TOPIC 2. DNA-seq: applications

Applications of DNA-seq: genome sequencing, re-sequencing and variant calling.

Omics Techniques

Bachelor's Degree in Bioinformatics

Sònia Casillas, UAB

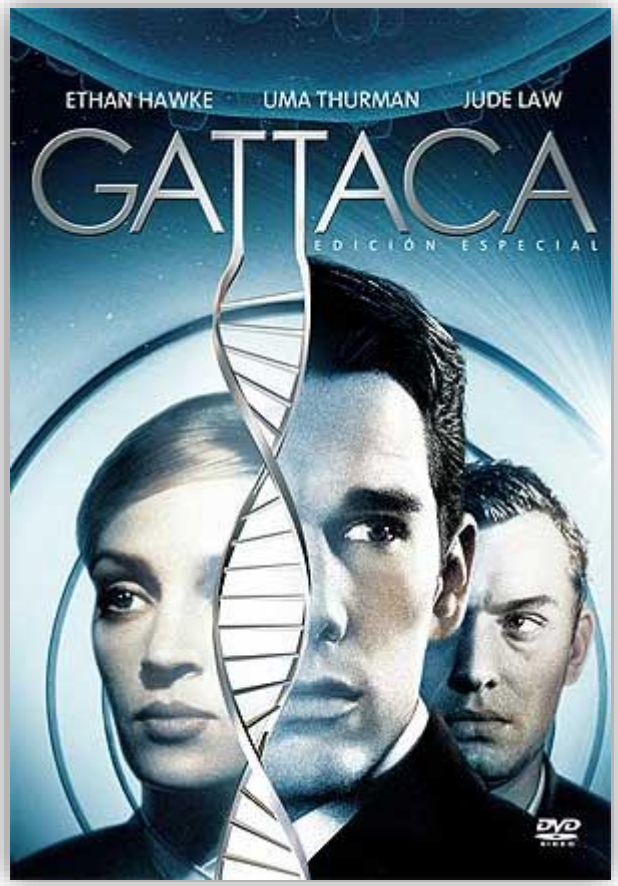
The (not too) distant future?

"The ideal device I want is, you put a blood or saliva sample in and in one hour you have your report out, and everything is fully integrated."

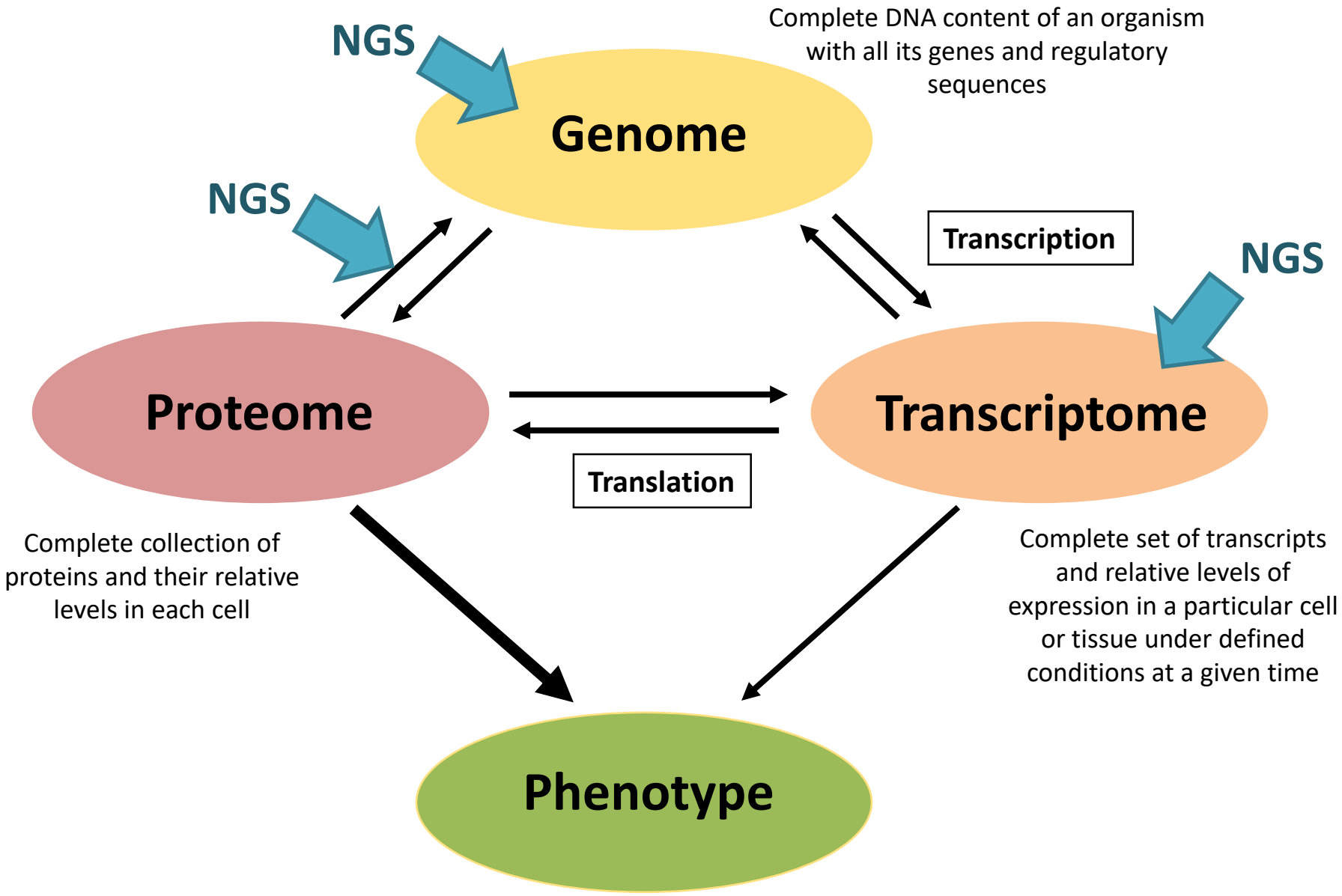
Mostafa Fonaghi

Illumina's chief technology officer

Nature Biotechnology 30, 2012



From the Genome to the Phenotype



NGS applications

Table 2 Applications of next-generation sequencing

Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46–52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56–63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67–70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multiple individuals	61,71

Table 2. Shendure and Ji (2008) Nature Biotechnology 26: 1135 - 1145.

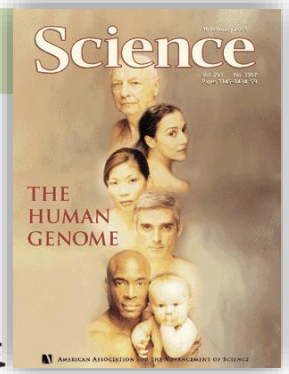
DNA-seq applications

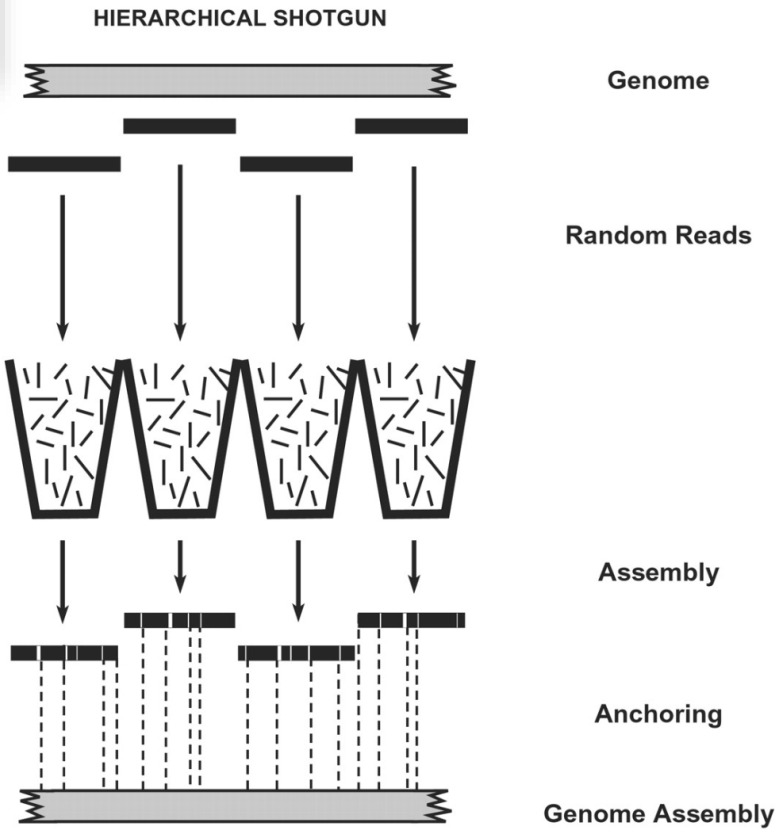
1. **De-novo genome sequencing**
2. Complete genome resequencing
3. Reduced representation sequencing
4. Targeted genome resequencing
5. Paired-end sequencing
6. Metagenomic sequencing
7. Molecular barcoding

Sequencing strategies

Human Genome Project

Celera Genomics

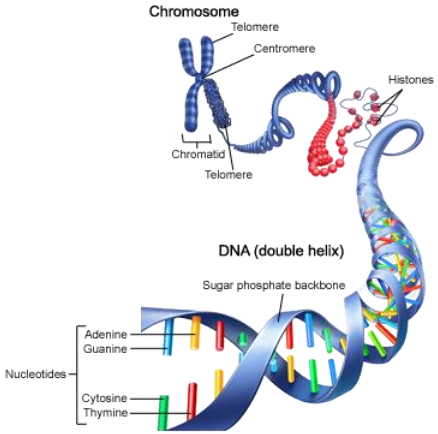




Waterston *et al*. (2002) *PNAS* 99: 3712-3716.

De-novo assembly

From the cells…



… to the sequencers

AAGTATTAGCCAAAAAATTA
TTAGCAAGTAACAATTA
AAGTATTAAGCCAAATTA
AAGTCATTAGCCAAA
AAGTCCAAAATTAGAAATTAGA
AAAAATTAAGTATTAGCCAA
AACAAAAAGTATTAGC
AAGTATTAGCCAAAATTAGCCATTA
AAGTATTAGCCAAAATTAGCCATTA
AACAAAAAGTATTAGC
AAGTATTAAGCCAAATTA
AAAAATTAAGTATTAGCCAA
AAGTCATTAGCCAAA
AAGTCCAAAATTAGAAATTAGA

De-novo assembly

No reference!

AAAAATTAAGTATTA
AATTAAGTATTAGCCTTAAG
GTATTAGCCTTAAGAAATT
CCTTAAGAAATTAAGTATTAGAA
AAATTAAGTATTAGAA

Assembly of the reads based on the overlap of their extremes

Each read is compared to all other reads

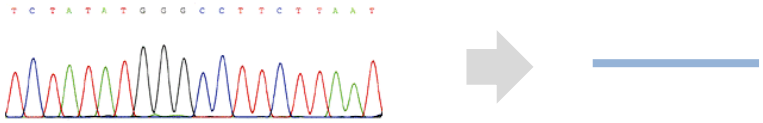
It allows detecting the real structure of the genome

Much more complex and slow, requires high computational power

Sequence assembly

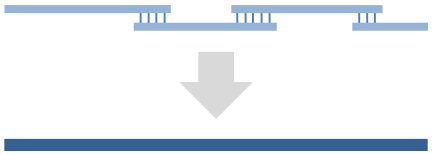
read

Sequence fragment obtained from a sequencing reaction



contig

Set of reads that overlap in their extremes and generate a longer contiguous sequence



Sequence assembly – quality measures

Quality of a base – phred score (Q)

$$Q = -10 \log_{10} P$$

P = Error probability of each base

Q > 20 reliable base (P = 0,01)

Redundancy

Average number of reads spanning each base of the assembly

$$R = \frac{N \cdot L}{G}$$

N = number of reads

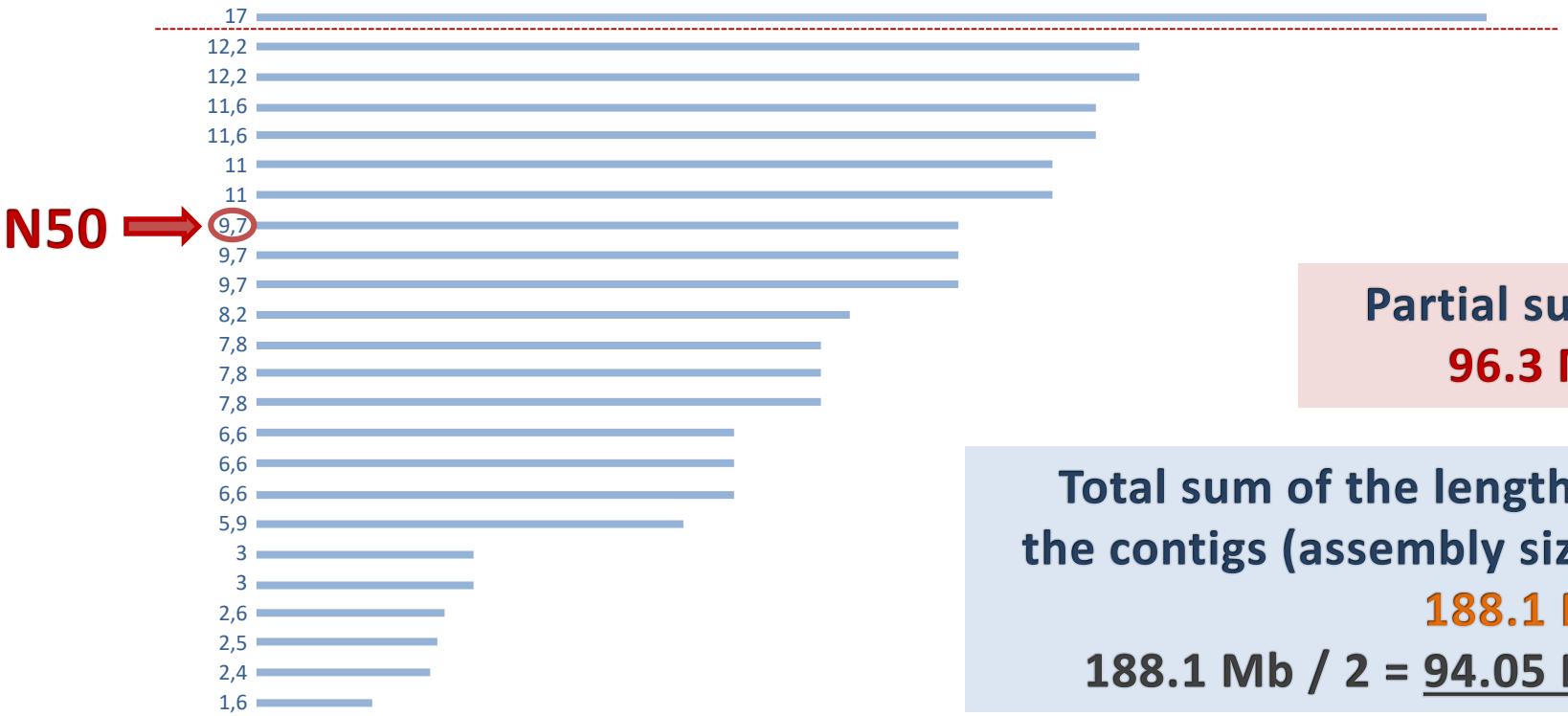
L = average read length

G = genome size

Sequence assembly – quality measures

N50 contig length

Contig length L such that 50% of the bases of the assembly are in contigs of length ≥ L



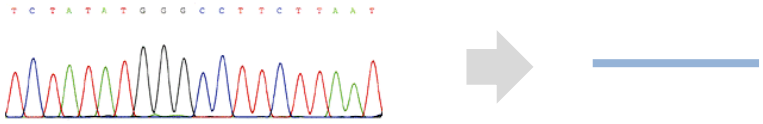
Partial sum:
96.3 Mb

Total sum of the length of the contigs (assembly size):
188.1 Mb
188.1 Mb / 2 = 94.05 Mb

Sequence assembly

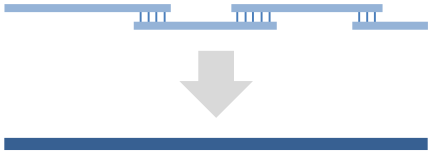
read

Sequence fragment obtained from a sequencing reaction



contig

Set of reads that overlap in their extremes and generate a longer contiguous sequence



scaffold

Ordered and oriented contigs based on information from paired-end reads. Contain *gaps* or undetermined bases

