Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

1er Cuat. **2020**

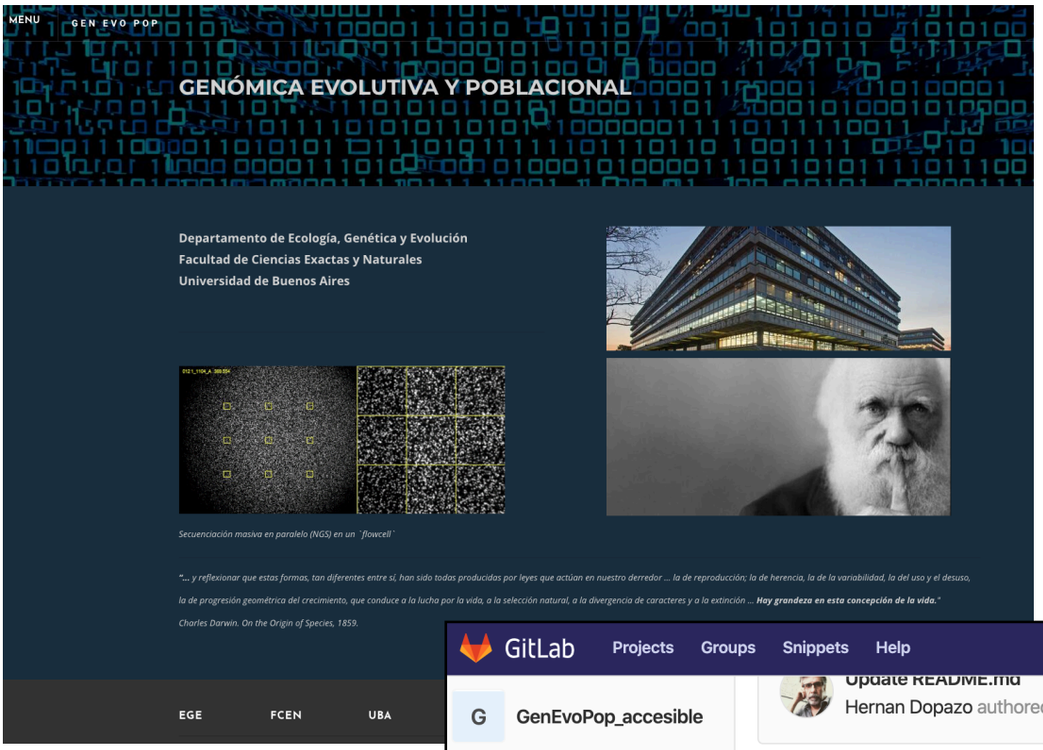
Genómica Evolutiva y Poblacional

Introducción al Curso

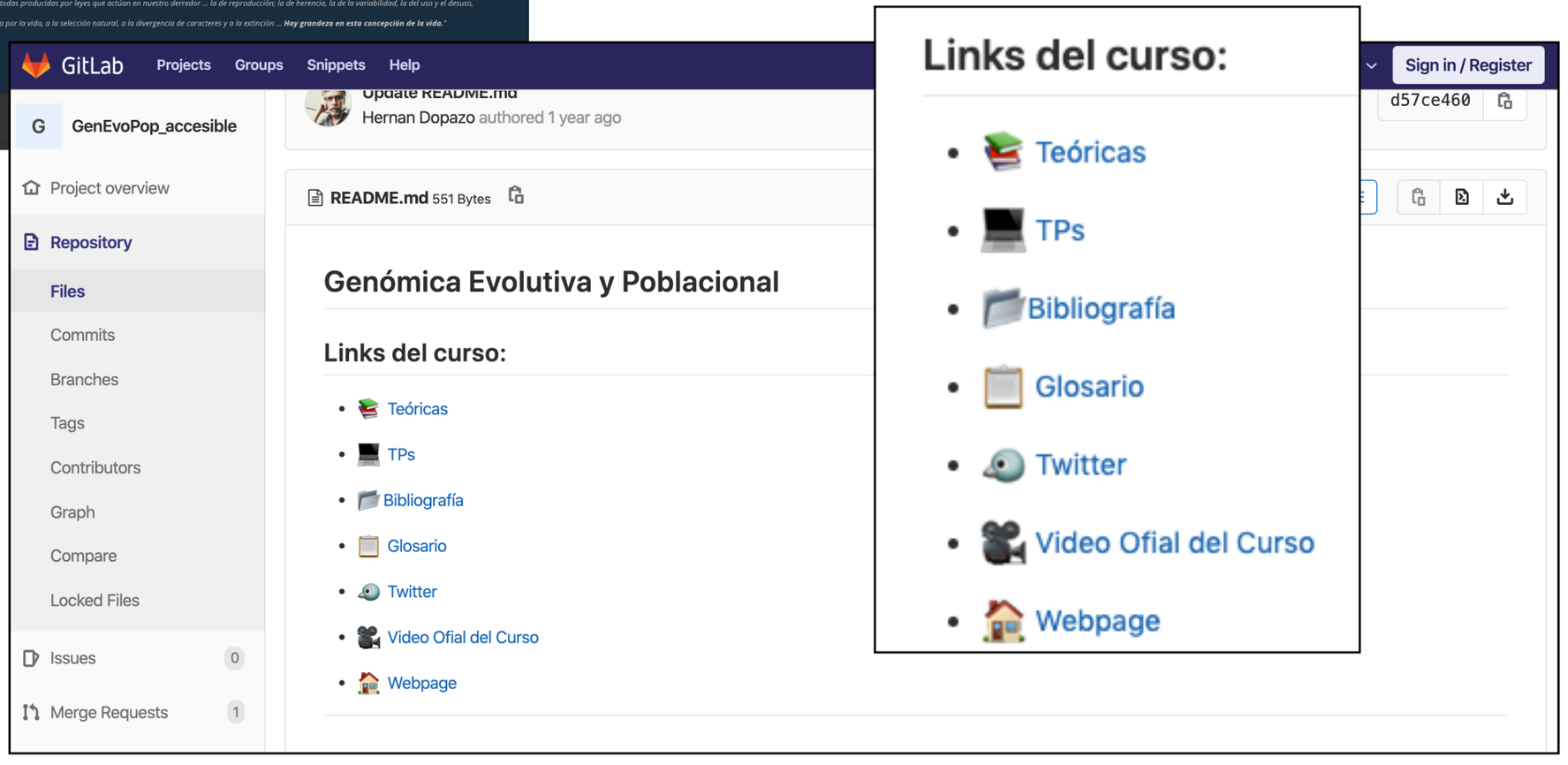


Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

genevopop.net/grado







OBJETIVO

Que el alumno conozca los principales avances técnicos y conceptuales de la genómica y se enfrente a los desafíos prácticos que impone el análisis masivo de estos datos, en un marco de referencia mayoritariamente evolutivo y poblacional.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Página Web del Curso

- ## Programa de Teóricas

  - ### Módulo I - Genómica Humana, Funcional y Comparativa

    - **Tema 1. Genomas Humanos, Resecuenciación y Tecnologías**
    - **Tema 2. Genómica Funcional**
    - **Tema 3. Genómica Comparativa**

  - ### Módulo II - Variabilidad Genética y Estructura Poblacional

    - **Tema 4. Variabilidad Genómica de Poblaciones Humanas**
    - **Tema 5. Demografía y Estructura Poblacional Humana**

  - ### Módulo III - Poblamiento Continental y Selección Natural

    - **Tema 6. Poblamiento Continental**
    - **Tema 7. Neutralidad y Selección en Genomas Completos**

  - ### Módulo IV - Genómica y Medicina de Precisión

    - **Tema 8. Genómica de Enfermedades**
    - **Tema 9. Medicina de Precisión I**
    - **Tema 10. Medicina de Precisión II**

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Página Web del Curso

- ## Programa de Trabajos Prácticos

  - **TP 1. Exploración de Datos Genómicos en Genome-Browsers**
  - **TP 2. Introducción al Sistema Linux y su Uso por Consola**
  - **TP 3. Control de Calidad y Preprocesamiento de Lecturas NGS**
  - **TP 4. Ensamblado Genómico**
  - **TP 5. Análisis de Datos de Expresión Génica. RNAseq**

  - **TP 6. Análisis de Variantes Estructurales**
  - **TP 7. Análisis de Datos de Genómica de Poblaciones Humanas**
  - **TP 8. Análisis de Estructura Poblacional**
  - **TP 9. Análisis de Selección en el Genoma Humano**
  - **TP 10. GWAS y Anotación de Variantes Genéticas**

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

JTP y Colaboradores



JTP EGE-FCEN
Dra Francisca Almeida



Ayt 1 EGE-FCEN
Lic Johnma Rondon



JTP EGE-FCEN
Lic Nicolas Moreyra

JEREMIAS ZUBRZYCKI



DIRECTOR LABORATORIO BIOCÓDICES

Dr. Biotecnología

Tecnologías Genómicas

JUAN MANUEL BERROS

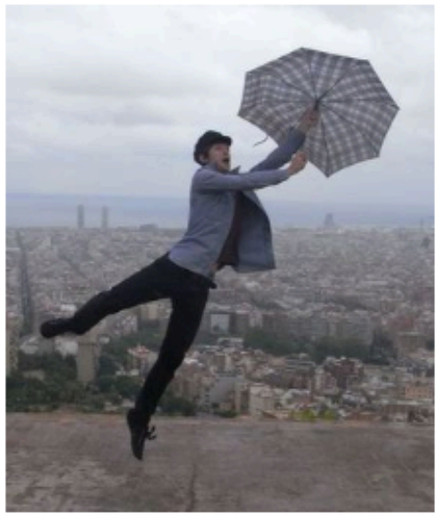


BECARIO DOCTORAL BIOCÓDICES-CONICET

Lic. Biología

Bioinformática y Medicina Genómica

PIERRE LUISI



BECARIO POSTDOCTORAL CONICET

Dr. Biología

Genómica Poblacional y Epidemiología

DIEGO DE PANIS



BECARIO POSTDOC CONICET

Dr. Biología

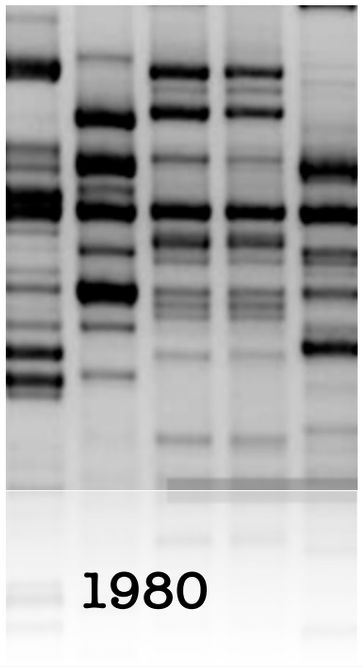
Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Technological Paradigm Shift

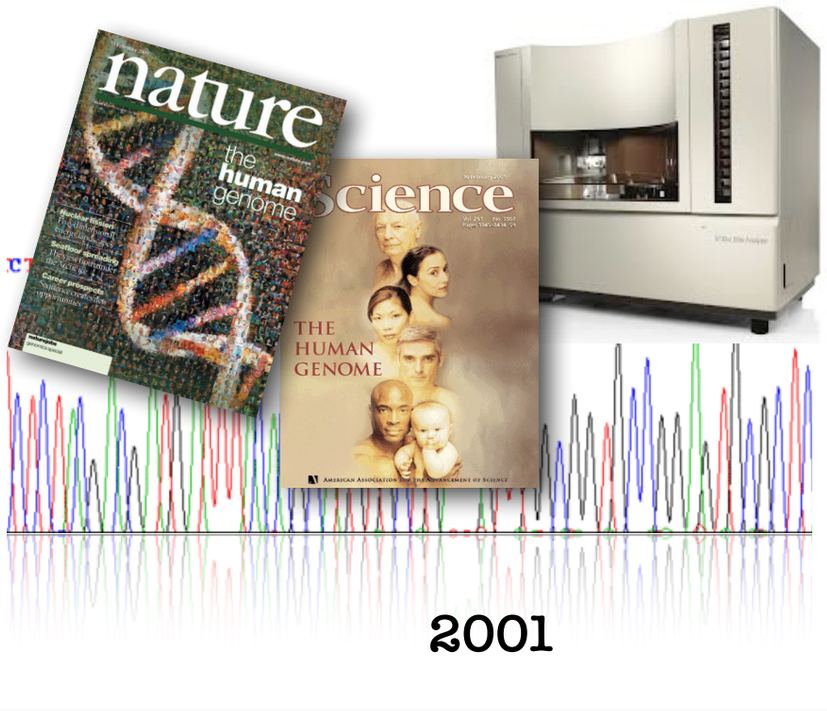
Produccion Artesanal de Datos Genéticos >> Producción Industrial de Datos Genómicos



1953



1980



2001





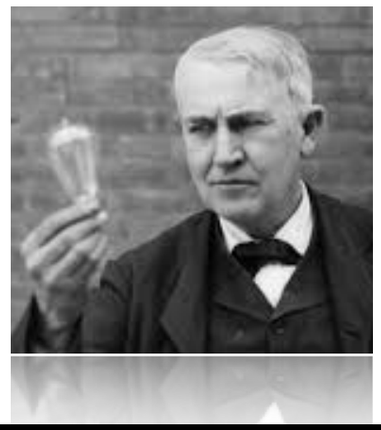


2019

Miniaturización, Paralelismo, Automatzación, Producción Masiva de Datos, Big Data











Biotecnología adquirió características de la 1ra, 2da, 3ra y 4ta Revolución Industrial

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Human Genomics - Main Topics

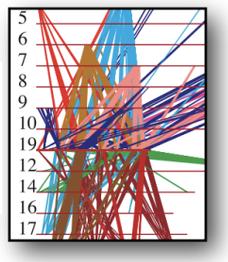
Proyecto Genoma Humano



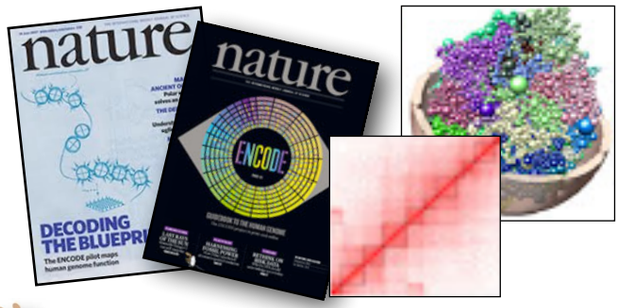
GWAS



Variación Estructural



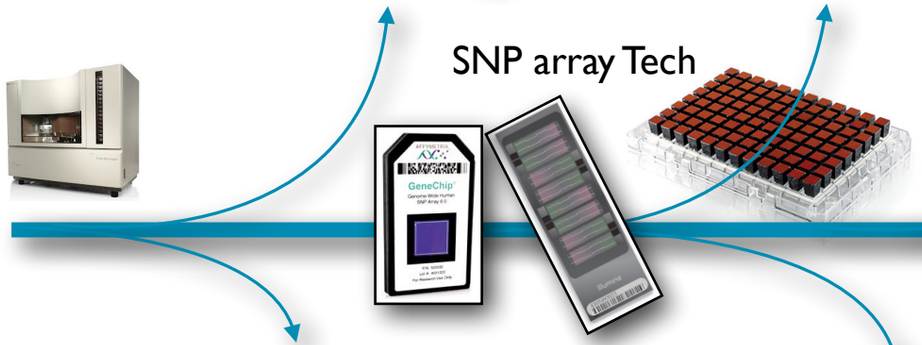
Transcriptómica, Genómica Funcional y Estructural



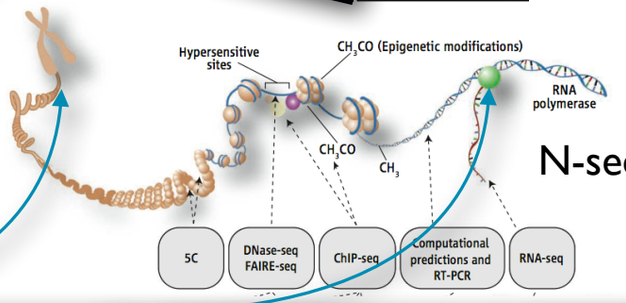
Genomas de Enfermedades y Microbiomas



SNP array Tech



N-seq Tech





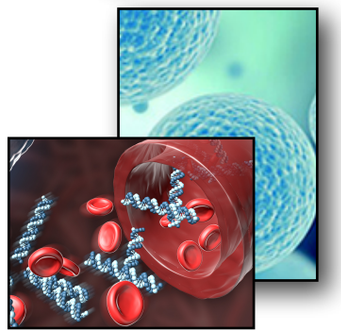
Estudios Poblacionales



Genómica Comparativa Paleogenómica



DNA Libre de Células, Genómica de Células Únicas



Sistema de Salud Consorcios de Buenas Prácticas



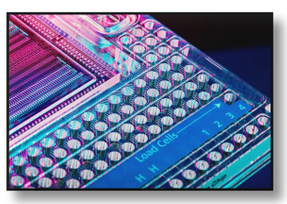






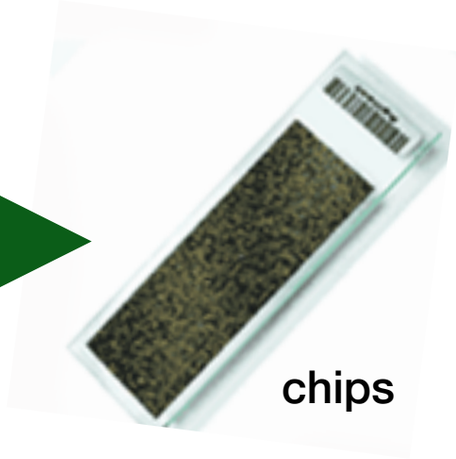
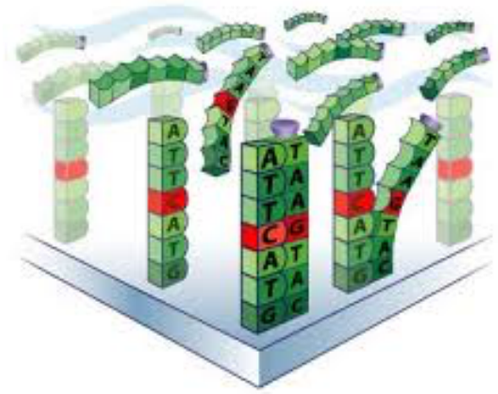
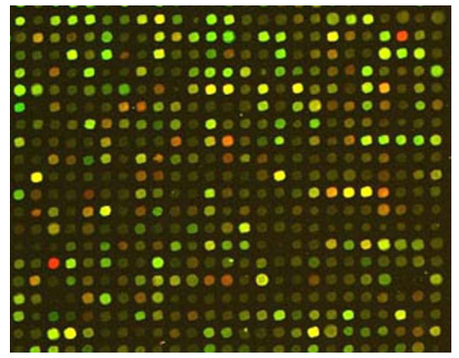
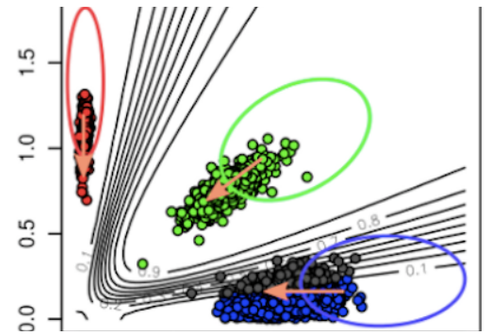
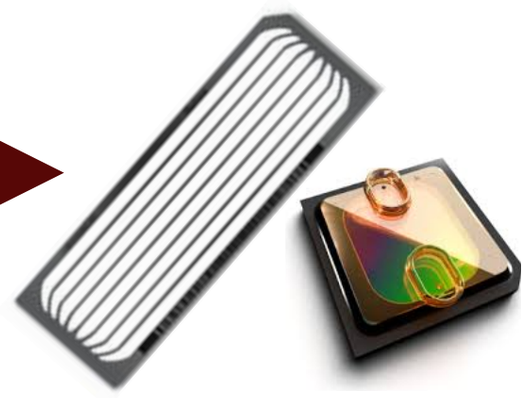
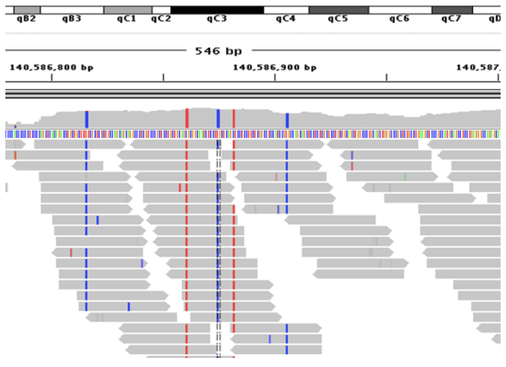
Tecnología





Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Genomics - Technologies

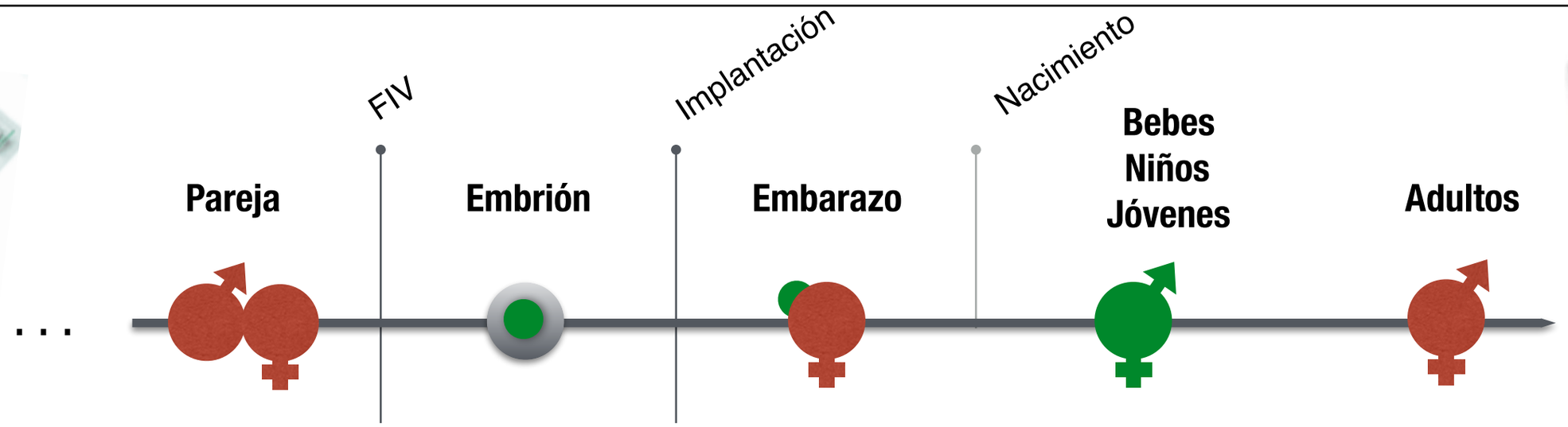
TECNOLOGÍA	SOPORTE	BASE BIOQUÍMICA	TIPO DE DATO
Arrays – CGH	 chips	 Hibridación Competitiva Sondas	 Analógicos Dispersos
Arrays – SNPs	 chips	 Extensión Específica Alelos	 Digitales Dispersos
NGS – cfDNA, RNAseq, Paneles, Exomas, WGS	 Flowcells & Chips	 Secuenciación Masiva en Paralelo	 Digitales Secuenciales

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

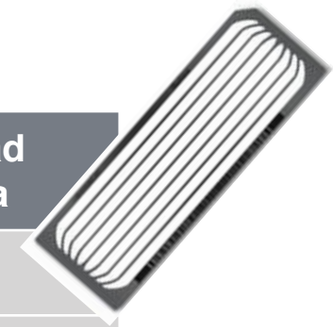
Medical Genomics







Tecnología / Test		Portadores Heterocigotas	Preimplantación	Prenatal	Postnatal	Enfermedad Específica
Arrays	CGH			+++	+++	
	SNPs	+++		+	++	++
NGS	cfDNA		+++	+++		+++
	Aleatorio		+++			
	Panel	+++	++		+++	+++
	Exoma				+++	+++
	Genoma		+		+	+
Productos		Carrier-code	PGS/PGD Karyomapping	Prenatal-code NIPT	Síndromes Polimalformativos, Autismo, etc	Onco- Risk/code Cardio/Neuro-code Nutri-code
		Genómica Reproductiva			Genómica Clínica	

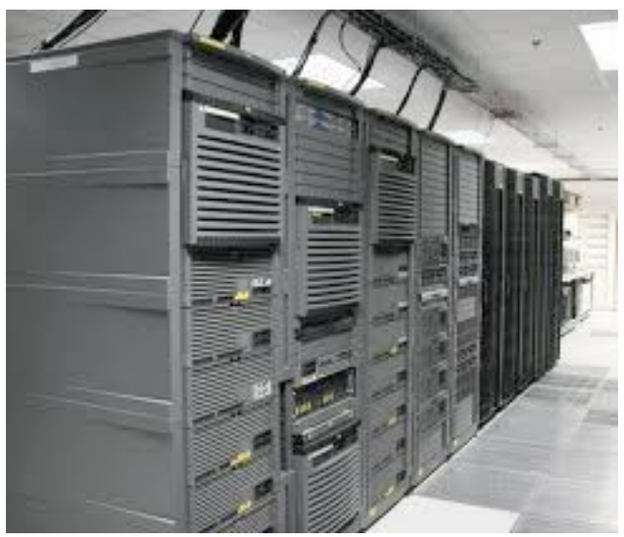




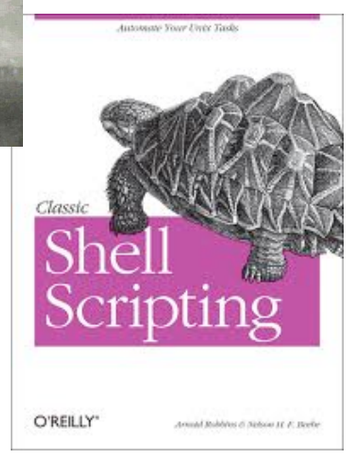


Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Programming-Statistical Toolkit





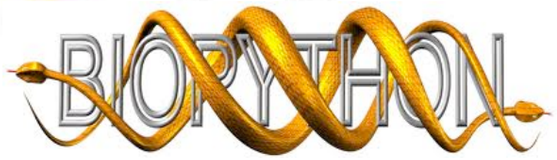




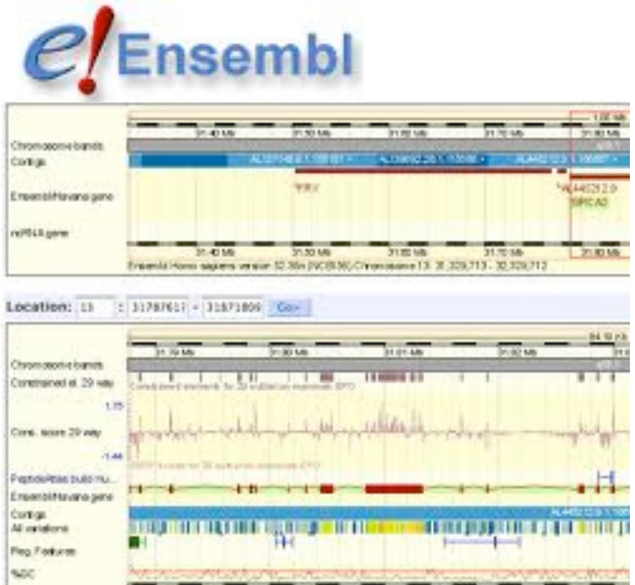


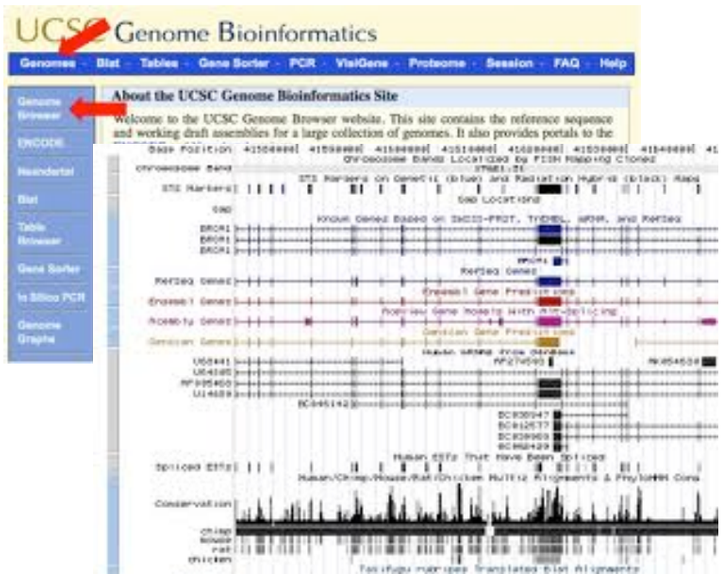
python

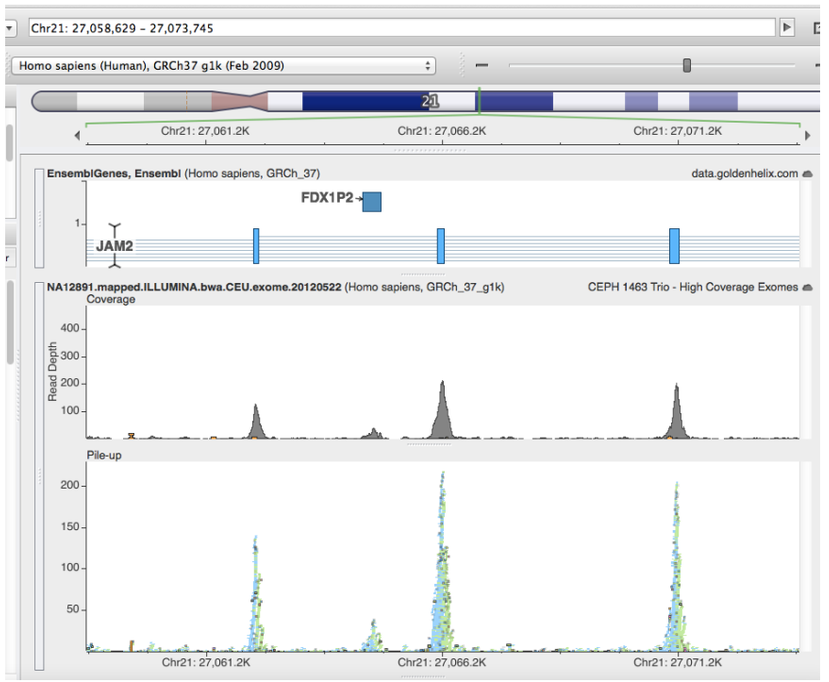
SAM tools





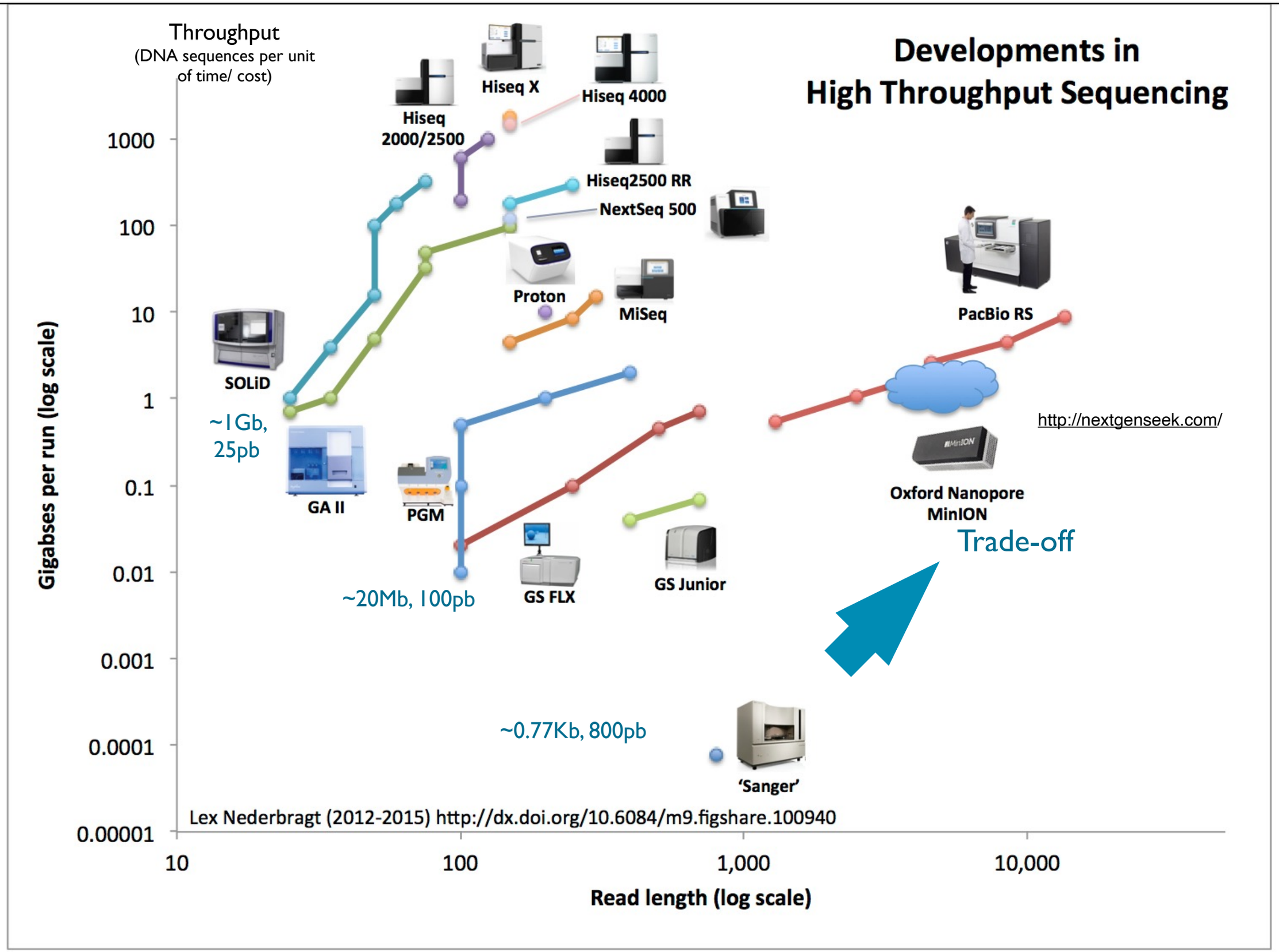






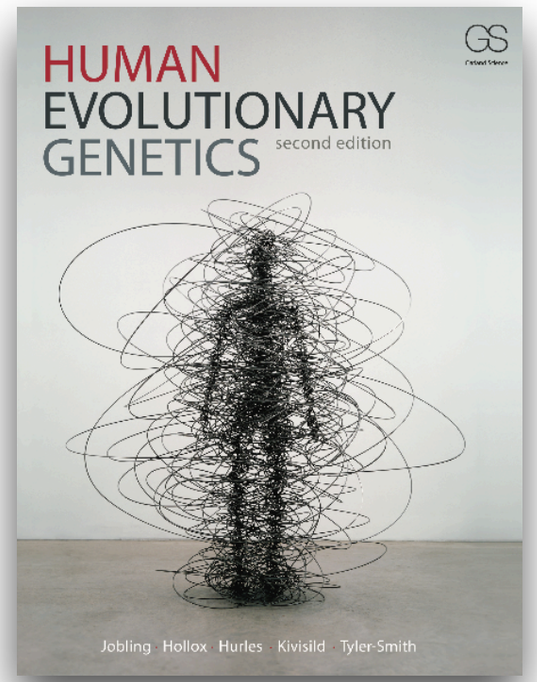
Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

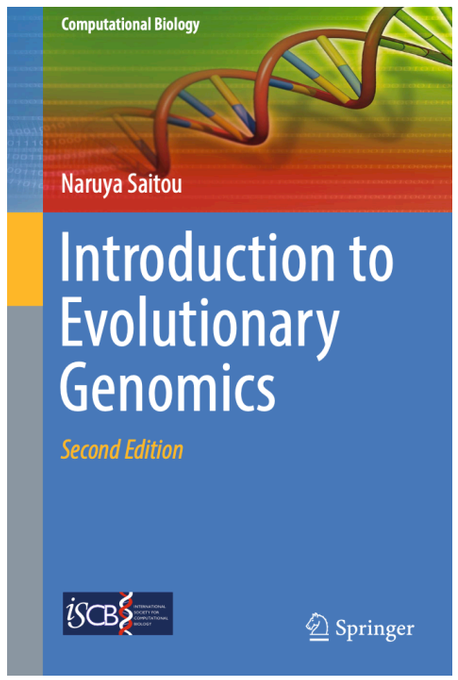
Evolution of NGS Technology

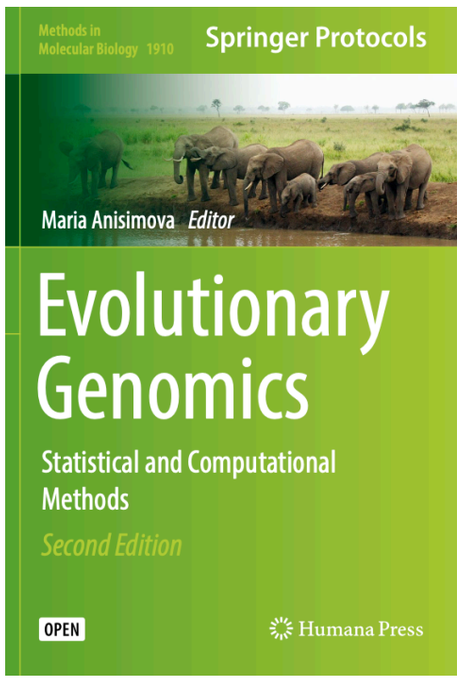


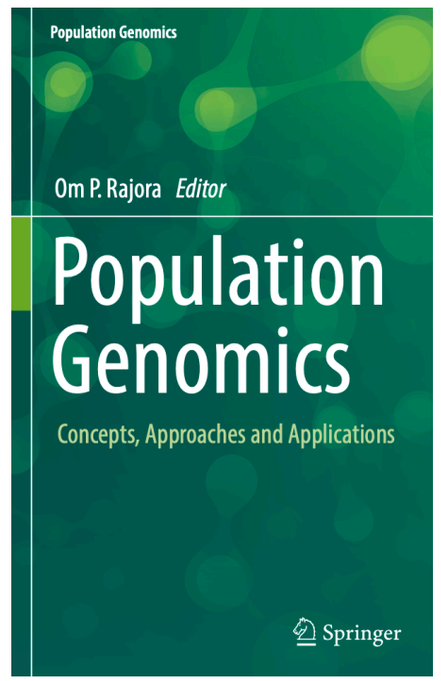
Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN. CONICET
www.genevopop.net
@hernandopazo

