Exploring the 1000 Genomes Project

GENE360: tutorial 4

A/Prof Mik Black
Department of Biochemistry
University of Otago

Overview

- 1000 Genomes Project
  - Inception & initial results (pilot)
  - Phase 1 and 3
- Fringe benefits
  - Methodology
  - Data
- Additional topics
  - Population diversity and Admixture
  - Human disease

http://www.1000genomes.org

1000 Genomes Project

- International collaboration
  - Detailed catalogue of human variation
  - Pilot launched Jan 2008 (180 samples, 3 populations)
  - 2500 genomes, 26 populations (main phase: 2009-2015)
- Funding
  - Wellcome Trust Sanger Institute
  - Beijing Genomics Institute
  - National Human Genome Research Institute
- In 2015 the International Genome Sample Resource (IGSR) was established to ensure the ongoing data access.

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

IGSR and the 1000 Genomes Project

Populations: - African; - American; - East Asian; - European; - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.

http://www.1000genomes.org/

Population Code	Population Description	Super Population Code
CHB	Han Chinese in Bejing, China	EAS
JPT	Japanese in Tokyo, Japan	EAS
CHS	Southern Han Chinese	EAS
CDX	Chinese Dai in Xishuangbanna, China	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
CEU	Utah Residents (CEPH) with Northern and Western Ancestry	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian Population in Spain	EUR
YRI	Yoruba in Ibadan, Nigeria	AFR
LWK	Luhya in Webuye, Kenya	AFR
GWD	Gambian in Western Divisions in the Gambia	AFR
MSL	Mende in Sierra Leone	AFR
ESN	Esan in Nigeria	AFR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Caribbeans in Barbados	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
PEL	Peruvians from Lima, Peru	AMR
GIH	Gujarati Indian from Houston, Texas	SAS
PJL	Punjabi from Lahore, Pakistan	SAS
BEB	Bengali from Bangladesh	SAS
STU	Sri Lankan Tamil from the UK	SAS
ITU	Indian Telugu from the UK	SAS

- 26 populations
- 5 super-populations
  - EAS (East Asian)
  - EUR (European)
  - AFR (African)
  - AMR (American)
  - SAS (South Asian)
- Roughly 100 individuals per population

http://www.1000genomes.org/

1000 Genomes project

A new international research consortium that aims to sequence the genomes of at least 1,000 people has just been set up. The project is expected to cost between $30 million and $50 million, and its aim is to uncover more detailed genetic factors involved in human health and disease. The consortium will sequence genomes from at least 1,000 volunteers worldwide to ensure representation of African, Asian and European populations. Support will come from several international institutions, including the Wellcome Trust Sanger Institute in the UK, the Beijing Genomics Institute, Shenzhen, in China, and the US National Human Genome Research Institute (NHGRI), which is part of the National Institutes of Health in Bethesda, Maryland. NHGRI will support and fund three of the large genome centers in the US which will primarily be responsible for producing sequence data for the project. Adam Felsenfeld, NHGRI's director, says that "the project goals are explicit: we want to produce a catalog of human variation down to variants that occur at 1% frequency or less over the genome, and 0.5–0.1% in genes." He adds that the intention of the project "is to provide a resource that will greatly increase the ability of scientists to do genetic studies on common human disease. If that happens, any of the causal variants thus found would be a significant advance." *NS*

NEWS

IN brief

JAMA The Journal of the American Medical Association

1000 Genomes Project Promises Closer Look at Variation in Human Genome

Bridget M. Kuehn

A NEW, LARGE-SCALE PUBLIC SCIence project is developing a more detailed picture of variations in the human genome that may one day aid scientists' understanding of the genetic basis of disease.

Building on the data and technology generated in previous "big science" projects, such as the Human Genome Project and the HapMap (an effort aimed at describing the common patterns of genetic variation in humans), investigators for the 1000 Genomes Project plan to develop an extensive catalog of variation in the human genome by sequencing the genomes of at least 1000 individuals from around the world. The project is being carried out by an international consortium of researchers, including scientists from the National Human Genome Research Institute in Bethesda, Md; the Wellcome Trust Sanger Institute in Hinxton, England, and the Beijing Genomics Institute in Shenzen, China. The first official data from the project will be released in January 2009, said David Altshuler, MD, PhD, cochair of the consortium and professor of genetics and medicine at Harvard Medical School in Boston, Mass.

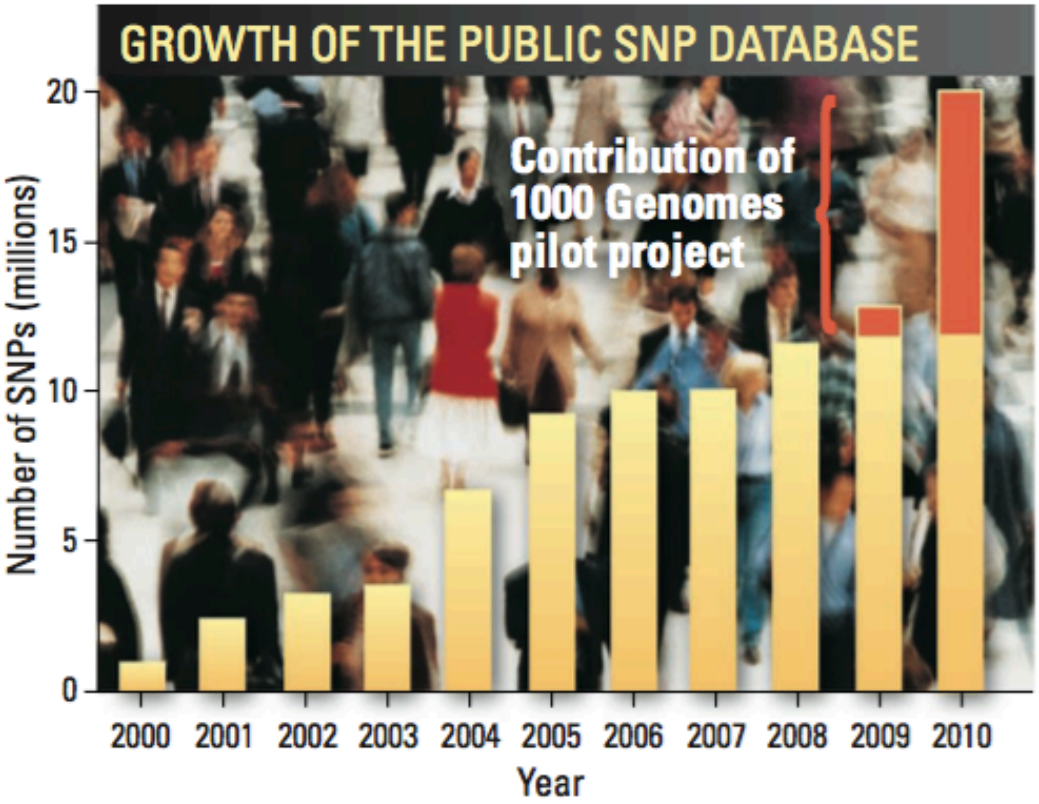
While it took years to sequence a single human genome during the Human Genome Project, new techniques and technologies are allowing the researchers to sequence DNA with much greater speed and at much lower cost. In fact, the project will be able to sequence approximately 8.2 billion bases per day—the equivalent of about 2 human genomes every 24 hours. Although the cost of the project has been estimated to be between $30 and $50 million, Altshuler noted that a precise estimate of cost is not currently available because the technological advances continue to reduce the costs of sequencing.

The project will reveal variation in the human genome at a higher resolution than previously possible, documenting single nucleotide polymorphisms and larger genetic variations, such as copy number variations. It also will produce a database that scientists will be able to search for variations in regions of the genome that have been linked to disease. Researchers will then be able to further "interrogate" those data to identify disease-causing variations and elucidate the underlying disease pathways, explained David Valle, MD, director of the Institute of Genetic Medicine at Johns Hopkins School of Medicine in Baltimore.

For example, Valle said, retinitis pigmentosa has been linked to variations in more than 100 genes; the fact that all these genes are involved in the phototransduction cascade suggests that targeting this pathway might be a fruitful avenue to pursue for potential interventions. □

GENOMICS

1000 Genomes Project Gives New Map Of Genetic Diversity



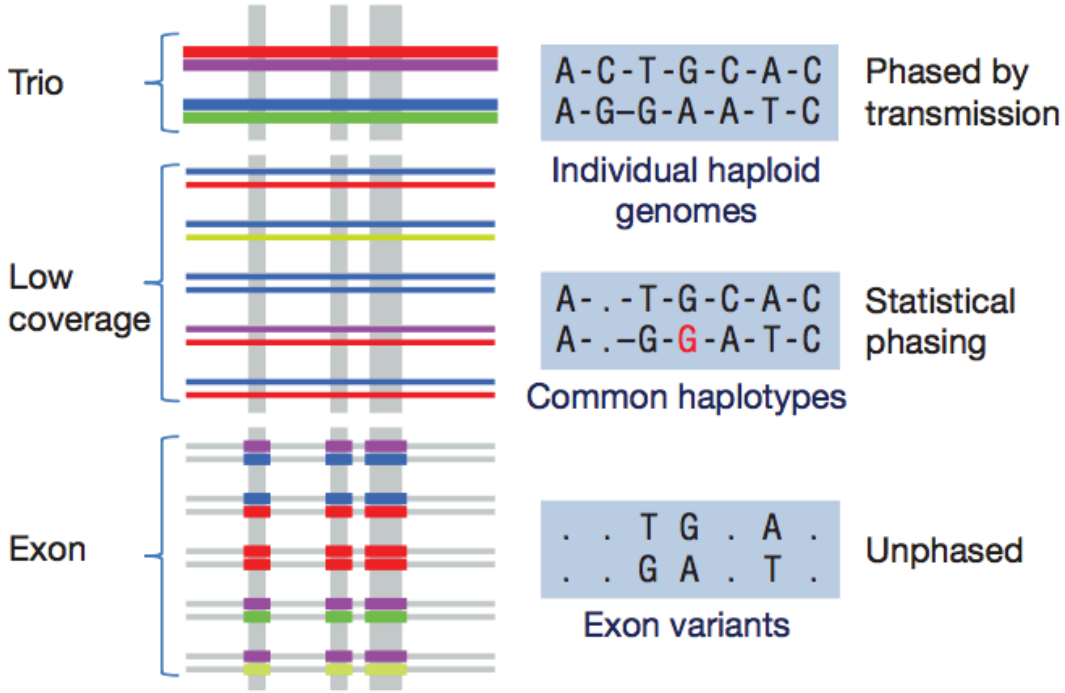
Pinning down differences. The 1000 Genomes Project has greatly increased the number of known single-base differences that can exist among people.