Paired-end reads

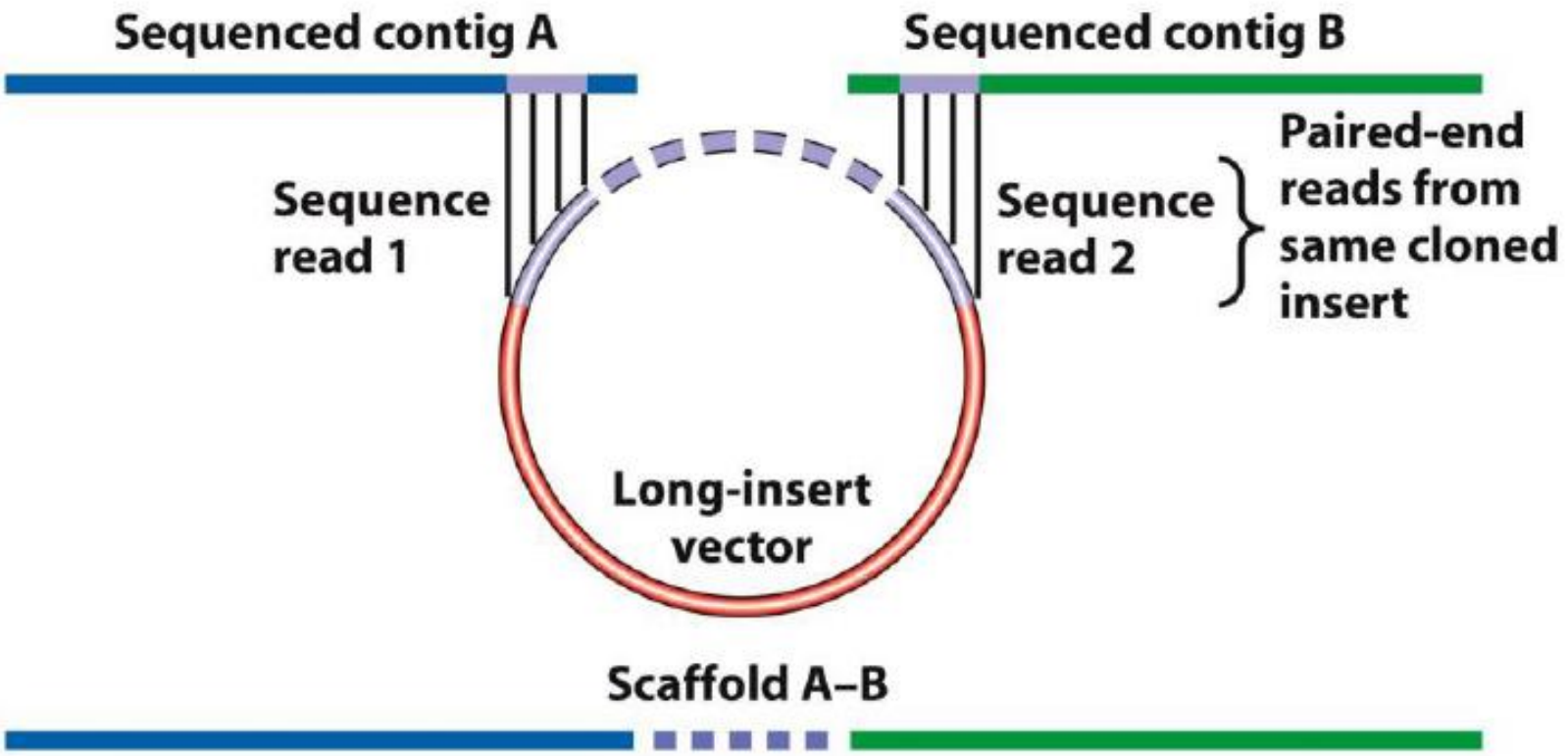
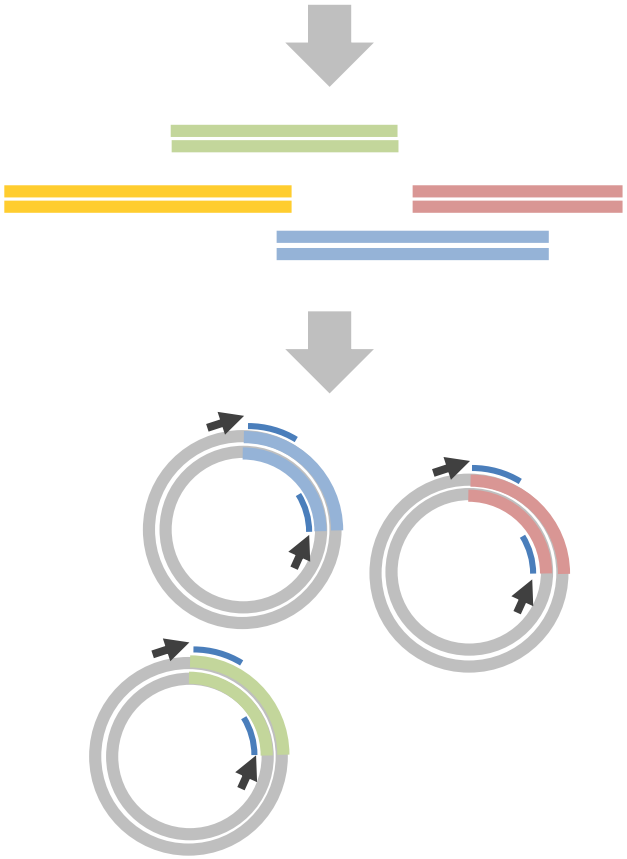


Figure 13-5
Introduction to Genetic Analysis, Ninth Edition

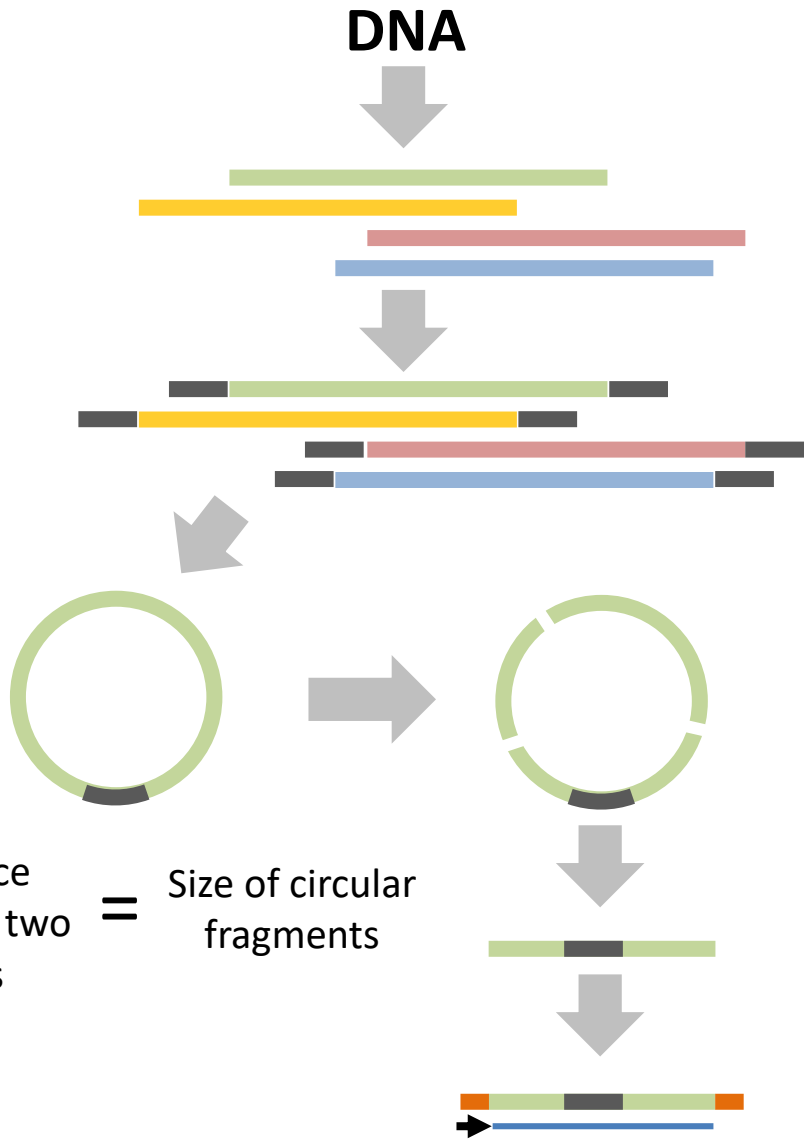
Paired-end reads

Sanger



Distance between two reads = Size of cloned fragments

Tecnologies nova generació



Distance between two reads = Size of circular fragments

Assembly steps: reads, contigs, scaffolds

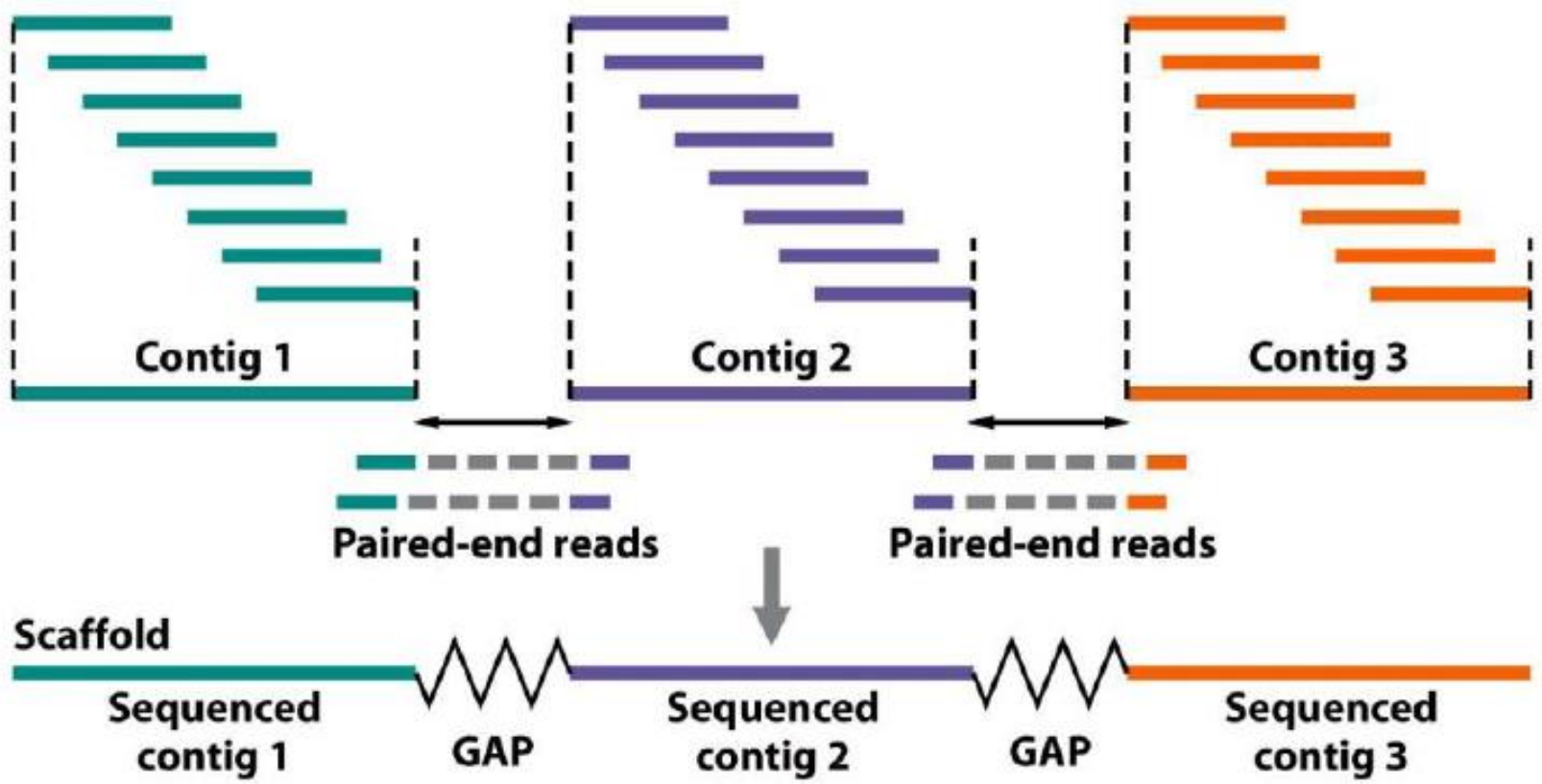


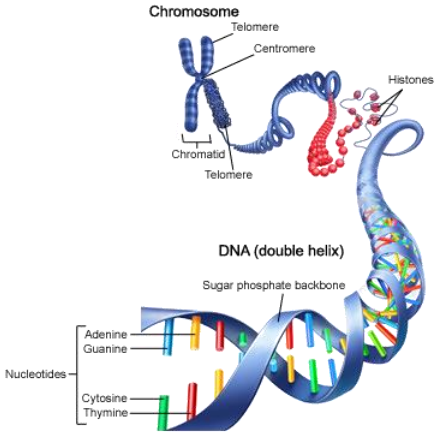
Figure 13-6. *Introduction to Genetic Analysis*. Griffiths *et al*. 9th edition (2008)

DNA-seq applications

1. De-novo genome sequencing
2. **Complete genome resequencing**
3. Reduced representation sequencing
4. Targeted genome resequencing
5. Paired-end sequencing
6. Metagenomic sequencing
7. Molecular barcoding

Complete genome resequencing

From the cells…



… to the sequencers

AAGTATTAGCCAAAAAATTA
TTAGCAAGTAACAATTA
AAGTATTAAGCCAAATTA
AAGTCATTAGCCAAA
AAGTCCAAAATTAGAAATTAGA
AAAAATTAAGTATTAGCCAA
AACAAAAAGTATTAGC
AAGTATTAGCCAAAATTAGCCATTA
AAGTATTAGCCAAAATTAGCCATTA
AACAAAAAGTATTAGC
AAGTATTAAGCCAAATTA
AAAAATTAAGTATTAGCCAA
AAGTCATTAGCCAAA
AAGTCCAAAATTAGAAATTAGA

De-novo assembly

No reference!

AAAAATTAAGTATTA
AATTAAGTATTAGCCTTAAG
GTATTAGCCTTAAGAAATT
CCTTAAGAAATTAAGTATTAGAA
AAATTAAGTATTAGAA

Assembly of the reads based on the overlap of their extremes

Each read is compared to all other reads

It allows detecting the real structure of the genome

Much more complex and slow, requires high computational power

The era of personal genomes

HUMAN GENOME

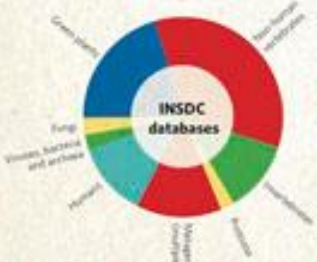
THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for 'finished' sequence: GenBank, run by the US National Center for Biotechnology Information; the DNA Databank of Japan; and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

See Editorial, page 649, and human genome special at www.nature.com/humangenome

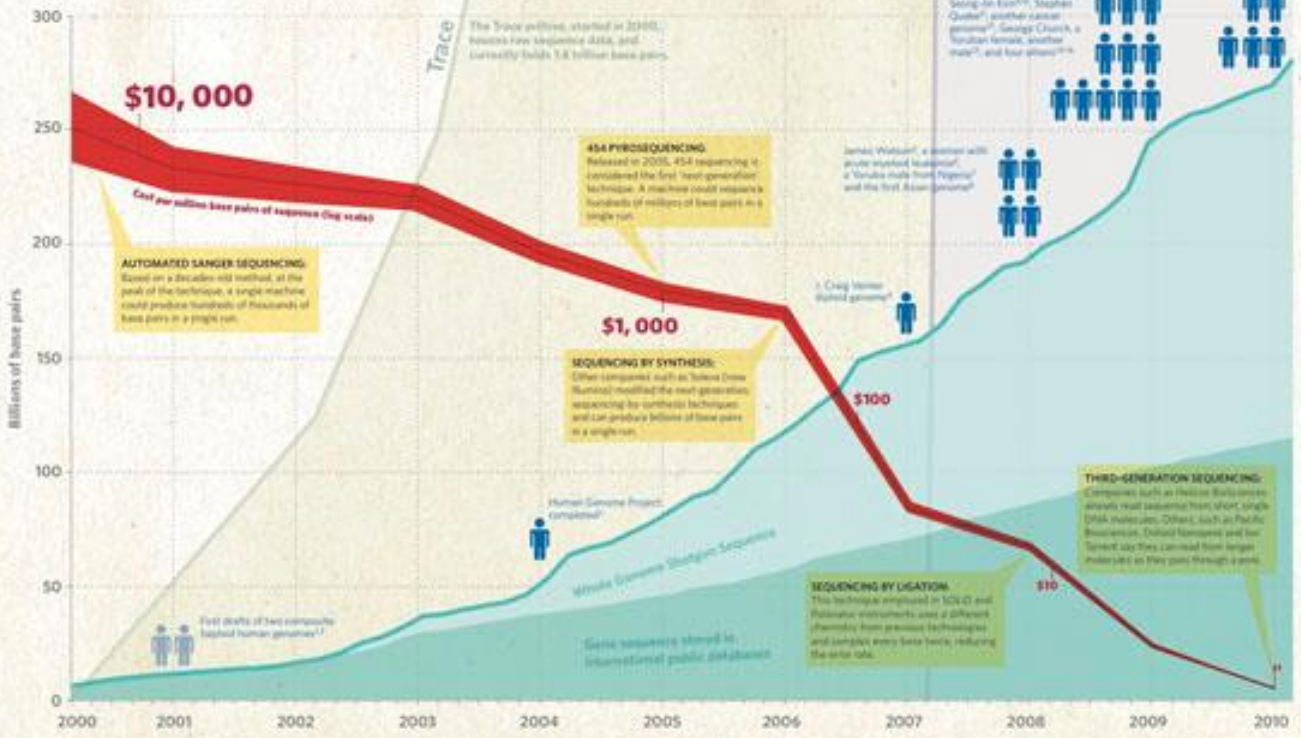
DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of 'finished' sequence span a wide range of organisms, representing the many priorities of scientists worldwide.



