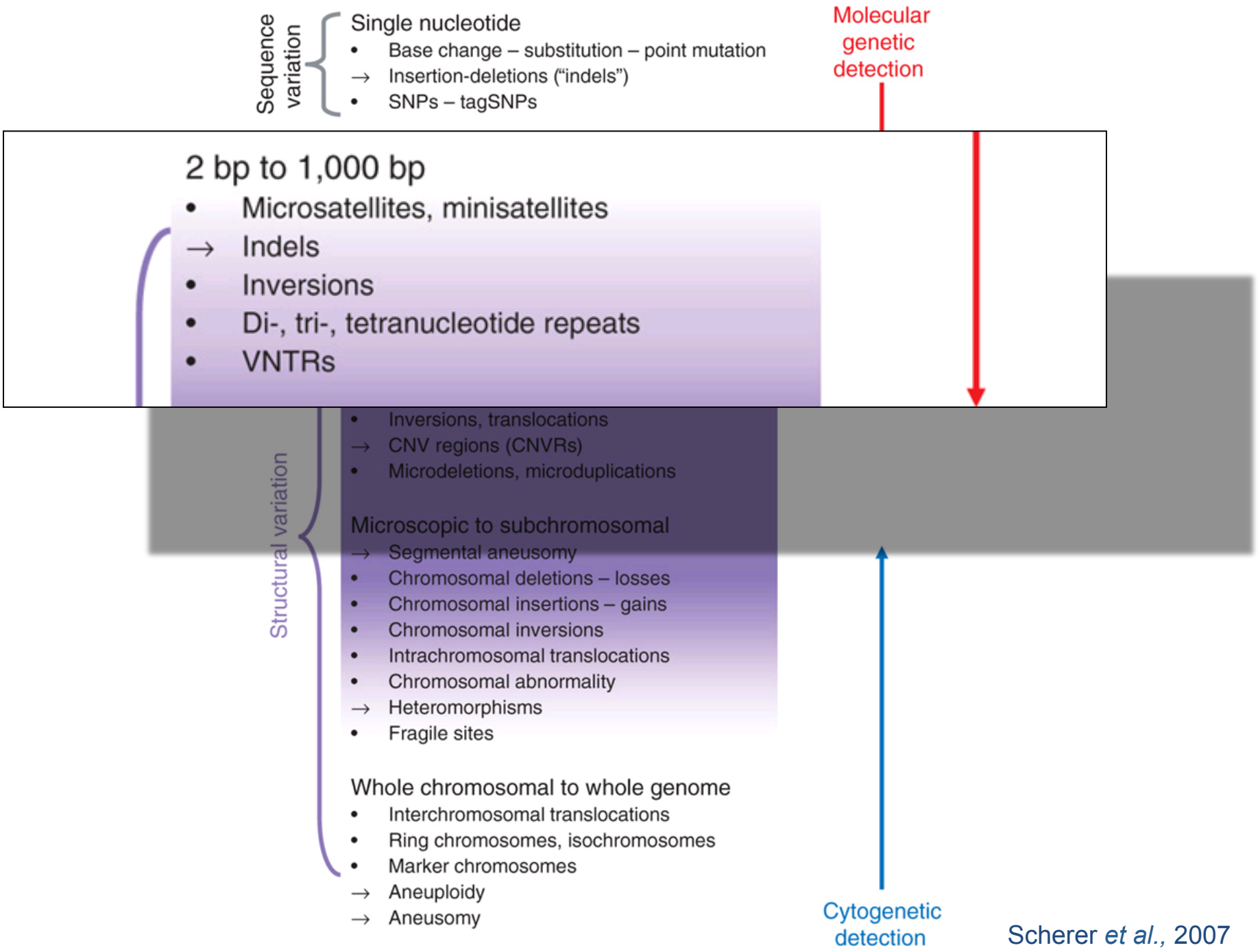




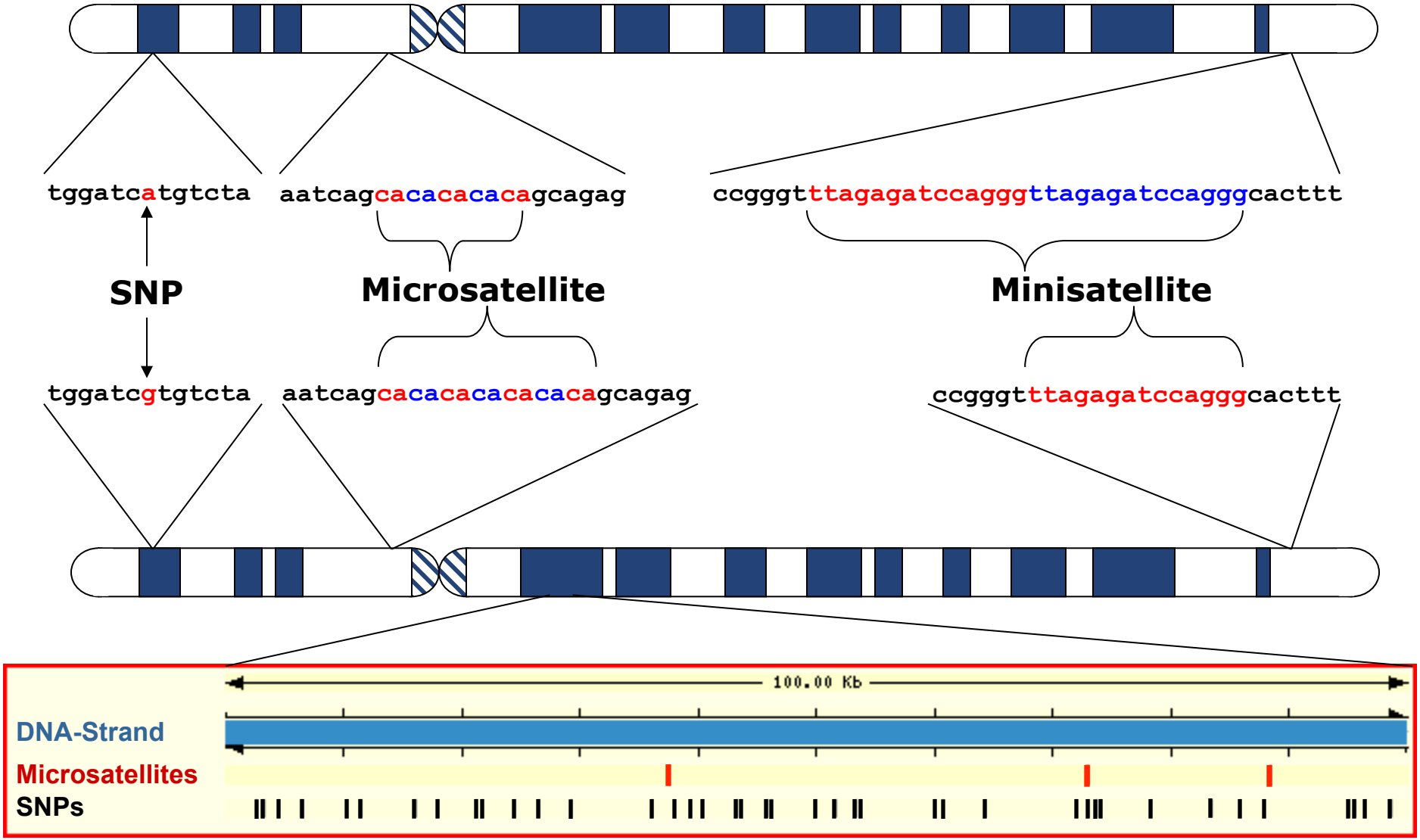
Read-out of the SNP-Arrays