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

BOX 1

The 1000 Genomes pilot projects

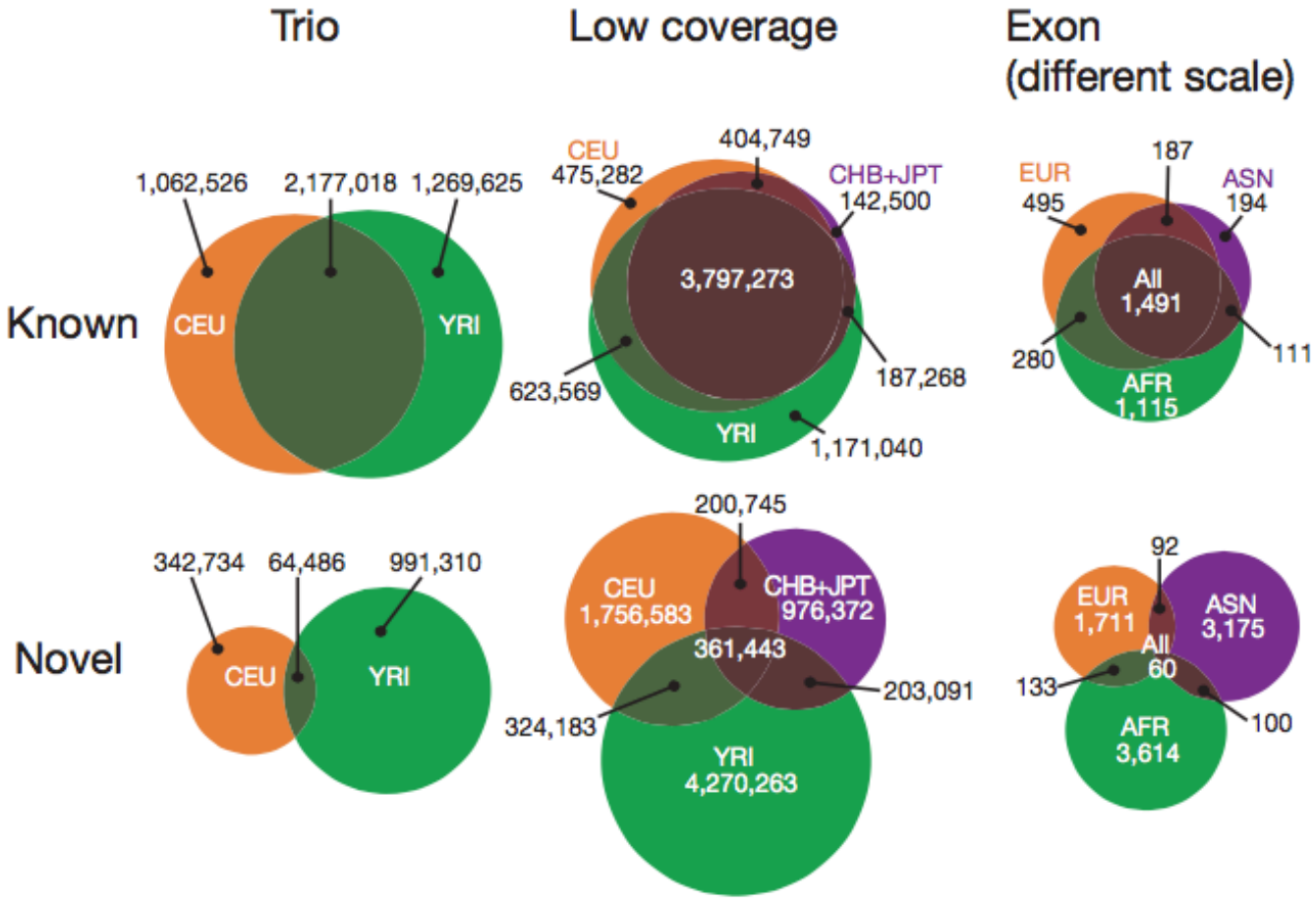


Pilot (low coverage)
4 populations

- CEU: 60
- YRI: 59
- CHB: 30
- JPT: 30

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*



A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

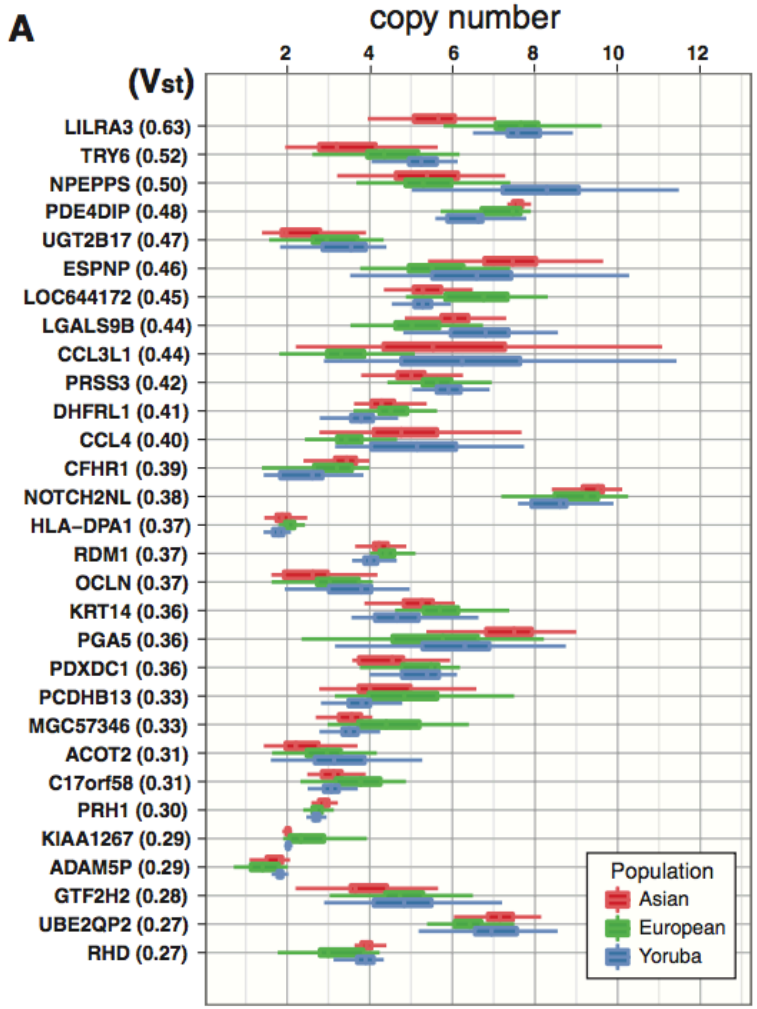
BOX 2

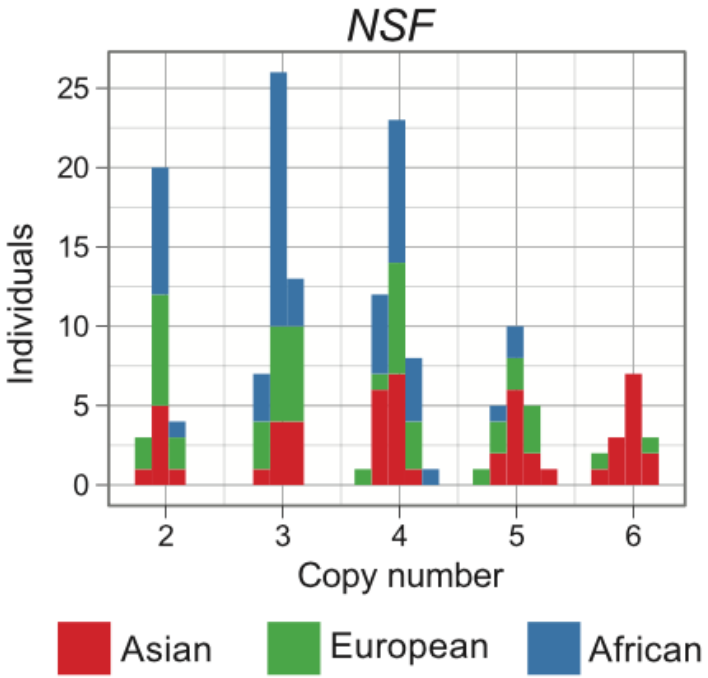
Design of the full 1000 Genomes Project

The production phase of the full 1000 Genomes Project will combine low-coverage whole-genome sequencing, array-based genotyping, and deep targeted sequencing of all coding regions in 2,500 individuals from five large regions of the world (five population samples of 100 in or with ancestry from each of Europe, East Asia, South Asia and West Africa, and seven populations totalling 500 from the Americas; Supplementary Table 9). We will increase the low-coverage average depth to over 4× per individual, and use blood-derived DNA where possible to minimize somatic and cell-line false positives.

Diversity of Human Copy Number Variation and Multicopy Genes

Peter H. Sudmant,[1]* Jacob O. Kitzman,[1]* Francesca Antonacci,[1] Can Alkan,[1] Maika Malig,[1]
Anya Tsalenko,[2] Nick Sampas,[2] Laurakay Bruhn,[2] Jay Shendure,[1]
1000 Genomes Project,† Evan E. Eichler[1,3]‡





Fringe benefits

- The 1000 Genomes Project has been hugely beneficial for bioinformatics
  - Standardized data formats
  - Improved mapping and variant calling algorithms
  - Access to large amounts of genotype data
  - Protocols developed for storage, management and sharing of large data sets
- Also taught us A LOT about sequence data, and the characteristics of different sequencing technologies.

BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 21 2011, pages 2987–2993
doi:10.1093/bioinformatics/btr509

Sequence analysis

Advance Access publication September 8, 2011

A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data

Heng Li

Medical Population Genetics Program, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Jeffrey Barrett

The 1000 Genomes Project: data management and community access

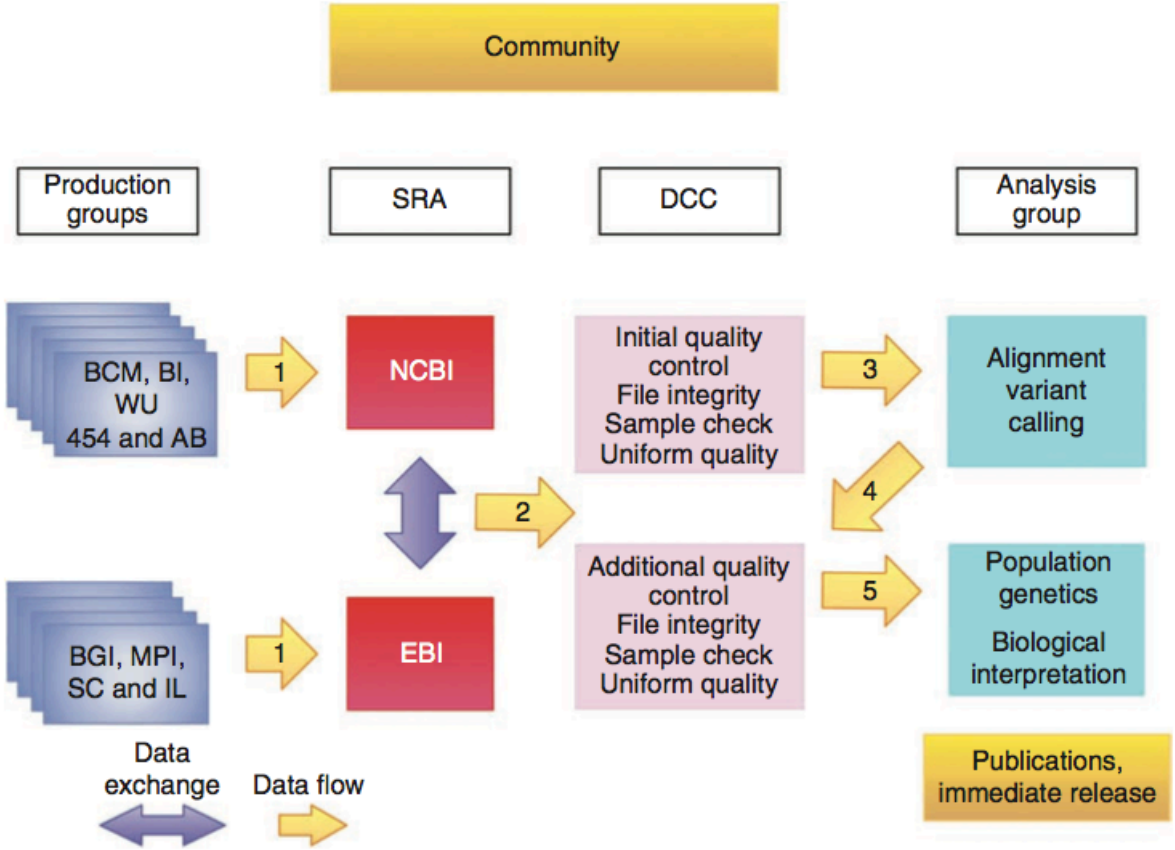
Laura Clarke[1], Xiangqun Zheng-Bradley[1], Richard Smith[1], Eugene Kulesha[1], Chunlin Xiao[2], Iliana Toneva[1], Brendan Vaughan[1], Don Preuss[2], Rasko Leinonen[1], Martin Shumway[2], Stephen Sherry[2], Paul Flicek[1] & The 1000 Genomes Project Consortium[3]

The 1000 Genomes Project was launched as one of the largest distributed data collection and analysis projects ever undertaken in biology. In addition to the primary scientific goals of creating both a deep catalog of human genetic variation and extensive methods to accurately discover and characterize variation using new sequencing technologies, the project makes all of its data publicly available. Members of the project data coordination center have developed and deployed several tools to enable widespread data access.

NATURE METHODS | VOL.9 NO.5 | MAY 2012

The 1000 Genomes Project: data management and community access

Laura Clarke[1], Xiangqun Zheng-Bradley[1], Richard Smith[1], Eugene Kulesha[1], Chunlin Xiao[2], Iliana Toneva[1], Brendan Vaughan[1], Don Preuss[2], Rasko Leinonen[1], Martin Shumway[2], Stephen Sherry[2], Paul Flicek[1] & The 1000 Genomes Project Consortium[3]

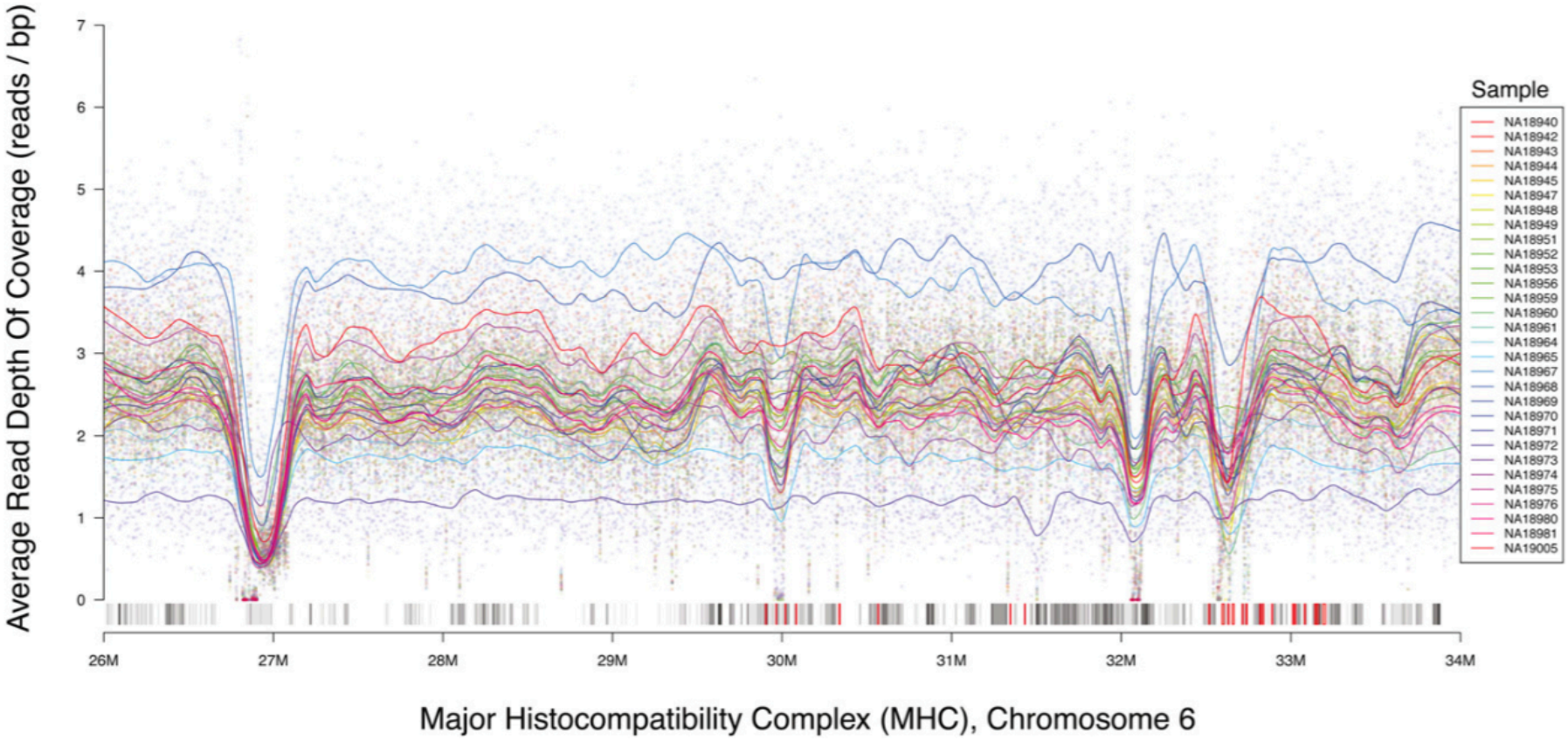


NATURE METHODS | VOL.9 NO.5 | MAY 2012

The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data

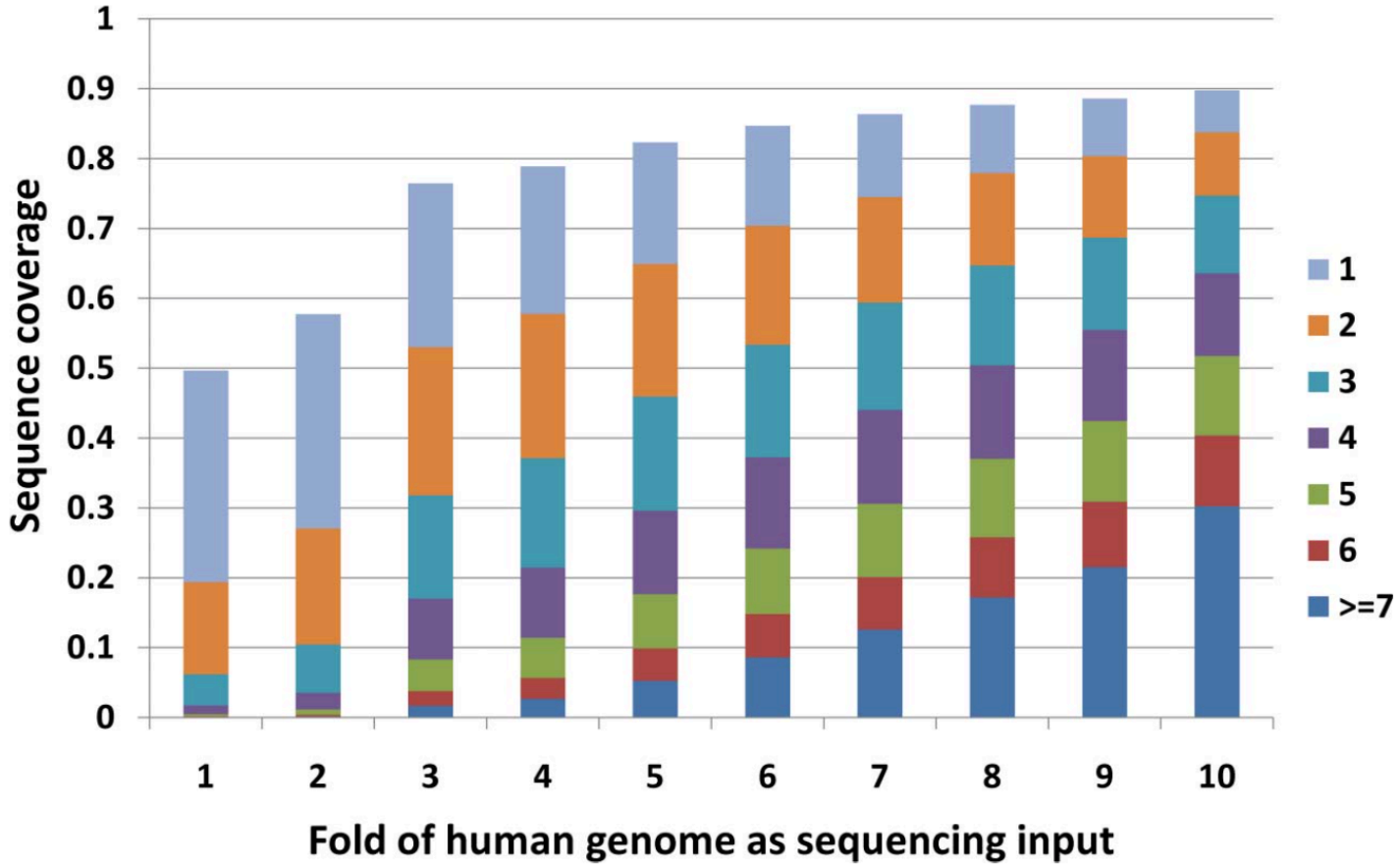
Aaron McKenna,[1] Matthew Hanna,[1] Eric Banks,[1] Andrey Sivachenko,[1] Kristian Cibulskis,[1] Andrew Kernytsky,[1] Kiran Garimella,[1] David Altshuler,[1,2] Stacey Gabriel,[1] Mark Daly,[1,2] and Mark A. DePristo[1,3]