Trace Archive: Developed to house the raw output of high-throughput sequencers built in the late 1990s, the trace archive spans a wide range of taxa.

Sequence Read Archive: Houses raw data from next-generation sequencers. Dominated by human sequence, including multiple coverage for more than 170 people.



HOW MANY HUMAN GENOMES?

The graphic shows all published, fully sequenced human genomes since 2000, including nine from the first quarter of 2010. Some are resequencing efforts on the same person and the list does not include unpublished completed genomes.

1. Venter, J. C. et al. Science **291**, 1304–1351 (2001).
2. International Human Genome Sequencing Consortium Nature **409**, 860–921 (2001).
3. International Human Genome Sequencing Consortium Nature **431**, 931–945 (2004).
4. Levy, S. et al. PLoS Biol. **5**, e254 (2007).
5. Wheeler, D. A. et al. Nature **452**, 872–876 (2008).
6. Ley, T. J. et al. Nature **456**, 66–72 (2008).
7. Bentley, D. R. et al. Nature **456**, 53–59 (2008).
8. Wang, J. et al. Nature **456**, 60–65 (2008).
9. Ahn, S.-M. et al. Genome Res. **19**, 1622–1629 (2009).
10. Kim, J.-I. et al. Nature **460**, 1011–1015 (2009).
11. Pushkarev, D., Neff, N. F. & Quake, S. R. Nature Biotechnol. **27**, 847–850 (2009).
12. Margulies, E. H. et al. N. Engl. J. Med. **10**, 1058–1066 (2009).
13. Drmanac, R. et al. Science **327**, 78–81 (2009).
14. McKernan, K. J. et al. Genome Res. **19**, 1527–1541 (2009).
15. Pleasance, E. D. et al. Nature **463**, 191–196 (2010).
16. Pleasance, E. D. et al. Nature **463**, 184–190 (2010).
17. Clark, M. J. et al. PLoS Genet. **6**, e1000832 (2010).
18. Rasmussen, M. et al. Nature **463**, 757–762 (2010).
19. Schuster, S. C. et al. Nature **463**, 943–947 (2010).
20. Lupski, J. R. et al. N. Engl. J. Med. doi 10.1056/NEJMoa1001965 (2010).
21. Roach, J. C. et al. Science doi 10.1126/science.1186802 (2010).

Page size by comparison

Table 2 | **Sequencing statistics on personal genome projects**

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

Whole-genome sequencing projects

- **Multiple individual genome sequencing projects**

  1000 Genomes Project, 1001 Genomes Project, DGRP, etc

  - Complete characterization of variants in normal populations
  - Association of genetic variants to phenotypic traits
  - Quantification of mutation rates
- **Sequencing of tumor and normal genomes**

  International Cancer Genome Consortium, The Cancer Genome Atlas

  - Identification of genetic variants predisposing to cancer
  - Identification of somatic mutations involved in cancer progression
- **Sequencing of extinct genomes**

  Neanderthal, Denisovan, ancient Eurasians

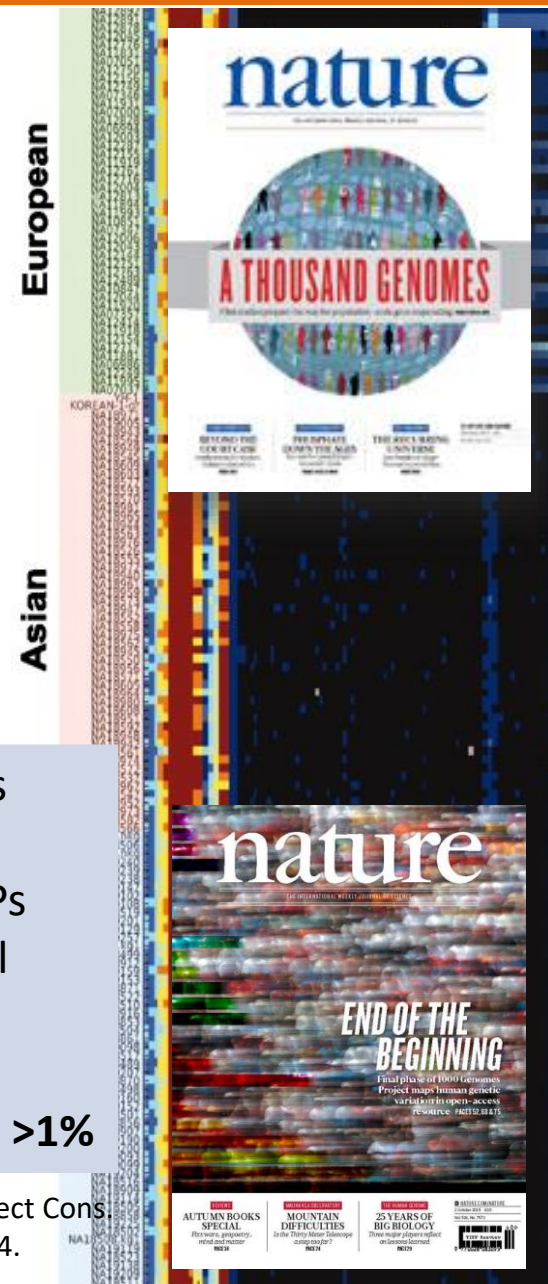
  - Identification of changes important in evolution
- **Sequencing of pathogens**

  Germany E. coli *strain outbreak, SARS/Ebola virus outbreak sequencing*

  - Identification of particular pathogens & therapeutic targets

The 1000 Genomes Project

The 1000 Genomes Project (http://www.1000genomes.org, 2008) aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the role of genetic variation in human history, evolution and disease. The genomes of 2,504 unidentified people from 26 populations around the world will be sequenced using next-generation sequencing technologies.



Pilot Phase

179 individuals
4 populations
15 million SNPs
1 million small indels
20,000 SVs
>95% SNPs freq>5%

The 1000 Genomes Project Cons. (2010) *Nature* 467: 1061–1073

Phase I

1092 individuals
14 populations
38 million SNPs
1.4 million small indels
14,000 deletions
98% SNPs freq>1%

The 1000 Genomes Project Cons. (2012) *Nature* 491: 56-65

Phase III

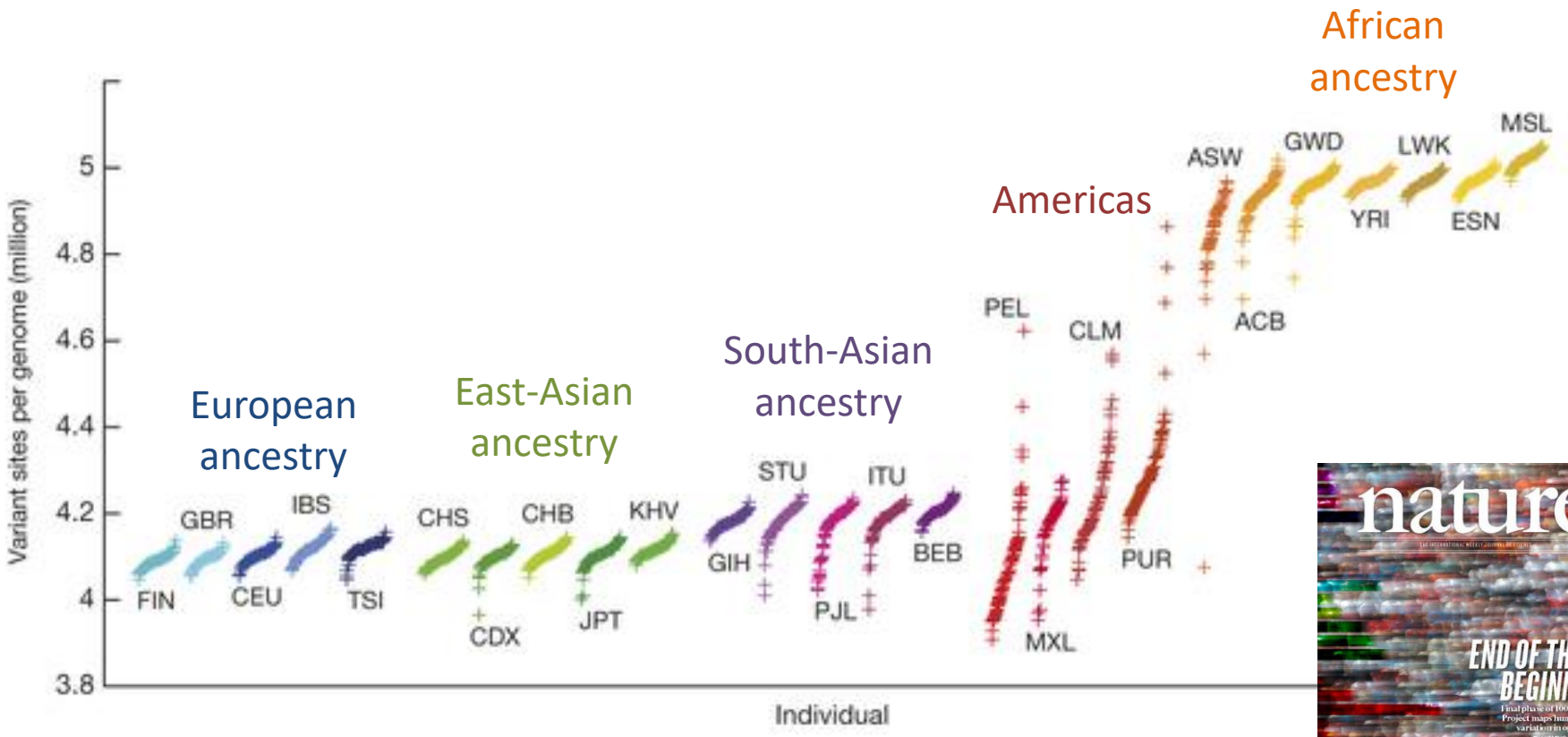
2504 individuals
26 populations
84.7 million SNPs
3.6 million small indels
60,000 SVs
>99% SNPs freq >1%

The 1000 Genomes Project Cons. (2015) *Nature* 526: 68-74.

The 1000 Genomes Project

A **typical genome** differs from the reference by:

- 4.1 – 5.0 million SNPs



The 1000 Genomes Project Cons. (2015) *Nature* 526: 68-74.

UK10K

Rare Genetic Variants in Health and Disease

What is UK10K?

The UK10K project will enable researchers in the UK and beyond to better understand the link between low-frequency and rare genetic changes, and human disease caused by harmful changes to the proteins the body makes.

Although many hundreds of genes that are involved in causing disease have already been identified, it is believed that many more remain to be discovered. The UK10K project aims to help uncover them by studying the genetic code of 10,000 people in much finer detail than ever before.

Wellcome Library, London

Project Design

Not all genetic changes are harmful or lead to disease, so the project is taking a two-pronged approach to identify rare variants and their effects:

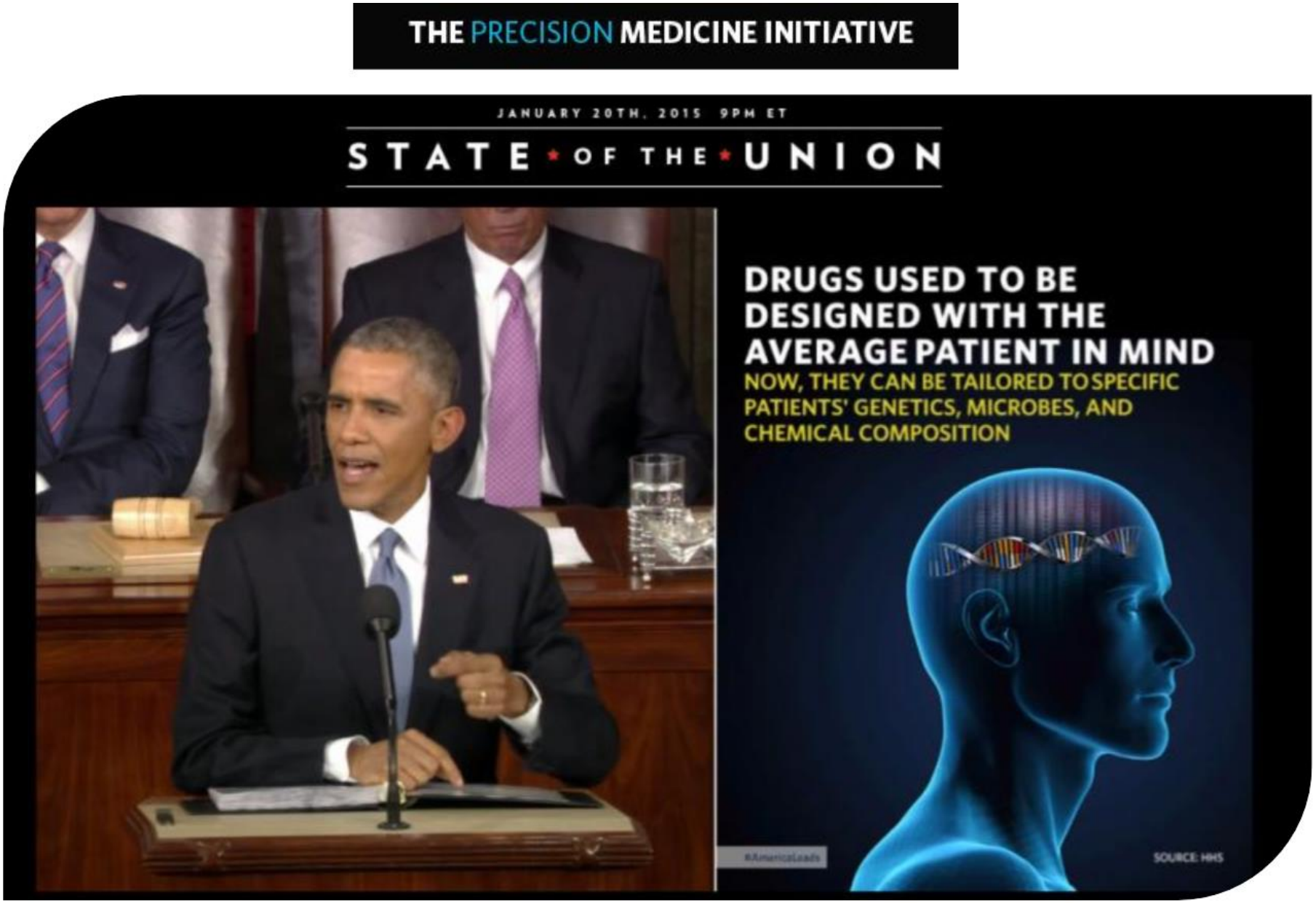
- by studying and comparing the DNA of 4,000 people whose physical characteristics are well documented, the project aims to identify those changes that have no discernible effect and those that may be linked to a particular disease;
- by studying the changes within protein-coding areas of DNA that tell the body how to make proteins of 6,000 people with extreme health problems and comparing them with the first group, it is hoped to find only those changes in DNA that are responsible for the particular health problems observed.

The project received a £10.5 million funding award from the Wellcome Trust in March 2010 and sequencing started in late 2010. For more information, please use the links on the right hand side.