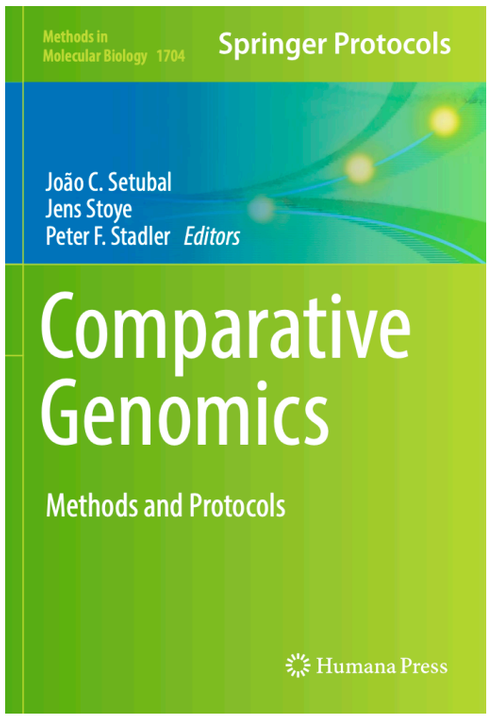
http://gen.lib.rus.ec/

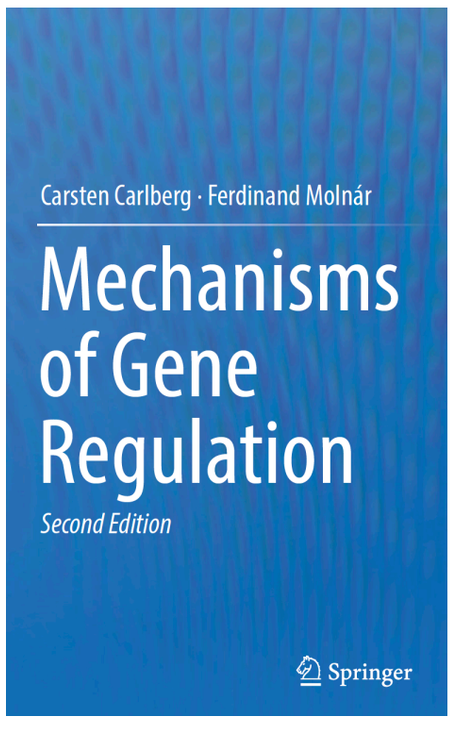


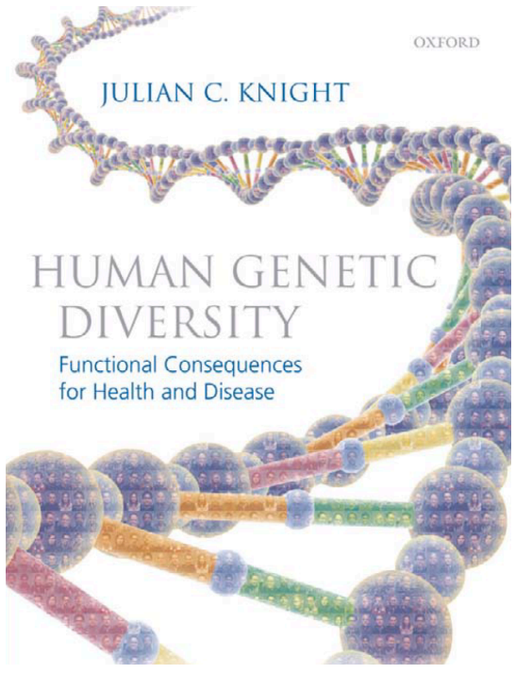


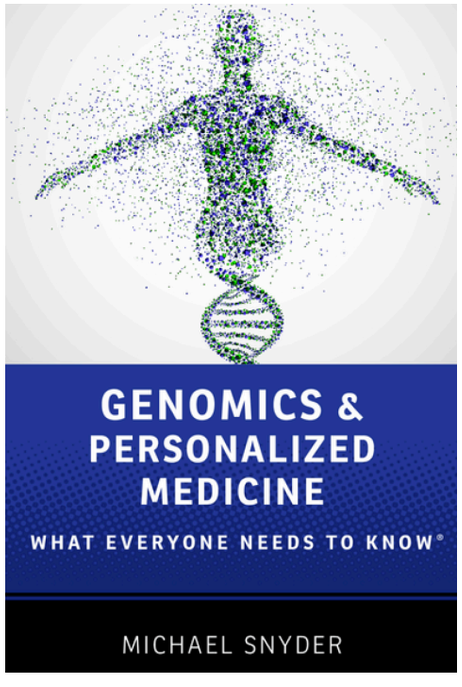












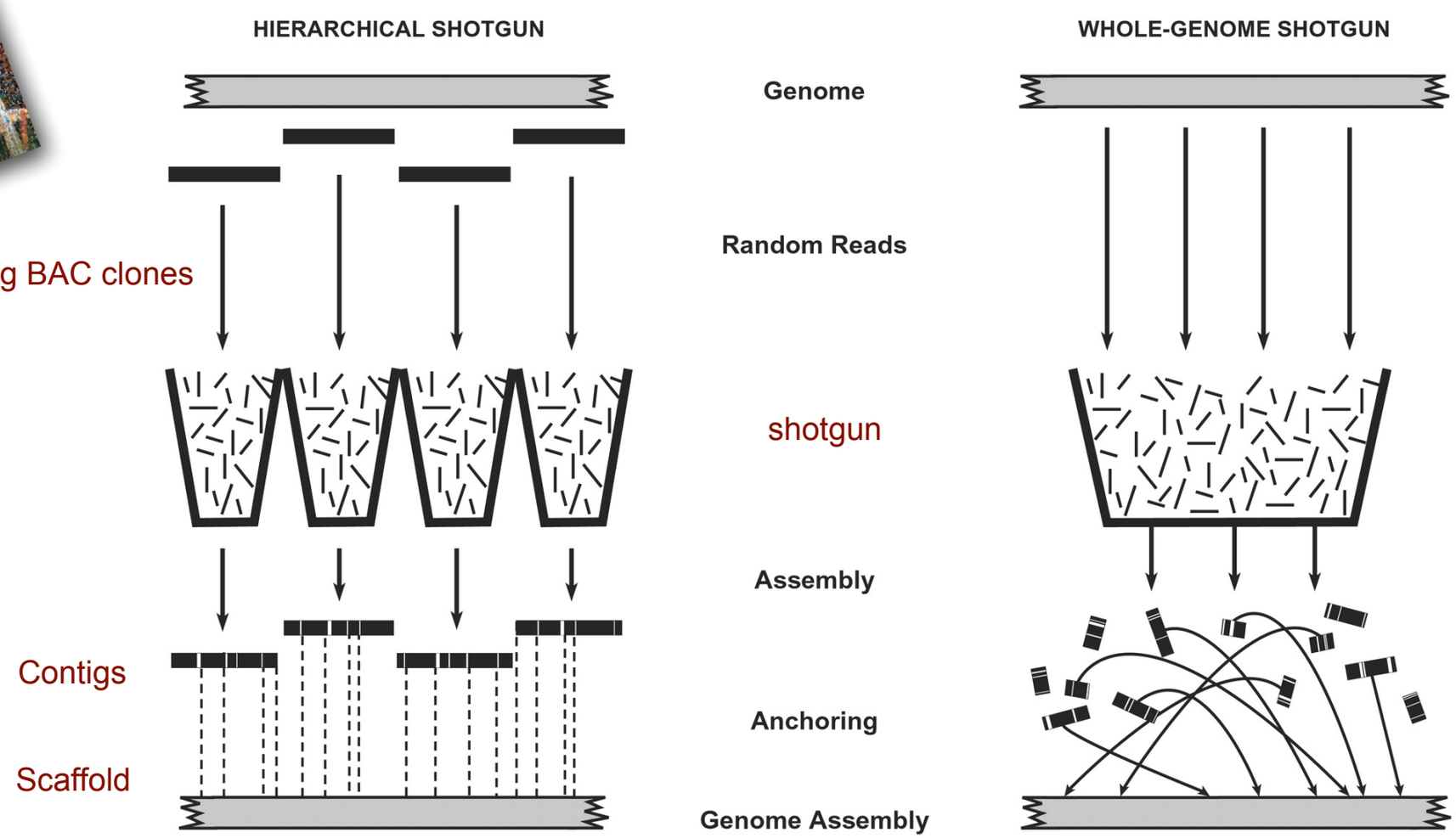
Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

The Human Genome Project





Sequencing strategies.



Waterston R H et al. PNAS 2002;99:3712-3716

PNAS

Sequencing strategies. (Left) The hierarchical shotgun (HS) strategy involves decomposing the genome into a tiling path of overlapping BAC clones, performing shotgun sequencing on and reassembling each BAC, and then merging the sequences of adjacent clones (CONTIGS). The method has the advantage that all sequence contigs and scaffolds derived from a BAC belong to a single compartment with respect to anchoring to the genome. (Right) Whole-genome shotgun (WGS) strategy involves performing shotgun sequencing on the entire genome and attempting to reassemble the entire collection. With the WGS method, each contig and scaffold is an independent component that must be anchored to the genome. In general, many scaffolds may not be anchored without directed efforts.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

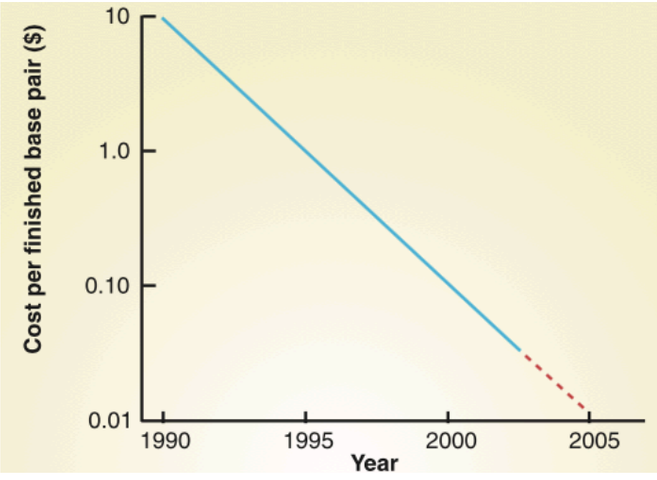
HGP goals & dates of achivements

The Human Genome Project: Lessons from Large-Scale Biology. Collins, F. S. (2003). Science, 300(5617), 286–290.

Area	Goal	Achieved	Date
Genetic map	2- to 5-cM resolution map (600 to 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical map	30,000 sequence-tagged sites (STSs)	52,000 STSs	October 1998
DNA sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	>98% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and cost of finished sequence	Sequence 500 Mb/year at <$0.25 per finished base	Sequence >1,400 Mb/year at <$0.09 per finished base	November 2002
Human sequence variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene identification	Full-length human cDNAs	15,000 full-length human cDNAs	March 2003
Model organisms	Complete sequences of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*	Finished sequences of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, plus whole-genome drafts of several others, including *C. briggsae*, *D. pseudoobscura*, mouse, and rat	April 2003
Functional analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis	1994
		DNA microarrays	1996
		Normalized and subtracted cDNA libraries	1996
		Eukaryotic, whole-genome knockouts (yeast)	1999
		Scale-up of two-hybrid mapping	2002

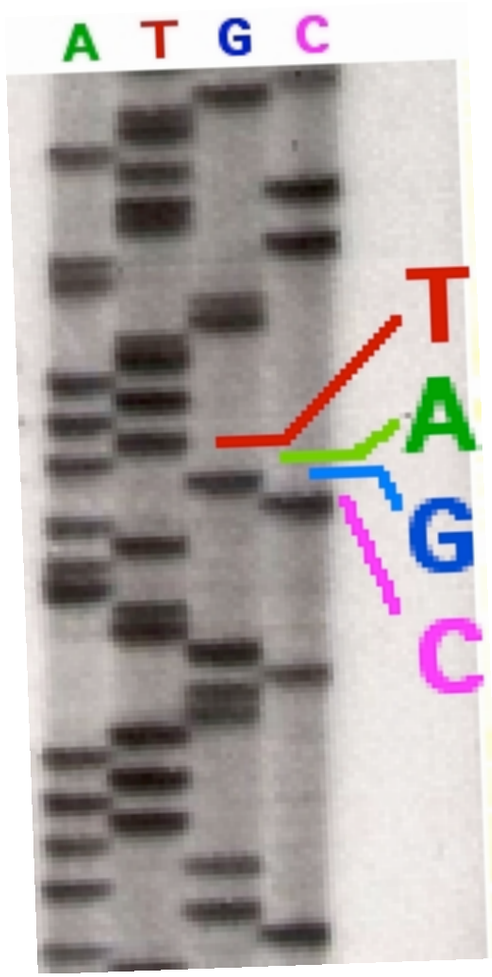
Table 2. Comparison of prices of large government projects circa 1990 with their projected useful life-span.

Proposed project	Projected cost ($ billion)	Target completion date	Estimated life-span (years)
Space Station Freedom	30.0	1999	30
Earth Observing System	17.0	2000	15
Superconducting Super Collider	11.0	1999	30
Human Genome Project	**3.0**	**2005**	**Perpetual**
Hubble Space Telescope	1.5	1990	15 to 20



Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

The Human Genome Project

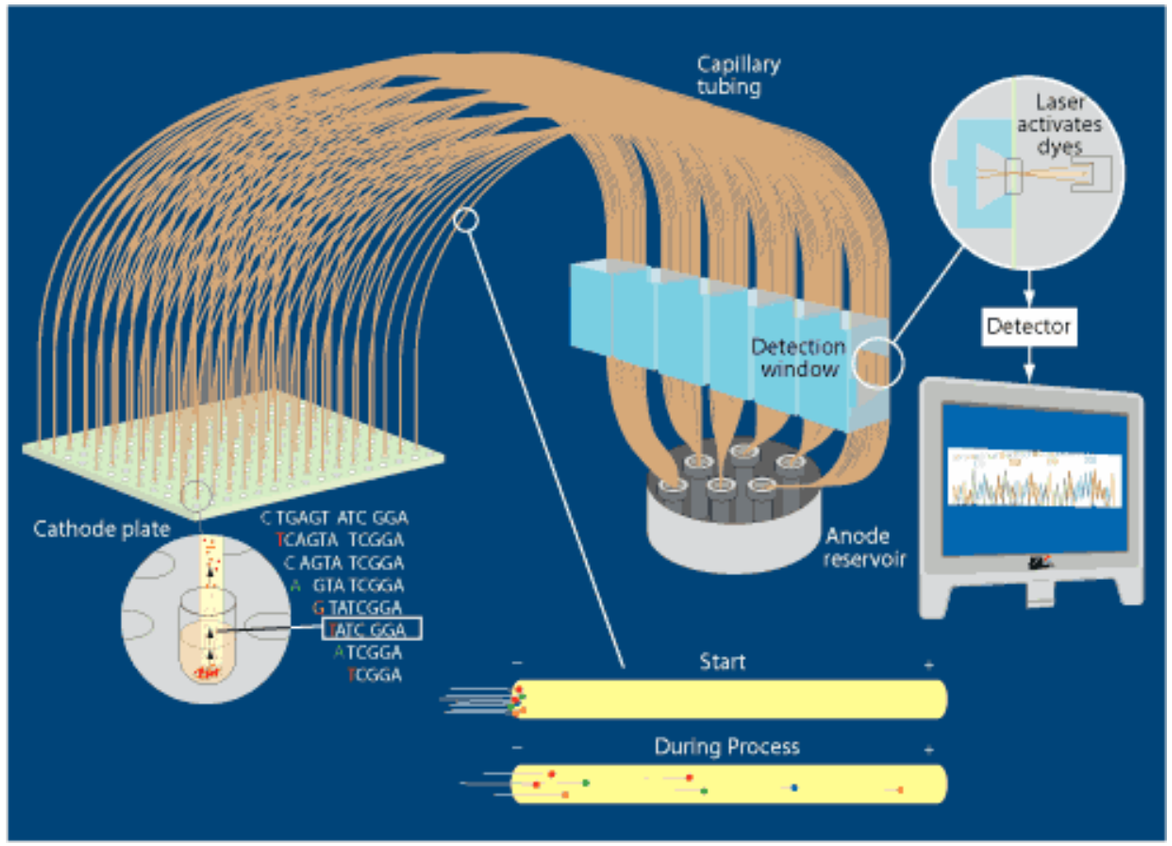


~1980 Manual sequencing
- X ray film (500 bases, four lanes, 30 cm gel)

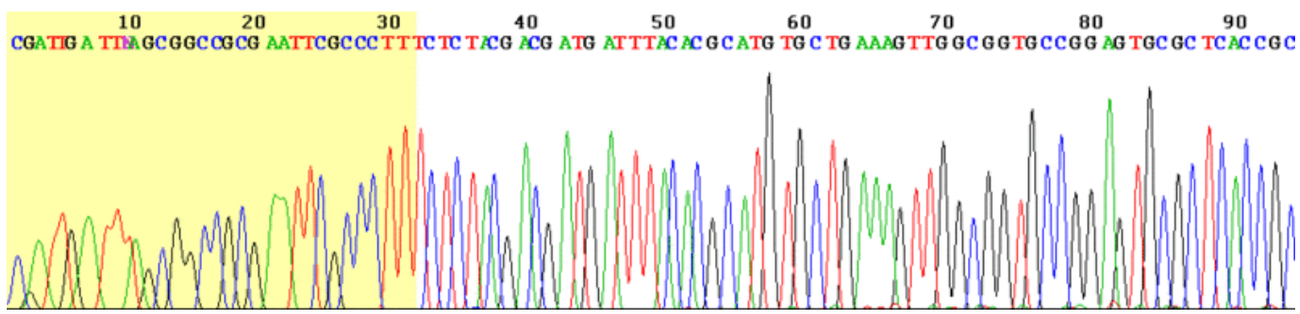
Three runs to sequence fragments longer than 1,000 pb



PHRED: automathed software for base-calling. It converts computer-generated traces into base sequences and asses probable accuracy for each base-call.
From the peak heights and widths, a Phred score is assigned to each individual base. A high Phred score indicates a high certainty as to the identity of that particular base.



~1990 High-Throughput sequencing by fluorescence automated Sanger sequencing.
Applied Biosystems introduced the capillary electrophoresis system (ABI Prism 3700), and advanced the Human Genome project by nearly five years ahead of schedule producing each machine ~2Mb per day



Ewing B, Hillier L, Wendl MC, Green P (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment". *Genome Res.* **8** (3): 175–185.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

About the Number of Human Genes

Pertea, M., & Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. Genome Biology, 11(5), 206.