[1]*Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA;* [2]*Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, Massachusetts 02114, USA*



Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions

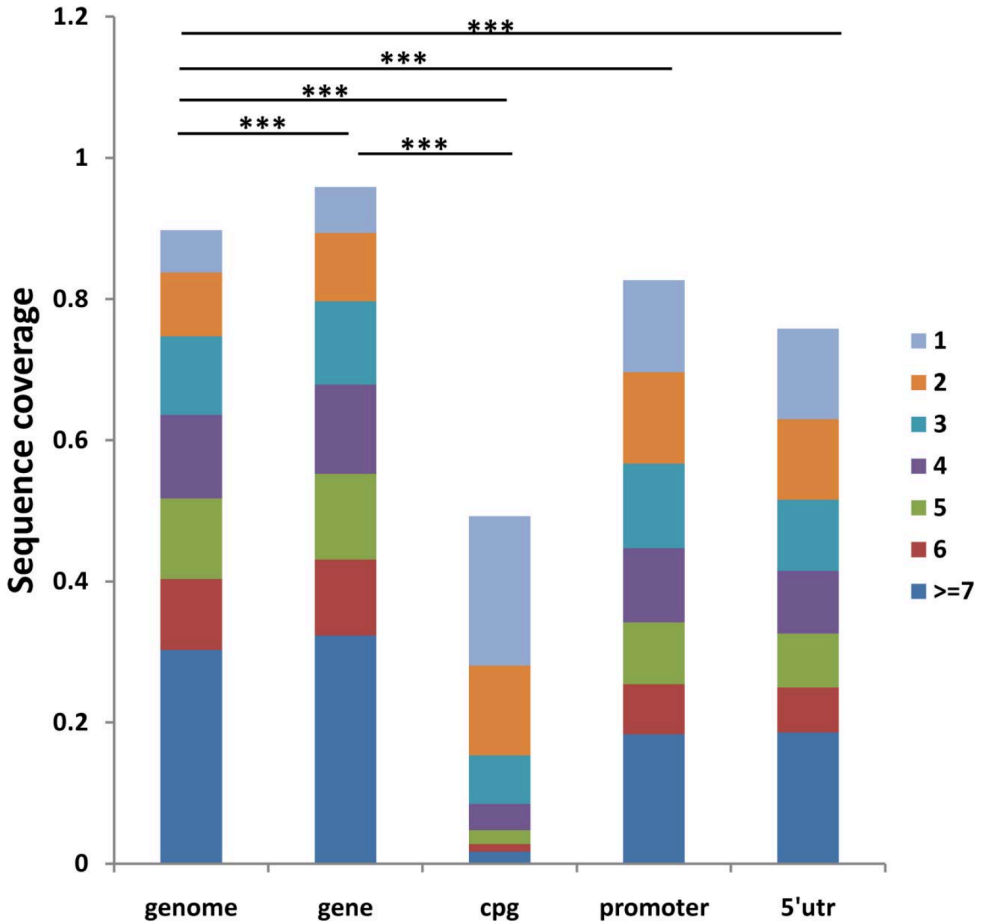
Weixin Wang[1], Zhi Wei[2], Tak-Wah Lam[3] & Junwen Wang[1]



SCIENTIFIC REPORTS | 1 : 55 | DOI: 10.1038/srep00055

Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions

Weixin Wang[1], Zhi Wei[2], Tak-Wah Lam[3] & Junwen Wang[1]



SCIENTIFIC REPORTS | 1 : 55 | DOI: 10.1038/srep00055

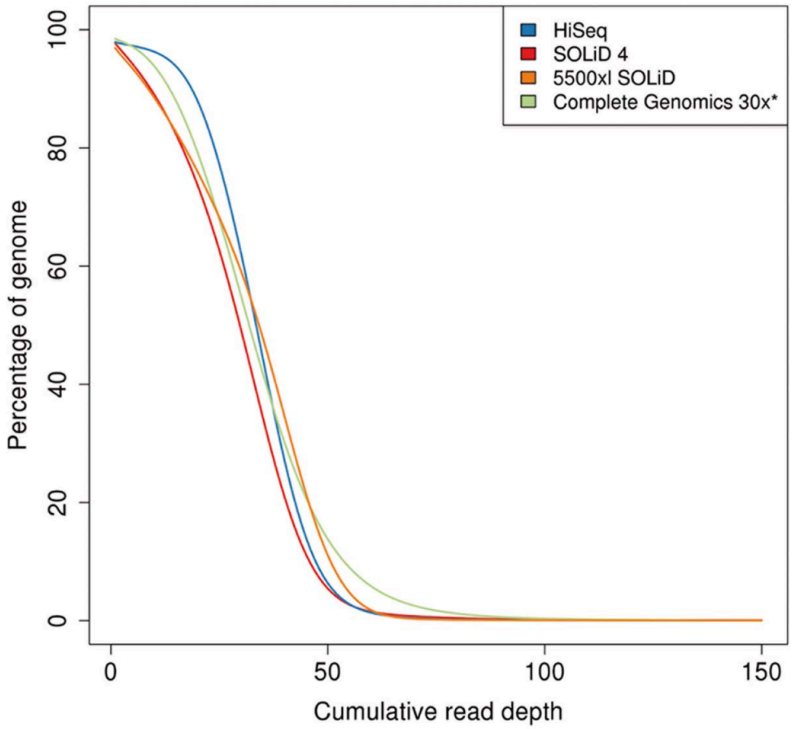
OPEN ACCESS Freely available online

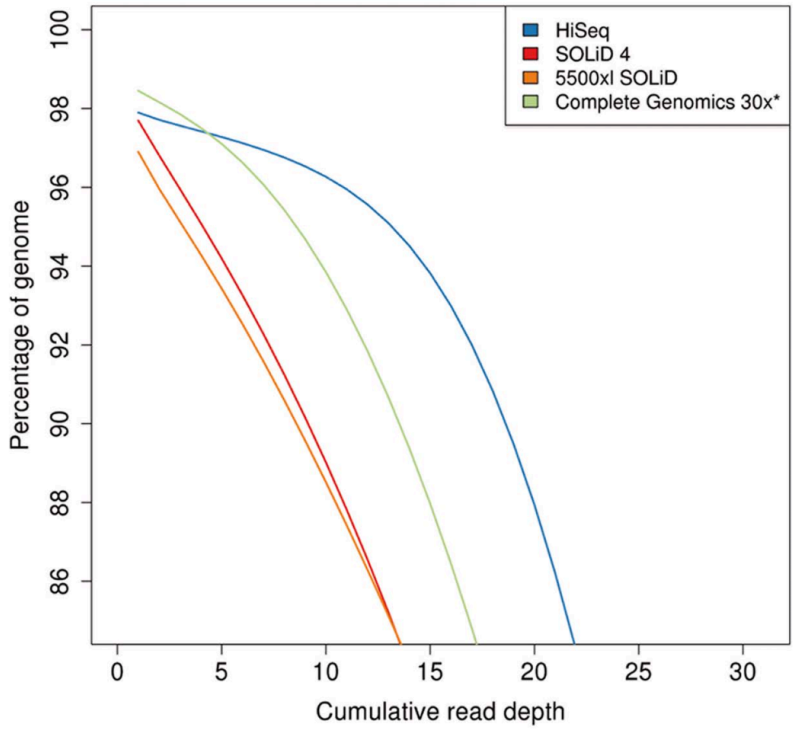
PLOS ONE

Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies

Nora Rieber[1☯], Marc Zapatka[2☯], Bärbel Lasitschka[3], David Jones[4], Paul Northcott[5], Barbara Hutter[1], Natalie Jäger[1], Marcel Kool[4], Michael Taylor[5,6], Peter Lichter[2], Stefan Pfister[4,7], Stephan Wolf[3], Benedikt Brors[1], Roland Eils[1,8]*

1 Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany, **2** Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany, **3** Genomics and Proteomics Core Facility, High Throughput Sequencing Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany, **4** Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany, **5** The Arthur and Sonia Labatt Brain Tumor Research Centre, The Hospital for Sick Children Research Institute, University of Toronto, Ontario, Canada, **6** Division of Neurosurgery, The Hospital for Sick Children, University of Toronto, Ontario, Canada, **7** Department of Pediatric Hematology and Oncology, Heidelberg University Hospital, Heidelberg, Germany, **8** Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, and Bioquant, University of Heidelberg, Heidelberg, Germany