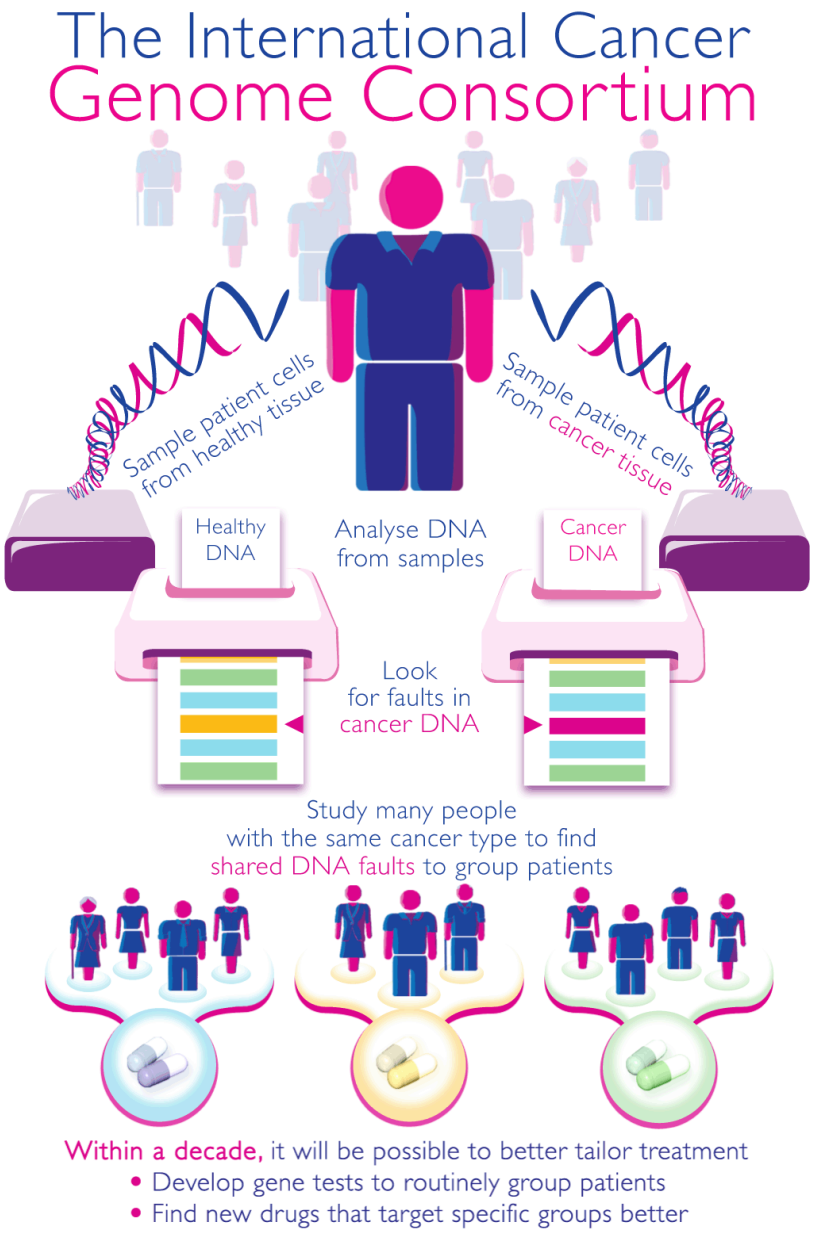
Aims

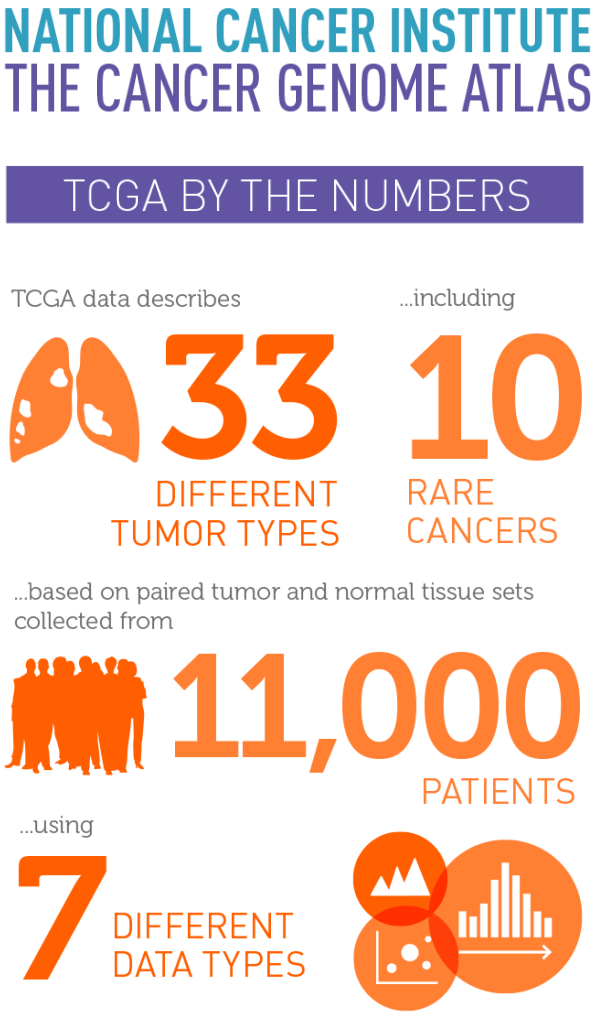
- Elucidate singleton variants by maximising variation detected
- Directly associate genetic variations to phenotypic traits
- Uncover rare variants contributing to disease
- Assign uncovered variations into genotyped cohort and case/control collections
- Provide a sequence variation resource for future studies

Precision medicine



Cancer genomics

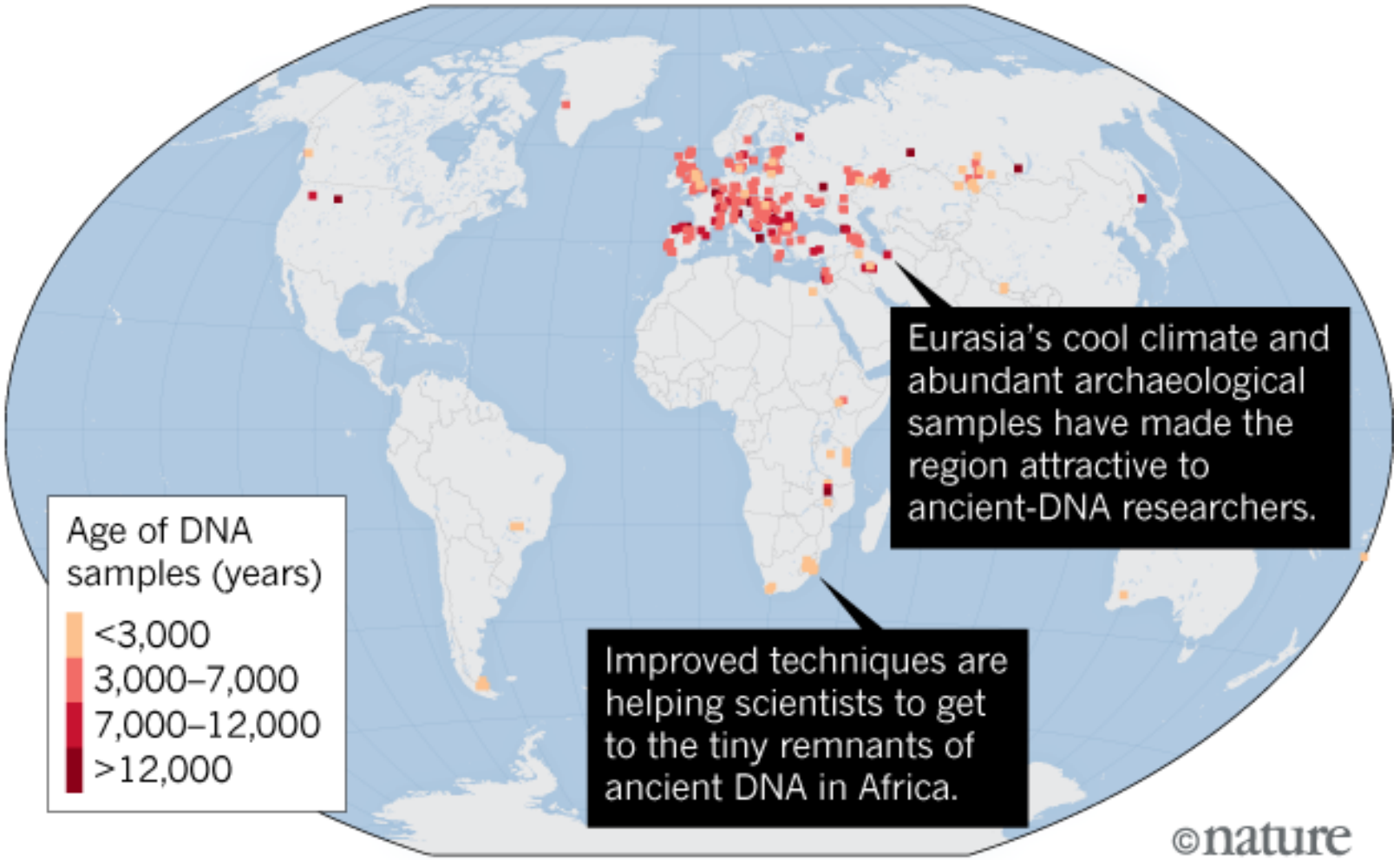




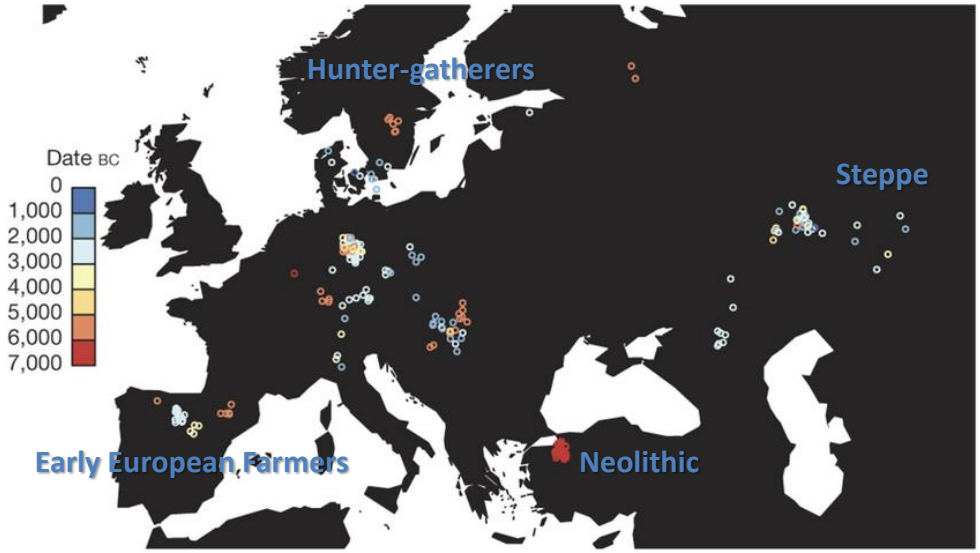
Paleogenomics / Ancient genomics

ANCIENT GENOMES AROUND THE GLOBE

Most of the ancient DNA that has been sequenced has come from individuals who lived in Eurasia.



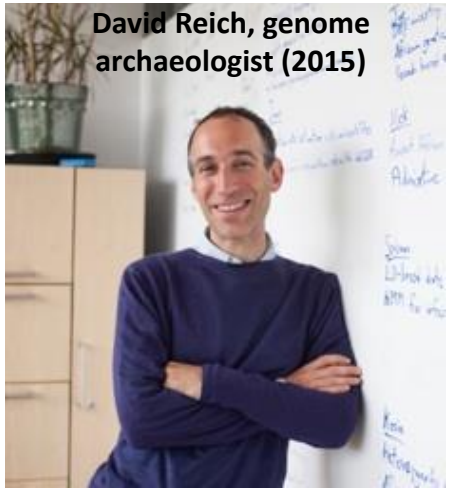
Genome-wide patterns of selection in 230 ancient Eurasians



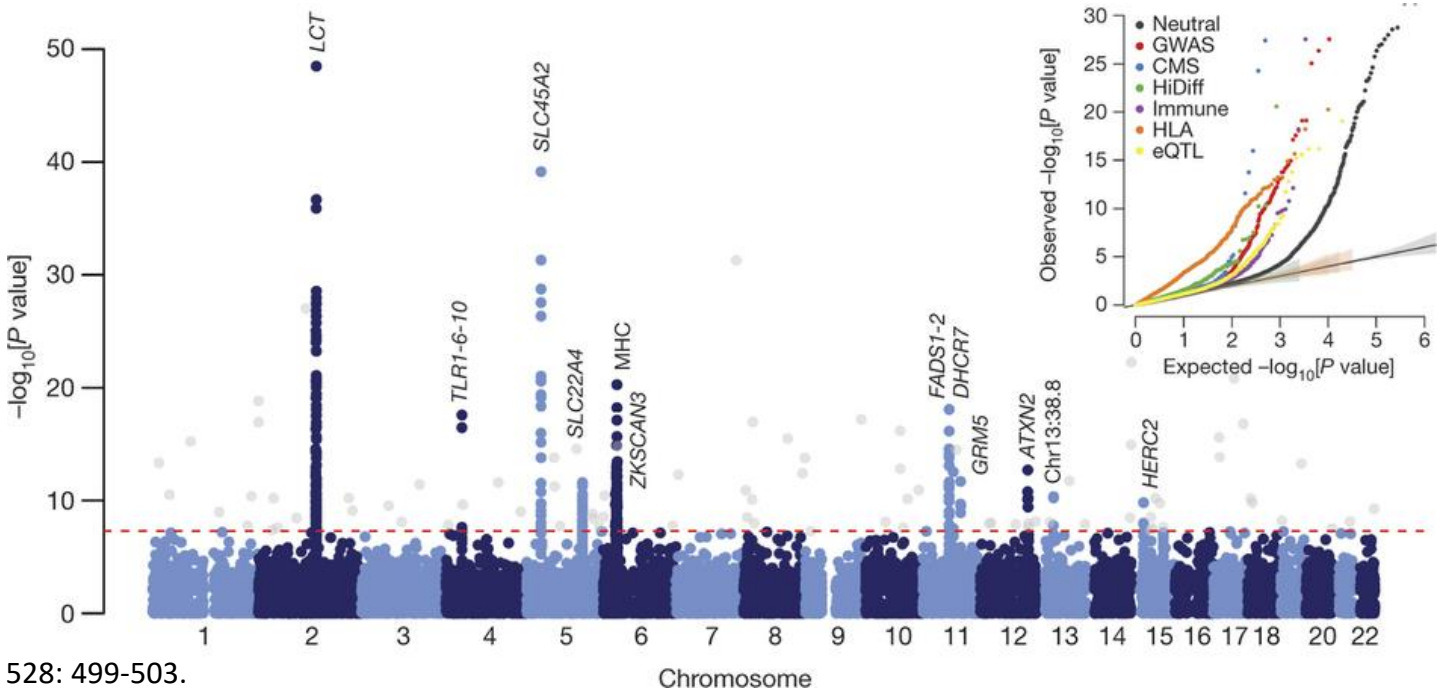
- Ancient DNA sequencing allows pinpointing the action of natural selection in samples right before, during and after the adaptation process
- 230 West Eurasians, 6500-300 BC
- Selection detected in loci associated with diet, pigmentation and immunity, and two independent episodes of selection on height

365 days: *Nature*'s 10

Ten people who mattered this year.

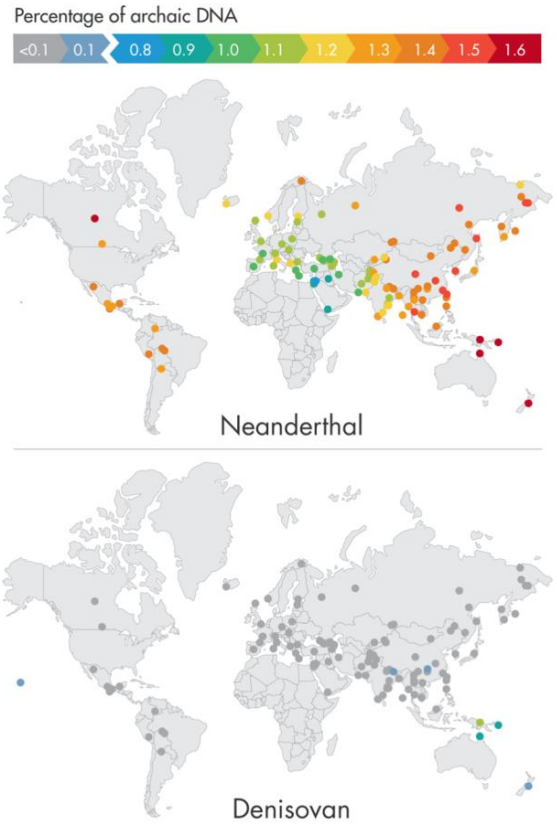


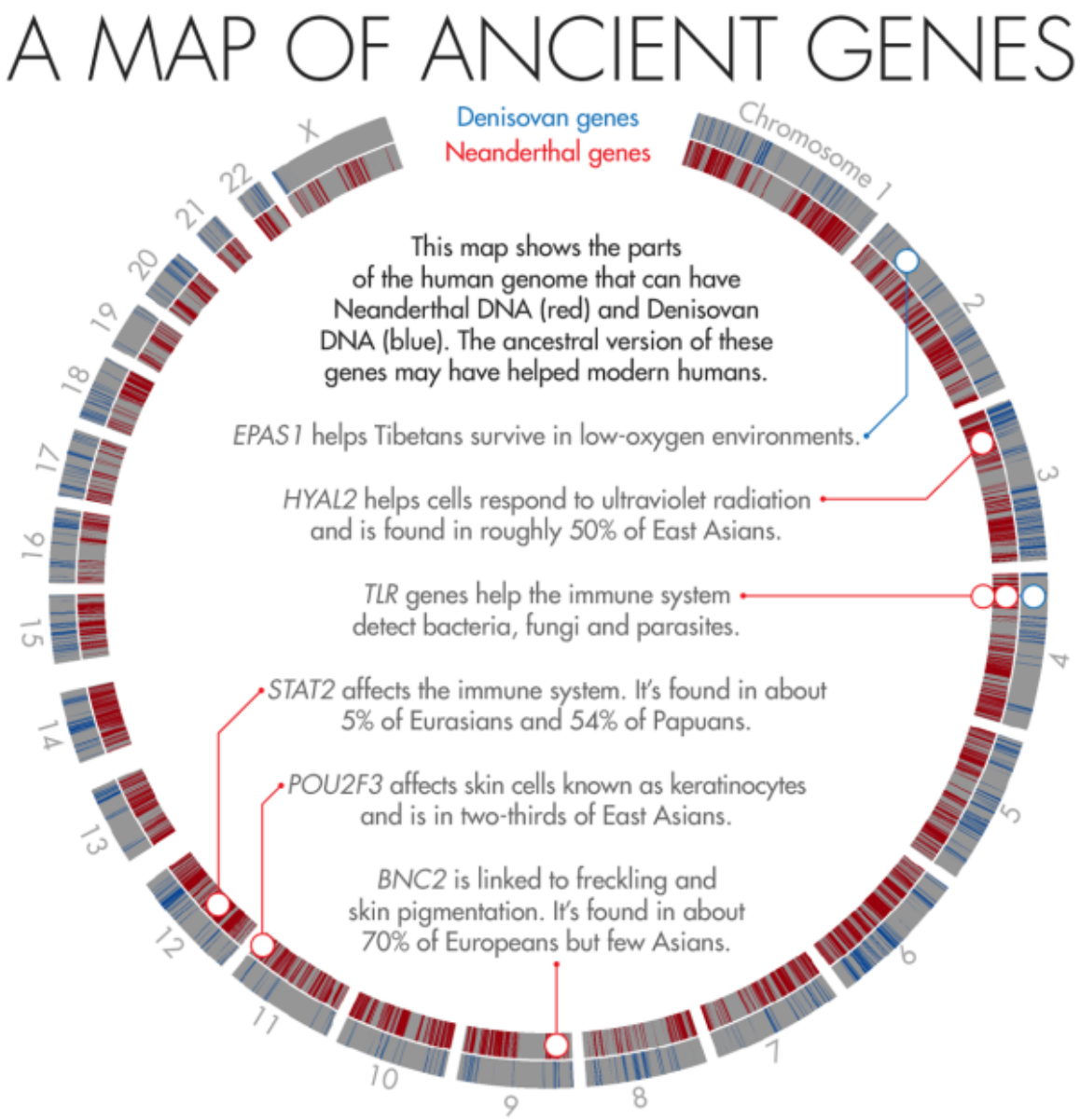
David Reich, genome archaeologist (2015)



