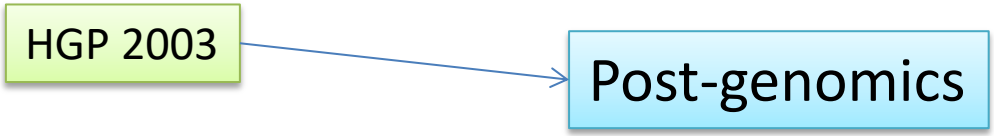
L1.2

today’s sequencing

Human genome variation



Genetics

Comparative (phylogenetic conservation indicates conserved function)

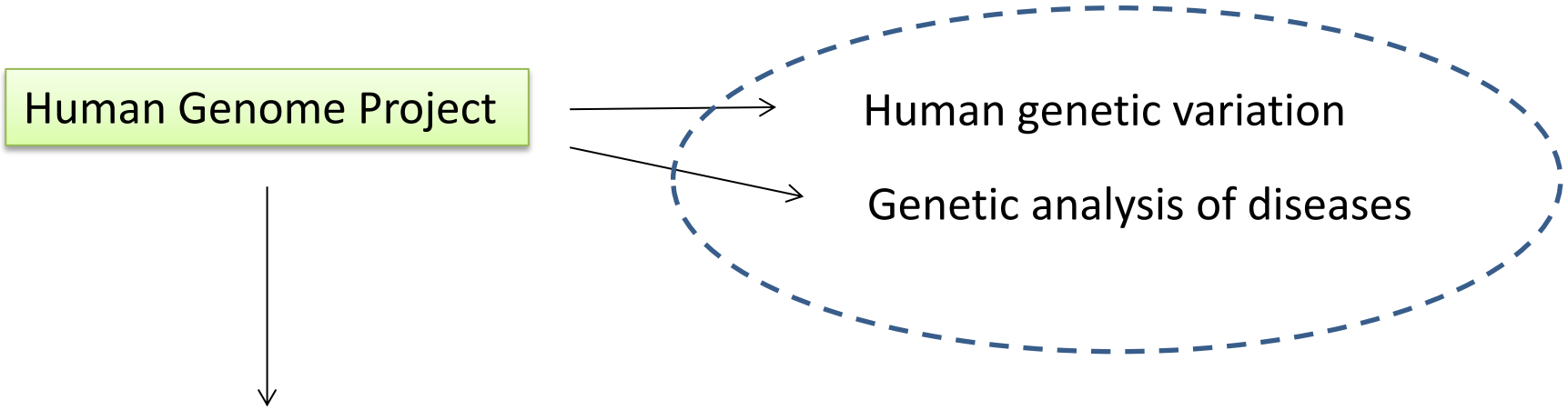
Human Genetic Variation (1000 Human Genomes - HapMap)

GWAS – Genome variations – phenotype correlation

Gene expression and phenotype

Functional Genomics (ENCODE – FANTOM)

Epigenomics:	CpG methylation
	Histone modifications (PTMs)
	Chromatin status
	Protein-DNA mapping (e.g. transcription factors)
Transcriptomics:	Coding and noncoding RNAs



Functional annotation of the Human Genome (and others, e.g. mouse)

The Encyclopedia of DNA Elements (ENCODE) + other similar projects (e.g. FANTOM)

The idea was to obtain functional information for every single nucleotide of the human genome

Started in 2000 using automated Sanger sequencing on 1% human genome (ca. 30 Mb), completed in 2006

With the advent of Next Generation Sequencing Technology, first draft completed in 2012

Genetics

Individual genomes display **variants**

SNP – single nucleotide polymorphisms

Indels – insertions and deletions

CNV – copy number variations

TE – transposable elements number and position

Variants are associated to more or less evident **phenotypes**

Some variants are clearly associated to specific **pathologies**.

Other variants are associated only weakly with a phenotype but require other variants (often in other loci) to become significantly associated (combinatorial association).