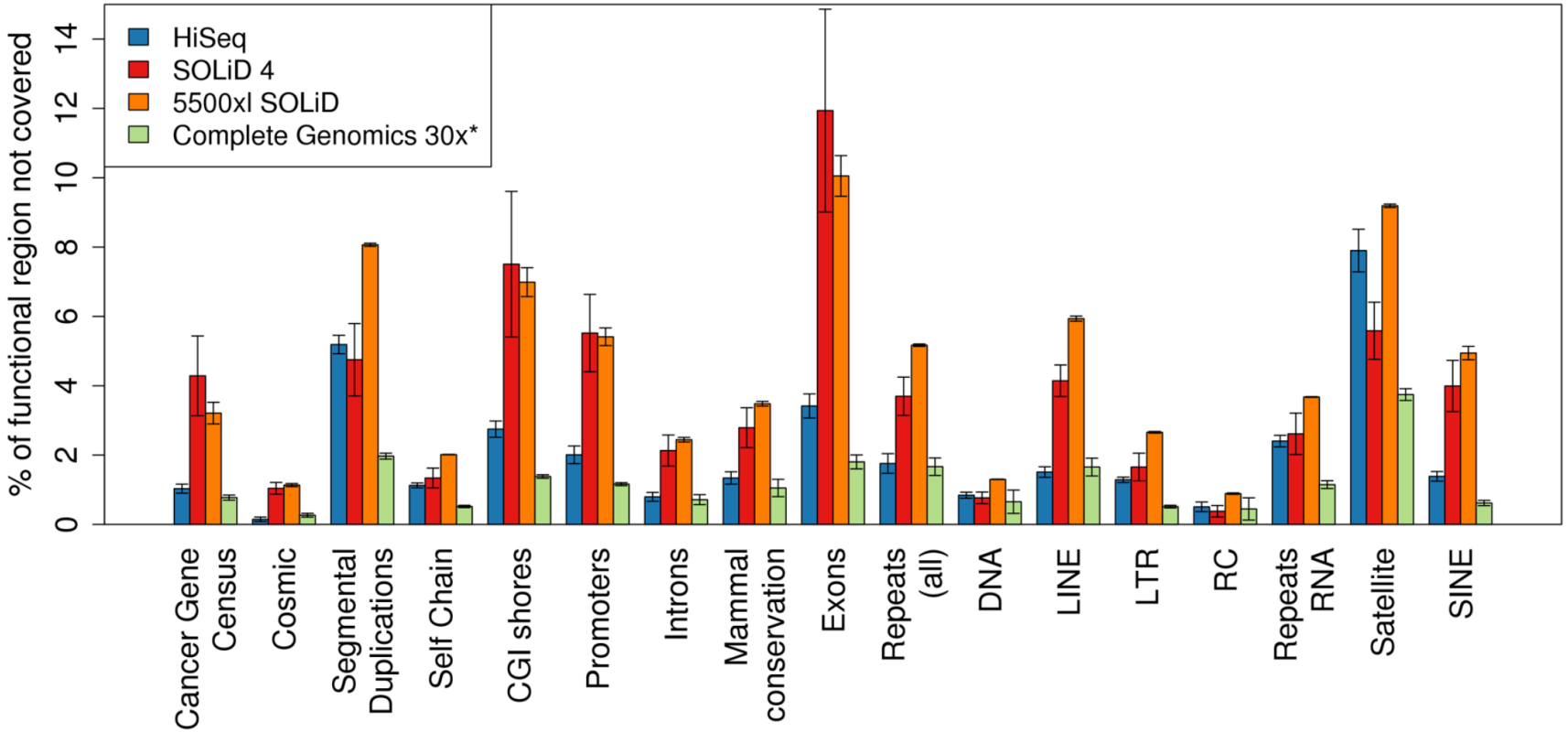
OPEN ACCESS Freely available online

PLOS ONE

Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies

Nora Rieber[1☯], Marc Zapatka[2☯], Bärbel Lasitschka[3], David Jones[4], Paul Northcott[5], Barbara Hutter[1], Natalie Jäger[1], Marcel Kool[4], Michael Taylor[5,6], Peter Lichter[2], Stefan Pfister[4,7], Stephan Wolf[3], Benedikt Brors[1], Roland Eils[1,8]*

1 Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany, **2** Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany, **3** Genomics and Proteomics Core Facility, High Throughput Sequencing Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany, **4** Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany, **5** The Arthur and Sonia Labatt Brain Tumor Research Centre, The Hospital for Sick Children Research Institute, University of Toronto, Ontario, Canada, **6** Division of Neurosurgery, The Hospital for Sick Children, University of Toronto, Ontario, Canada, **7** Department of Pediatric Hematology and Oncology, Heidelberg University Hospital, Heidelberg, Germany, **8** Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, and Bioquant, University of Heidelberg, Heidelberg, Germany



1000 Genomes – phase 1

- Phase 1: low coverage and exome data analysis
- Results of phase 1 published in late 2012
  - 14 populations, 1092 individuals

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

Table 1 | Summary of 1000 Genomes Project phase I data

	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K
Novelty rate†	58%	77%	50%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K	199/293/6.3 K
Average no. SNPs per sample	3.60 M	105 K	24.0 K
Indels			
No. sites overall	1.38 M	59 K	1,867
Novelty rate†	62%	73%	54%
No. inframe/frameshift	NA	19/14	719/1,066
Average no. indels per sample	344 K	13 K	440
Genotyped large deletions			
No. sites overall	13.8 K	432	847
Novelty rate†	54%	54%	50%
Average no. variants per sample	717	26	39

NA, not applicable.

* Autosomal genes only.

† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

1000 Genomes – phase 3

- Data available at the “end” of the 1000 Genomes project (2015):

Available data

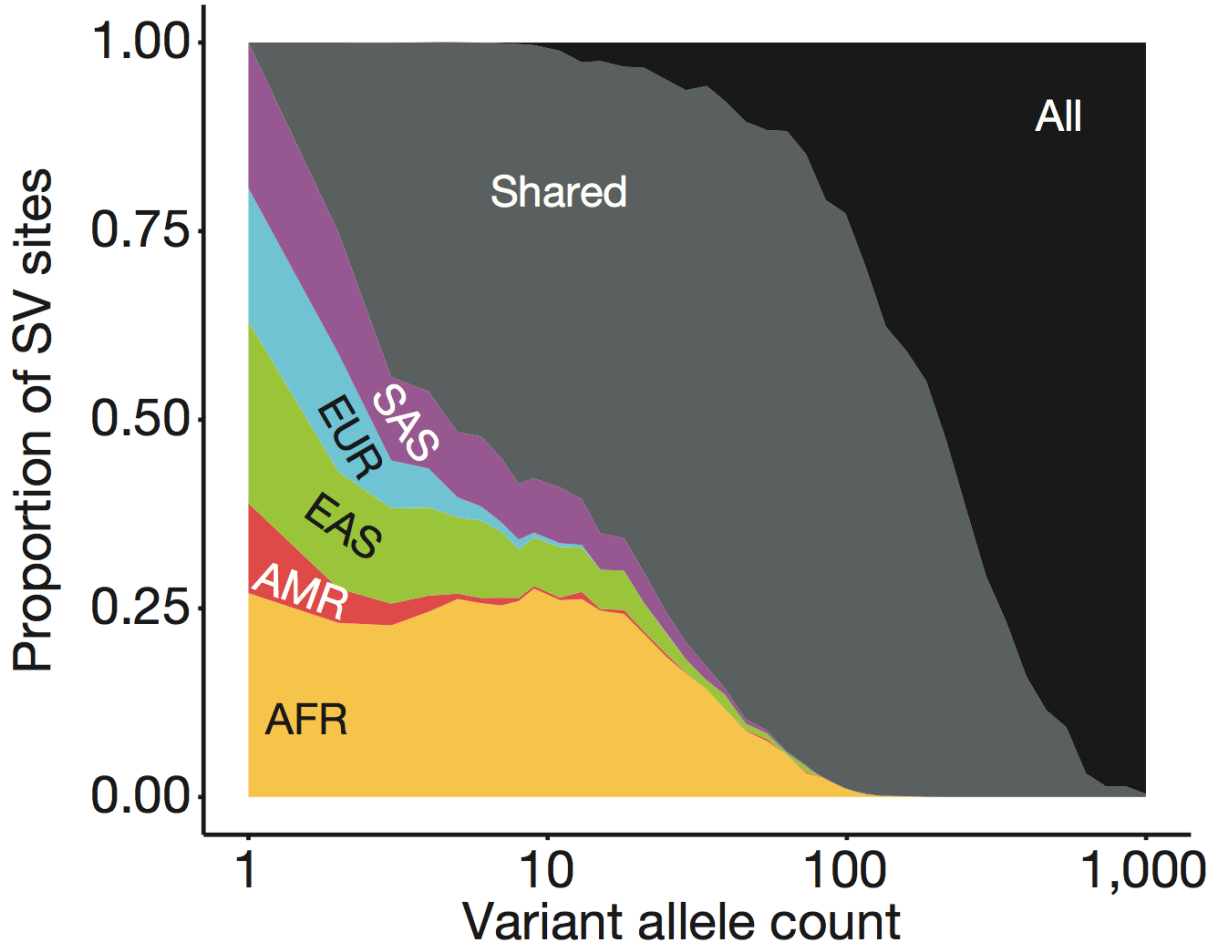
1000 Genomes Project

1000 Genomes Release	Variants	Individuals	Populations	VCF	Alignments	Supporting Data
Phase 3	84.4 million	2504	26	VCF	Alignments	Supporting Data
Phase 1	37.9 million	1092	14	VCF	Alignments	Supporting Data
Pilot	14.8 million	179	4	VCF	Alignments	Supporting Data

http://www.1000genomes.org/data

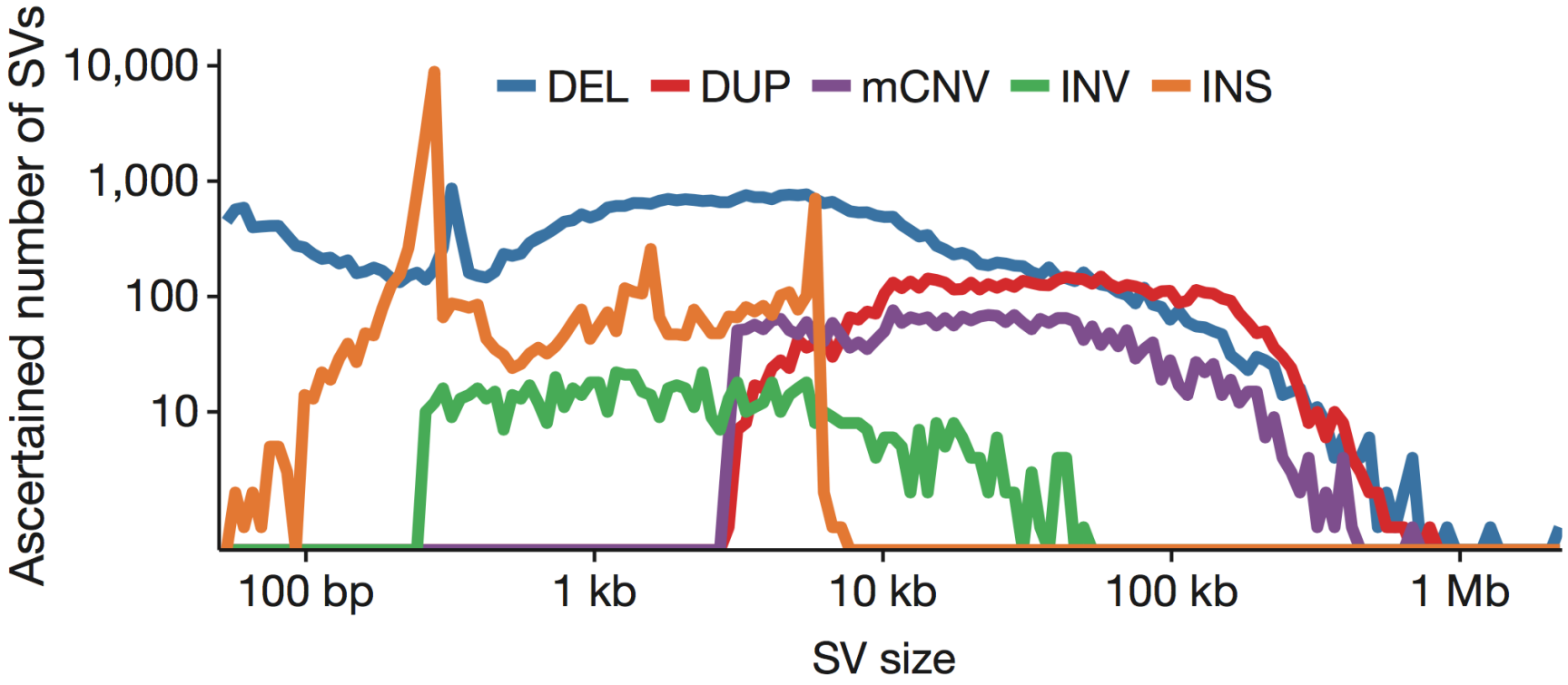
An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.