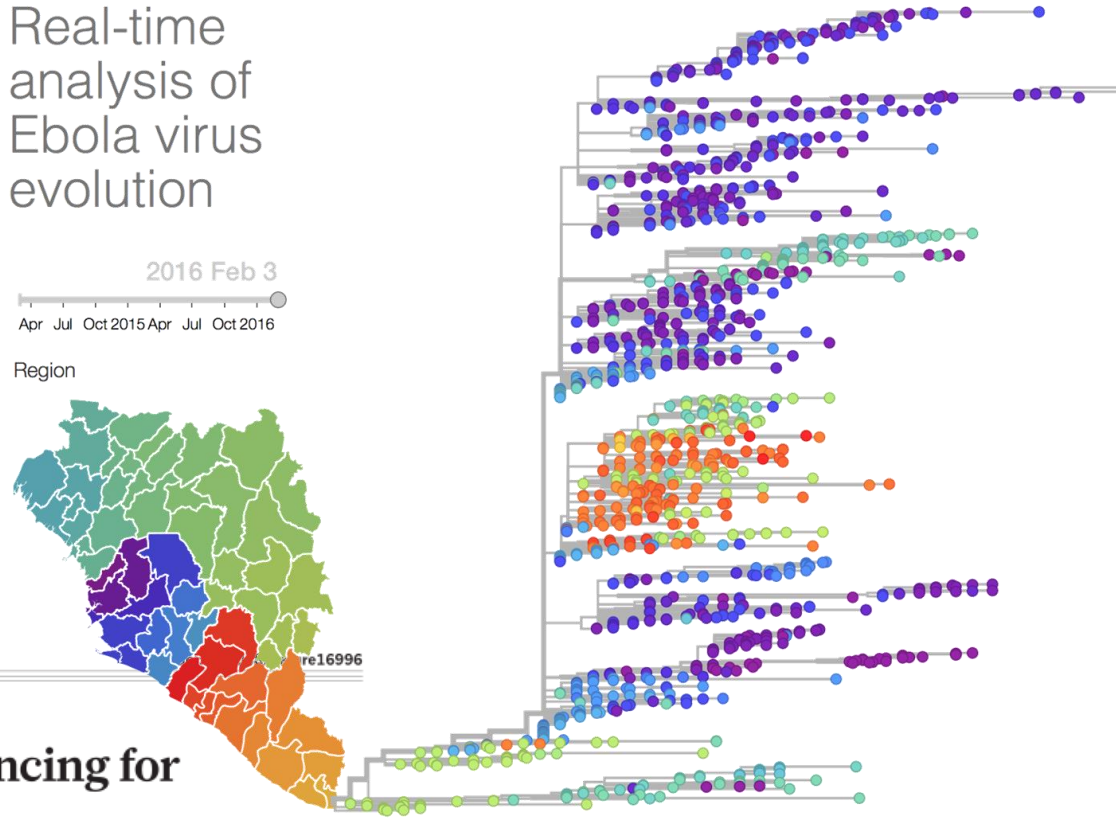
Iain Mathieson et al. (2015) *Nature* 528: 499-503.

We are ~1.5% Neanderthals





Monitoring the transmission of Ebola with MinION



LETTER

Real-time, portable genome sequencing for Ebola surveillance

The Ebola virus disease epidemic in West Africa is the largest on record, responsible for over 28,599 cases and more than 11,299 deaths[1]. Genome sequencing in viral outbreaks is desirable to characterize the infectious agent and determine its evolutionary rate. Genome sequencing also allows the identification of signatures of host adaptation, identification and monitoring of diagnostic targets, and characterization of responses to vaccines and treatments. The Ebola virus (EBOV) genome substitution rate in the Makona strain has been estimated at between 0.87×10^{-3} and 1.42×10^{-3} mutations per site per year. This is equivalent to 16–27 mutations in each genome, meaning that sequences diverge rapidly enough to identify distinct sub-lineages during a prolonged epidemic[2–7]. Genome sequencing provides a high-resolution view of pathogen evolution and is increasingly sought after for outbreak surveillance. Sequence data may be used to guide control measures, but only if the results are generated quickly enough to inform interventions[8]. Genomic surveillance during the epidemic has been sporadic owing to a lack of local sequencing capacity coupled with practical difficulties transporting samples to remote sequencing facilities[9]. To address this problem, here we devise a genomic surveillance system that utilizes a novel nanopore DNA sequencing instrument. In April 2015 this system was transported in standard airline luggage to Guinea and used for real-time genomic surveillance of the ongoing epidemic. We present sequence data and analysis of 142 EBOV samples collected during the period March to October 2015. We were able to generate results less than 24 h after receiving an Ebola-positive sample, with the sequencing process taking as little as 15–60 min. We show that real-time genomic surveillance is possible in resource-limited settings and can be established rapidly to monitor outbreaks.

Conventional sequencing technologies are difficult to deploy in developing countries, where availability of continuous power and cold chains, laboratory space, and trained personnel is restricted. In addition, some genome sequencer instruments, such as those using optical

DNA-seq applications

1. De-novo genome sequencing
2. Complete genome resequencing
3. **Reduced representation sequencing**
4. Targeted genome resequencing
5. Paired-end sequencing
6. Metagenomic sequencing
7. Molecular barcoding

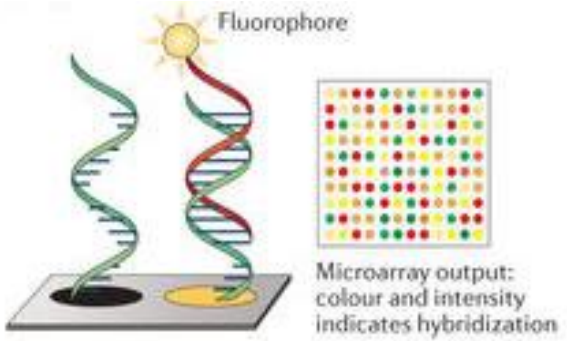
DNA microarrays

Collection of microscopic DNA spots attached to a solid surface. They allow genotyping multiple regions of a genome simultaneously.

Each DNA spot contains picomoles of a specific DNA sequence, known as probes. These can be short sections of a gene or other DNA element that are used to hybridize a sample, called target, under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-labeled targets.

(DNA) Applications:

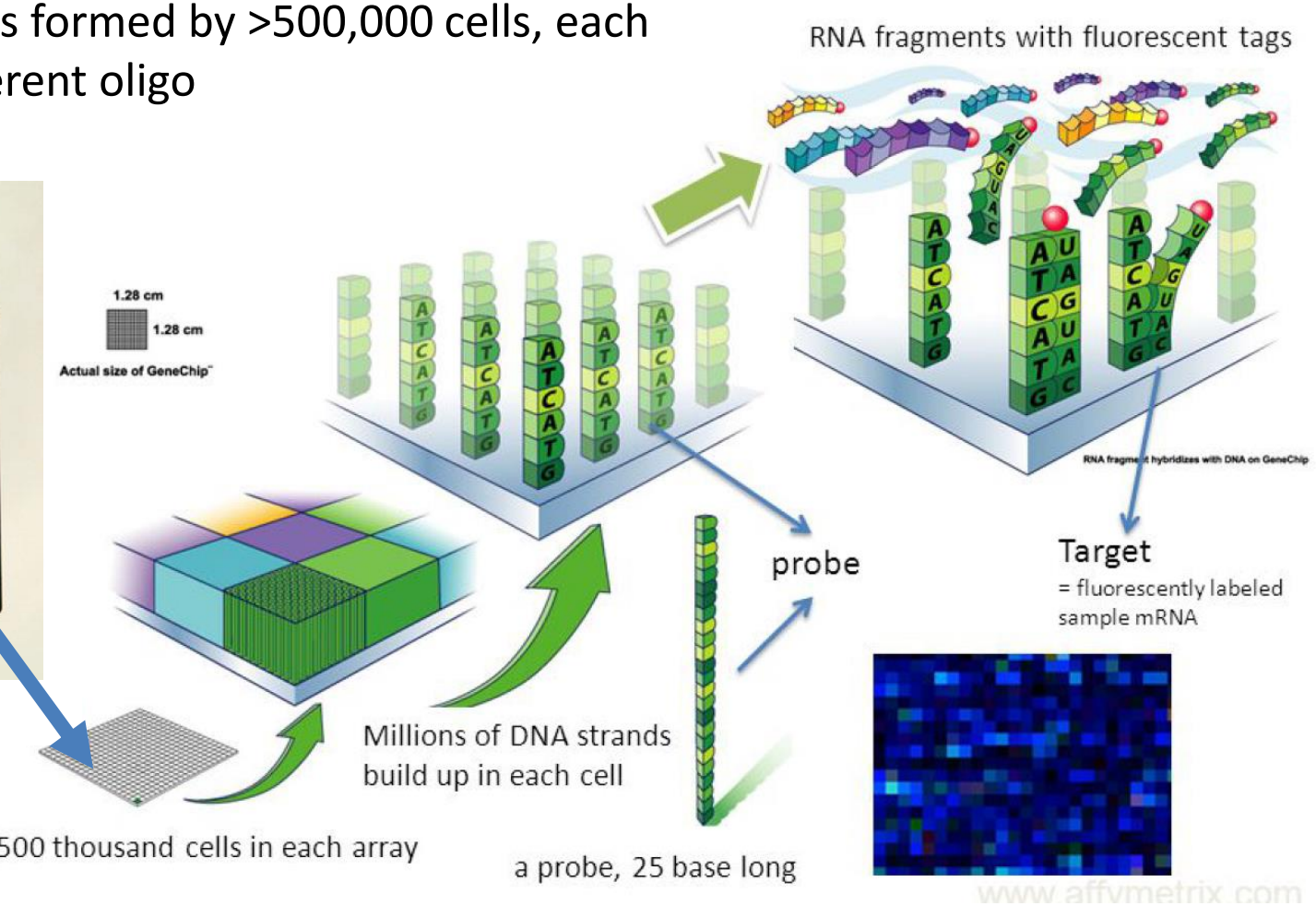
- Re-sequencing and SNP genotyping
- Estimating DNA copy number (CGH-arrays)



Affymetrix oligonucleotide arrays

The array elements are a series of 25-mer oligos designed from known sequence and synthesized directly on the surface

The entire array is formed by >500,000 cells, each containing a different oligo



Genome-Wide Association studies (GWAS)

Identification of genes associated with 7 common human diseases

2,000 individuals for each disease + 3,000 controls

>500,000 genotyped SNPs

24 associated regions
- 1 bipolar disorder
- 1 coronary artery disease
- 9 Crohn's disease
- 3 rheumatoid arthritis
- 7 type 1 diabetes
- 3 type 2 diabetes

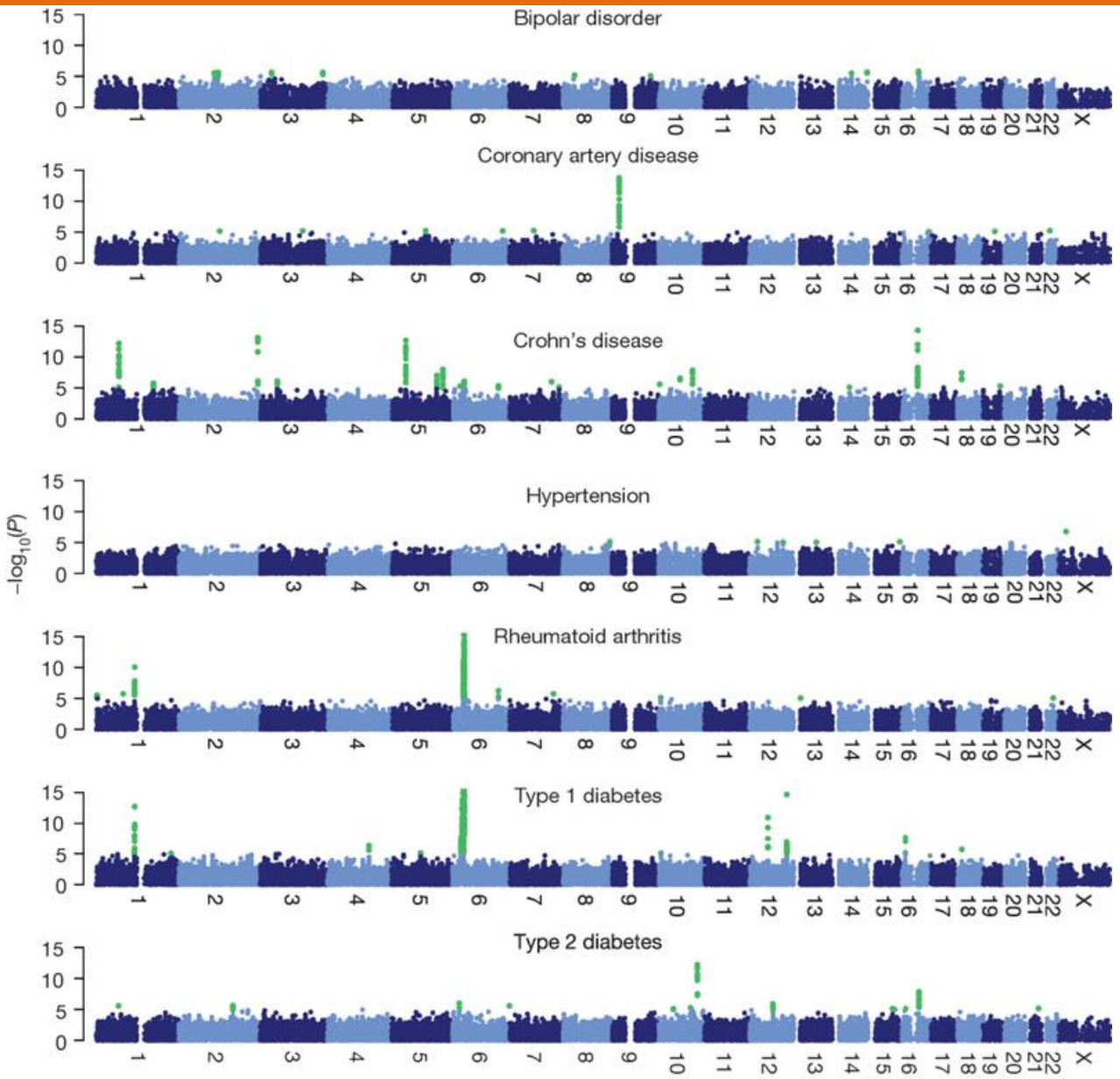


Fig. 4. The Wellcome Trust Case Control Consortium (2007) *Nature* 447: 661-678

First catalog of human genetic variation: the HapMap

The HapMap project (http://www.hapmap.org, 2002) is an International effort to catalog common human genetic variation with the aim to identify genes that affect health, disease and drug/environmental response. It encompasses >4M SNPs in 1,184 individuals from 11 human populations.

FASE 1
270 samples
4 populations
>1,000,000 SNPs
(1 SNP / 5 kb)
MAF > 0.05

The International HapMap Consortium (2005) *Nature* 437: 1299-1320

FASE 2
270 samples
4 populations
>3,100,000 SNPs
(1 SNP / 1 kb)

The International HapMap Consortium (2007) *Nature* 449: 851-861

FASE 3
1184 samples
11 populations
>4,000,000 SNPs

The International HapMap Consortium (2010) *Nature* 467: 52-58

23andMe genotyping service

23andMe

HumanOmniExpress-24 chip

- 24 samples per chip
- >750,000 common SNPs
- ~30,000 additional SNPs of particular interest

MEETS FDA REQUIREMENTS

Genetic Health Risks*

Learn how your genetics can influence your risk for certain diseases.