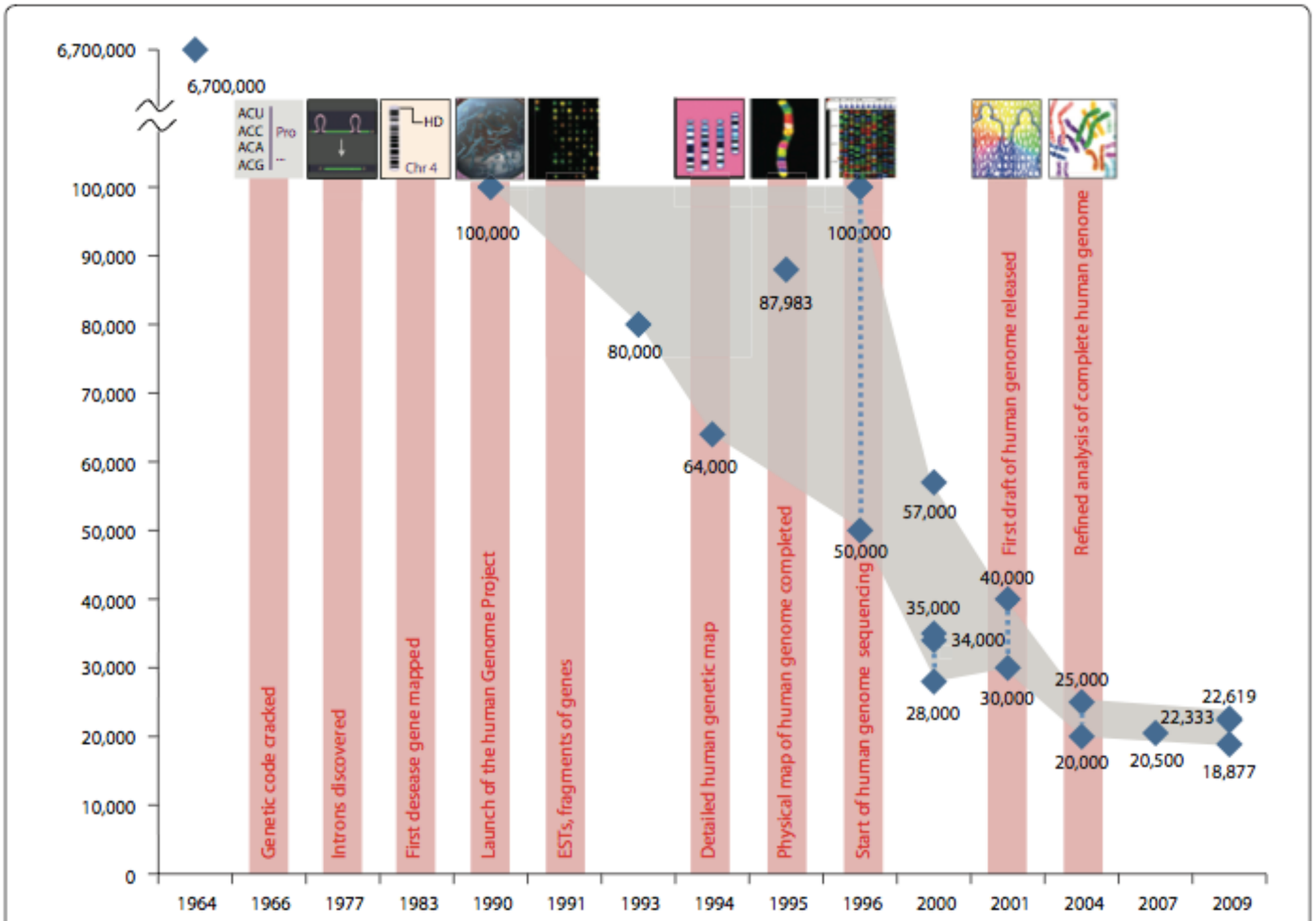


Figure 2. The trend of human gene number counts together with human genome-related milestones. Individual estimates of the human gene count are shown as blue diamonds. The range of estimates at different times is shown by the two vertical blue dotted lines. Note how this range has narrowed in recent years.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

About the Number of Human Genes

Pertea, M., & Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. Genome Biology, 11(5), 206.

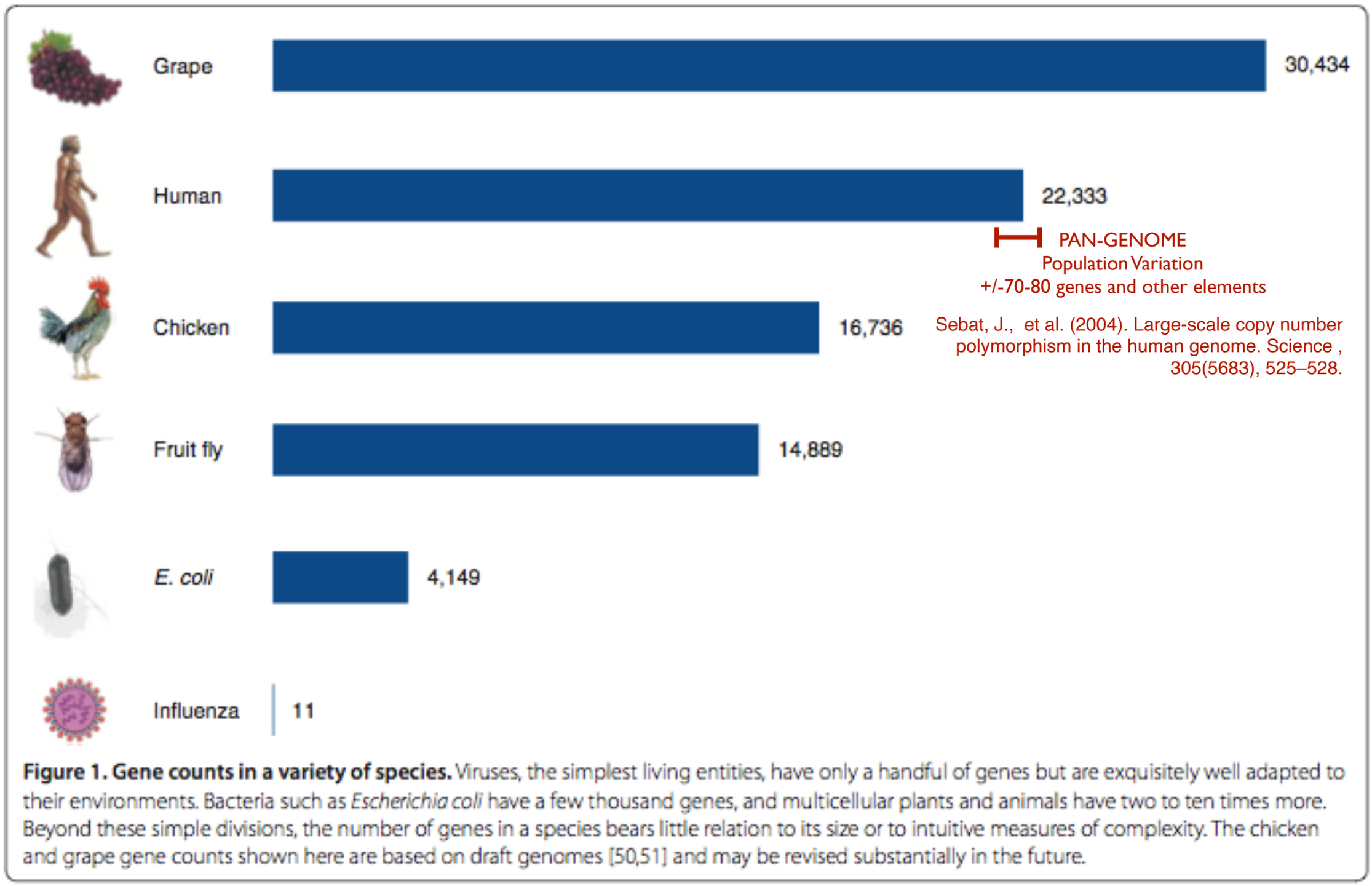


Figure 1. Gene counts in a variety of species. Viruses, the simplest living entities, have only a handful of genes but are exquisitely well adapted to their environments. Bacteria such as *Escherichia coli* have a few thousand genes, and multicellular plants and animals have two to ten times more. Beyond these simple divisions, the number of genes in a species bears little relation to its size or to intuitive measures of complexity. The chicken and grape gene counts shown here are based on draft genomes [50,51] and may be revised substantially in the future.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Data Storage Problem

Bytes(8 Bits)

- A "bit" is atomic: the smallest unit of storage
- A bit stores just a 0 or 1
- A bit is too small to be much use
- Group 8 bits together to make 1 byte
- e.g. 0 1 0 1 1 0 1 0

Kilobyte (1.0 E+03 Bytes)

- 1 Kilobyte: A very short story

Megabyte (1.0 E+06 Bytes)

- 1 Megabyte: A small novel
- 4 Mebagytes: A file containing all Genome Reference's Differences
- 750 Mb: The space needed to store the entire Human Genome (1X)

Gigabyte (1.0 E+09 Bytes)

- 1 Gigabyte: A pickup truck filled with paper
- 220 Gigabytes: A Full Human Genome (30X+ fastq format)

Terabyte (1.0 E+12 Bytes)

- 1 Terabyte: All the X-ray films in a large technological hospital
- 2 Terabytes: A confortable size to manage a Human Genome Analysis

Petabyte (1.0 E+15 Bytes)

- 1 Petabytes: 5 years of Earth Science Observation Data (NASA)
- 200 Petabytes: All printed material

Exabyte (1.0 E+18 Bytes)

- 5 Exabytes: All words ever spoken by human beings.
- From the beginning of recorded time until 2003, humans created 5 exabytes of data

Zettabyte (1.0 E+21 Bytes)

- 1.9 zettabytes in 2007 (this is the equivalent to every person on earth receiving 174 newspapers per day)

What is Big Data?

Think for a sec... what does "Big Data" mean?

There are huge volumes of data in the world:

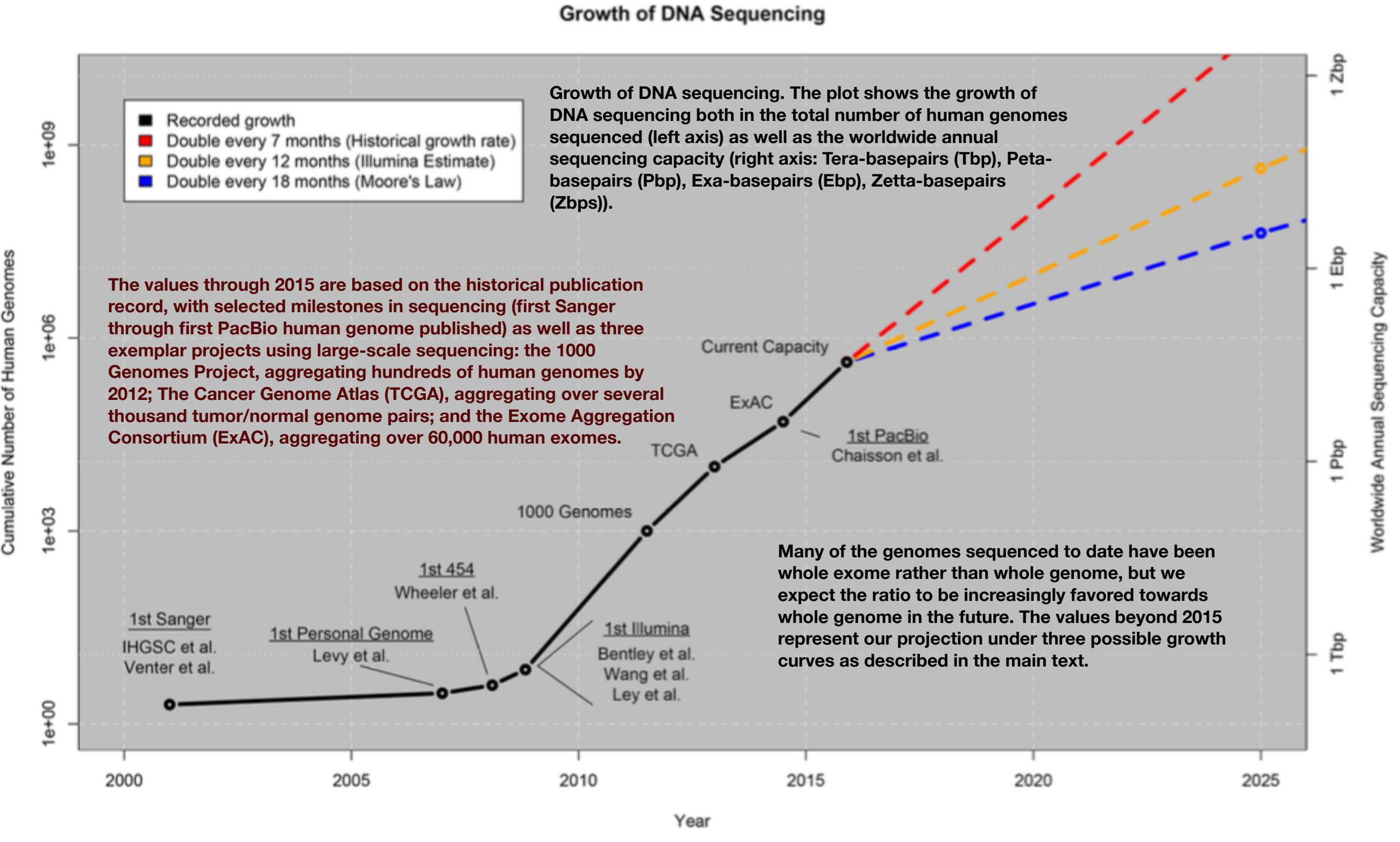
- 5 exabytes of data.
- In 2011, the same amount was created every two days
- In 2013, every 10 minutes.
- **90% of the data in the world today has been created in the last two years alone**. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

This is Big Data

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Big Data - Genomics

Stephens, Z. D., et al. (2015). Big Data: Astronomical or Genomical? PLOS Biol, 13(7), e1002195.



Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Big Data - Genomics & Cloud Computing

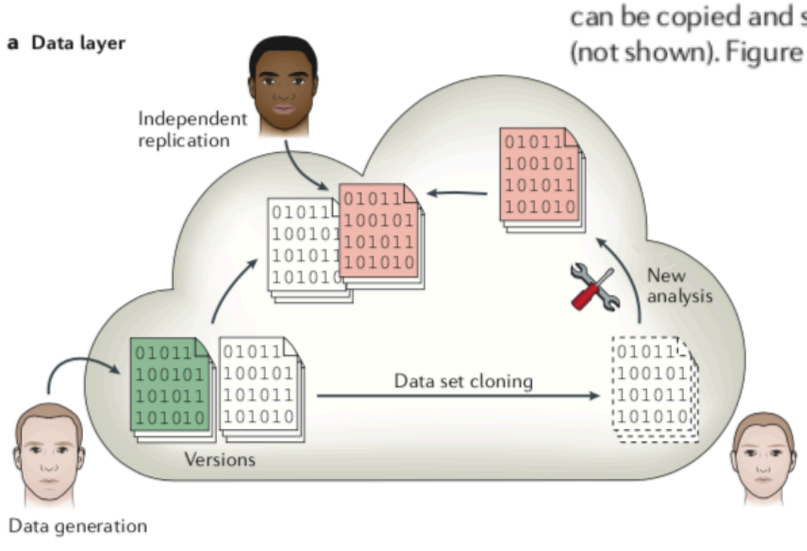
Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

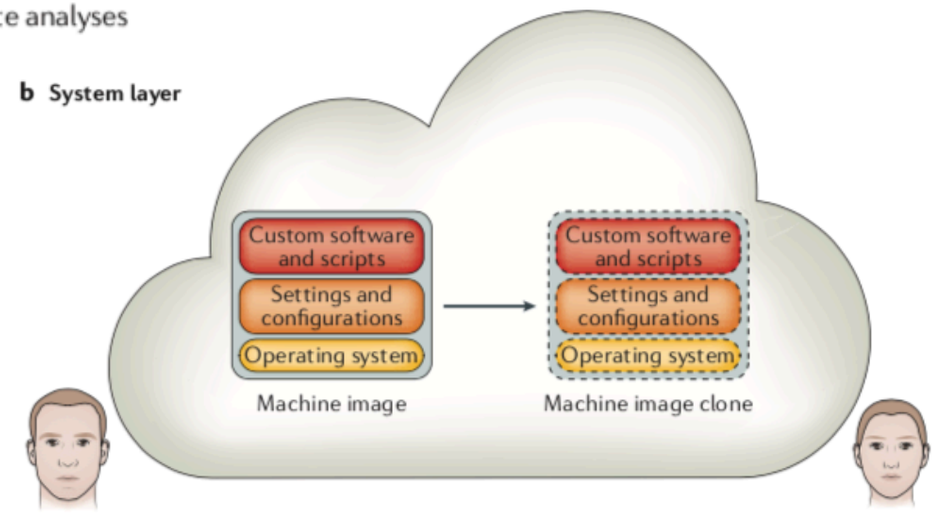
Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Fig. 3 | **Cloud reproducibility.** The cloud fosters reproducibility by enabling investigators to publish data sets to the cloud, including different versions thereof, without loss or modification of the previous data set. Other investigators, situated near or far geographically, can clone data sets within the cloud and apply customized software to perform their own analyses and derive new results. Independent investigators can copy original primary data sets, software and published results within the cloud to replicate published analyses (part **a**). Collaborating investigators can set up cloud-based virtual machine images that contain the software, configurations and scripts needed for specific computational analyses (part **b**). Customized machine images can be copied and shared with other investigators within the cloud for replicate analyses (not shown). Figure adapted from REF. [114], Macmillan Publishers Limited.