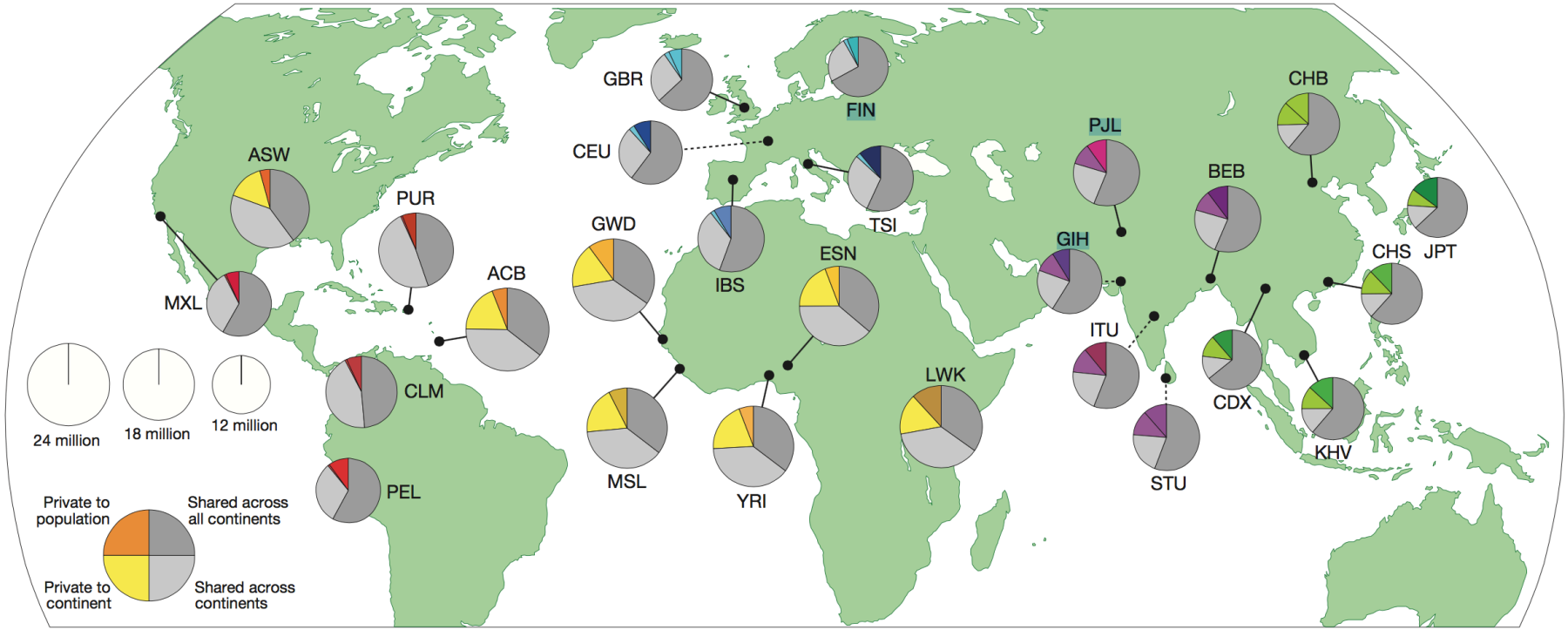
An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.



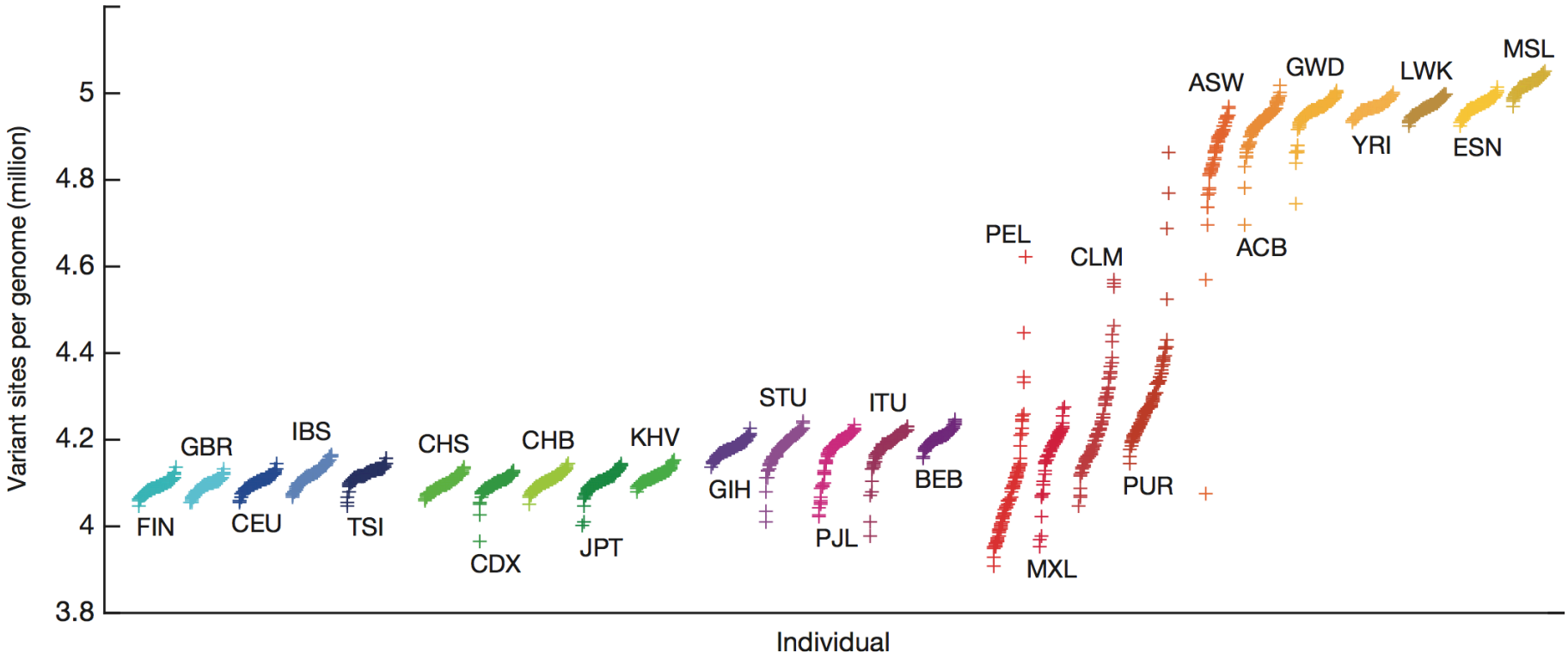
A global reference for human genetic variation

The 1000 Genomes Project Consortium*



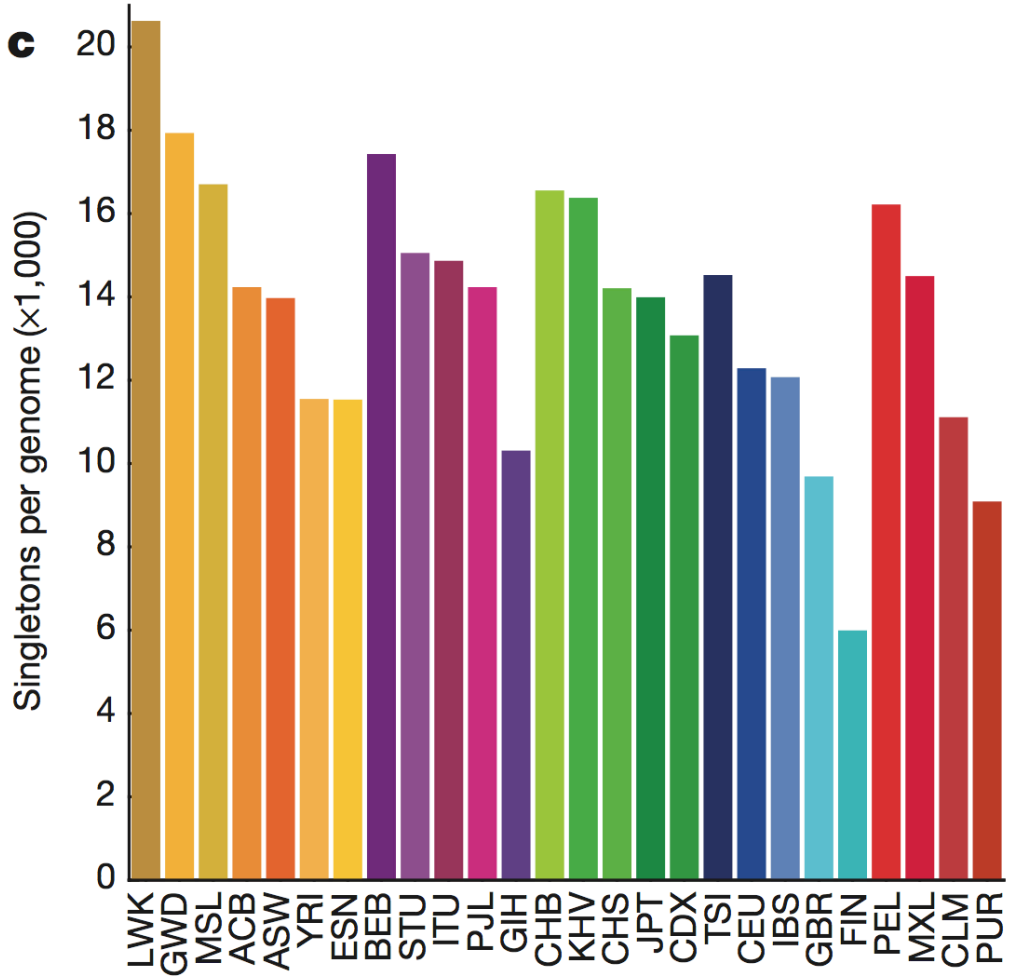
A global reference for human genetic variation