Ancestry

Discover where your DNA is from out of 150+ regions worldwide - and more.

Wellness

Learn how your genes play a role in your well-being and lifestyle choices.

MEETS FDA REQUIREMENTS

Carrier Status*

If you are starting a family, find out if you are a carrier for certain inherited conditions.

Traits

Learn how your DNA influences your facial features, taste, smell and other traits.

23andMe genotyping service

23andMe

Late-Onset Alzheimer's Disease

Jamie, you **do not have** the ε4 variant we tested.

0 variants detected
in the APOE gene

MEETS FDA REQUIREMENTS

Genetic Health Risks*

Learn how your genetics can influence your risk for certain diseases.

Ancestry

Discover where your DNA is from out of 150+ regions worldwide - and more.

Wellness

Learn how your genes play a role in your well-being and lifestyle choices.

MEETS FDA REQUIREMENTS

Carrier Status*

If you are starting a family, find out if you are a carrier for certain inherited conditions.

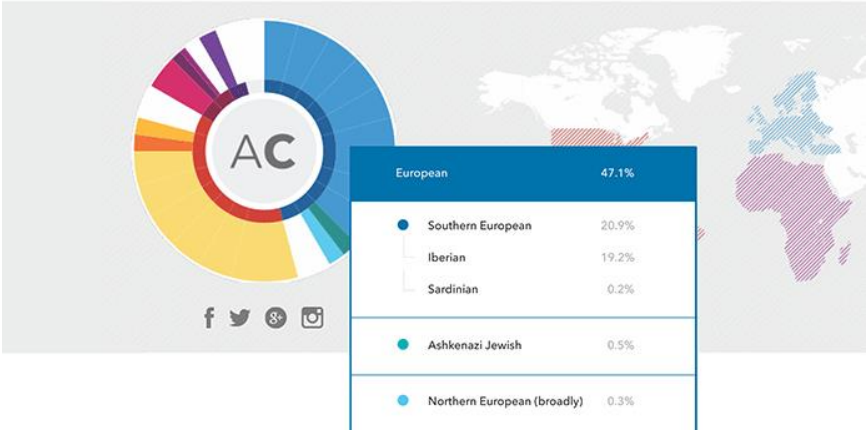
Traits

Learn how your DNA influences your facial features, taste, smell and other traits.

23andMe genotyping service

23andMe

Ancestry Composition



MEETS FDA REQUIREMENTS

Genetic Health Risks*

Learn how your genetics can influence your risk for certain diseases.

Ancestry

Discover where your DNA is from out of 150+ regions worldwide - and more.

Wellness

Learn how your genes play a role in your well-being and lifestyle choices.

MEETS FDA REQUIREMENTS

Carrier Status*

If you are starting a family, find out if you are a carrier for certain inherited conditions.

Traits

Learn how your DNA influences your facial features, taste, smell and other traits.

23andMe genotyping service

23andMe

Your Wellness Result

Jamie, your genes predispose you to weigh about 6% less than average.

146 lbs

What is average?
For a woman who is **5'4"**

MEETS FDA REQUIREMENTS

Genetic Health Risks*

Learn how your genetics can influence your risk for certain diseases.

Ancestry

Discover where your DNA is from out of 150+ regions worldwide - and more.

Wellness

Learn how your genes play a role in your well-being and lifestyle choices.

MEETS FDA REQUIREMENTS

Carrier Status*

If you are starting a family, find out if you are a carrier for certain inherited conditions.

Traits

Learn how your DNA influences your facial features, taste, smell and other traits.

23andMe genotyping service

23andMe

Jamie, you do not have
the variant we tested.

0 variants detected
in the BLM Gene

MEETS FDA REQUIREMENTS

Genetic Health Risks*

Learn how your genetics can influence your risk for certain diseases.

Ancestry

Discover where your DNA is from out of 150+ regions worldwide - and more.

Wellness

Learn how your genes play a role in your well-being and lifestyle choices.

MEETS FDA REQUIREMENTS

Carrier Status*

If you are starting a family, find out if you are a carrier for certain inherited conditions.

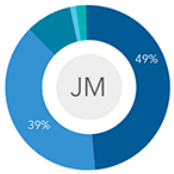
Traits

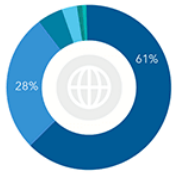
Learn how your DNA influences your facial features, taste, smell and other traits.

23andMe genotyping service

23andMe

Jamie, you are likely to have straight or wavy hair.





Jamie Maberry			All 23andMe customers		
	Big curls	2%	1%	Big curls	
	Small curls	1%	1%	Small curls	
	Very tight curls	<1%	4%	Very tight curls	
96%	Wavy	9%	5%	Wavy	89%
	Slight wavy	39%	28%	Slight wavy	

MEETS FDA REQUIREMENTS

Genetic Health Risks*

Learn how your genetics can influence your risk for certain diseases.

Ancestry

Discover where your DNA is from out of 150+ regions worldwide - and more.

Wellness

Learn how your genes play a role in your well-being and lifestyle choices.

MEETS FDA REQUIREMENTS

Carrier Status*

If you are starting a family, find out if you are a carrier for certain inherited conditions.

Traits

Learn how your DNA influences your facial features, taste, smell and other traits.

DNA-seq applications

1. De-novo genome sequencing
2. Complete genome resequencing
3. Reduced representation sequencing
4. **Targeted genome resequencing**
5. Paired-end sequencing
6. Metagenomic sequencing
7. Molecular barcoding

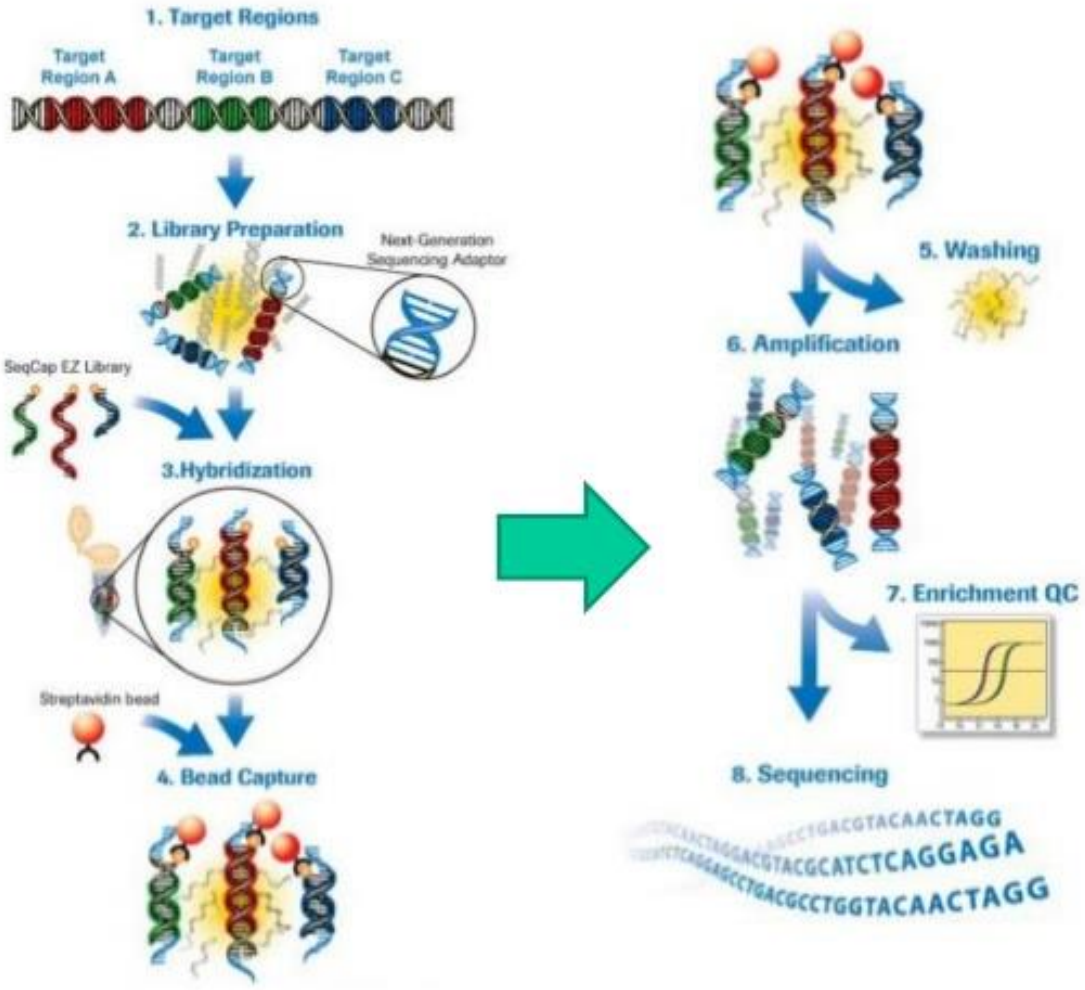
Targeted DNA sequencing

Targeted sequencing allows cheap and efficient sequencing of genes or regions of interest (polymorphism analysis, variant discovery, etc.):

- All exons in the genome (exome sequencing)
- Candidate regions for a disease
- Tumor genes
- Structural variants (CGH-arrays, optical mapping)

Exome sequencing – Targeted enrichment

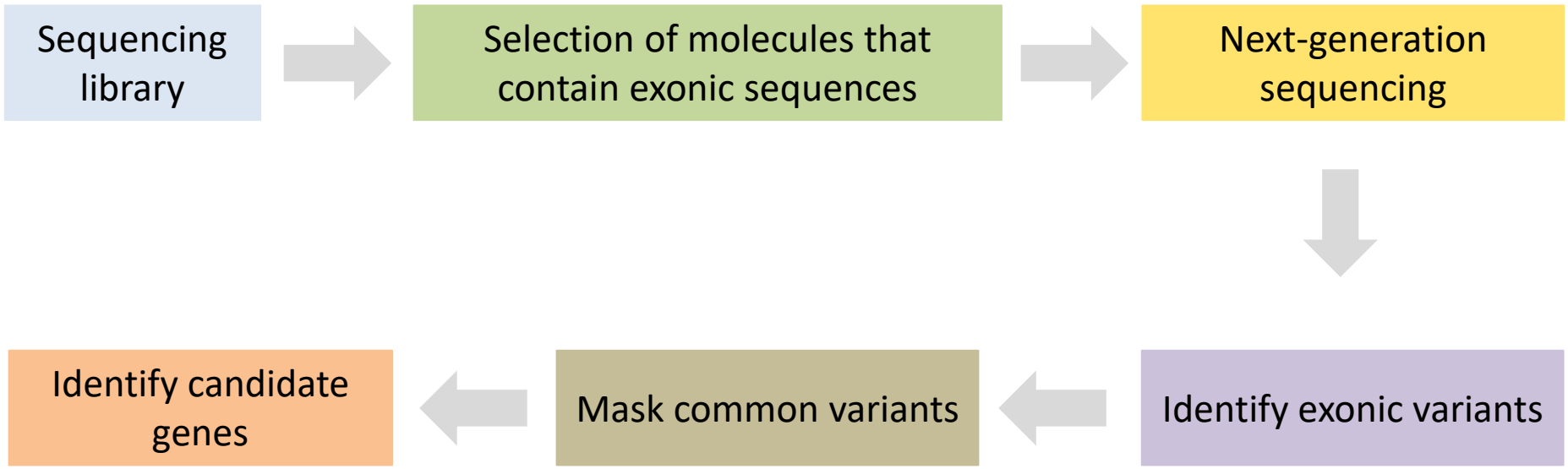
Focusing NGS effort on predefined targets :
« Target Enrichment » Technology (Capture Beads)



Exome sequencing

Sequencing exons of a human individual's genome

Application Determine causal gene of a genetic disease



Exome sequencing

- Millions of exomes have been sequenced already!
- Almost routine diagnostic method for rare mendelian diseases, cancer, …

Table 2. Mendelian disease gene identifications by exome or genome sequencing

Disorder	Inheritance	Gene identified	Scope	References
Congenital chloride diarrhea	Recessive	*SLC26A3*	Exome	Choi *et al.* [16]
Miller syndrome	Recessive	*DHODH*	Exome	Ng *et al.* [14]
Charcot-Marie-Tooth neuropathy	Recessive	*SH3TC2*	Genome	Lupski *et al.* [20]
Metachondromatosis	Dominant	*PTPN11*	Genome	Sobreira *et al.* [23]
Schinzel-Giedion syndrome	Dominant	*SETBP1*	Exome	Hoischen *et al.* [29]
Nonsyndromic hearing loss	Recessive	*GPSM2*	Exome	Walsh *et al.* [69]
Perrault syndrome	Recessive	*HSD17B4*	Exome	Pierce *et al.* [25]
Hyperphosphatasia mental retardation syndrome	Recessive	*PIGV*	Exome	Krawitz *et al.* [68]
Sensenbrenner syndrome	Recessive	*WDR35*	Exome	Gilissen *et al.* [26]
Cerebral cortical malformations	Recessive	*WDR62*	Exome	Bilguvar *et al.* [70]
Kaposi sarcoma	Recessive	*STIM1*	Exome	Byun *et al.* [71]
Spinocerebellar ataxia	Dominant	*TGM6*	Exome	Wang *et al.* [72]
Combined hypolipidemia	Recessive	*ANGPTL3*	Exome	Musunuru *et al.* [40]
Complex I deficiency	Recessive	*ACAD9*	Exome	Haack *et al.* [52]
Autoimmune lymphoproliferative syndrome	Recessive	*FADD*	Exome	Bolze *et al.* [73]
Amyotrophic lateral sclerosis	Dominant	*VCP*	Exome	Johnson *et al.* [74]
Nonsyndromic mental retardation	Dominant	Various	Exome	Vissers *et al.* [31]
Kabuki syndrome	Dominant	*MLL2*	Exome	Ng *et al.* [30]
Inflammatory bowel disease	Dominant	*XIAP*	Exome	Worthey *et al.* [18]
Nonsyndromic mental retardation	Recessive	*TECR*	Exome	Caliskan *et al.* [75]
Retinitis pigmentosa	Recessive	*DHDDS*	Exome	Züchner *et al.* [56]
Osteogenesis imperfecta	Recessive	*SERPINF1*	Exome	Becker *et al.* [53]
Dilated cardiomyopathy	Dominant	*BAG3*	Exome	Norton *et al.* [24]
Hajdu-Cheney syndrome	Dominant	*NOTCH2*	Exome	Simpson *et al.* [76]
Hajdu-Cheney syndrome	Dominant	*NOTCH2*	Exome	Isidor *et al.* [77]
Skeletal dysplasia	Recessive	*POP1*	Exome	Glazov *et al.* [78]
Amelogenesis	Recessive	*FAM20A*	Exome	O'Sullivan *et al.* [80]
Chondrodysplasia and abnormal joint development	Recessive	*IMPAD1*	Exome	Vissers *et al.* [80]
Progeroid syndrome	Recessive	*BANF1*	Exome	Puente *et al.* [81]
Infantile mitochondrial cardiomyopathy	Recessive	*AARS2*	Exome	Götz *et al.* [82]
Sensory neuropathy with dementia and hearing loss	Dominant	*DNMT1*	Exome	Klein *et al.* [49]
Autism	Dominant	Various	Exome	O'Roak *et al.* [32]

Gilissen *et al.* (2011)
Genome Biology 12: 228

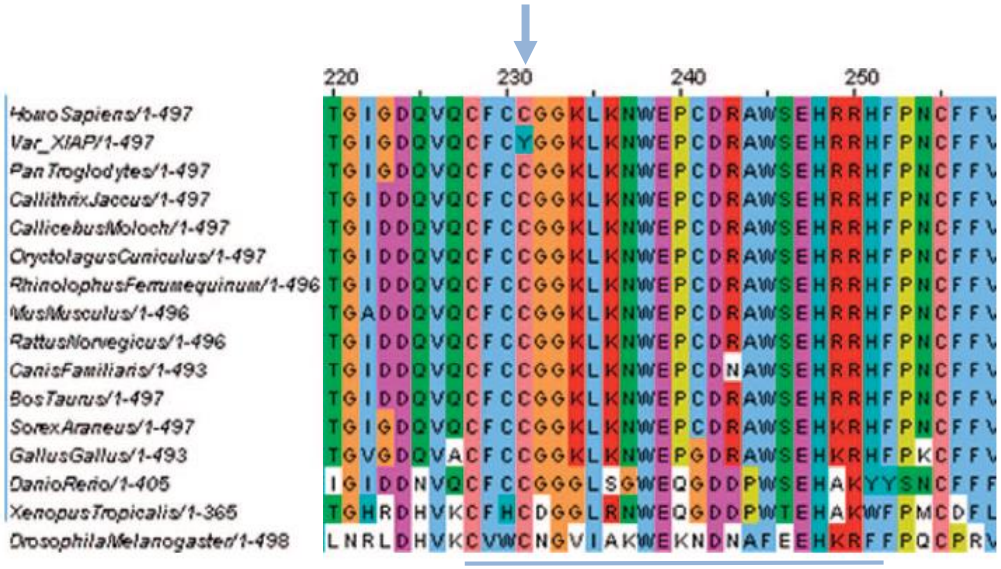
Exome clinical applications

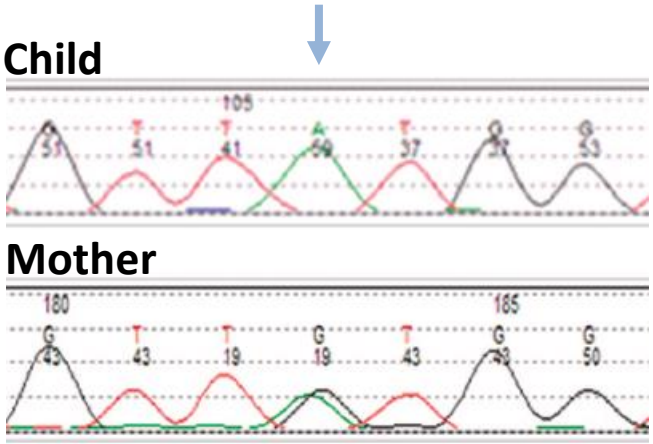
Nicholas Volker

Intractable inflammatory bowel disease

Analysis of 16,124 exome variants in Volker's exome identified a novel mutation in the X-linked inhibitor of apoptosis (*XIAP*) gene in hemizigosis