Projects are under way to describe all variants associated to risk of disease (GWAS: Genome Wide Association Studies)

SPECIAL FEATURE | COMMENTARY

Next-generation sequencing transforms today's biology

Stephan C Schuster

A new generation of non-Sanger-based sequencing technologies has delivered on its promise of sequencing DNA at unprecedented speed, thereby enabling impressive scientific achievements and novel biological applications. However, before stepping into the limelight, next-generation sequencing had to overcome the inertia of a field that relied on Sanger-sequencing for 30 years.

NGS

Fragment the DNA (or RNA) to be sequenced in smaller pieces

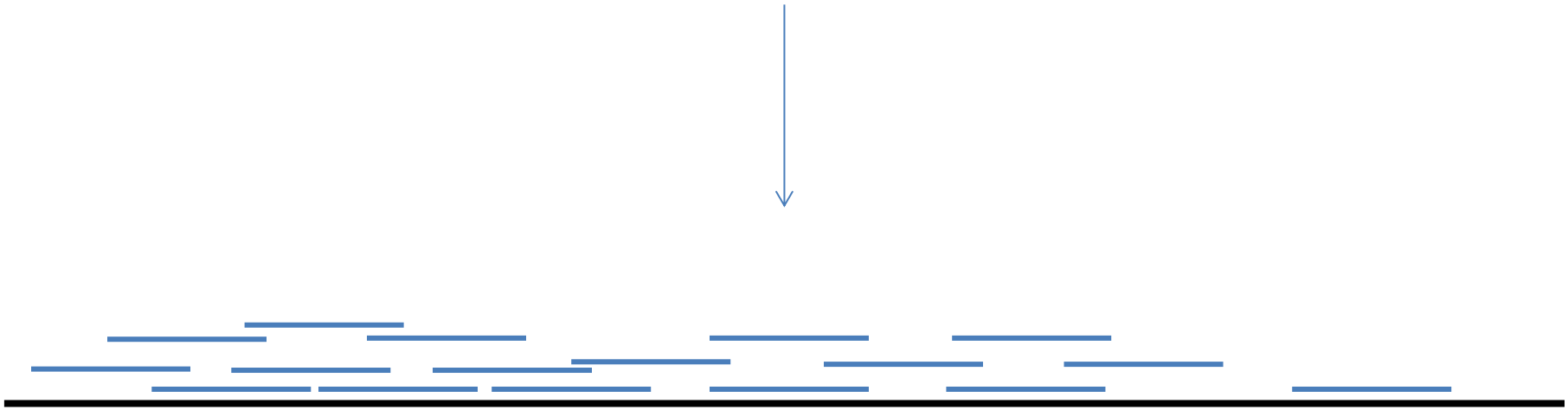
Physically separate the fragments

Highly-parallel sequencing of fragments, high-throughput

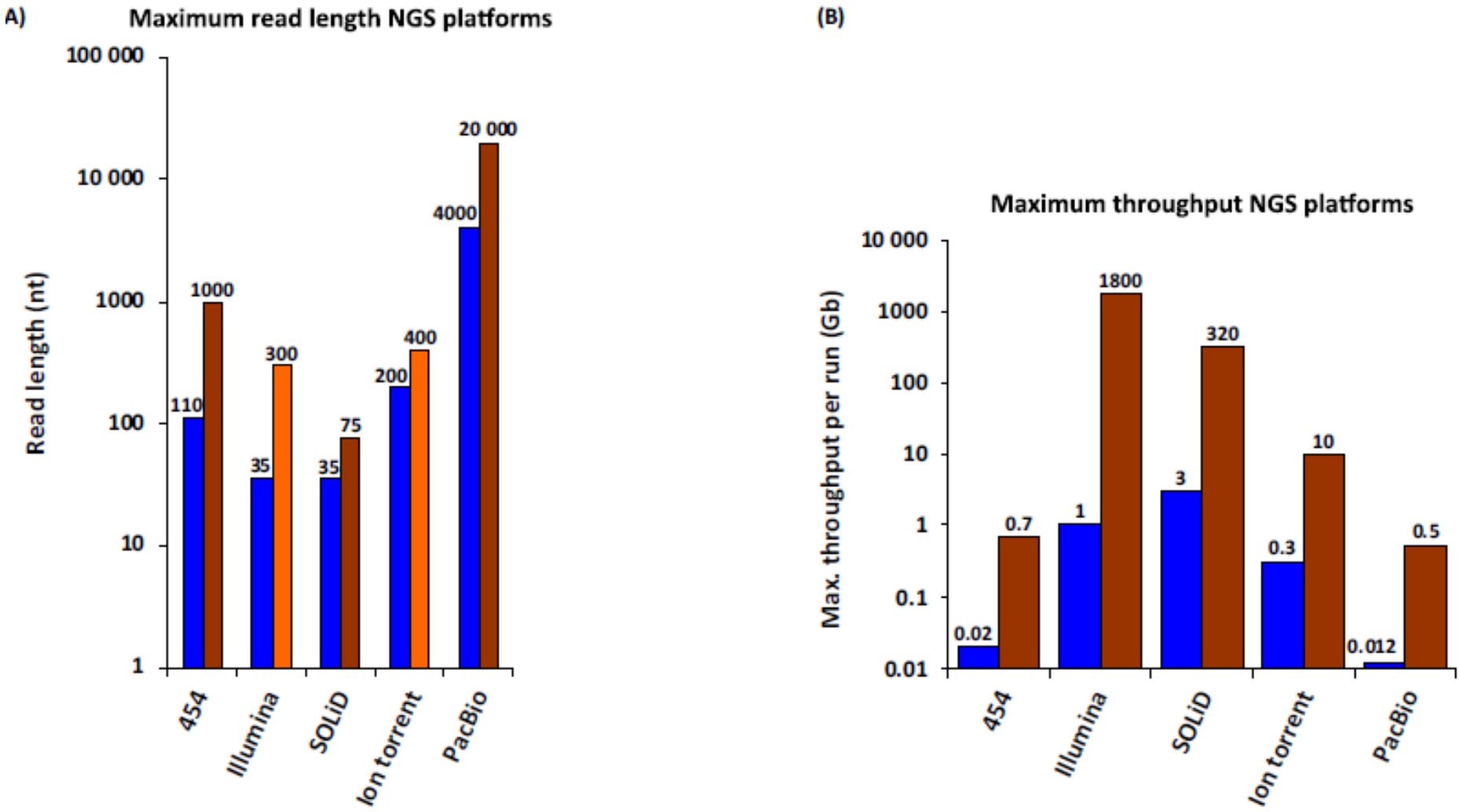
No cloning step required

NGS sequencing produces hundreds of millions of short «reads» per run

Reads are mapped to the reference genome



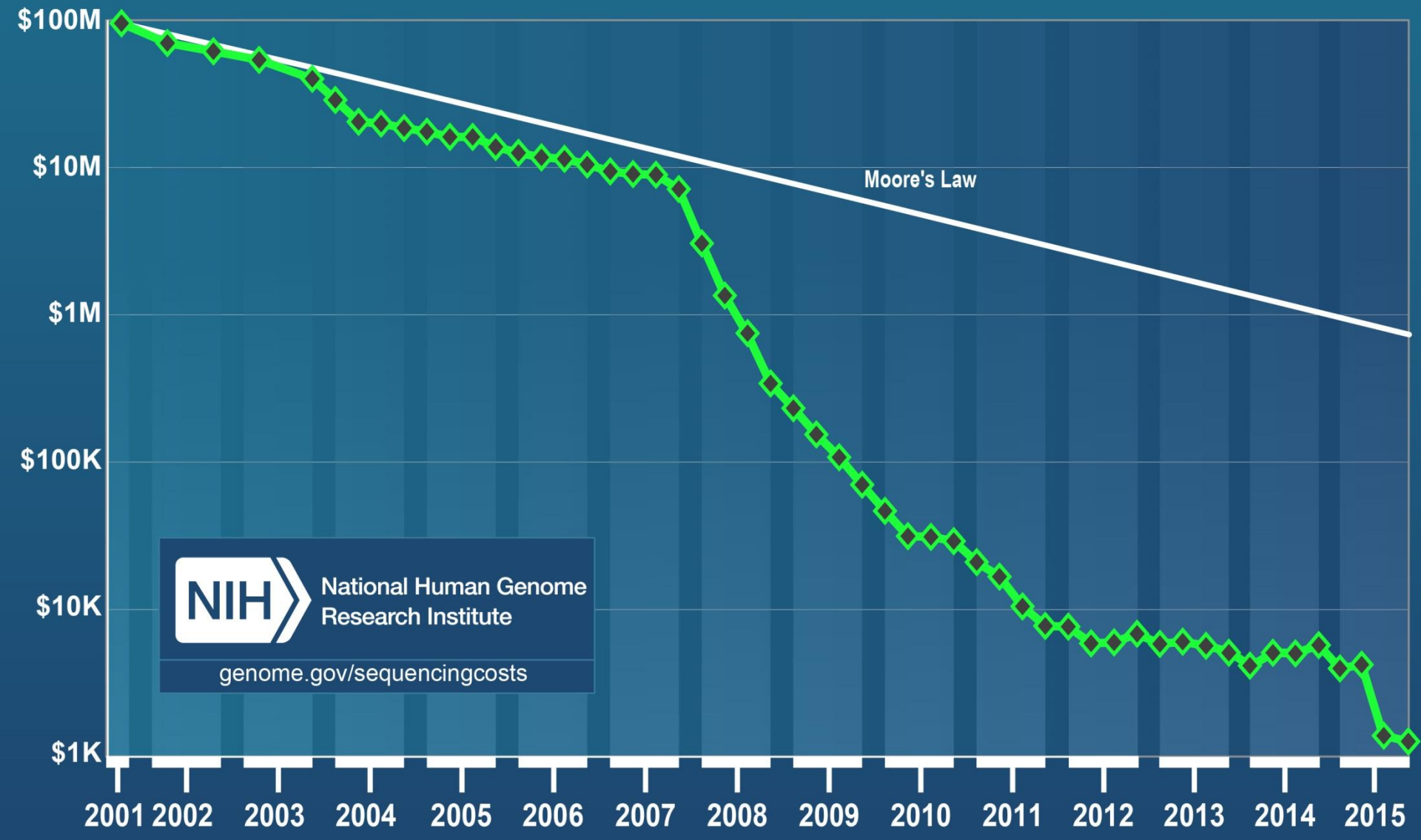
reference genome



In blue the first version of the instruments

From Van Dijk et al., 2014 (Textbook)

Cost per Genome



1000 Human Genomes, HapMap project
Describing variations among genomes of individuals

GWAS
Genome-wide association studies
Variations (SNPs, CNV, indels) studied in individuals as related to the occurence of a phenotype (pathology, risks, other features)

TCGA – The Cance Genome Atlas
Sequencing of tumor cell DNA to evidence mutations occurring in tumors.

The 1000 Genomes Project

http://www.internationalgenome.org/

Started immediately after the HGP but it was dramatically accelerated by introduction of NGS