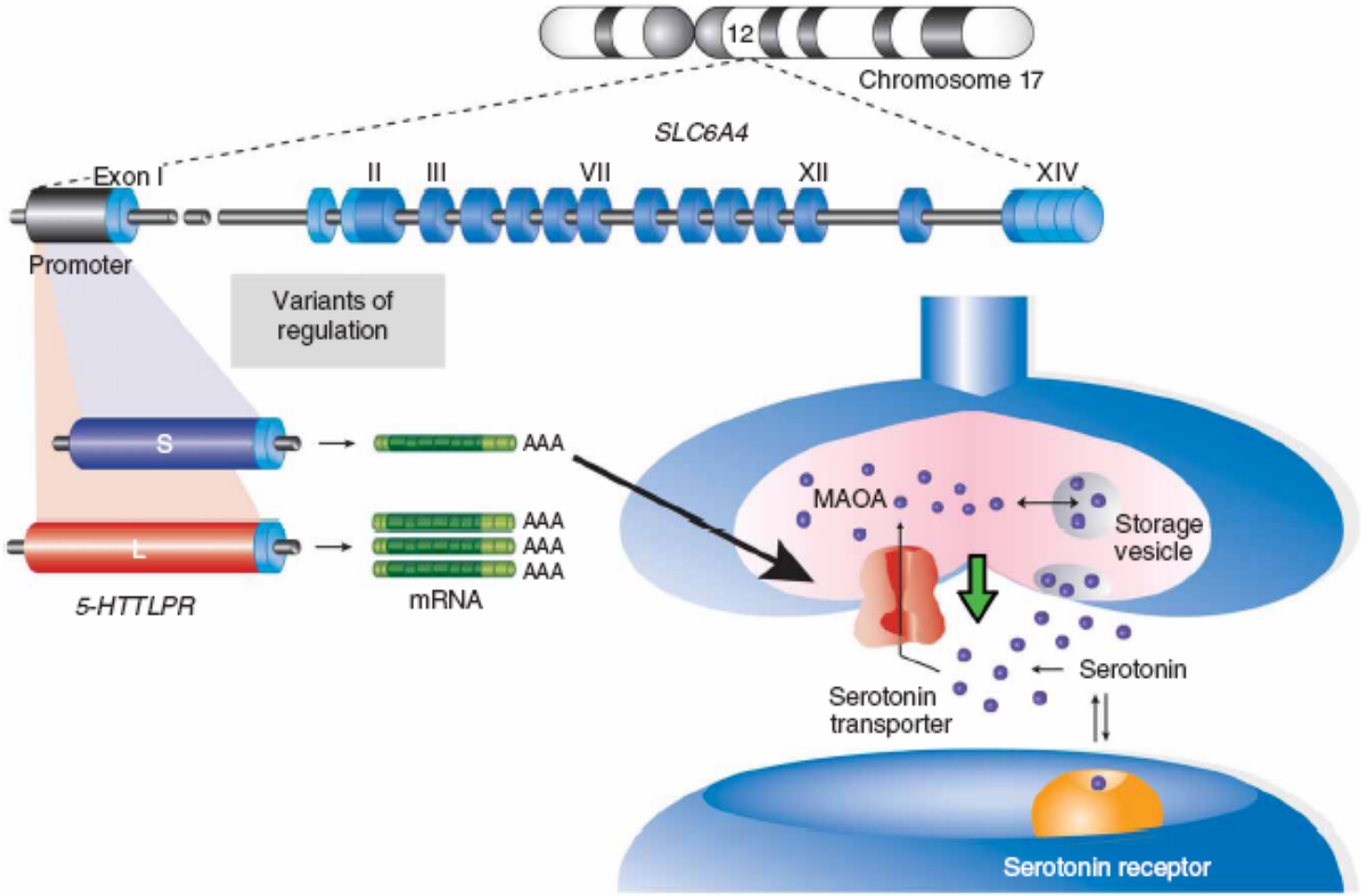
Scan of an individual‘s DNA with an array harbouring a genome wide set of 550,000 tag SNP markers (Illumina)