Allogenic hematopoietic progenitor cell transplant treated *XIAP* deficiency

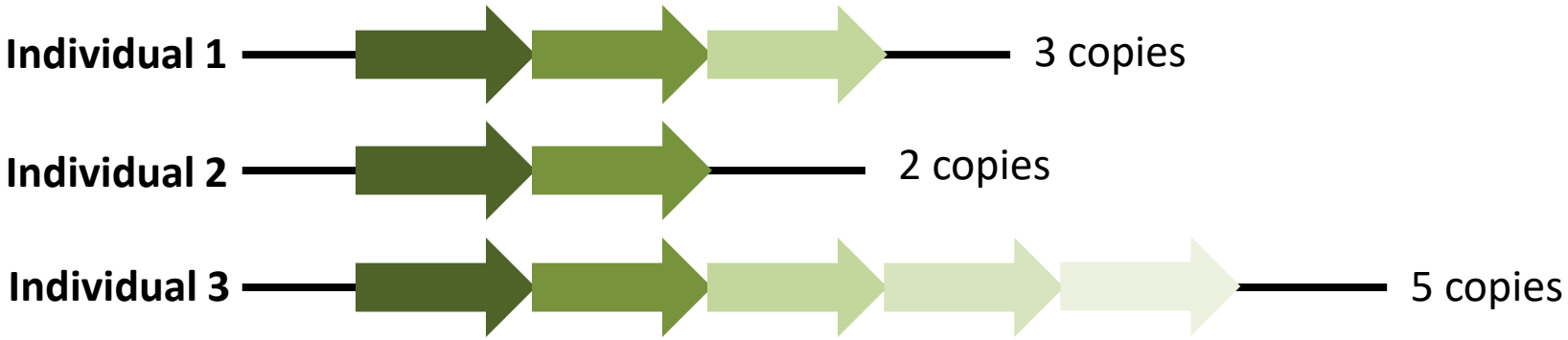




Worthey *et al*. (2011) *Genet Med*. 13:255-62.

Targeting copy number variants (CNV)

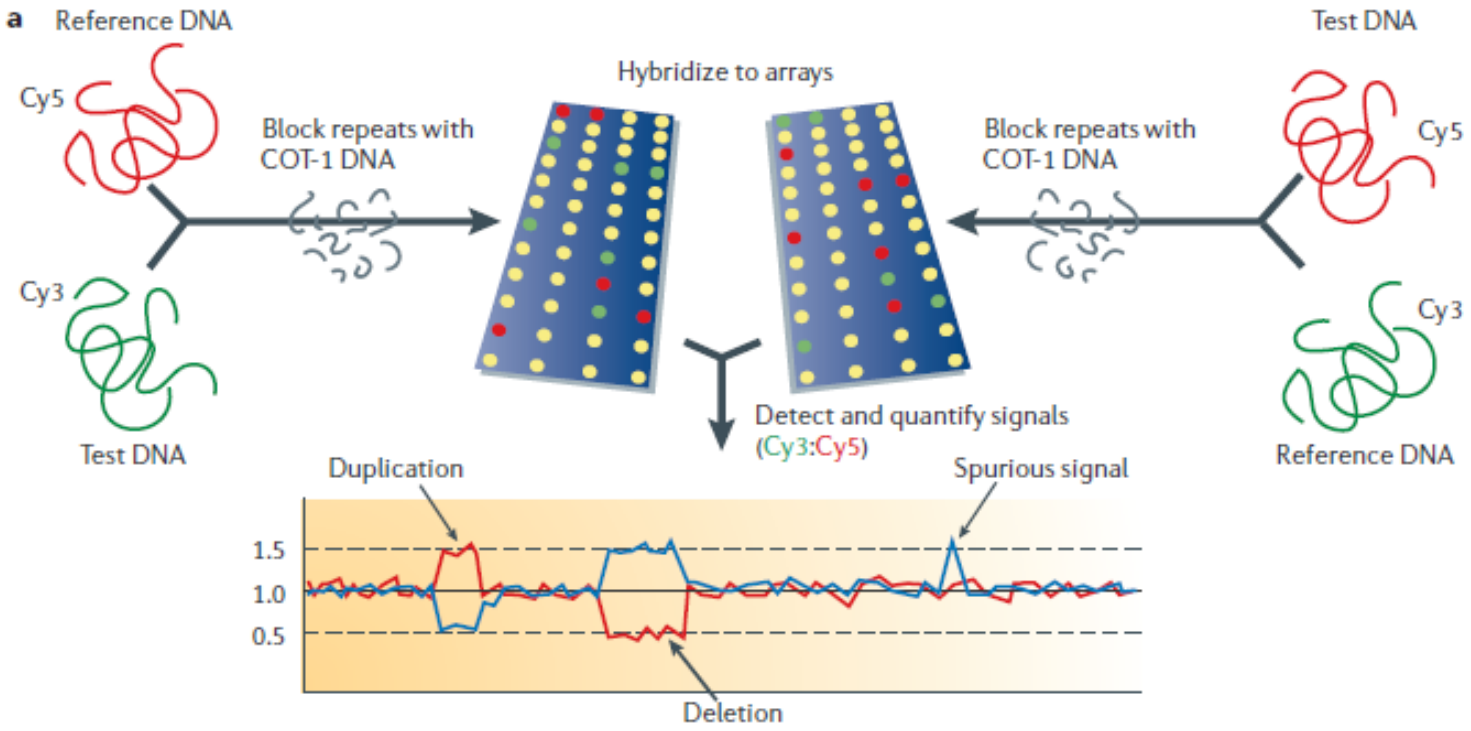
CNV DNA segment that is present in a variable number of copies in different genomes



- 8,599 validated CNVs span a total of 112.7 Mb (3.7% of the genome)
- Detected CNVs vary in size between 443 bp and 1.28 Mb (average size 2.9 kb) and might include genes
- Two random genomes differ in copy-number for 1,098 CNVs
- Some CNVs do not seem to influence the phenotype, but many others have been associated to disease

Conrad *et al*. (2010) *Nature* 464: 704-712

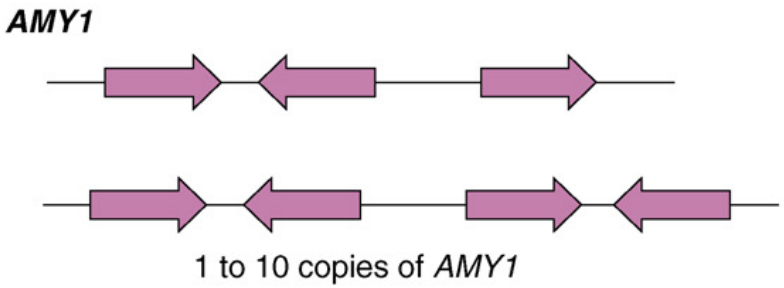
Comparative genomic hybridization (CGH) arrays



The ratio between the fluorescence intensity of test and reference DNAs indicate the difference in the number of copies in certain regions of the genome

Figure 2. Feuk *et al*. (2006) *Nature Reviews Genetics* 7: 85-97

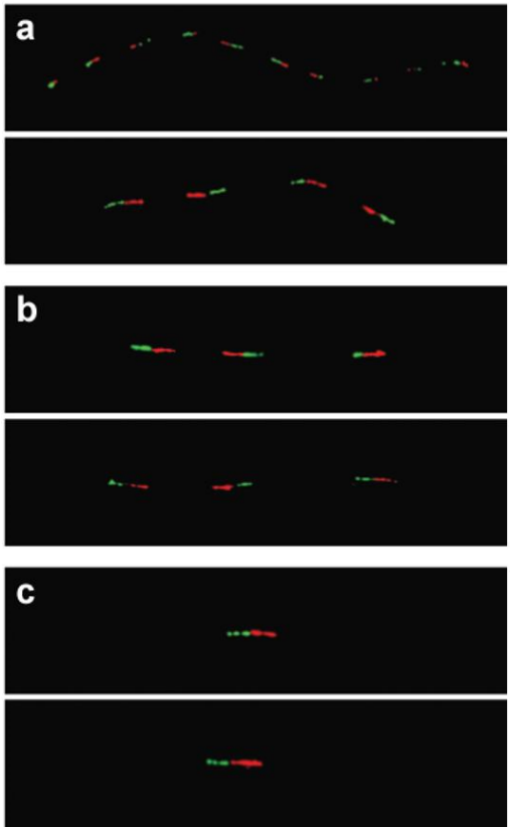
CNV example: the amilase gene

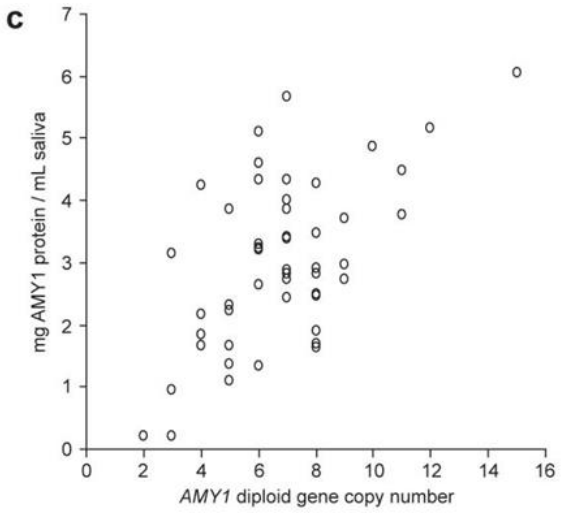


Japanese
High starch consumption
(14 copies)

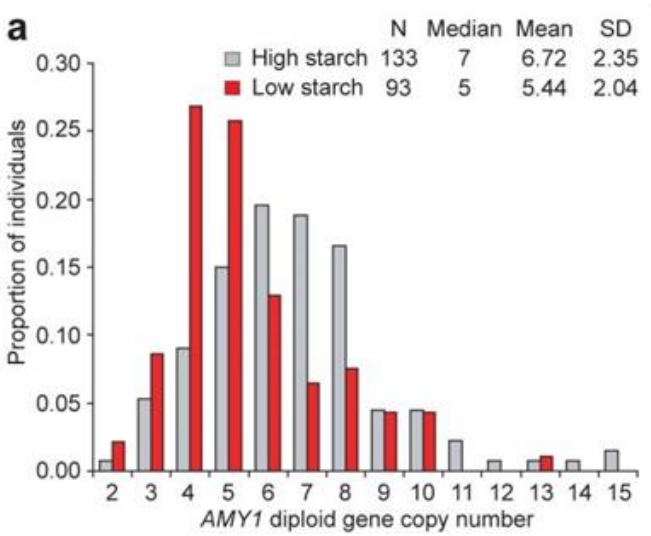
African
Low starch consumption
(6 copies)

Chimpanzee
Low starch consumption
(2 copies)





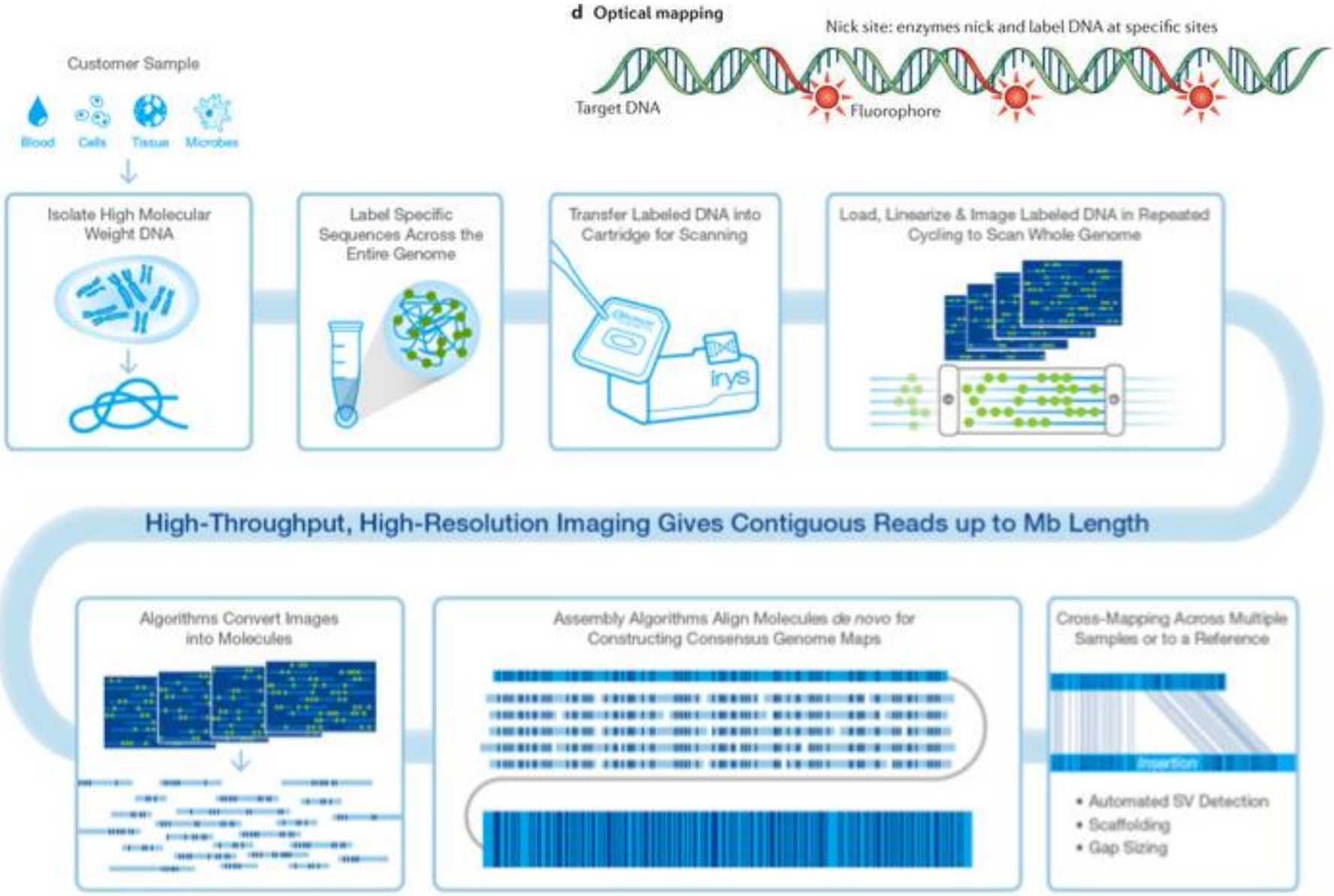
The amount of amylase protein in saliva is proportional to the number of copies of the *AMY1* gene.



Individuals living in populations with a high starch diet have on average more copies of *AMY1* than those with poor starch diets.