2010 report

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

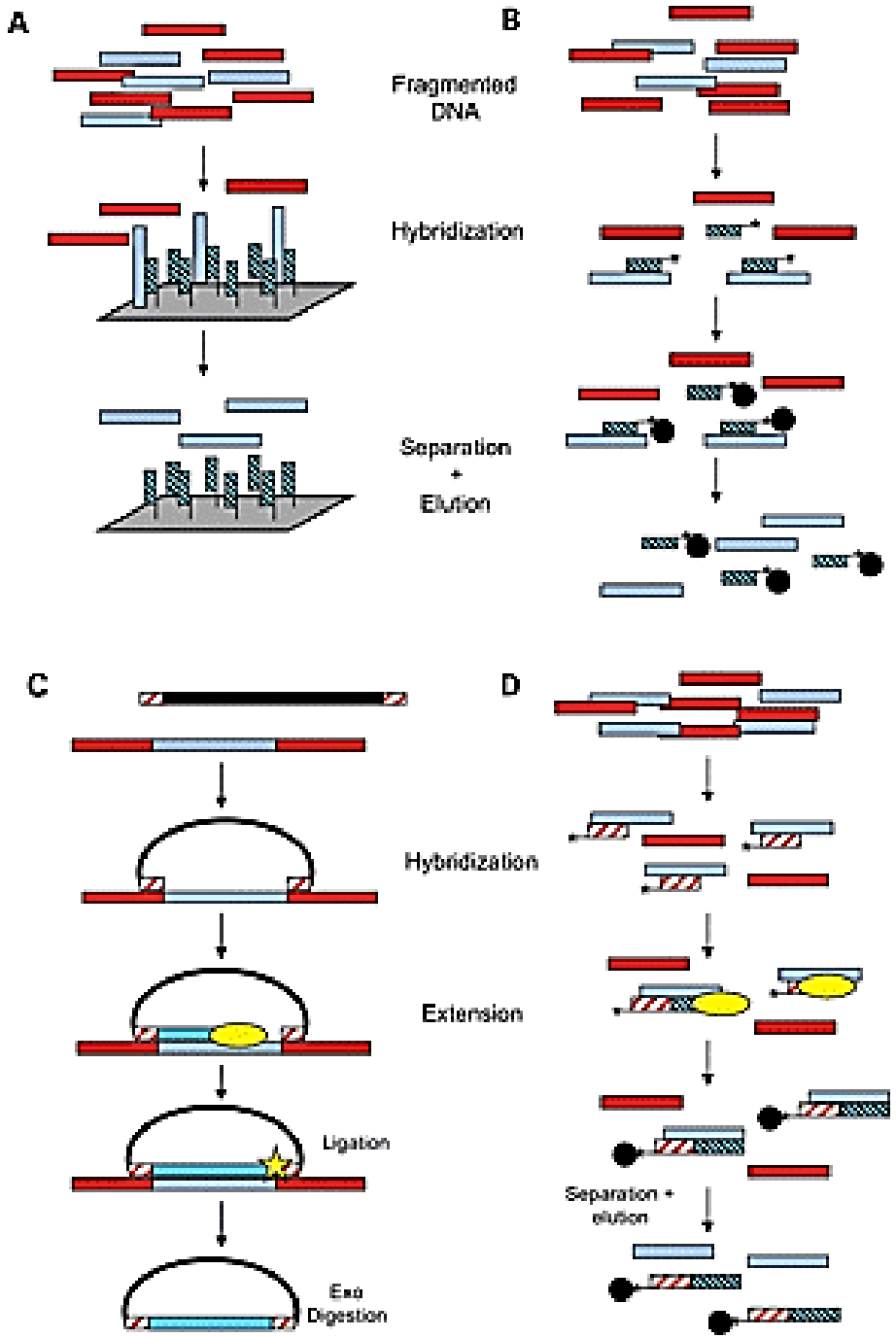
The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

Exome sequencing

Due to elevated costs, many studies were limited to the «**exome**»
Exome is the set of sequences that make up all known mRNAs.

Requires enrichment of exon sequences from a genomic DNA. This is obtained using different methods, as exemplified in these schemes.

From: Teer and Mullikin, 2010. Hum Mol Genet. 9(R2):R145-51



ARTICLE

OPEN
doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Abstract

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of **low-coverage whole-genome sequencing**, **deep exome sequencing**, and **dense microarray genotyping**. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all **phased onto high-quality haplotypes**. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