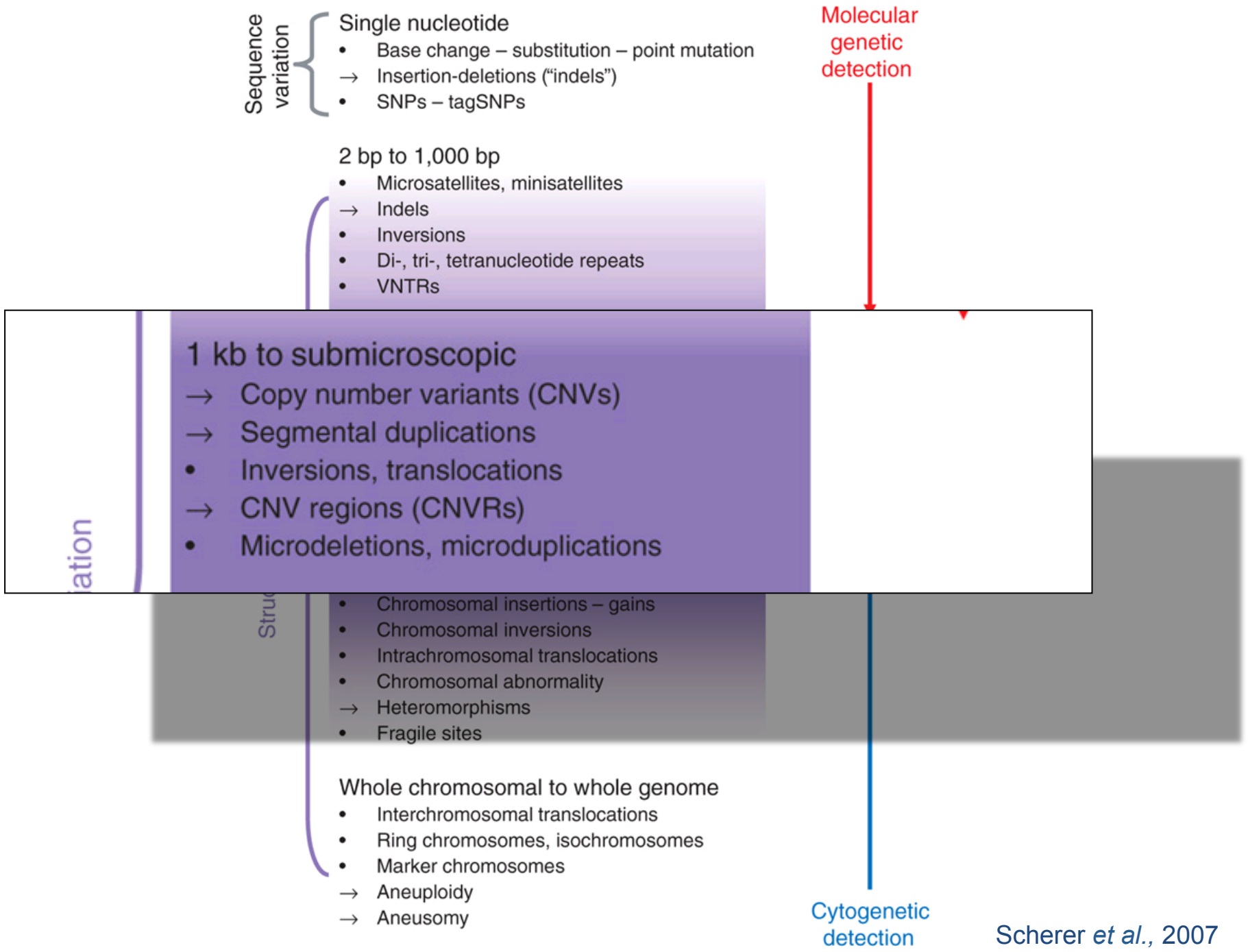


Scherer *et al.*, 2007


tggatcatgtcta
aatcagcacacacacagcagag
ccgggttagagatccagggttagagatccagggcacttt
SNP
Microsatellite
Minisatellite
tggatcgtgtcta
aatcagcacacacacacacagcagag
ccgggttagagatccagggcacttt
100.00 Kb
DNA-Strand
Microsatellites
SNPs

5-HTTLPR (serotonin-transporter-linked polymorphic region): a degenerate repeat polymorphism