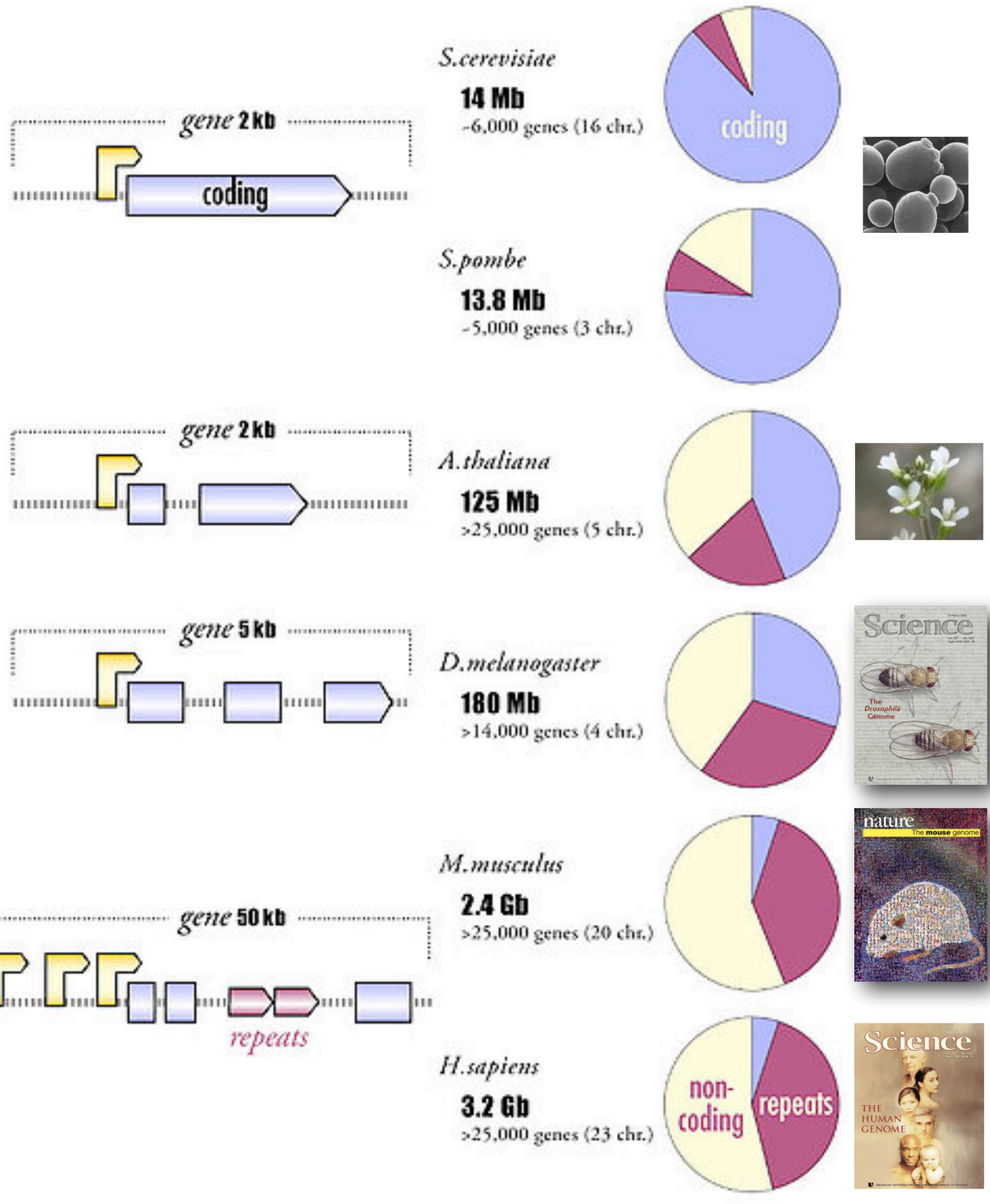
Ben Langmead, & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. Nature Review Genetics, 19(4), 1–12. http://doi.org/10.1038/nrg.2017.113

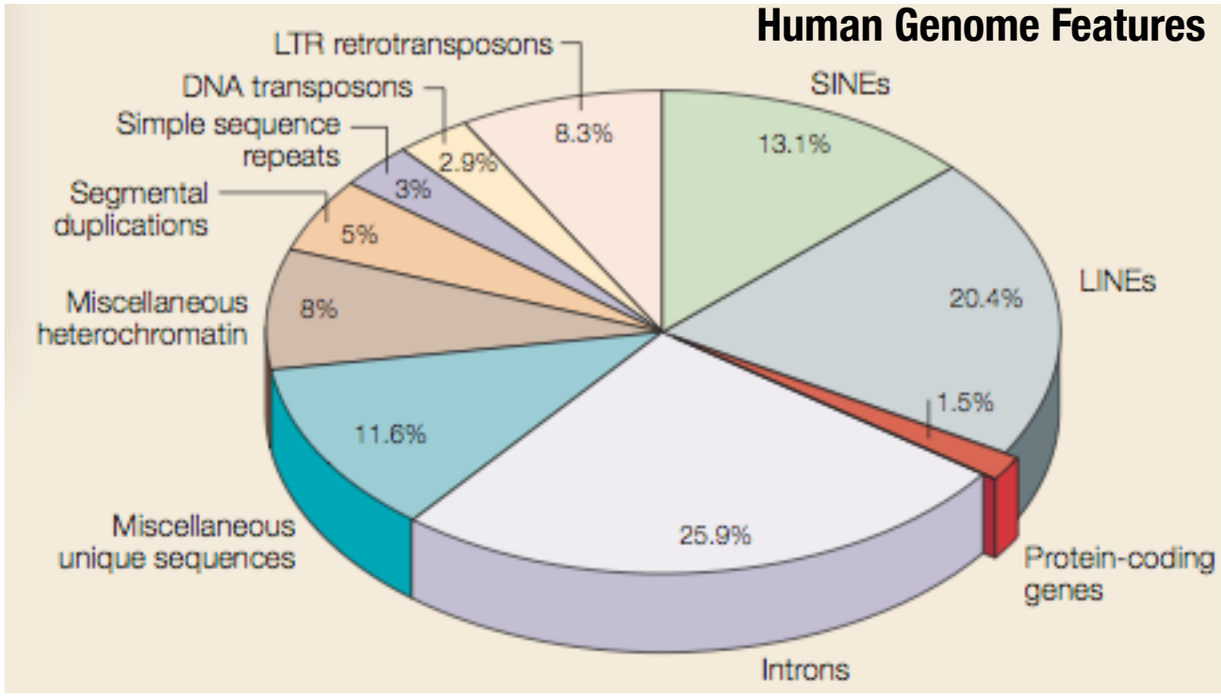


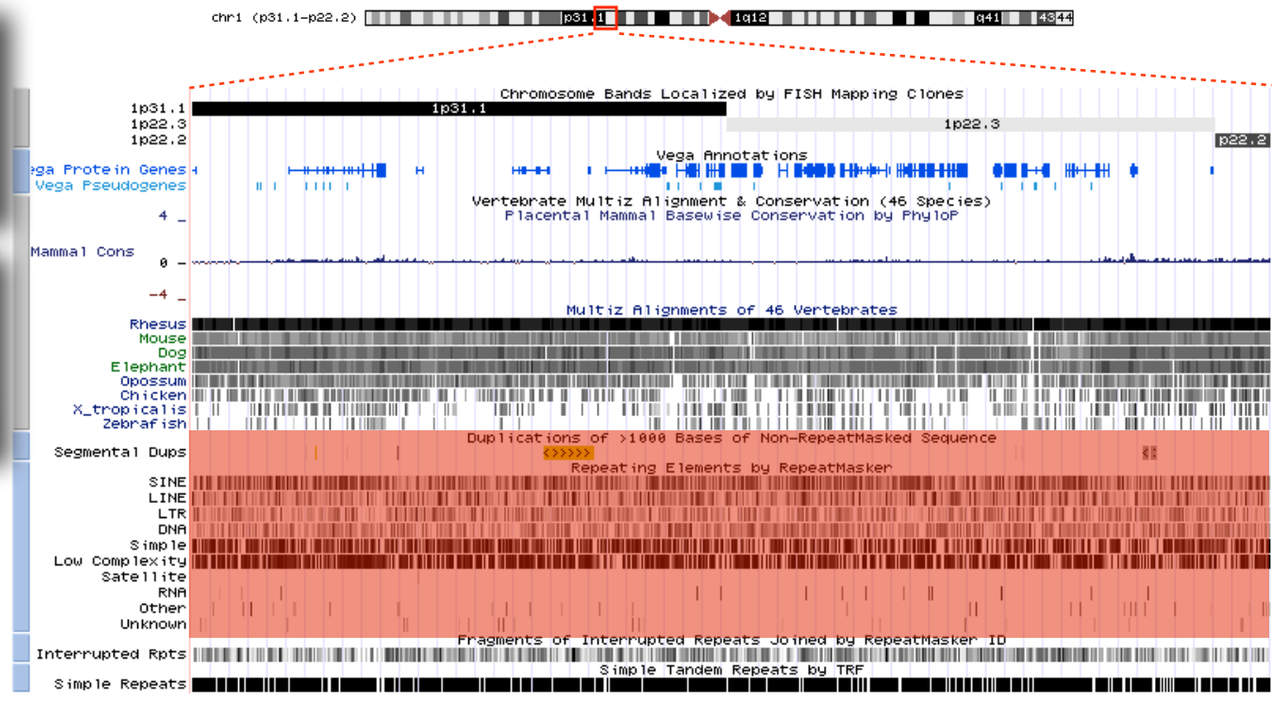


Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Genome Size & Complexity







Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Genome Architecture

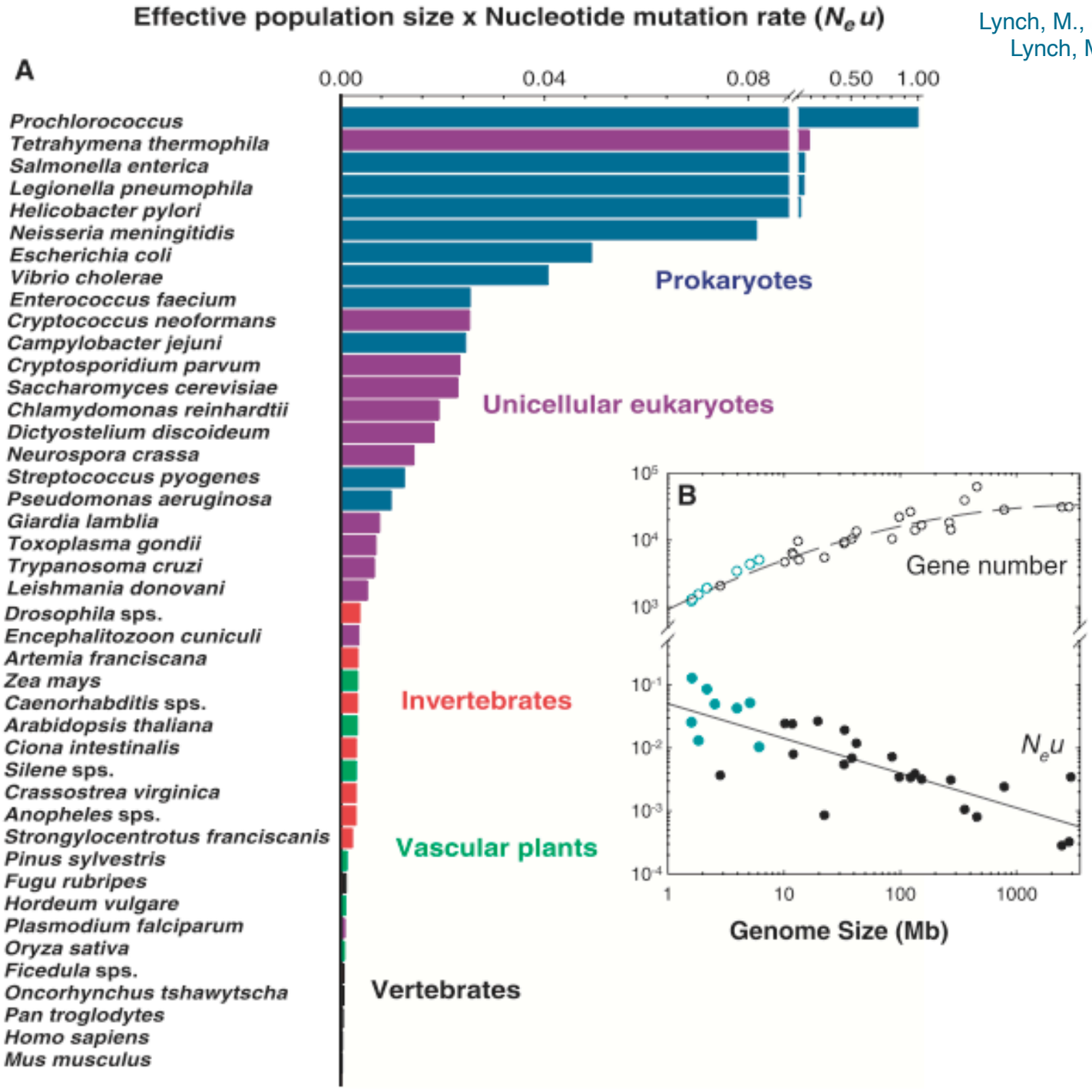


Fig. 1. (**A**) Estimates of the composite parameter $N_e u$ for a phylogenetically diverse assemblage of species. (**B**) The relationship between estimated $N_e u$, total gene number, and genome size. Data for prokaryotes are plotted in blue. The log-log regression of $N_e u$ versus genome size is highly significant, with an intercept of -1.30 ± 0.40, a slope of -0.55 ± 0.07, and $r^2 = 0.659$, df = 28 (*1*). The number of species plotted differs between graphs because genome structure information is not available for all species with $N_e u$ estimates.

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. Science, 302, 1401–4.
Lynch, M. (2006). The origins of eukaryotic gene structure, Mol Biol Evol 23(2), 450–468.

The resultant increase in the power of random genetic drift appears to be sufficient to overwhelm the weak mutational disadvantages associated with most novel aspects of the eukaryotic gene, supporting the idea that most such changes are simple outcomes of semi-neutral processes rather than direct products of natural selection.

However, by establishing an essentially permanent change in the population-genetic environment permissive to the genome-wide repatterning of gene structure, the eukaryotic condition also promoted a reliable resource from which natural selection could secondarily build novel forms of organismal complexity.

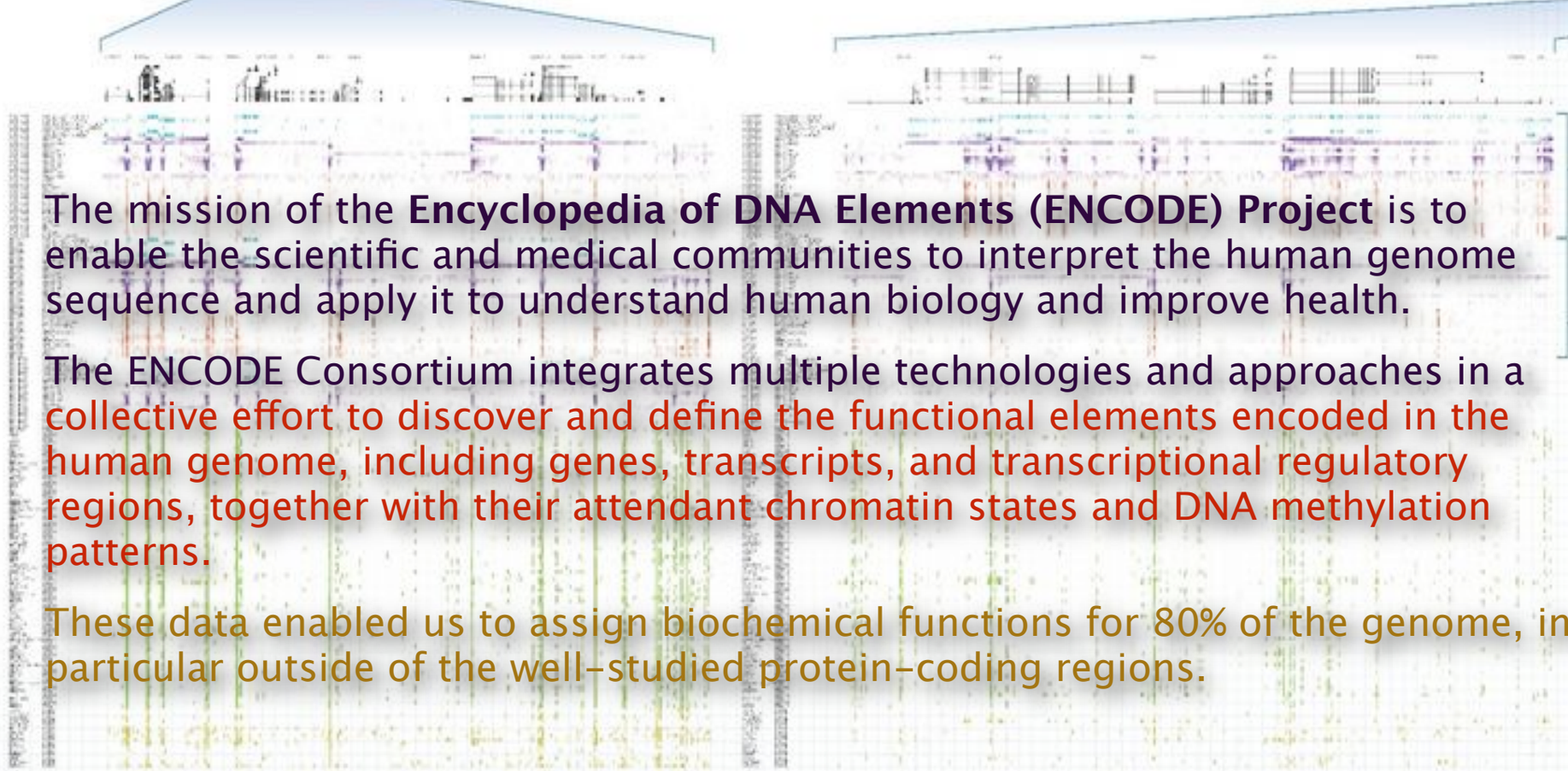
Under this hypothesis, arguments based on molecular, cellular, and/or physiological constraints are insufficient to explain the disparities in gene, genomic, and phenotypic complexity between prokaryotes and eukaryotes.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

ENCODE 2007 & 2012

the ENCODE project
ENCyclopedia Of DNA Elements

The human genome encodes the blueprint of life





The mission of the **Encyclopedia of DNA Elements (ENCODE) Project** is to enable the scientific and medical communities to interpret the human genome sequence and apply it to understand human biology and improve health.

The ENCODE Consortium integrates multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns.

These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions.

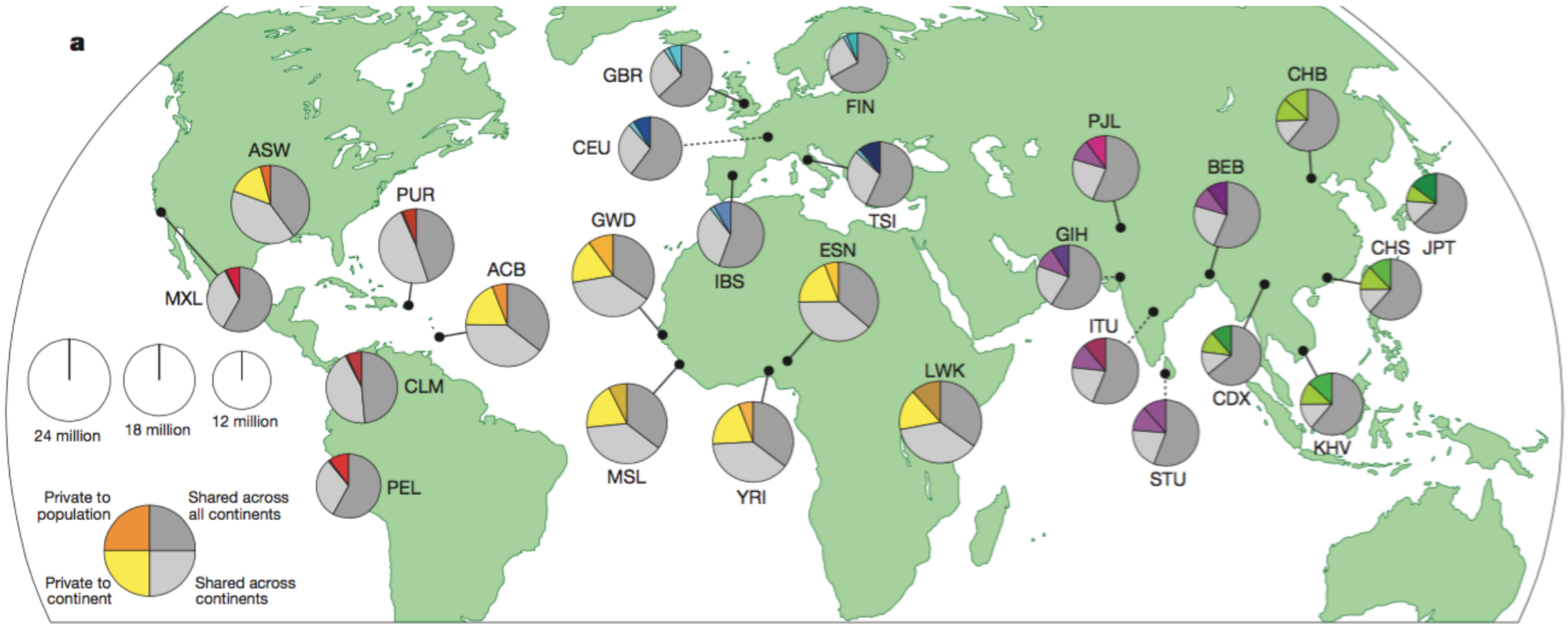
Produced in association with nature

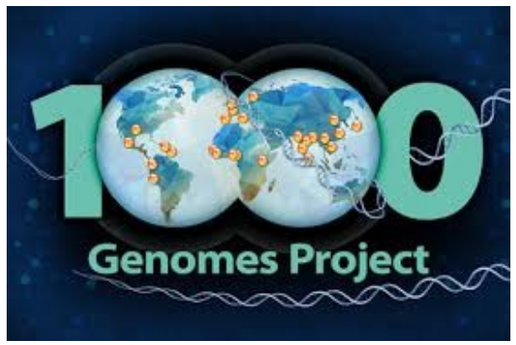
Produced with support from illumina

The poster shows a subset of the ENCODE data; on the top panel is chromosome 2, with the tracks showing the density of elements from RNA, histone modification, transcription factor binding and DNase I in 100-kilobase windows. Below this is a selection of protein-coding gene symbols for chromosome 2. Two regions are shown representing 115 out of the 1,648 experiments analysed in the 6 September 2012 publication of ENCODE. The cell type is shown on the left, followed by the experiment type. The intensity of the colour represents the strength of signal over a 400-base-pair window. If the entire genome was rendered at this scale, this poster would be 30 kilometres long and 16 metres wide to accommodate all of the data. The anatomical diagram shows the location of 47 of the 147 cell types analysed in the 6 September 2012 publication of ENCODE, and highlights GM12878 (a lymphoblastoid cell line) and HepG2 (a hepatocellular carcinoma that is widely used as a model for hepatocytes), two of the six focus cell types of ENCODE.
More information on the ENCODE project, including full data downloads, can be found at http://www.encodeproject.org/. Artistic credits: Darryl Leja, Evan Birney.