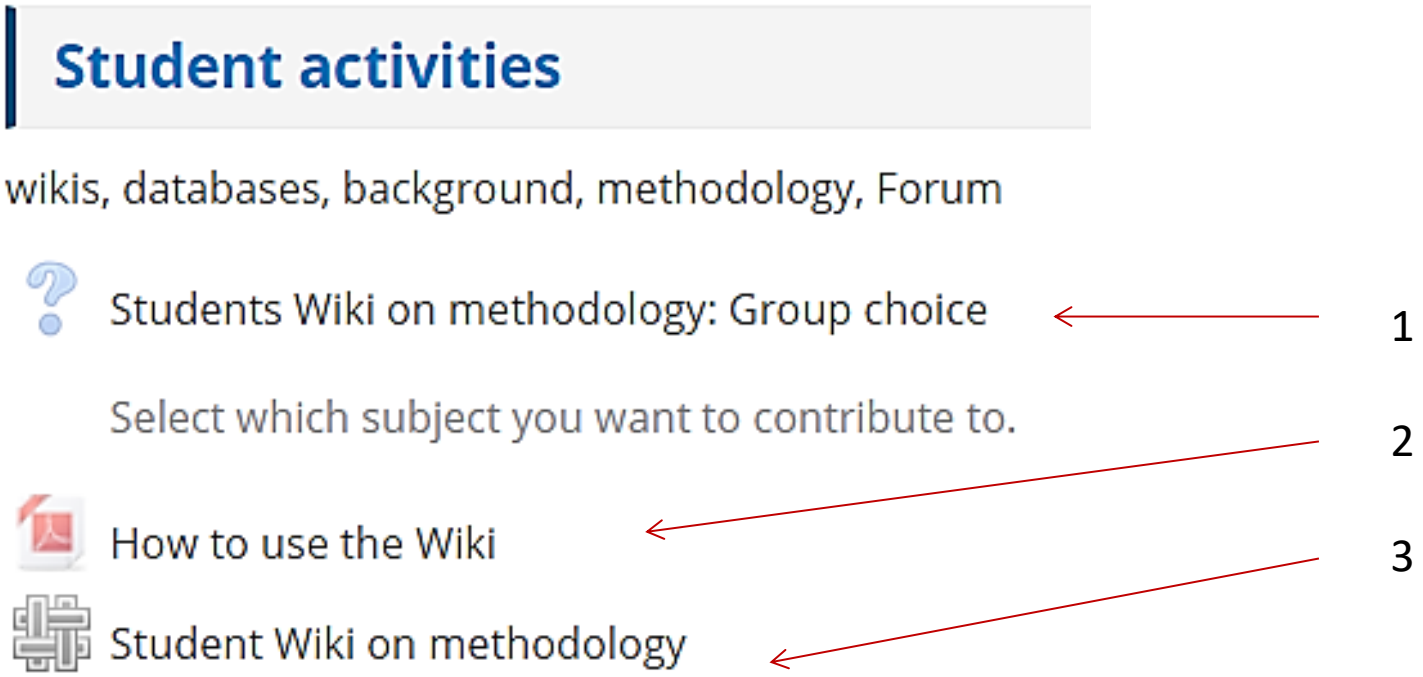
WGS

exome

µarray genotyping

Note:
orange buttons mean something that you should search for and contribute to in the Methodological Wiki on the Moodle site