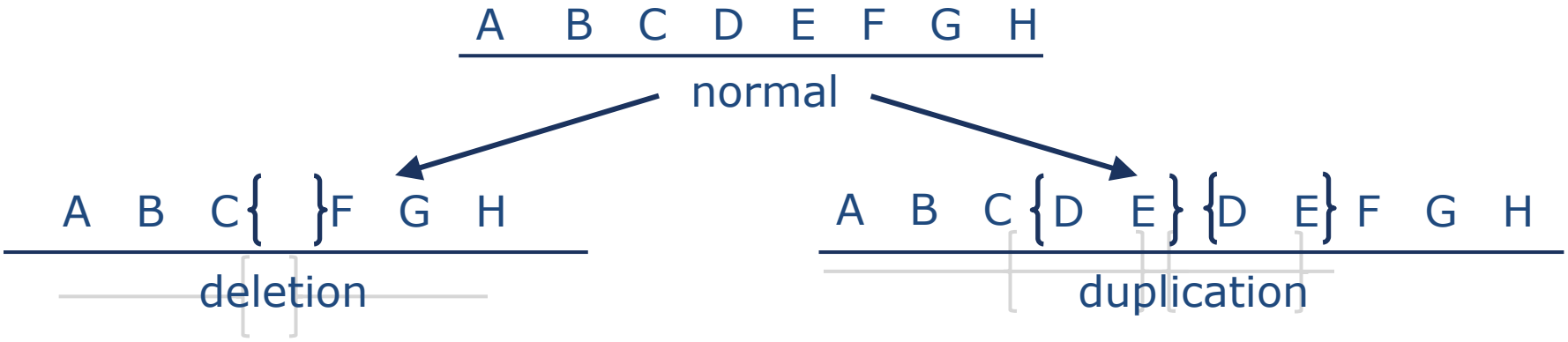


Scherer *et al.*, 2007

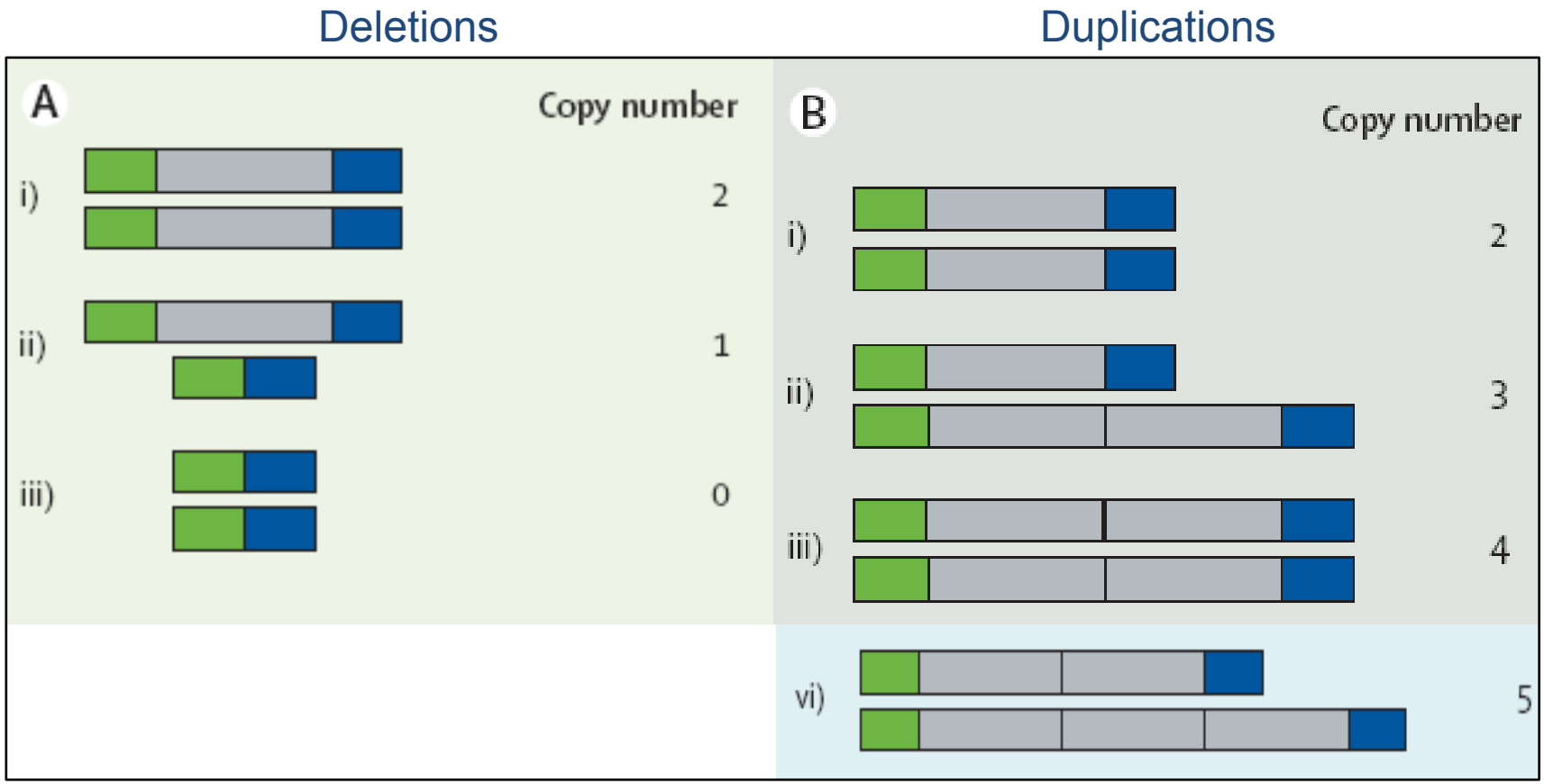
Copy number variants (CNVs) - Definition