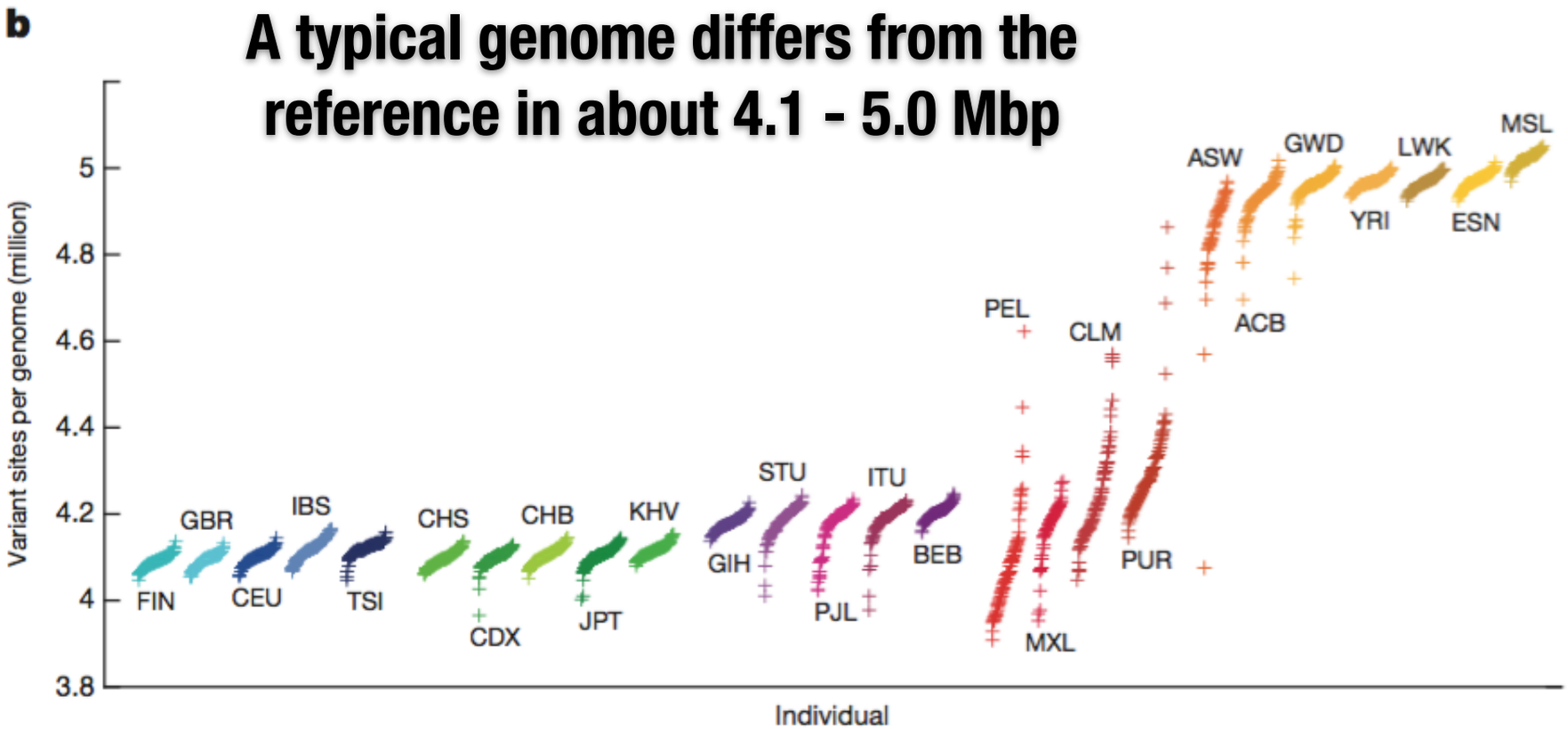
Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

1,000 Genomes Project





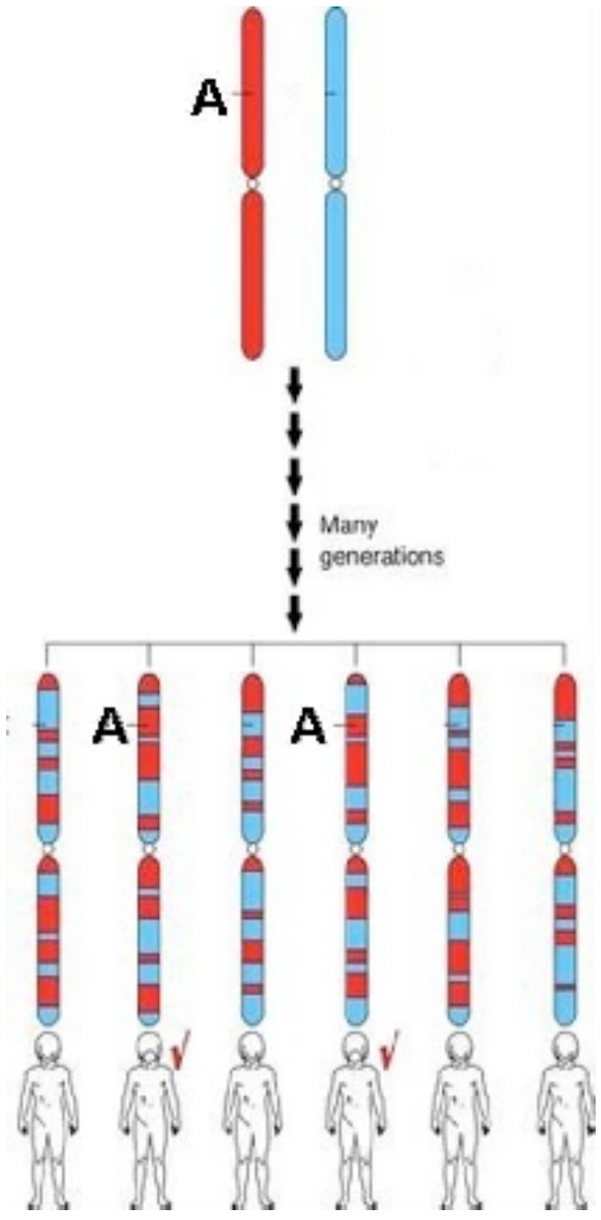
2.504 genomes
26 populations



1. (ACB) African Caribbean in Barbados
2. (ASW) African ancestry in the Ssouth-Western USA
3. (BEB) Bengali in Bangladesh
4. (CDX) Chinese Dai in Xishuangbanna, China
5. (CEU) Central Europe in USA
6. (CHB) Chinese in Beijing
7. (CHS) Han Chinese South
8. (CLM) Colombians in Medellín, Colombia
9. (ESN) Esan in Nigeria
10. (FIN) Finnish in Finland
11. (GBR) British from England and Scotland
12. (GIH) Gujarati Indians in Houston, Texas, USA
13. (GWD) Gambian in Western Division – Mandinka
14. (IBS) Iberian Populations in Spain
15. (ITU) Indian Telugu in the UK
16. (JPT) Japanese in Tokio
17. (KHV) Kinh in Ho Chi Minh City, Vietnam
18. (LWK) Luhya in Webuye, Kenya
19. (MSL) Mende in Sierra Leone
20. (MXL) Mexican ancestry in Los Angeles
21. (PEL) Peruvian in Lima, Peru
22. (PJL) Punjabi in Lahore, Pakistan
23. (PUR) Puerto Ricans in Puerto Rico
24. (STU) Sri Lankan Tamil in the UK
25. (TSI)Tuscans in Italy
26. (YBI) Yoruba in Ibadan

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

On Human Variation



Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Publishing Group 6: 95–108.

- **A haplotype is a sequence of alleles stretching along an extended segment of DNA – a sort of super allele!**
- **Haplotypes are usually inherited as a single unit from parents.**

This diagram shows **two ancestral chromosomes being scrambled through recombination over many generations** to yield different descendant chromosomes.

Adjacent to the variant marked by the A are many SNPs that can be used to identify the location of the variant.

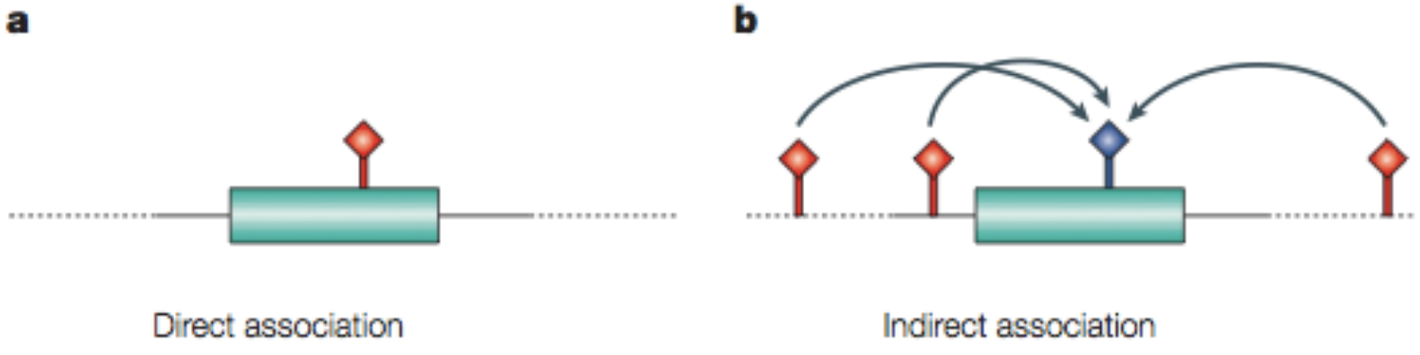
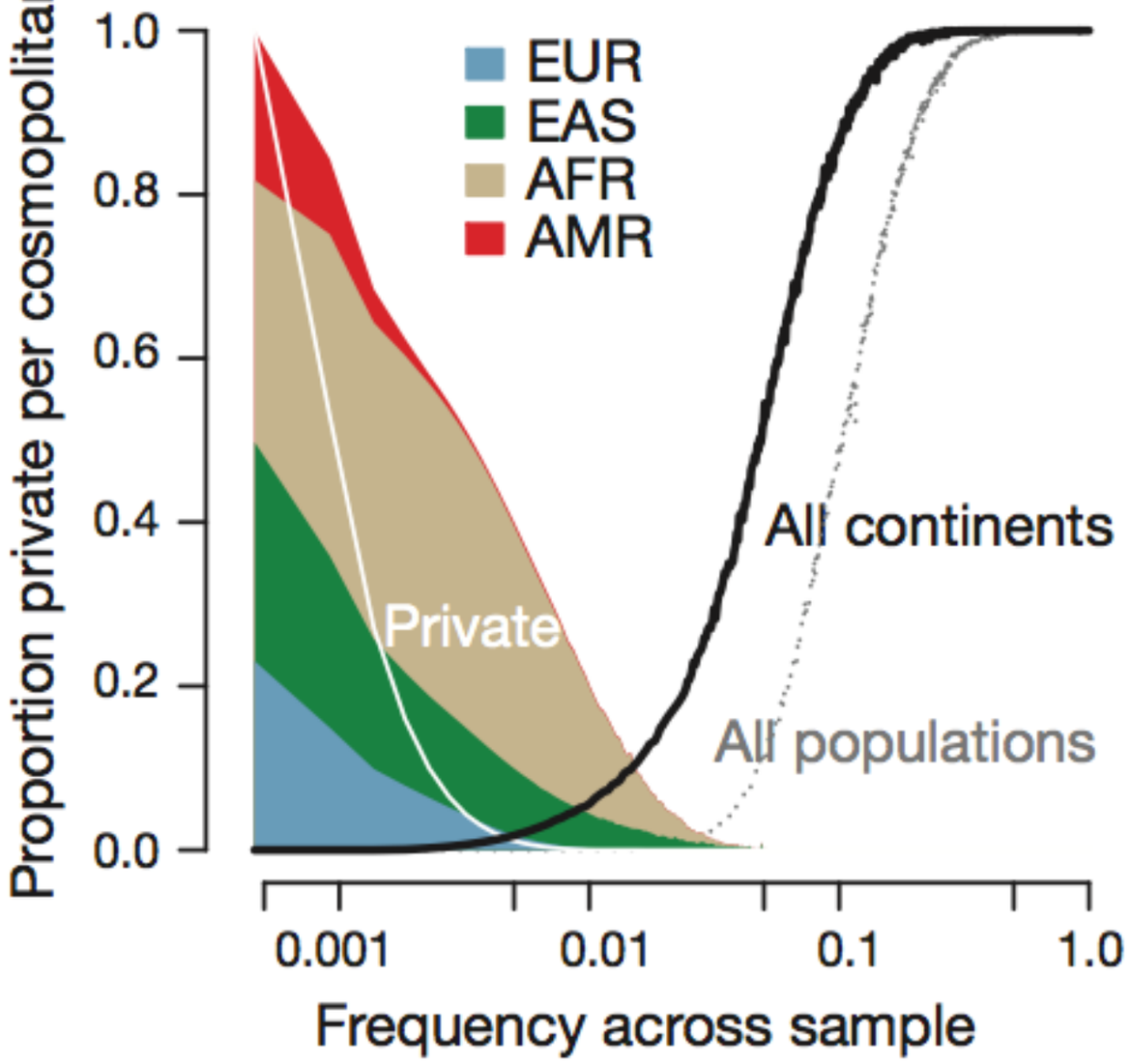


Figure 1 | **Testing SNPs for association by direct and indirect methods. a** | A case in which a candidate SNP (red) is directly tested for association with a disease phenotype. For example, this is the strategy used when SNPs are chosen for analysis on the basis of prior knowledge about their possible function, such as missense SNPs that are likely to affect the function of a candidate gene (green rectangle). **b** | The SNPs to be genotyped (red) are chosen on the basis of linkage disequilibrium (LD) patterns to provide information about as many other SNPs as possible. In this case, the SNP shown in blue is tested for association indirectly, as it is in LD with the other three SNPs. A combination of both strategies is also possible.

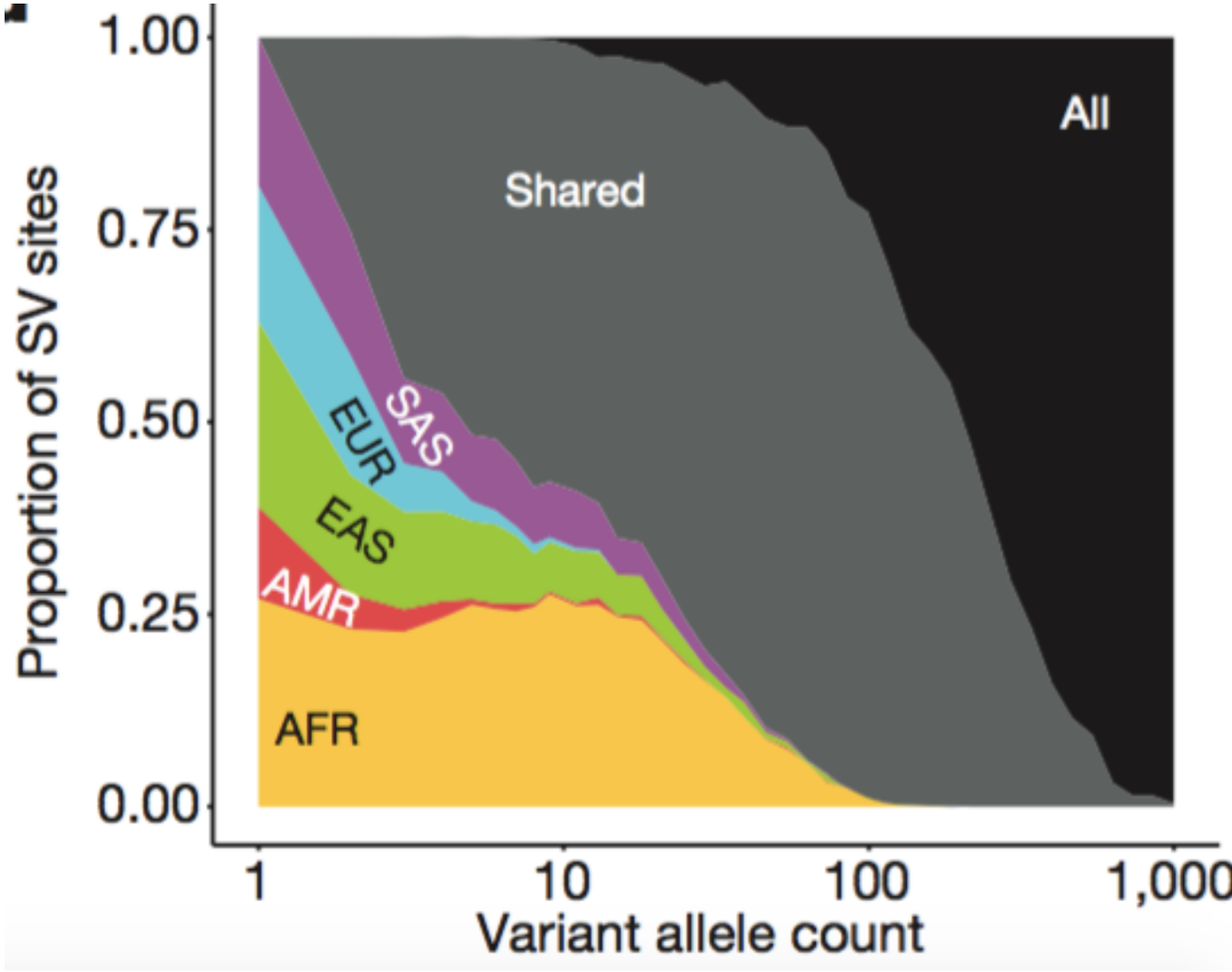
Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