The 1000 Genomes Project Consortium*



A global reference for human genetic variation

The 1000 Genomes Project Consortium*



Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences

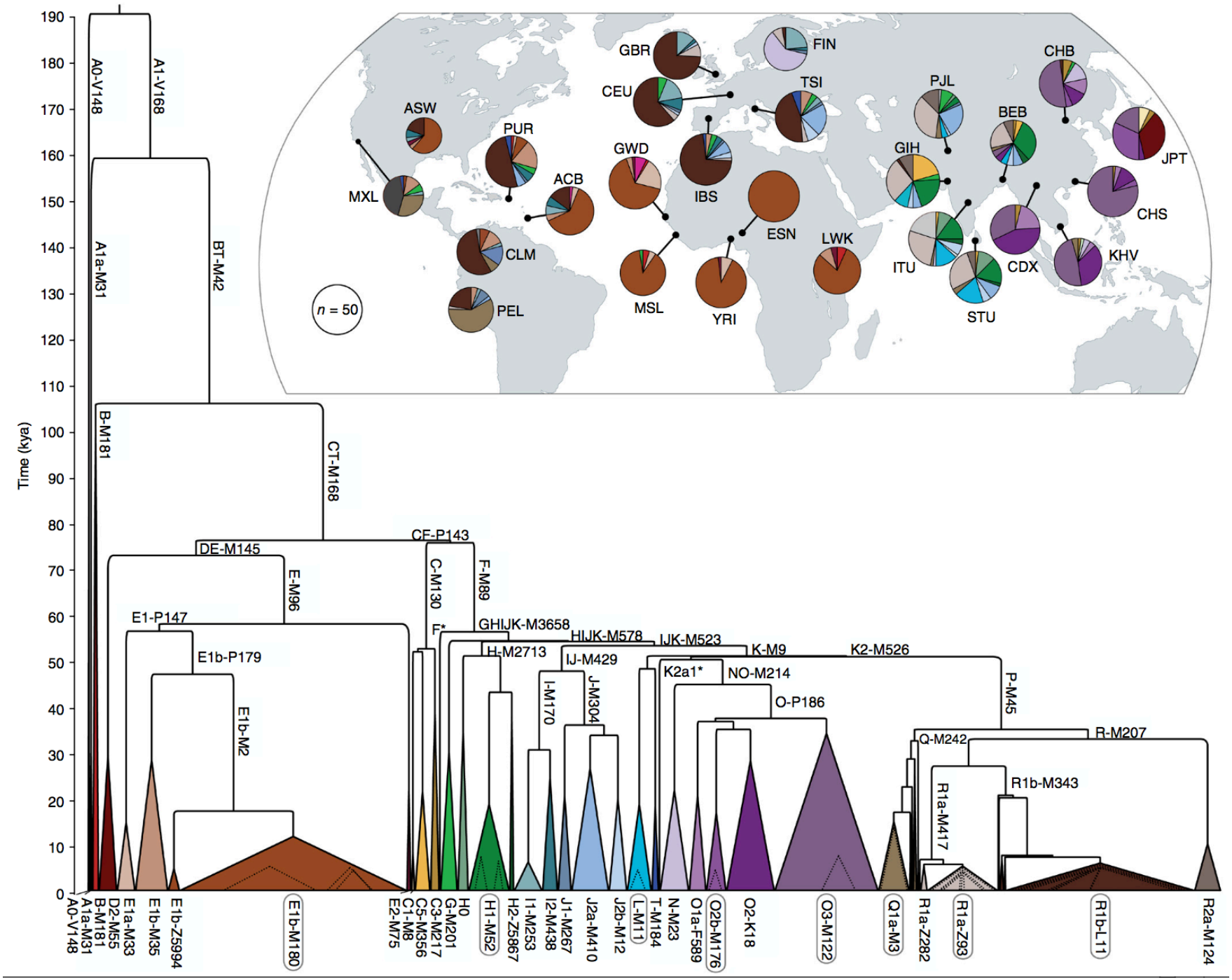
G David Poznik[1,2,25], Yali Xue[3,25], Fernando L Mendez[2], Thomas F Willems[4,5], Andrea Massaia[3], Melissa A Wilson Sayres[6,7], Qasim Ayub[3], Shane A McCarthy[3], Apurva Narechania[8], Seva Kashin[9], Yuan Chen[3], Ruby Banerjee[3], Juan L Rodriguez-Flores[10], Maria Cerezo[3], Haojing Shao[11], Melissa Gymrek[5,12], Ankit Malhotra[13], Sandra Louzada[3], Rob Desalle[8], Graham R S Ritchie[3,14], Eliza Cerveira[13], Tomas W Fitzgerald[3], Erik Garrison[3], Anthony Marcketta[15], David Mittelman[16,17], Mallory Romanovitch[13], Chengsheng Zhang[13], Xiangqun Zheng-Bradley[14], Gonçalo R Abecasis[18], Steven A McCarroll[19], Paul Flicek[14], Peter A Underhill[2], Lachlan Coin[11], Daniel R Zerbino[14], Fengtang Yang[3], Charles Lee[13,20], Laura Clarke[14], Adam Auton[15], Yaniv Erlich[5,21,22], Robert E Handsaker[9,19], The 1000 Genomes Project Consortium[23], Carlos D Bustamante[2,24] & Chris Tyler-Smith[3]

Table 1 Y-chromosome variants discovered in 1,244 males

Variant type	Number	FDR (%)	Concordance (%)
SNVs	60,555	3.9	99.6
Indels and MNVs	1,427	3.6	96.4
CNVs	110	2.7	86
STRs	3,253	NA	89–97

The concordance shown is with independent genotype calls, and the CNVs considered were those computationally inferred using GenomeSTRiP. FDR, false discovery rate; NA, not available.

Received 8 November 2015; accepted 1 April 2016; published online 25 April 2016; doi:10.1038/ng.3559

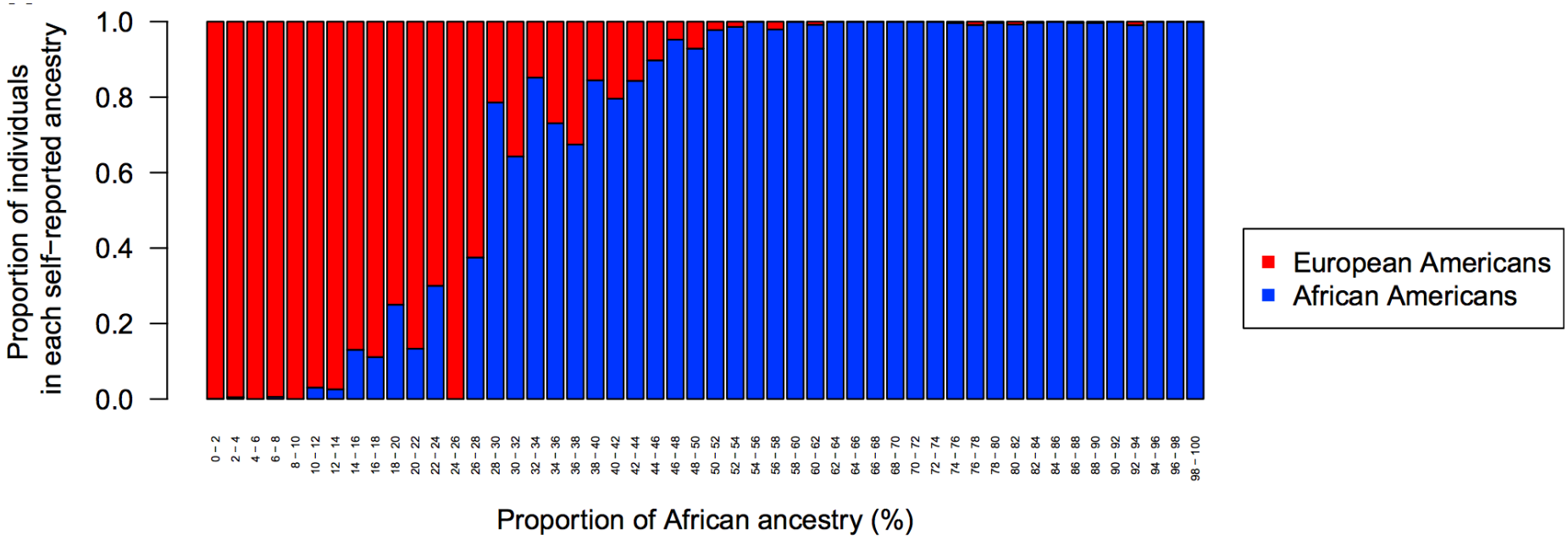

Time (kya)
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
A0-V148
A1-V168
A1a-M31
BT-M42
B-M181
CT-M168
DE-M145
CF-P143
E-M96
C-M130
F-M89
E1-P147
F*
GHIJK-M3658
E1b-P179
H-M2713
HIJK-M578
IJ-M429
IJK-M523
K-M9
K2-M526
E1b-M2
I-M170
J-M304
K2a1*
NO-M214
O-P186
P-M45
Q-M242
R-M207
R1a-M417
R1b-M343
A0-V148
A1a-M31
B-M181
D2-M55
E1a-M33
E1b-M35
E1b-Z5994
E1b-M180
C1-M8
C5-M356
C3-M217
G-M201
H0
H1-M52
H2-Z5867
I1-M253
I2-M438
J1-M267
J2a-M410
J2b-M12
L-M11
T-M184
N-M23
O1a-F589
O2b-M176
O2-K18
O3-M122
Q1a-M3
R1a-Z282
R1a-Z93
R1b-L11
R2a-M124
n = 50
ASW
MXL
PUR
ACB
CLM
PEL
GBR
CEU
GWD
IBS
MSL
YRI
FIN
TSI
ESN
LWK
PJL
GIH
ITU
BEB
CHB
JPT
CHS
CDX
KHV
STU