- Variable presence or absence of a DNA sequence >1000 bp

- Different types of CNVs:
  - Deletions
  - Duplications/Triplications
  - Insertions

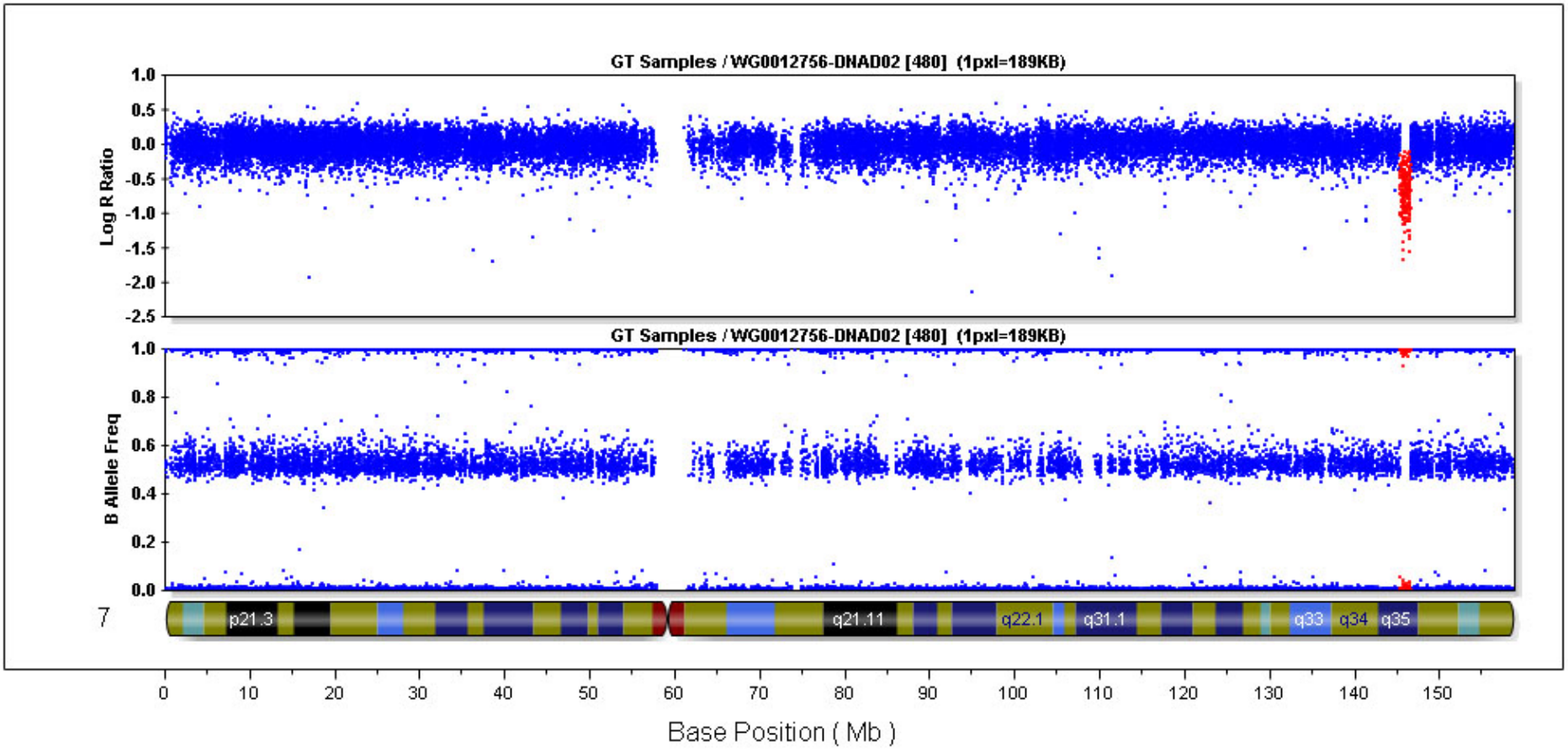


Copy number variants (CNVs) - Definition

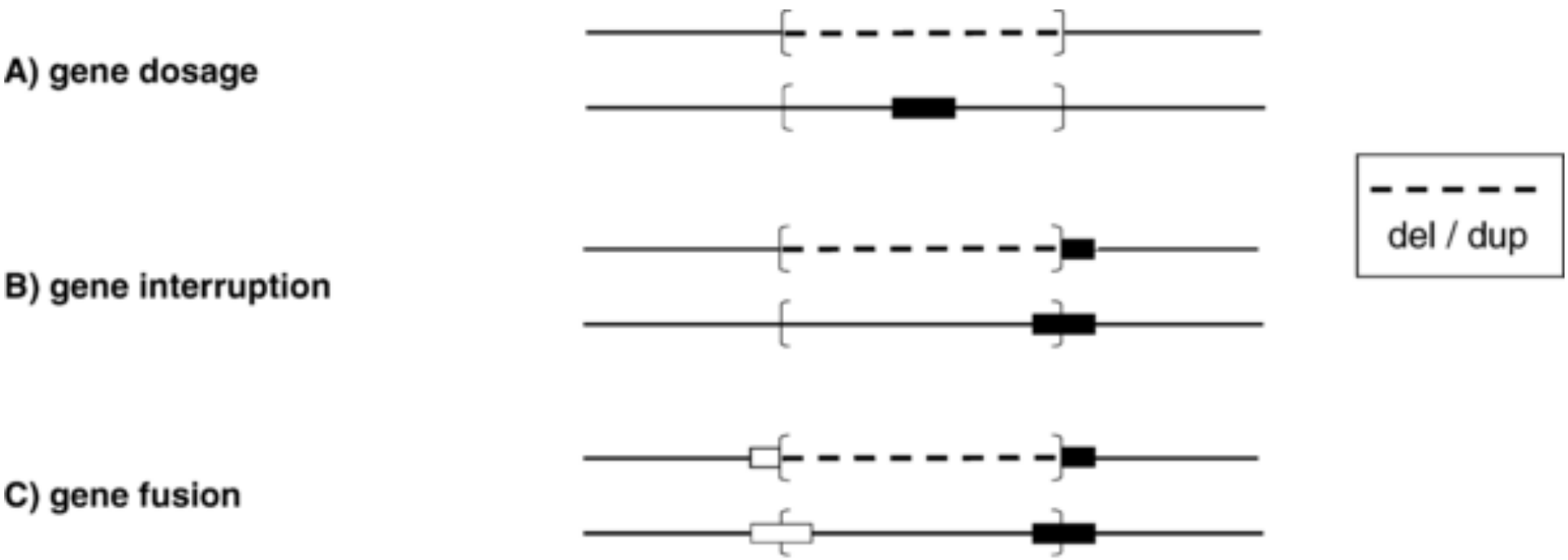