Figures 1, 2 and 3. Perry *et al*. (2007) *Nature Genetics* 39: 1256 - 1260

Optical mapping



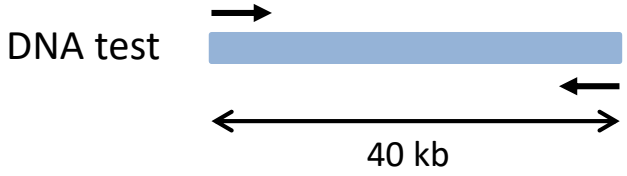
DNA-seq applications

1. De-novo genome sequencing
2. Complete genome resequencing
3. Reduced representation sequencing
4. Targeted genome resequencing
5. **Paired-end sequencing**
6. Metagenomic sequencing
7. Molecular barcoding

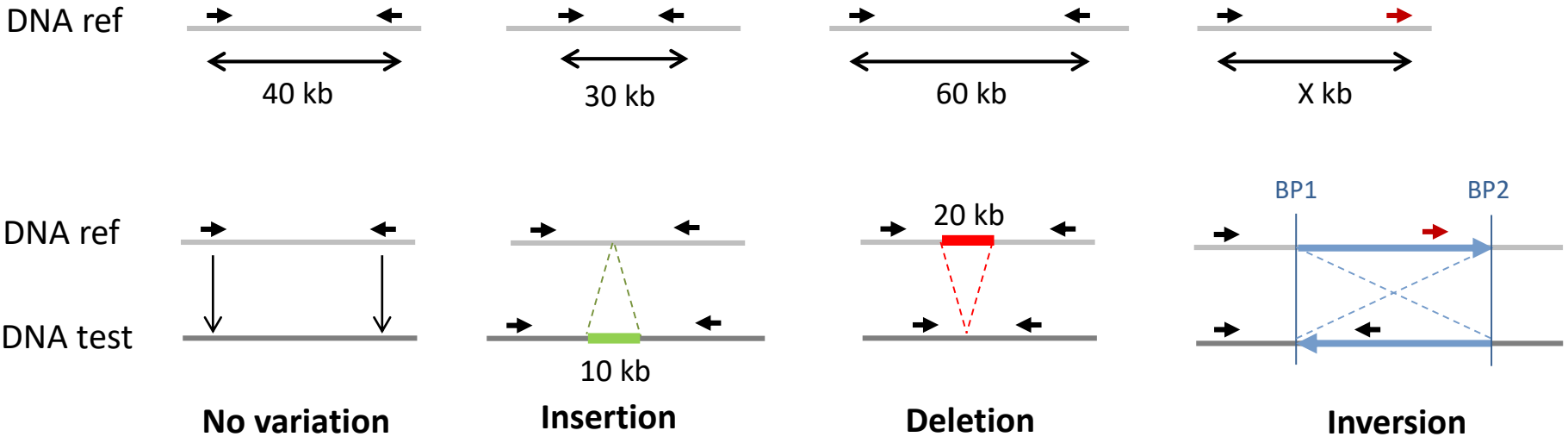
Paired-end mapping (PEM)

1. Construction of a genomic library of fragments of a certain size (DNA test)

2. Sequencing of the ends of these fragments (paired-end sequencing)



3. Mapping of paired-ends to a reference genome

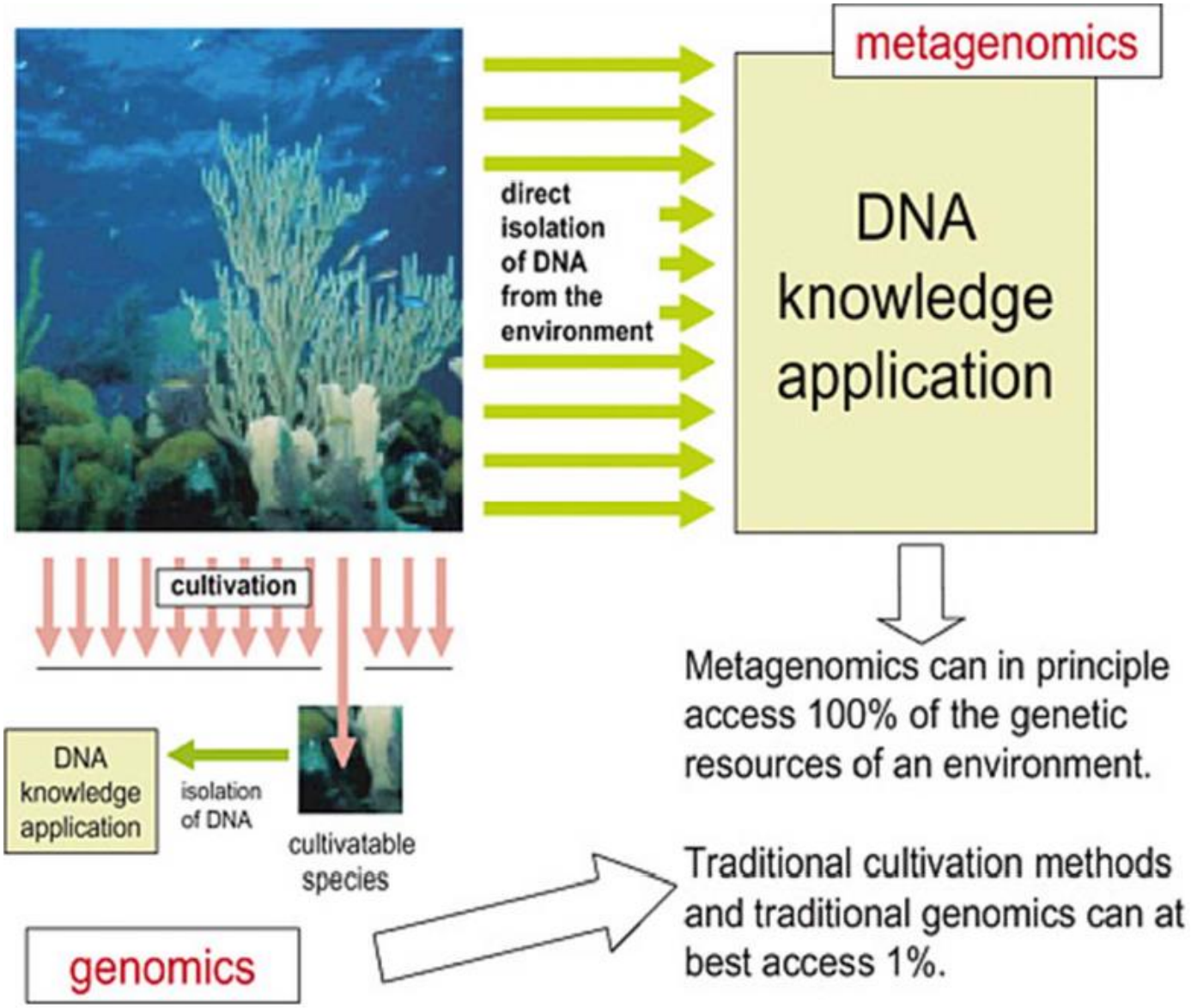


DNA-seq applications

1. De-novo genome sequencing
2. Complete genome resequencing
3. Reduced representation sequencing
4. Targeted genome resequencing
5. Paired-end sequencing
6. **Metagenomic sequencing**
7. Molecular barcoding

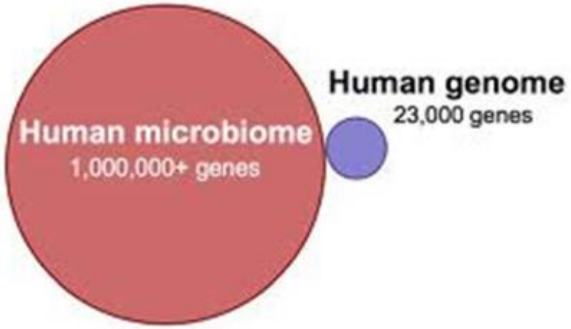
Metagenomics

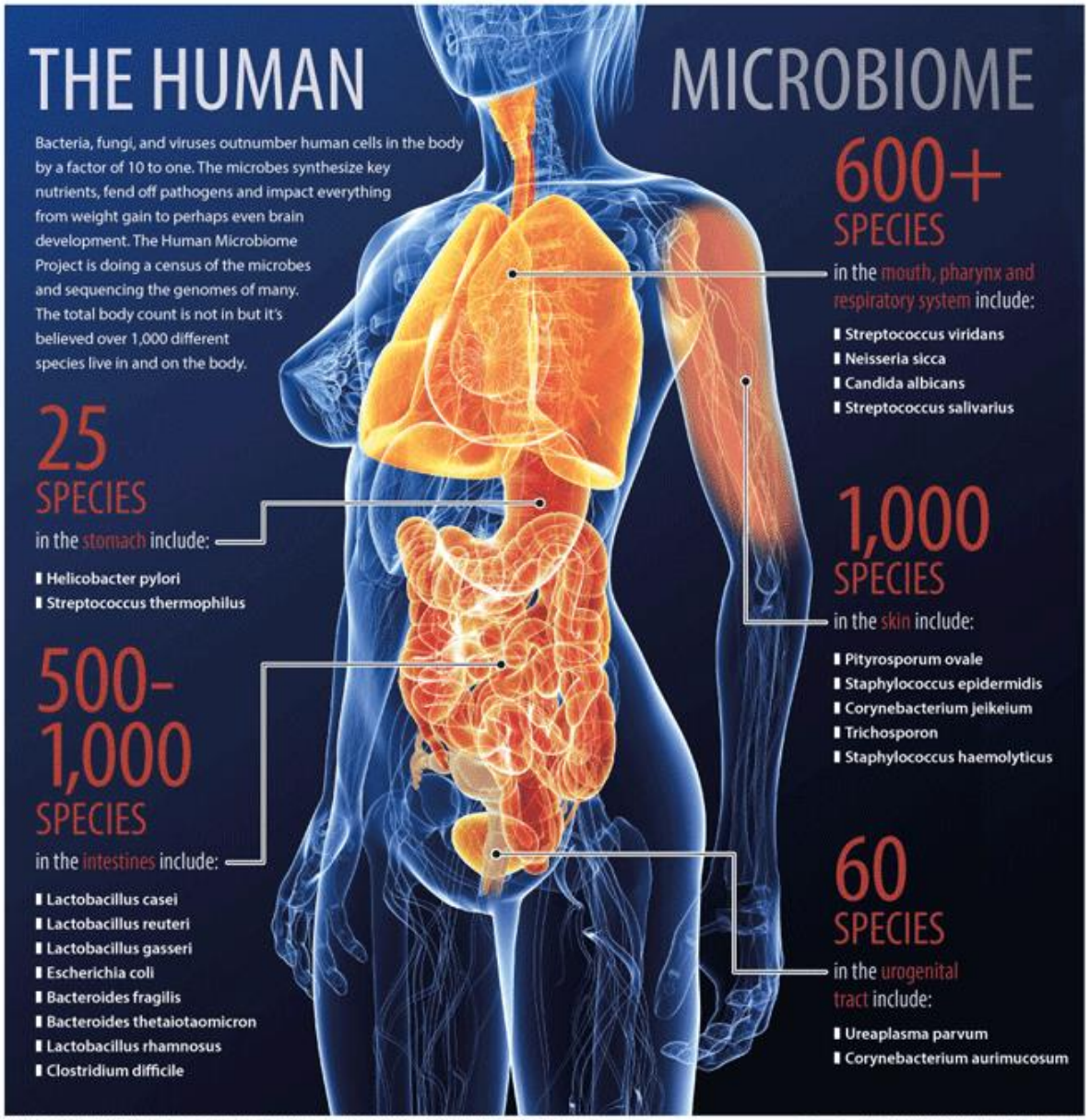
Sequencing of genomes directly from mixed environmental samples to obtain unbiased information of all genes of all members of the community



The human microbiome project

We are 98% non-human!

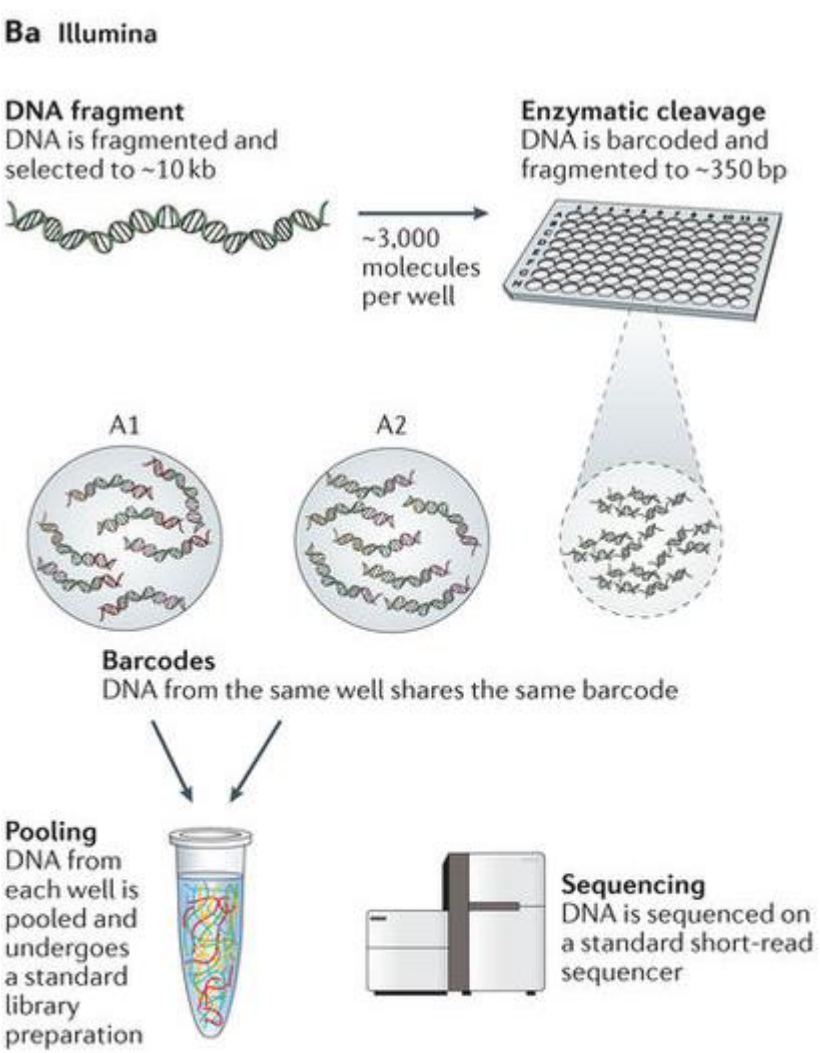


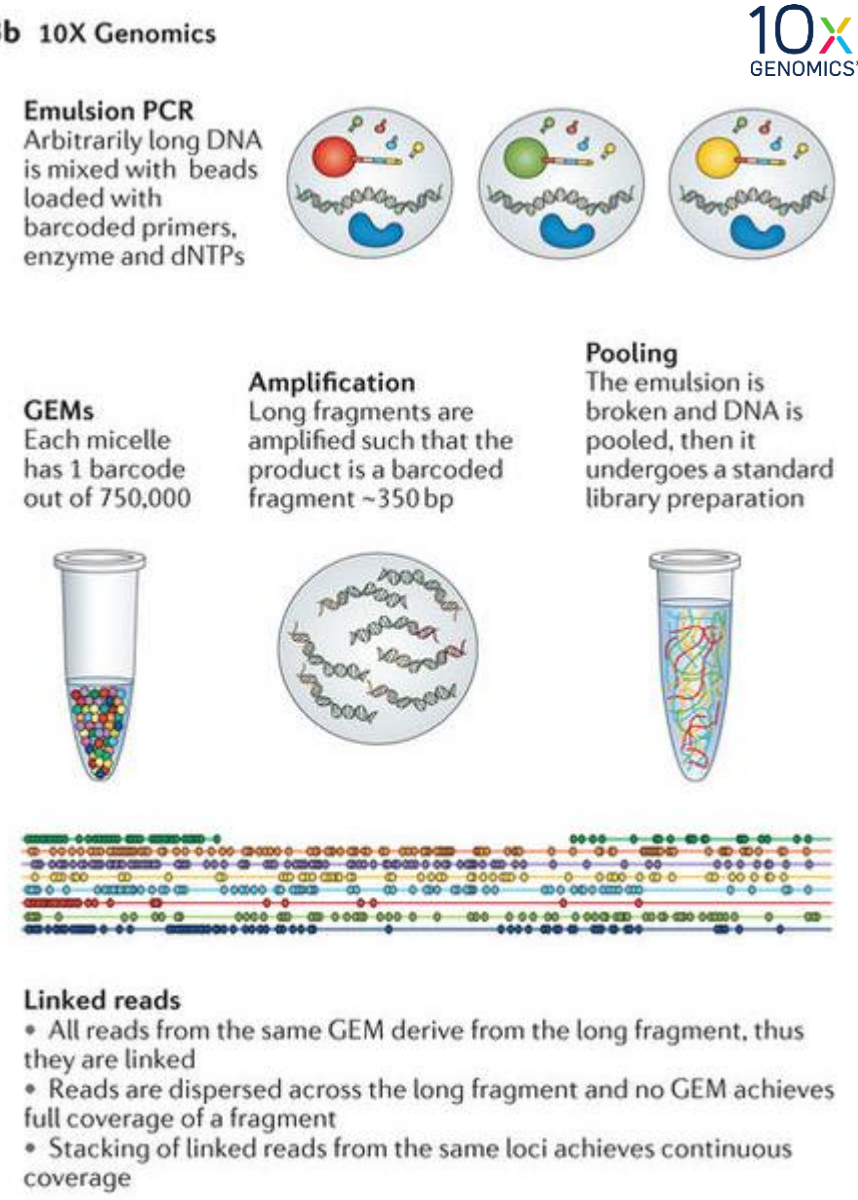


DNA-seq applications

1. De-novo genome sequencing
2. Complete genome resequencing
3. Reduced representation sequencing
4. Targeted genome resequencing
5. Paired-end sequencing
6. Metagenomic sequencing
7. **Molecular barcoding**

Synthetic long-read sequencing





Applications of synthetic long-read sequencing

- Study repetitive and complex regions of the genome (gap closure and extension)
- Call and phase SVs across >10 Mb haplotype blocks, even in regions inaccessible to short-read sequencers
- Pool multiple samples together