The 1000 Genomes Project

Chris Tyler-Smith

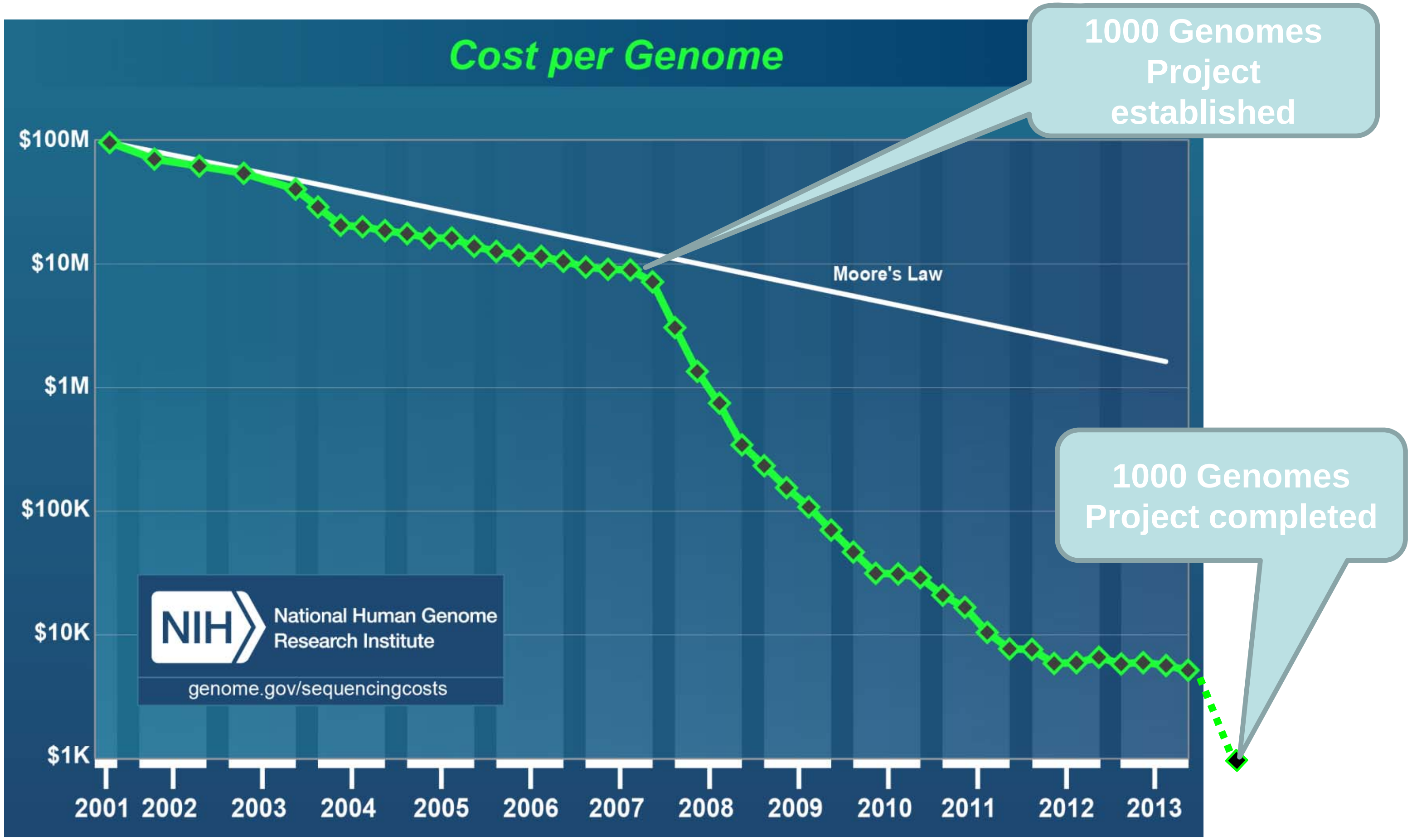
The Wellcome Trust Sanger Institute

Hinxton, CB10 1SA

Topics

- Why was the 1000 Genomes Project established?
- What has the project achieved?
- What is its importance and legacy?