Wain *et al.*, 2009

Larger CNVs (>50-100 kb) can be detected on SNP arrays

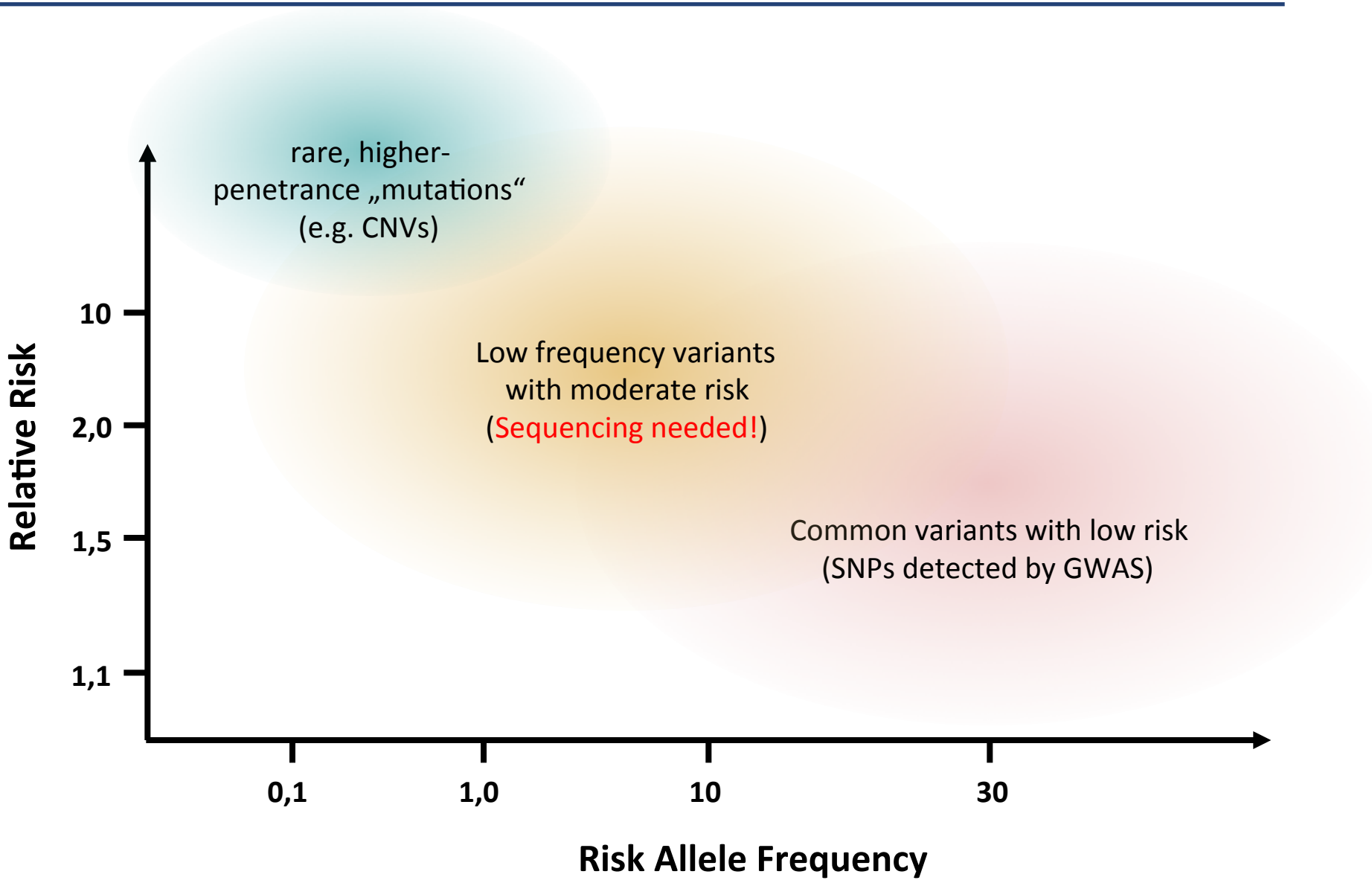


Molecular mechansims leading to CNV phenotype

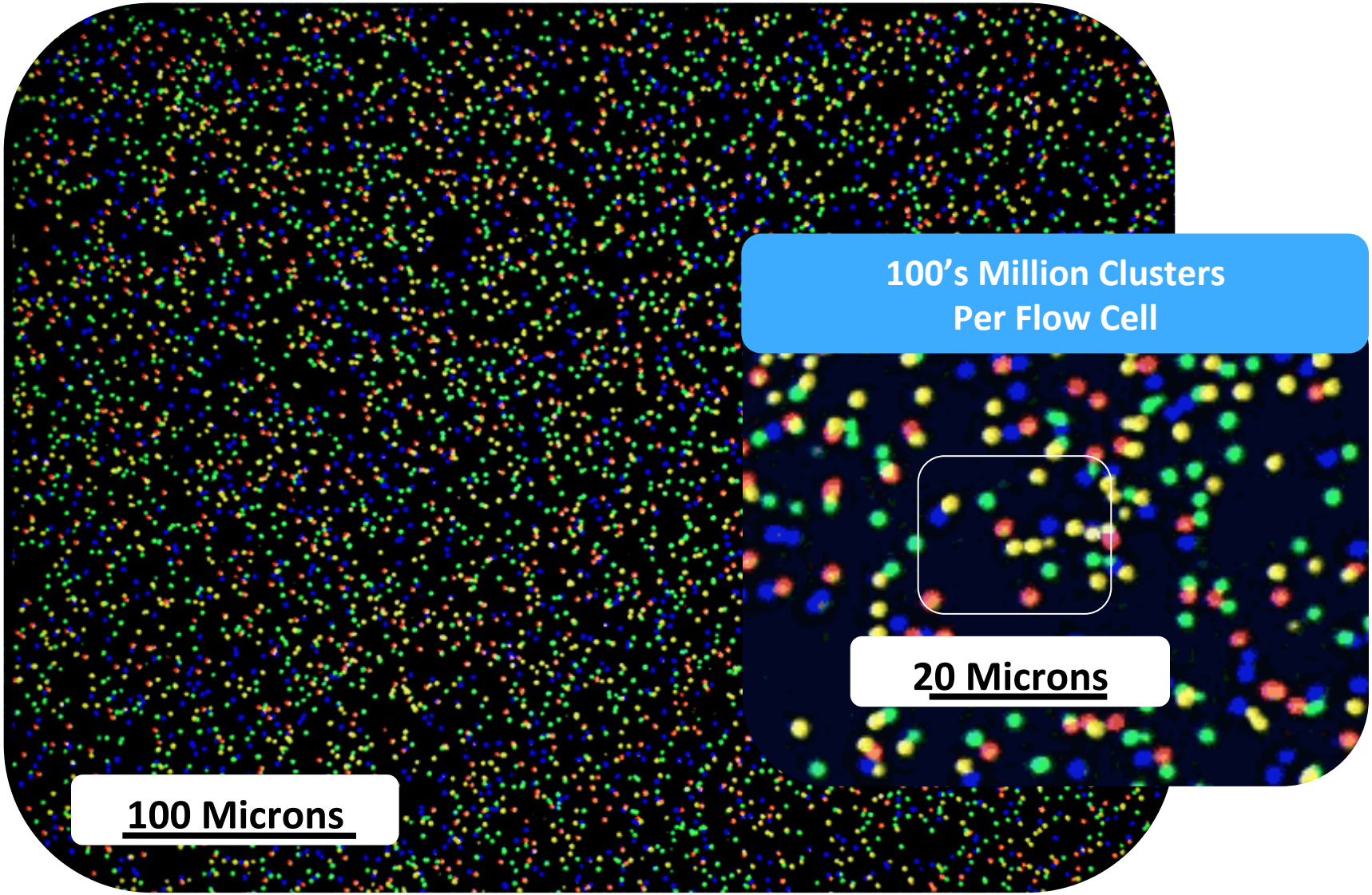


Lupski & Stankiewicz, 2005

Most imaging phenotypes will be explained by a spectrum of common and rare functional alleles



New sequencing technologies may help to identify variants relevant for imaging phenotypes not detected so far



Sequencing whole exomes identifies a lot of „neutral“ background variation – how to find the phenotype-relevant variants?