1,000 Genomes Project

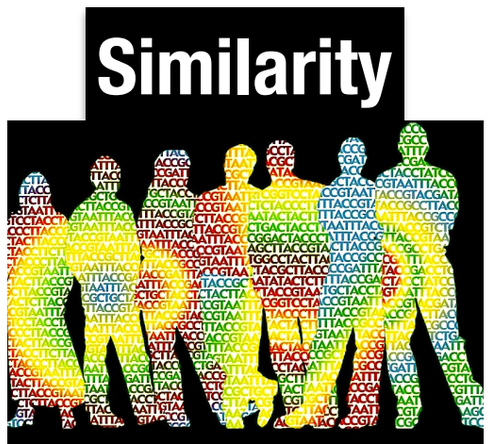
SNPs



SVs



Similarity



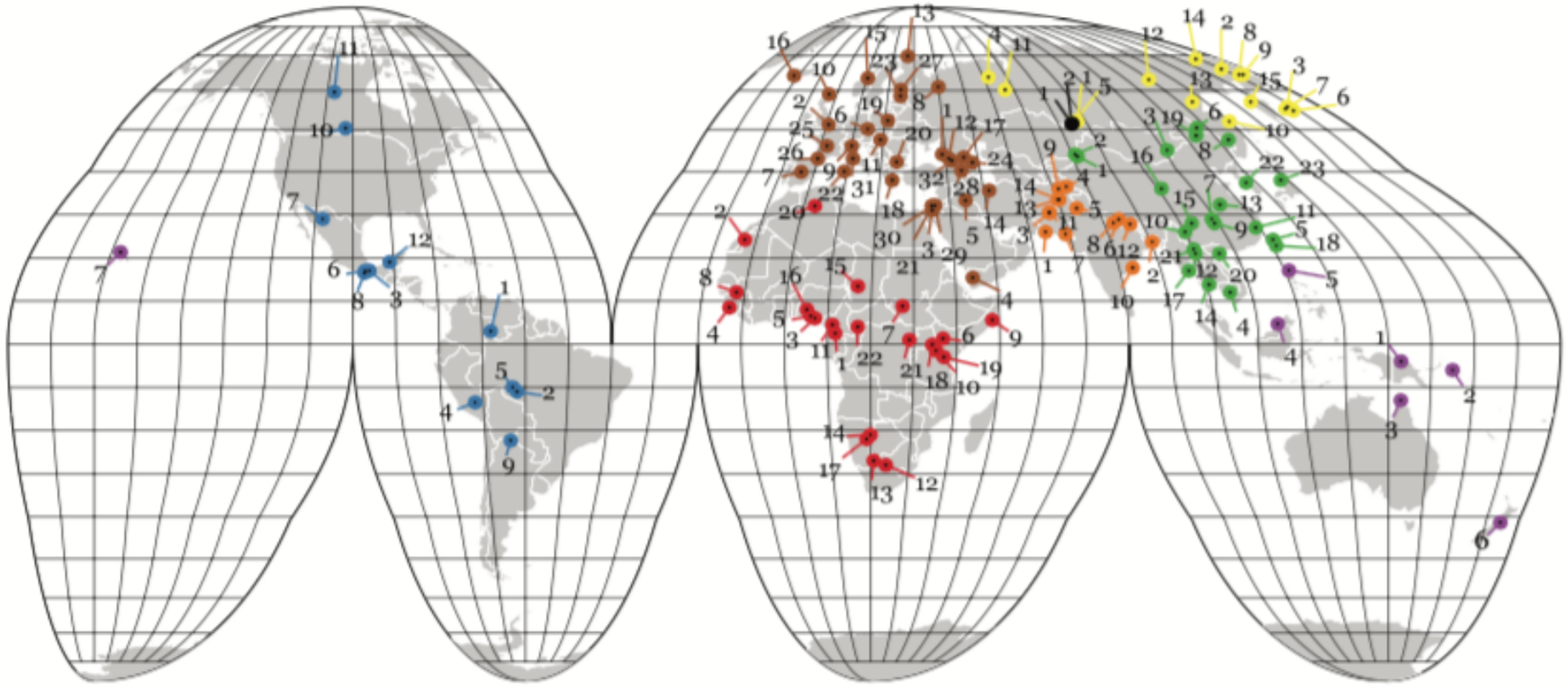
~ 99%

~ 95%

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Population Structural Variation

Sudmant, P. H., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. Science, 349(6253), aab3761.



1 - Neanderthal
2 - Denisova
1 - Piapoco
2 - Surui
3 - Mixe
4 - Quechua
5 - Karitiana
6 - Mixtec
7 - Pima
8 - Zapotec
9 - Chane
10 - Cree
11 - Chipewyan
12 - Maya
1 - Xibo
2 - Uygur
3 - Mongola
4 - Cambodian
5 - Atayal
6 - Oroqen
7 - Tujia
8 - Hezhen
9 - Miao
10 - Naxi
11 - She
12 - Dai
13 - Han
14 - Thai
15 - Yi
16 - Tu
17 - Burmese
18 - Ami
19 - Daur
20 - KinhKVH
21 - Lahu
22 - Korean
23 - Japanese
1 - Papuan
2 - Bougainville
3 - AustraliaECCAC
4 - Dusun
5 - Igorot
6 - Maori
7 - Hawaiian
1 - Adygei
2 - EnglandGBR
3 - Bedouin
4 - YemeniteJew
5 - IraqiJew
6 - Czechoslovakian
7 - SpainIBS
8 - Russian
9 - Bergamo
10 - Orcadian
11 - Hungarian
12 - Abkhasian
13 - Saami
14 - Iranian
15 - Norwegian
16 - Icelander
17 - Chechen
18 - Greek
19 - Polish
20 - Bulgarian
21 - Palestinian
22 - Sardinian
23 - Estonian
24 - Lezgin
25 - French
26 - Basque
27 - FinlandFIN
28 - Armenian
29 - Jordanian
30 - Druze
31 - Tuscan
32 - Georgian
1 - Kongo
2 - Saharawi
3 - Igbo
4 - MendeMSL
5 - EsanESN
6 - LuhyaLWK
7 - Dinka
8 - Mandenka
9 - Somali
10 - MasaiMKK
11 - Lemande
12 - BantuSETswana
13 - KhomaniSan
14 - San
15 - GambiaGWD
16 - Yoruba
17 - BantuSEHerero
18 - Luo
19 - BantuKenya
20 - Mozabite
21 - Mbuti
22 - Biaka
1 - Makrani
2 - BengaliBEB
3 - Balochi
4 - Burusho
5 - PunjabiPJL
6 - Tibetan
7 - Sindhi
8 - Kusunda
9 - Kalash
10 - Madiga
11 - Brahui
12 - Sherpa
13 - Hazara
14 - Pathan
1 - Tubalar
2 - EskimoNaukan
3 - AleutBeringIsland
4 - MansiSosvaRiver
5 - Altaian
6 - TlingitPreobrazhenskoye
7 - TlingitNikolskoye
8 - EskimoChaplino
9 - EskimoSireniki
10 - Ulchi
11 - MansiKondaRiver
12 - Yakut
13 - Even
14 - Chukchi
15 - Itelman

Fig. 1. Analysis of CNVs in several world populations. The geographical locations of the 125 human populations, including two archaic genomes, assessed in this study. Populations are colored by their continental population groups, and archaic individuals are indicated in black.

We sequenced to high coverage a panel of 236 human genomes representing 125 diverse human populations from across the globe

Sequencing was performed to a mean genome coverage of 41-fold from libraries prepared by using a standard polymerase chain reaction–free protocol on the HiSeq 2000 II- lumina (San Diego, CA) sequencing platform.

14,467 autosomal CNVs and 545 X-linked CNVs among individuals relative to the reference genome

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Human Explosive Growth

Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants

Alon Keinan[1]* and Andrew G. Clark[1,2]

Human populations have experienced recent explosive growth, expanding by at least three orders of magnitude over the past 400 generations. This departure from equilibrium skews patterns of genetic variation and distorts basic principles of population genetics. We characterized the empirical signatures of explosive growth on the site frequency spectrum and found that the discrepancy in rare variant abundance across demographic modeling studies is mostly due to differences in sample size. Rapid recent growth increases the load of rare variants and is likely to play a role in the individual genetic burden of complex disease risk. Hence, the extreme recent human population growth needs to be taken into consideration in studying the genetics of complex diseases and traits.

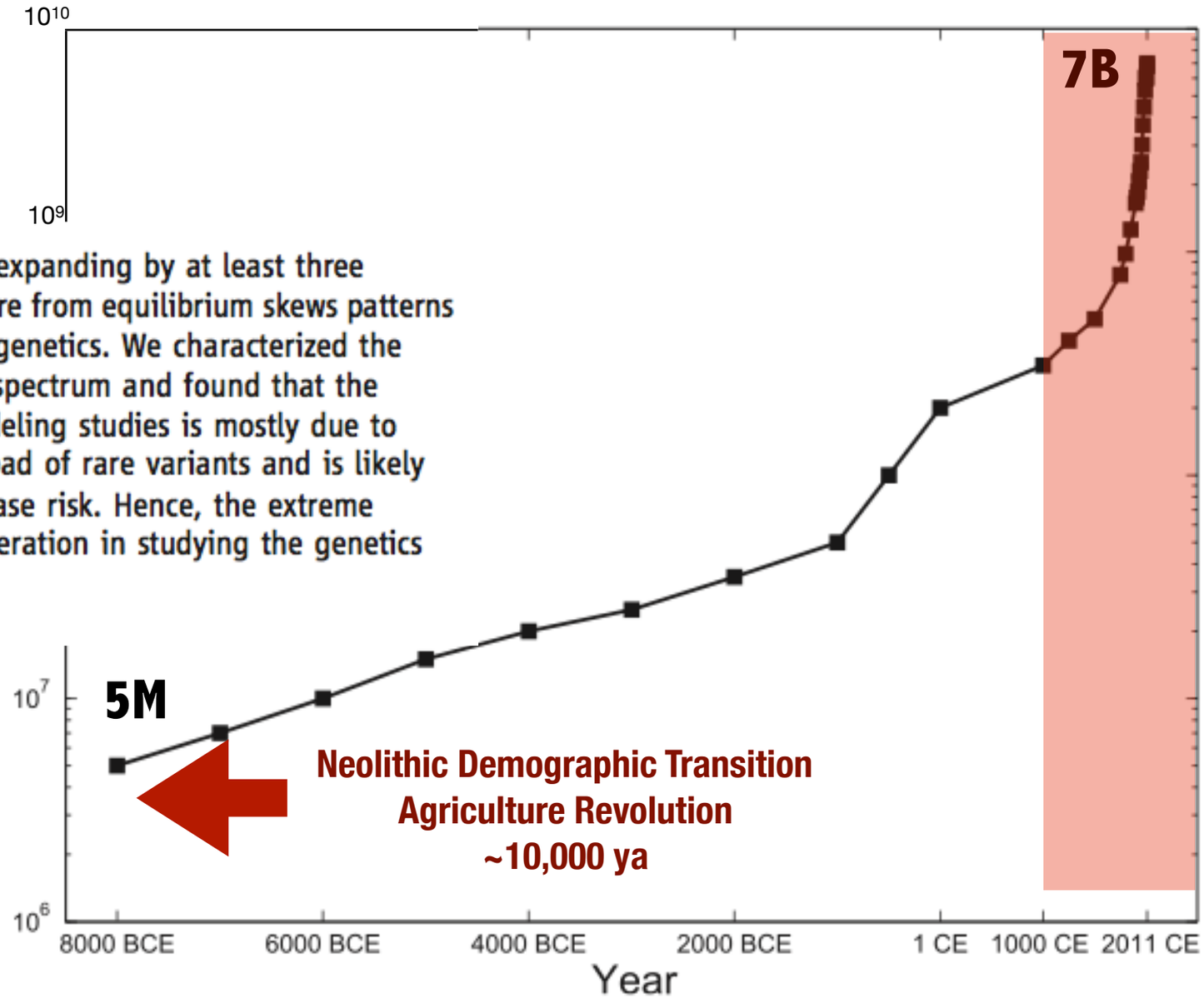
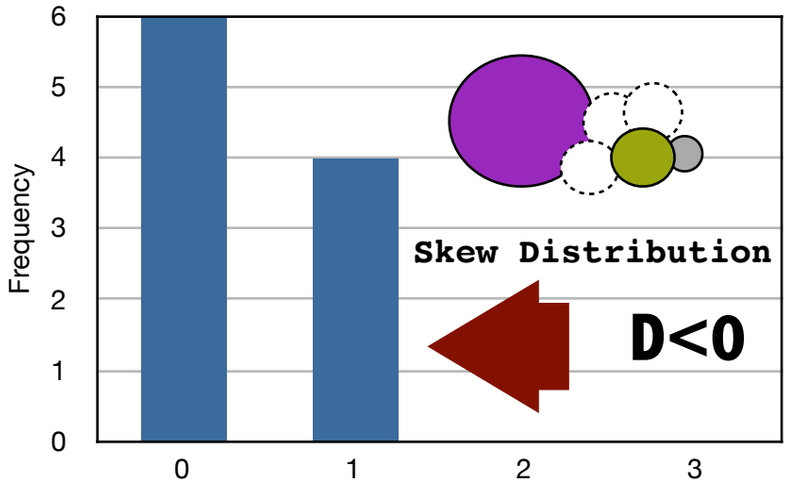


Fig. 1. Census (rather than effective) population size is presented on a logarithm scale over the past 10,000 years, from about 5 million at 8000 BCE to about 7 billion today from data in (*1*, *3*, *30*, *31*). The depicted linear increase (on the log scale) through most of the presented epoch denotes exponential growth of relatively constant percentage increase in population size per year. An acceleration of that increase starting in the Common Era is evident.

Selection Removing Variation (Selective Sweep), or Recent Population Expansion

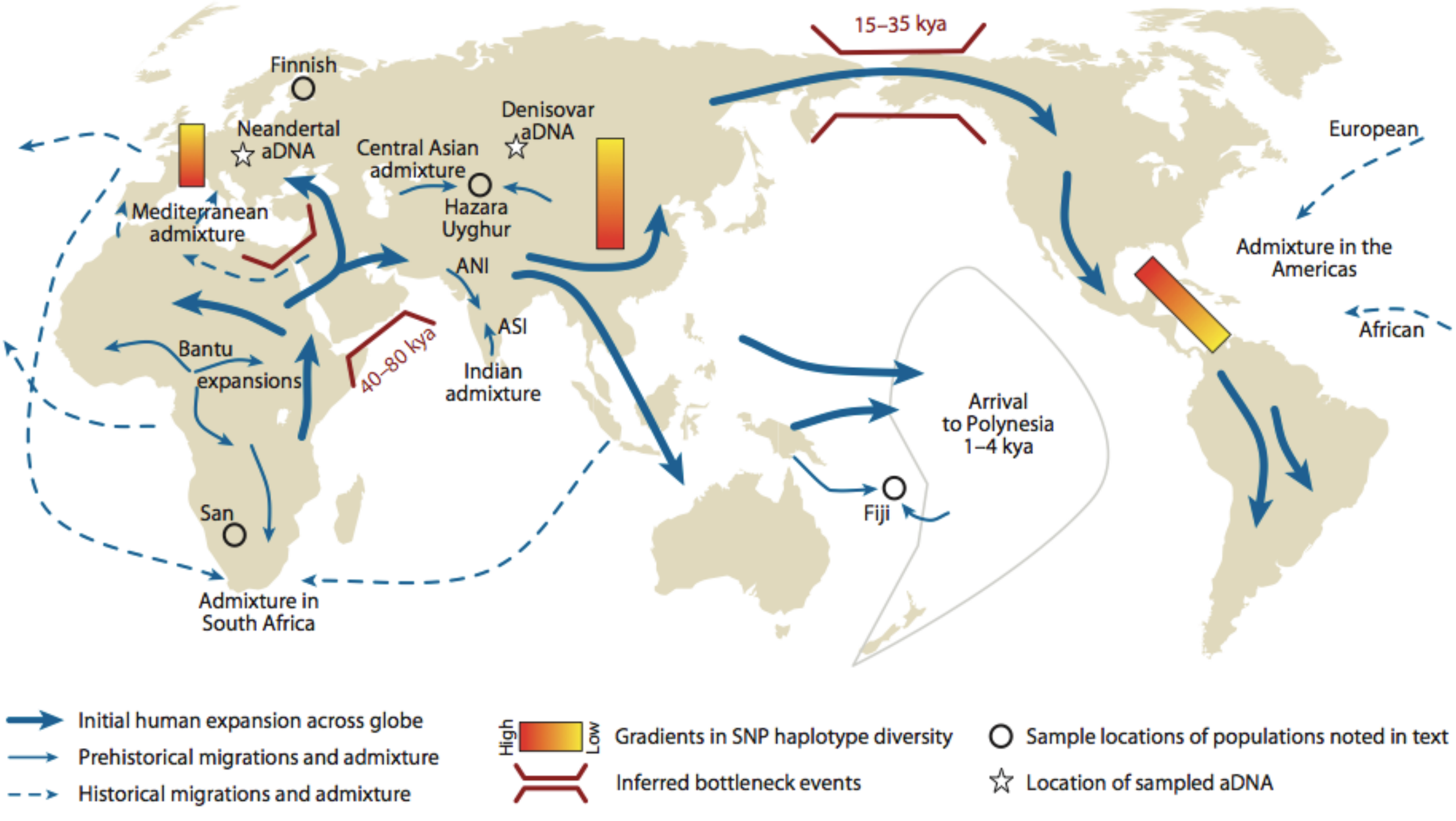


Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Human Population Structure

Novembre, J., & Ramachandran, S. (2011). Perspectives on Human Population Structure at the Cusp of the Sequencing Era. An. Rev. Genom & Hum Genet., 12(1), 245–274.