Additional topics

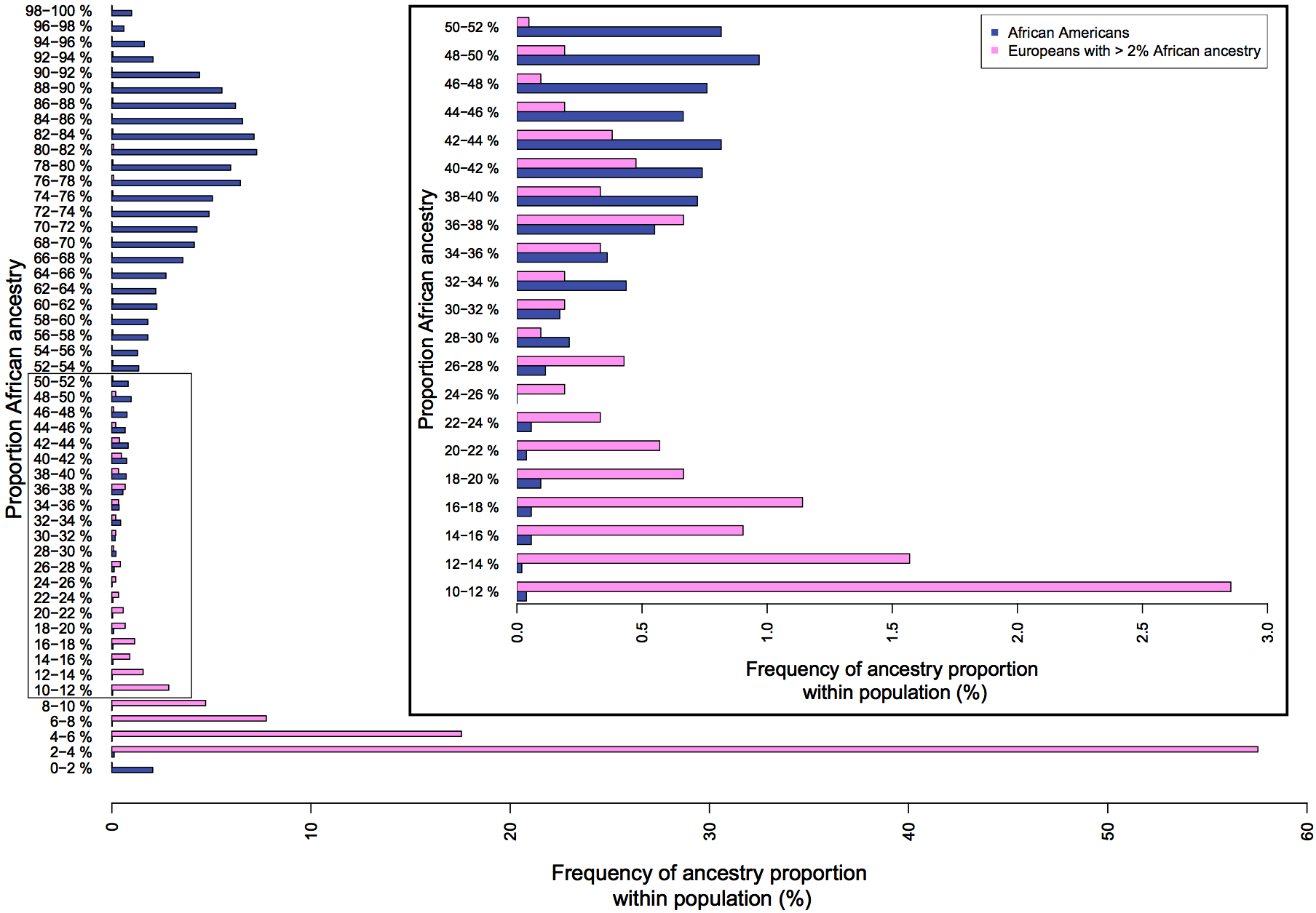
- Population diversity
  - Clear differences in SNP and SNV genotypes between populations
  - Strong evidence of intra-population variation (e.g., Y-chromosome study).
- Admixture
  - Populations are not homogenous, and "declared ancestry" may not reflect the underlying genetics.
  - "Blended" genomes for individuals of mixed ancestry

The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States

Katarzyna Bryc,[1,2,*] Eric Y. Durand,[2] J. Michael Macpherson,[3] David Reich,[1,4,5] and Joanna L. Mountain[2]



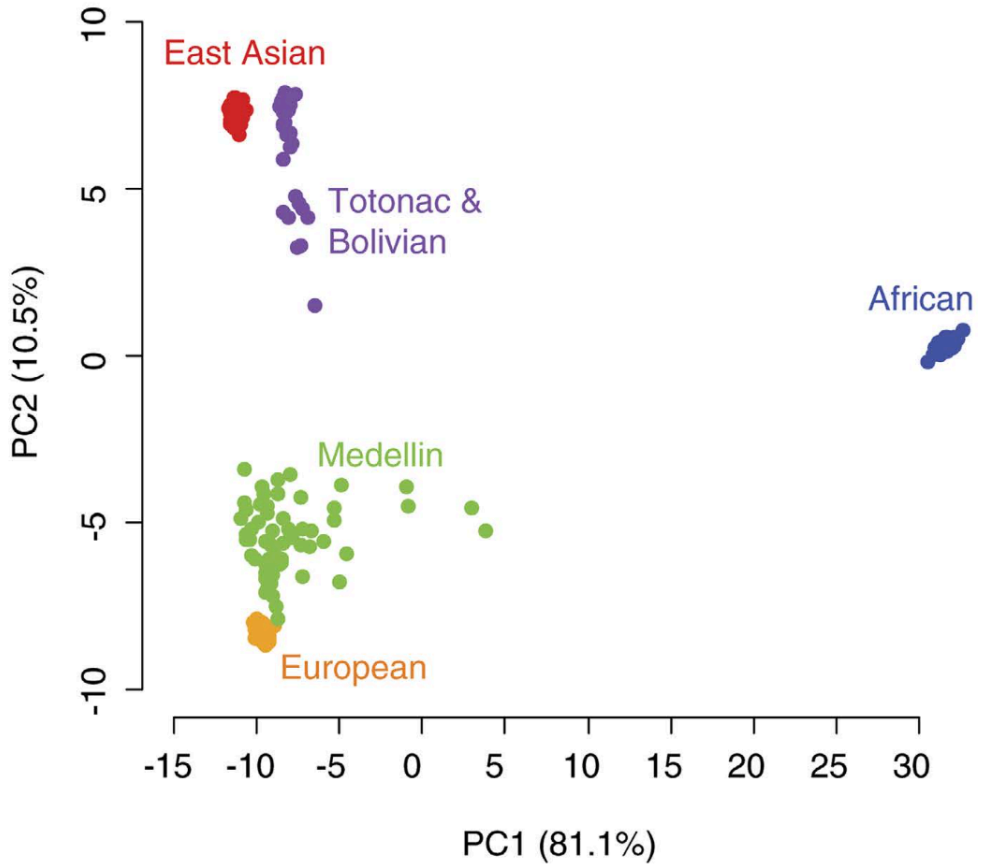
The American Journal of Human Genetics *96*, 37–53, January 8, 2015



The American Journal of Human Genetics *96*, 37–53, January 8, 2015

Ancestry, admixture and fitness in Colombian genomes

Lavanya Rishishwar[1,2,3,*], Andrew B. Conley[1,*], Charles H. Wigington[1], Lu Wang[1], Augusto Valderrama-Aguirre[2,4,5] & I. King Jordan[1,2,3]



- Principle components analysis (PCA) often used to investigate population structure and relatedness
- Admixture shows up as a "spread" of points in the PCA plot.

SCIENTIFIC REPORTS | 5:12376 | DOI: 10.1038/srep12376

Ancestry, admixture and fitness in Colombian genomes

Lavanya Rishishwar[1,2,3,*], Andrew B. Conley[1,*], Charles H. Wigington[1], Lu Wang[1], Augusto Valderrama-Aguirre[2,4,5] & I. King Jordan[1,2,3]

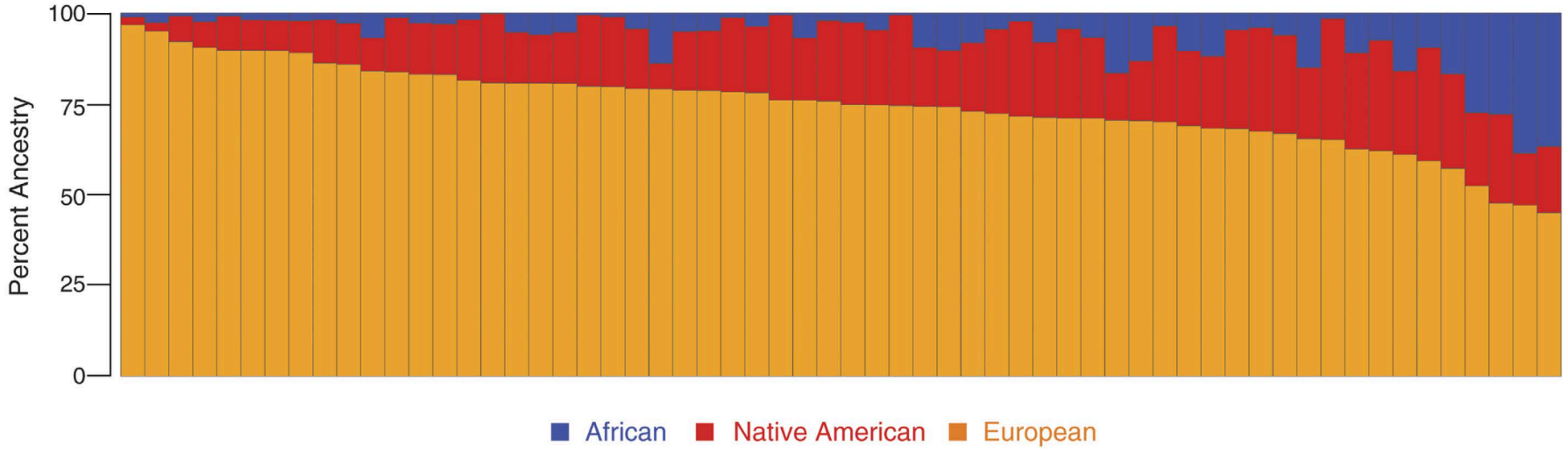


Figure 1. Ancestry and admixture patterns for Colombian genomes.

SCIENTIFIC REPORTS | 5:12376 | DOI: 10.1038/srep12376

Admixture and Clinical Phenotypic Variation

Hum Hered 2014;77:73–86

Laura H. Goetz[b] Liliana Uribe-Bruce[a] Danjuma Quarless[c, d] Ondrej Libiger[a]
Nicholas J. Schork[d]

[a]The Scripps Translational Science Institute, and [b]Scripps Clinic Medical Group, La Jolla, Calif., [c]Biomedical Sciences Graduate Program, University of California, San Diego, Calif., and [d]The J. Craig Venter Institute, La Jolla, Calif., USA

- This paper provides extensive summary of how admixture modifies disease risk:
  - Metabolic disease
  - Cardiovascular disease
  - Pulmonary disease
  - Cancer
- Need to take population diversity and admixture into account in human disease studies

Summary

- 1000 Genomes Project has provided valuable insights into genetic variation and population diversity
- Data generation and analysis has helped up-skill and inform the genomics community
- Data availability has helped to maximize the value of the project.

Next week

- Bring laptops with R and RStudio installed
  - R: https://cran.rstudio.com/
  - RStudio: https://www.rstudio.com/product/rstudio/download/
  - Who doesn't have a laptop to bring along?
- What will we be doing?
  - 1000 Genomes Project data
  - Population diversity
  - Admixture