- ~20,000 DNA variants in/near protein coding DNA
- ~200 rare missense variants
- ~100 loss-of-function variants (~20 rare or private)

A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes

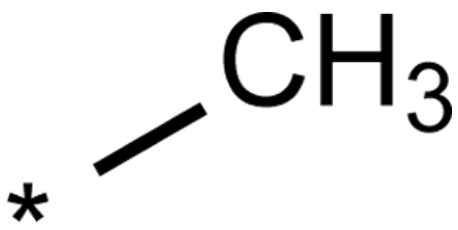
Daniel G. MacArthur,[1,2*] Suganthi Balasubramanian,[3,4] Adam Frankish,[1] Ni Huang,[1] James Morris,[1] Klaudia Walter,[1] Luke Jostins,[1] Lukas Habegger,[3,4] Joseph K. Pickrell,[5] Stephen B. Montgomery,[6,7] Cornelis A. Albers,[1,8] Zhengdong D. Zhang,[9] Donald F. Conrad,[10] Gerton Lunter,[11] Hancheng Zheng,[12] Qasim Ayub,[1] Mark A. DePristo,[13] Eric Banks,[13]

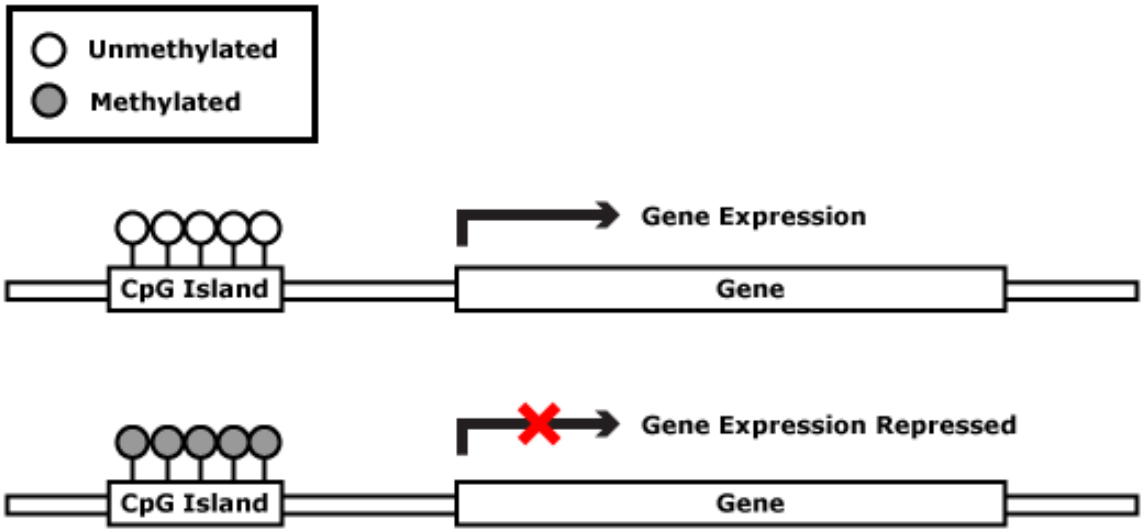
A few things are a bit more interpretable (obvious functionality), but not absolute proof in each case…

- ~1 *de novo* variant per exome (only ~5% LoF)
- <5% chance that an individual has a complete knockout of a single well-preserved* gene anywhere in the genome

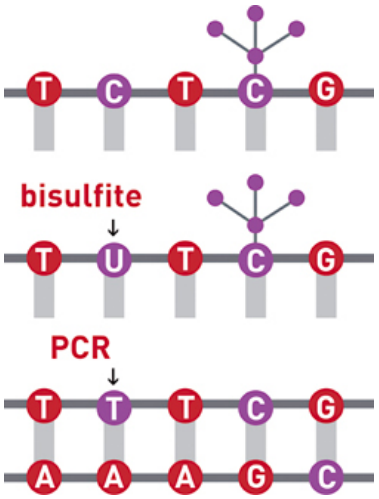
* well-preserved = 98-99% of genes without a common LoF mutation

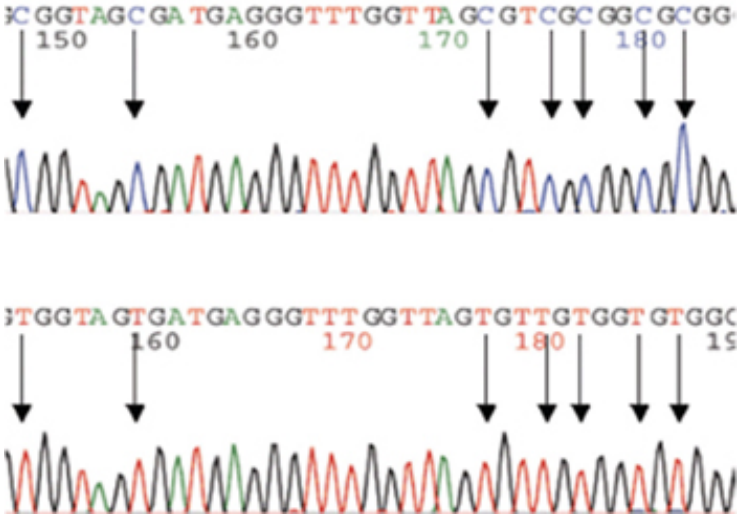
Functional annotation for variants is crucial:
influence on methylation of DNA (methylation quantitative trait loci – mQTL)





Methylated and unmethylated “C” bases can be distinguished after bisulfate sequencing





Summary

- The genome is highly variable.
- Especially common SNP variants as well as large structural variants (CNVs) can be tested using array-based technologies.
- The field is moving to whole-genome sequencing which allows also detection of rare SNPs and small CNVs/InDels
- Functional annotation of identified genetic variants that might play a role in brain phenotypes is of great importance:

  influence on gene regulation (incl. methylation), protein function