Schematic Human Demographic History (SNP's based studies)

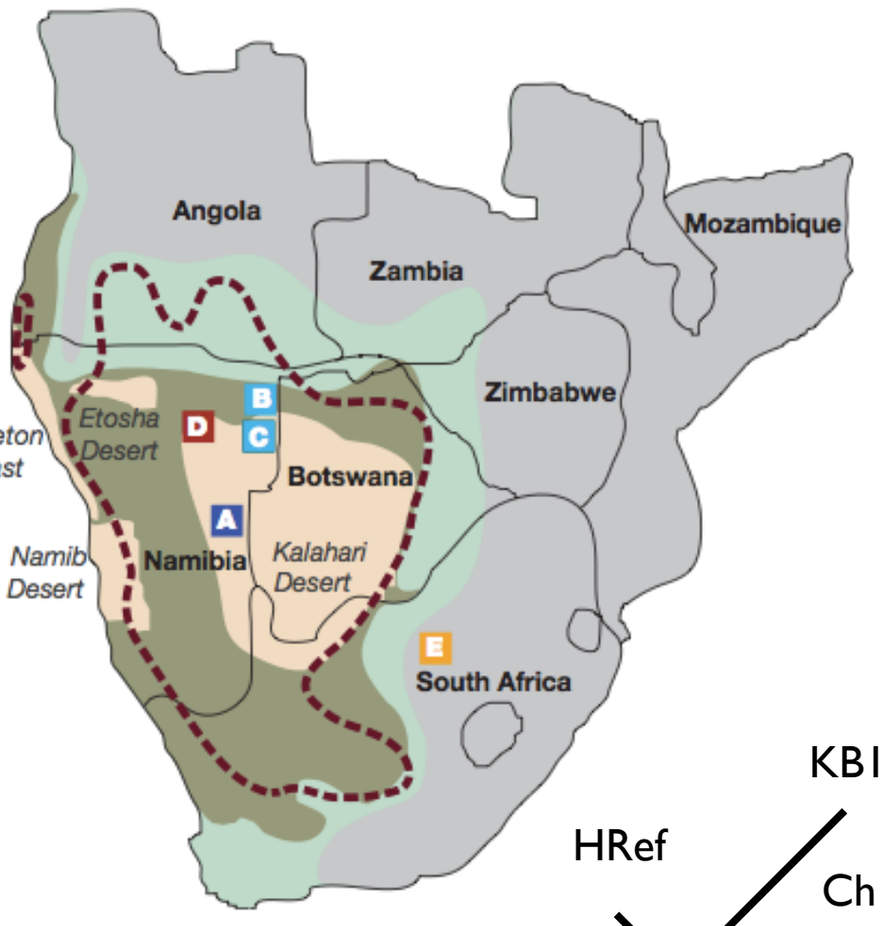


Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

The Oldest Humans -Population Structure

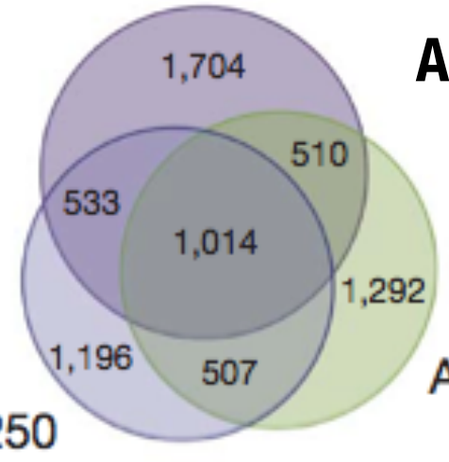


The Bushmen are the oldest inhabitants of southern Africa, where they have lived for at least 20,000 years. Their home is the vast expanse of the Kalahari desert. The Bushmen are remnants of the oldest cultural group in Africa. They are ~ 100,000 Indiv.

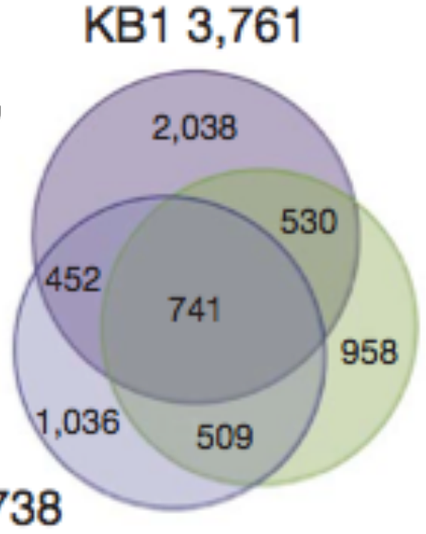


Bushmen show more differences among themselves than those showed by Europeans and Asians

Africans



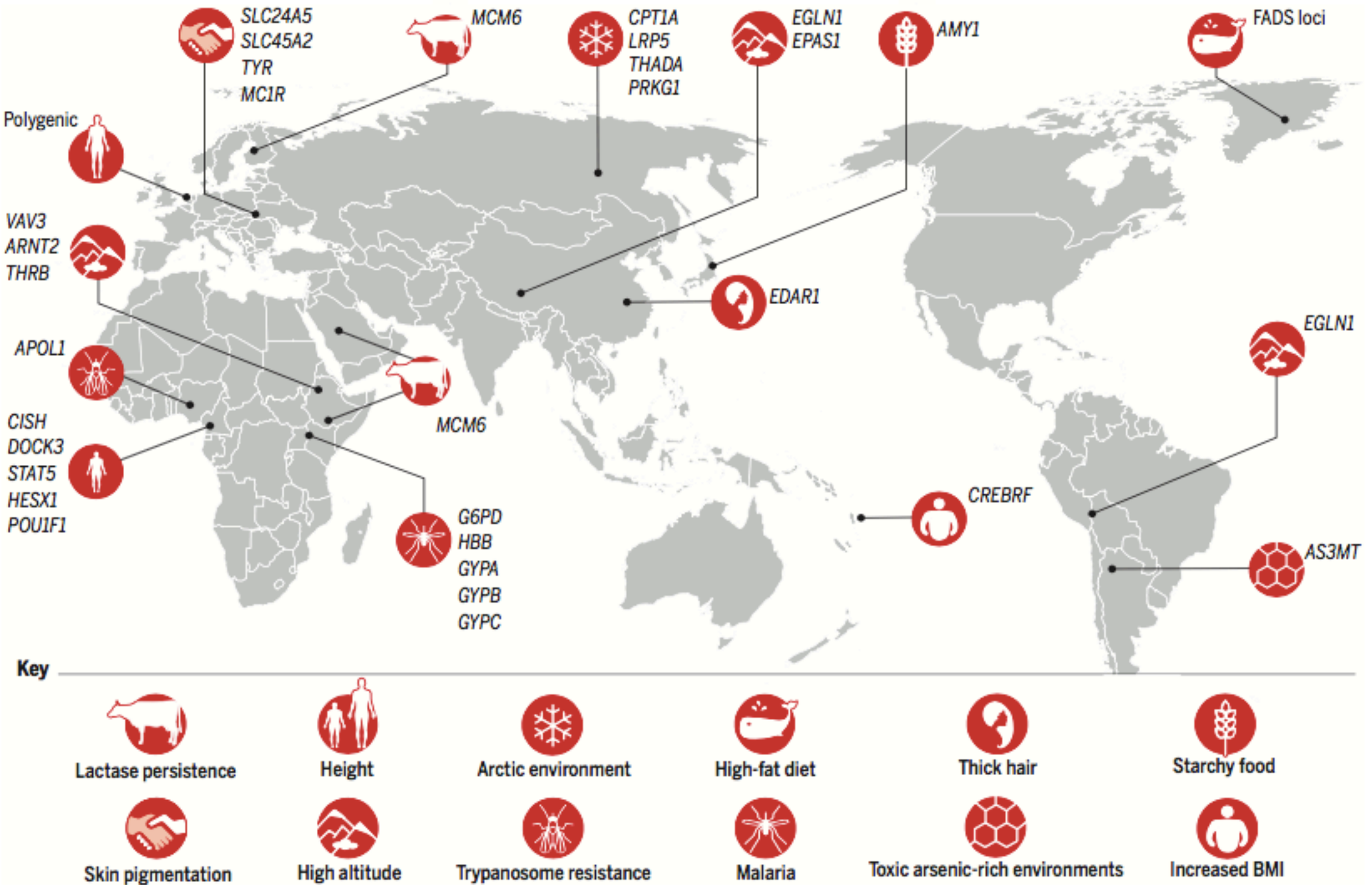
Africans, Asians, Europeans



Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Signatures of Adaptive Selection

Fan, S., Hansen, M. E. B., Lo, Y., & Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. Science, 354(6308), 54–59.

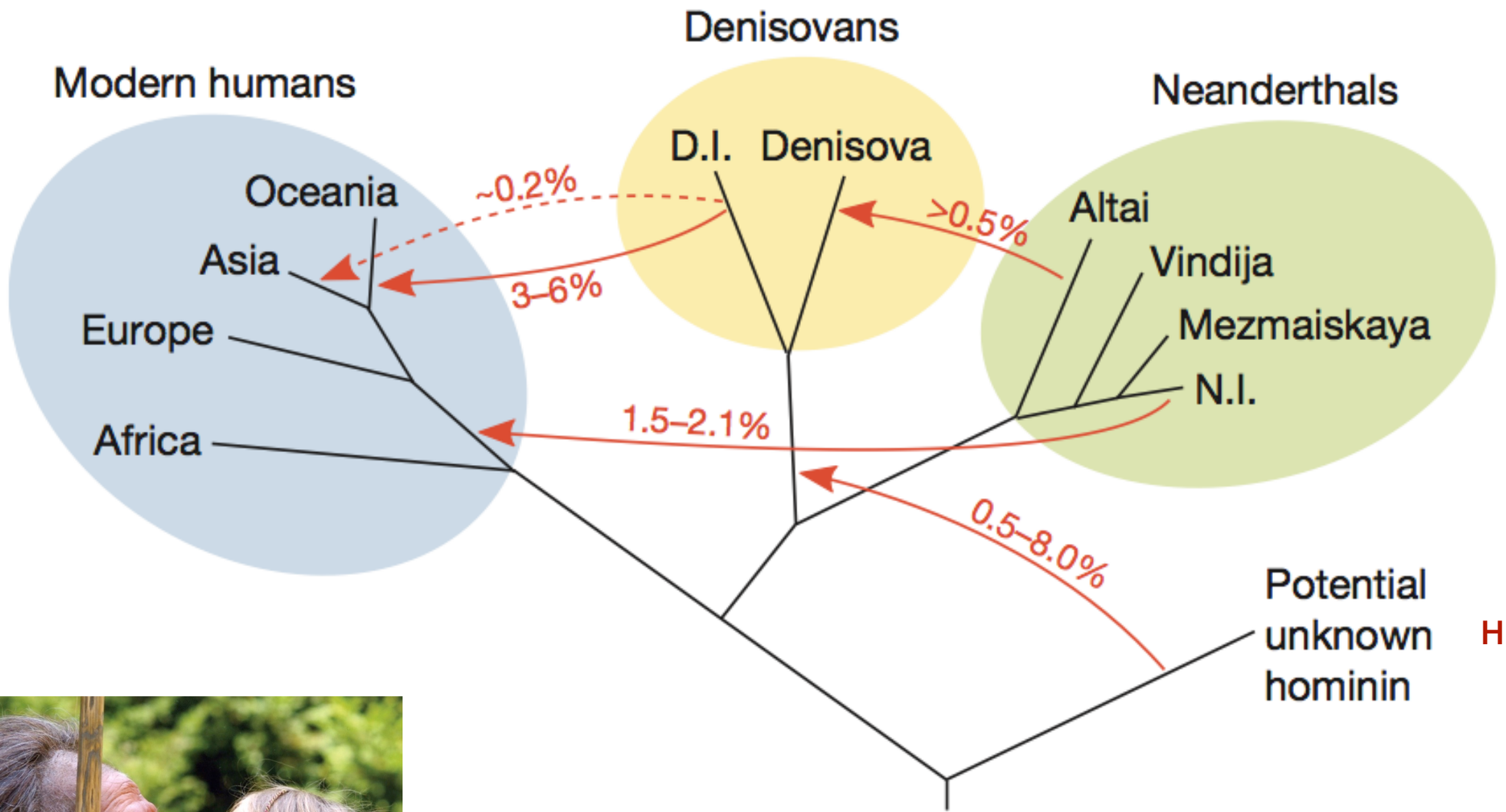


Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Hominid Ancestral Genomes

Prüfer, K., , et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature, 505(7481), 43–49.

A possible model of gene flow events in the Late Pleistocene





All genomes analysed show evidence of a reduction in population size that occurred sometime before 1.0 million years ago. Subsequently, the population ancestral to present-day humans increased in size, whereas the Altai and Denisovan ancestral populations decreased further in size. It is thus clear that the demographic histories of both archaic populations differ substantially from that of present-day humans.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

Genetic Architecture of Diseases

Manolio, T. A., et al. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753.

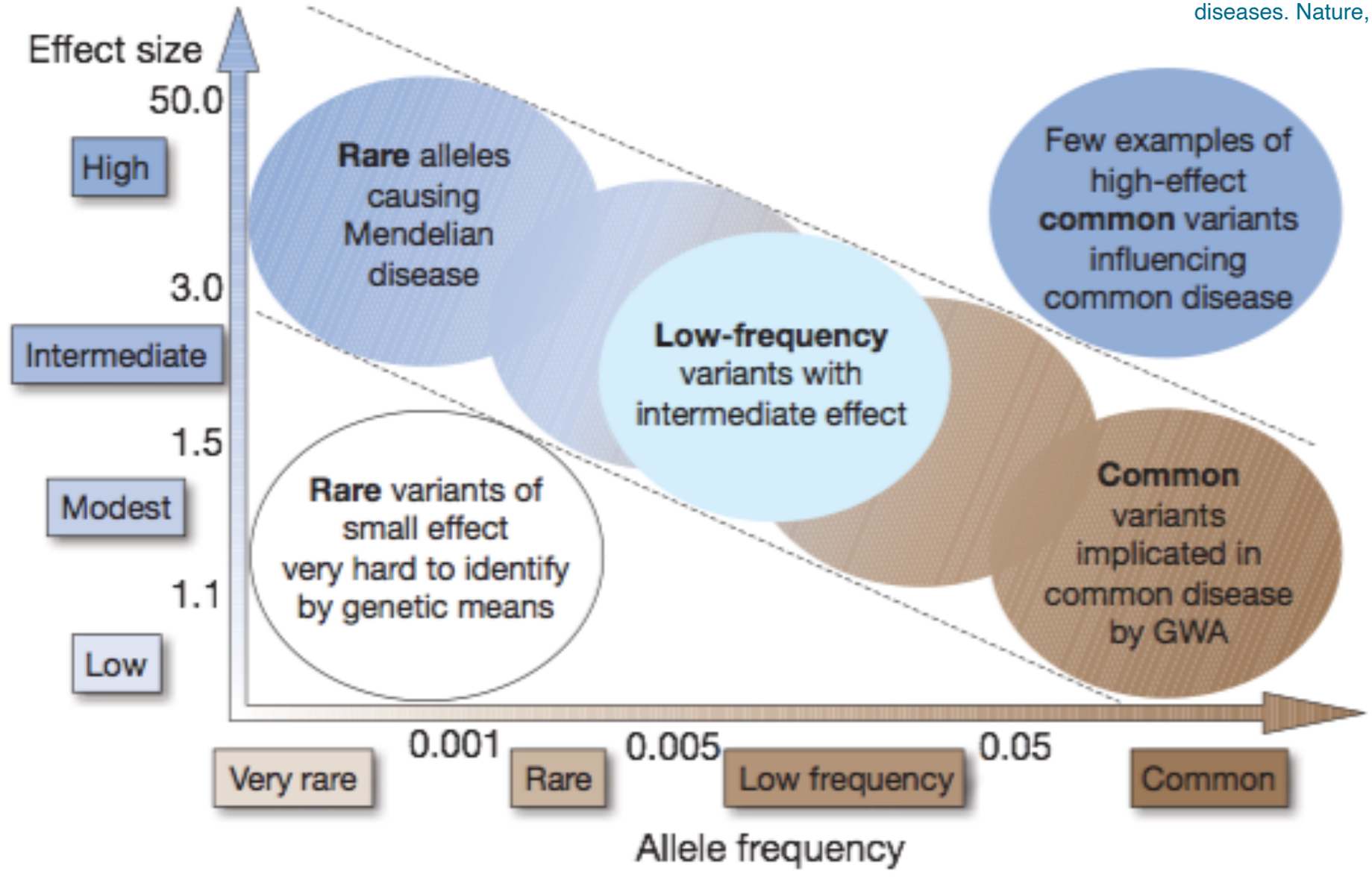
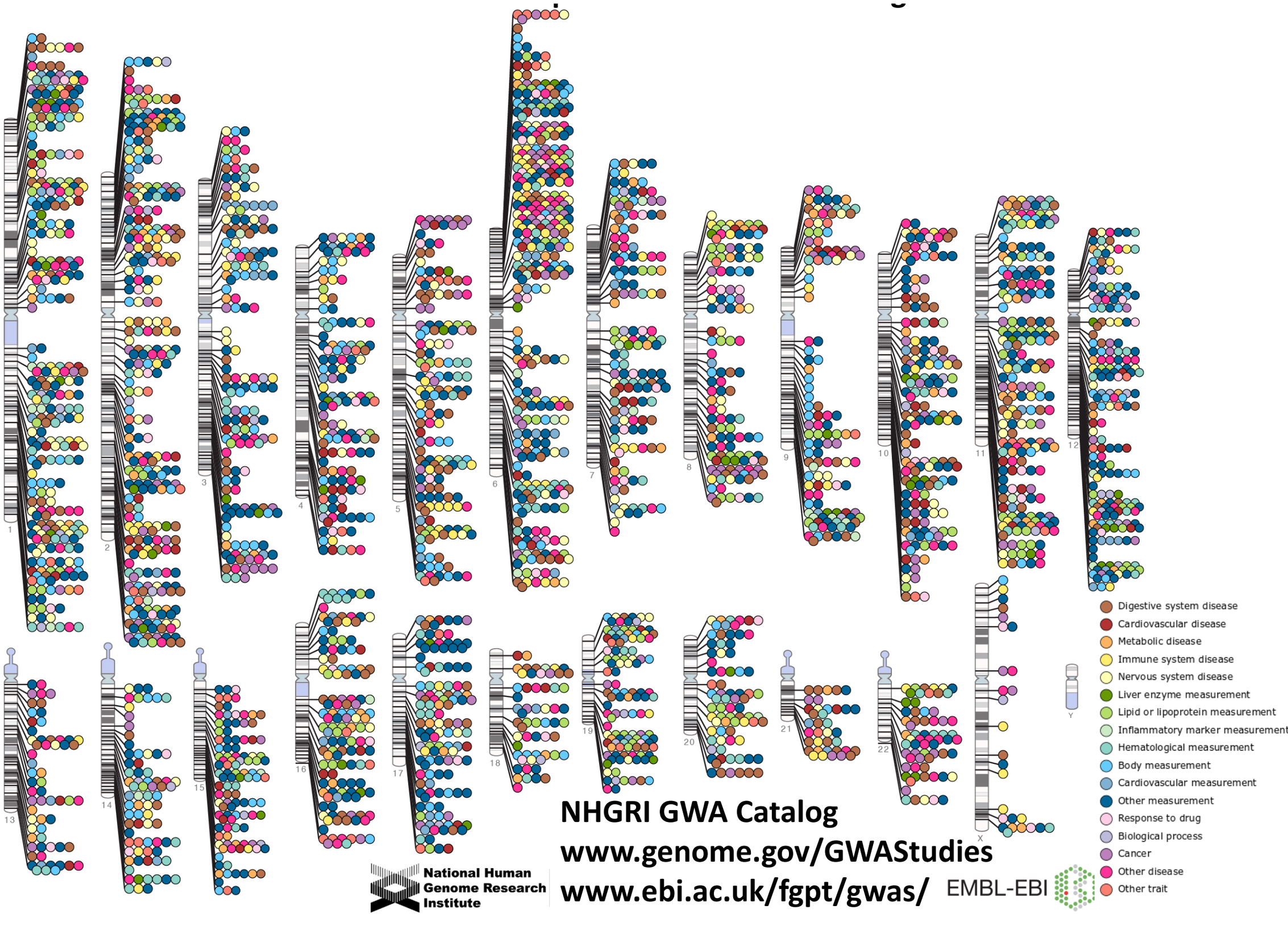


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Prof. Hernán Dopazo
Genómica Evolutiva y Poblacional
EGE-FCEN-CONICET
genevopop.net
@hernandopazo

GWAS Studies



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

National Human Genome Research Institute

EMBL-EBI