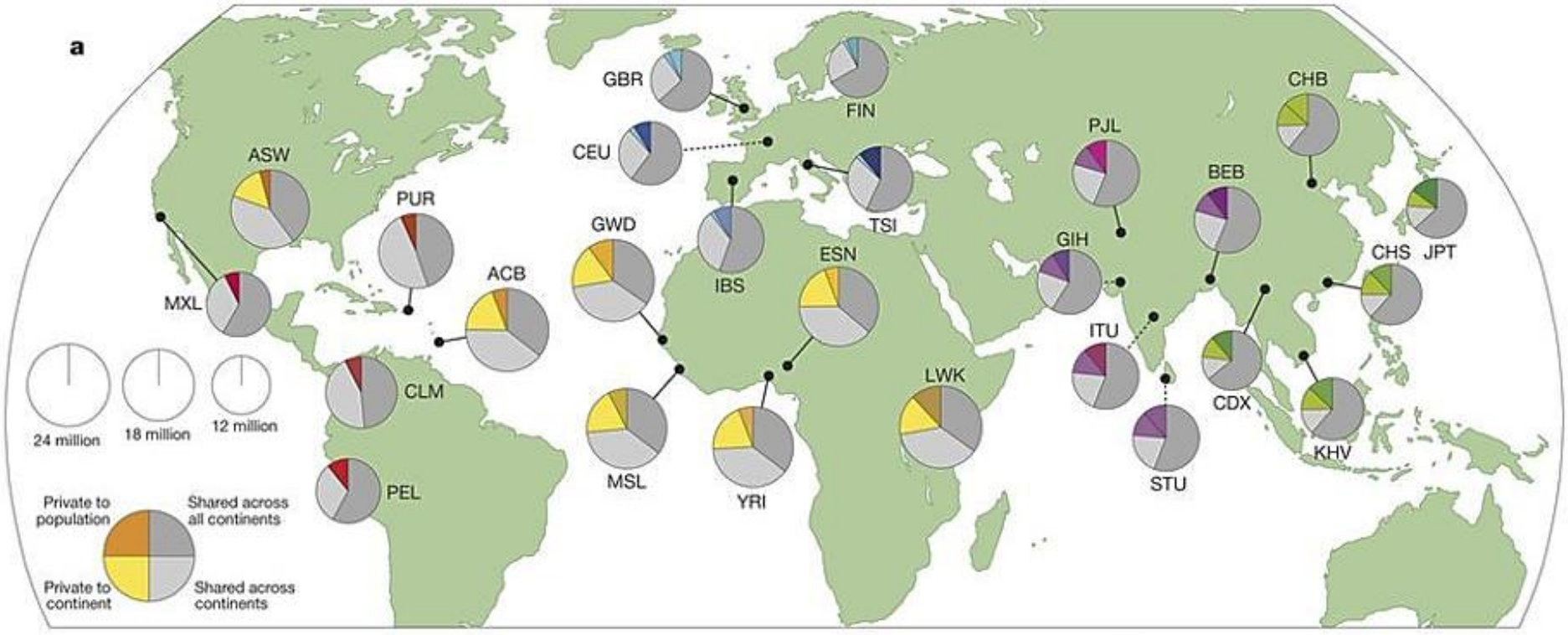
something

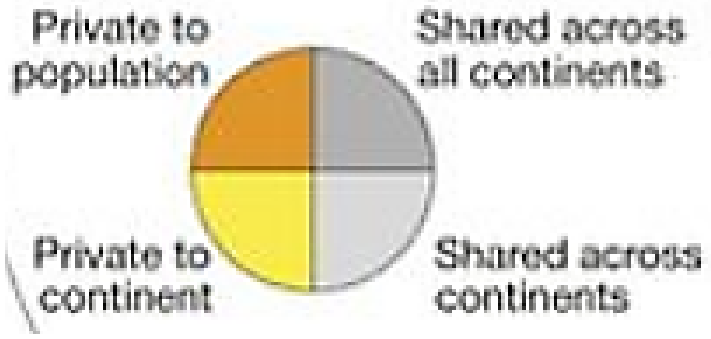


nature

68 | NATURE | VOL 526 | 1 OCTOBER 2015

Population sampling.





Graphs sow the fraction of variants that are either shared among all Humans or private to continents or populations.

The 1000 Genomes Project Consortium

A typical genome

We find that a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites (Fig. 1b and Table 1).
Although **>99.9% of variants consist of SNPs and short indels**, **structural variants affect more bases**: the typical genome contains an estimated **2,100 to 2,500 structural variants** (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), affecting ~20 million bases of sequence.

NUMT=nuclear mitochondrial DNA segment

The 1000 Genomes Project Consortium

The majority of variants in the data set are **rare**: ~64 million autosomal variants have a frequency <0.5%, ~12 million have a frequency between 0.5% and 5%, and only ~8 million have a frequency >5%

Nevertheless, the majority of variants observed in a single genome are common: just 40,000 to 200,000 of the variants in a typical genome (1–4%) have a frequency <0.5%

Genome Browser for variants, with data from 1000HGP is available:

https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/