Human genome sequencing costs 2001-2014



1000 Genomes Project established

1000 Genomes Project completed

wellcome trust sanger institute

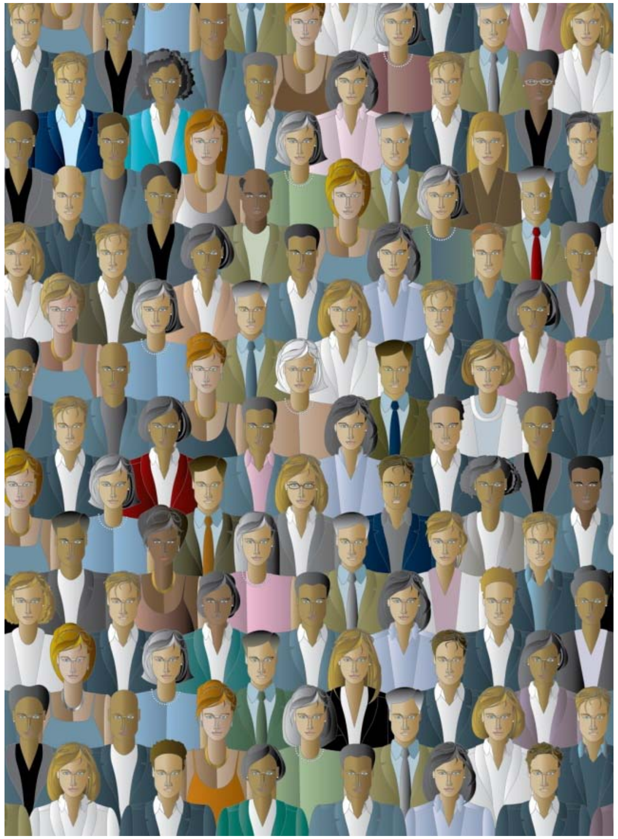
SME Bioinformatics, Hinxton

6th March 2014

The view in 2007

- Major developments in sequencing technology on the horizon
- Becoming possible to sequence multiple whole human genomes
- No single group thought they could do this alone at scale
- Need to establish an international collaborative project

Aims of the 1000 Genomes Project

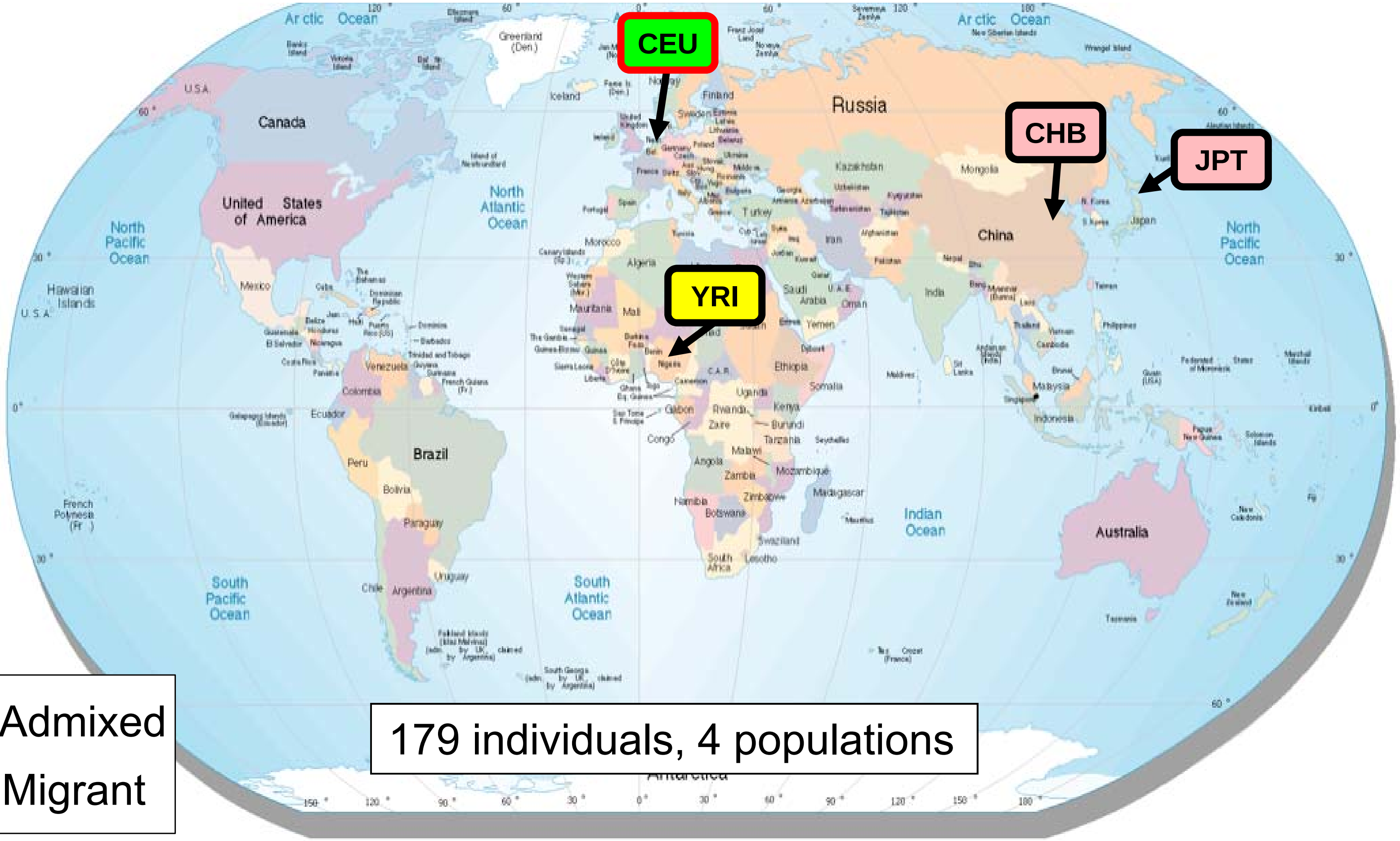


- Primary goal: to develop a public resource of genetic variation to support the next generation of medical association studies
- Find all accessible variants ≥1% across the genome and 0.1-0.5% in gene regions
- Estimate allele frequencies, identify haplotype backgrounds, etc.

Sample choice

- Consent for full free web release of sequence data needed
- When the project began, only HapMap samples met these criteria
- Consent process developed, additional samples recruited
- All individuals are anonymous adults, able to consent, with no phenotype information

Pilot project samples (2010)

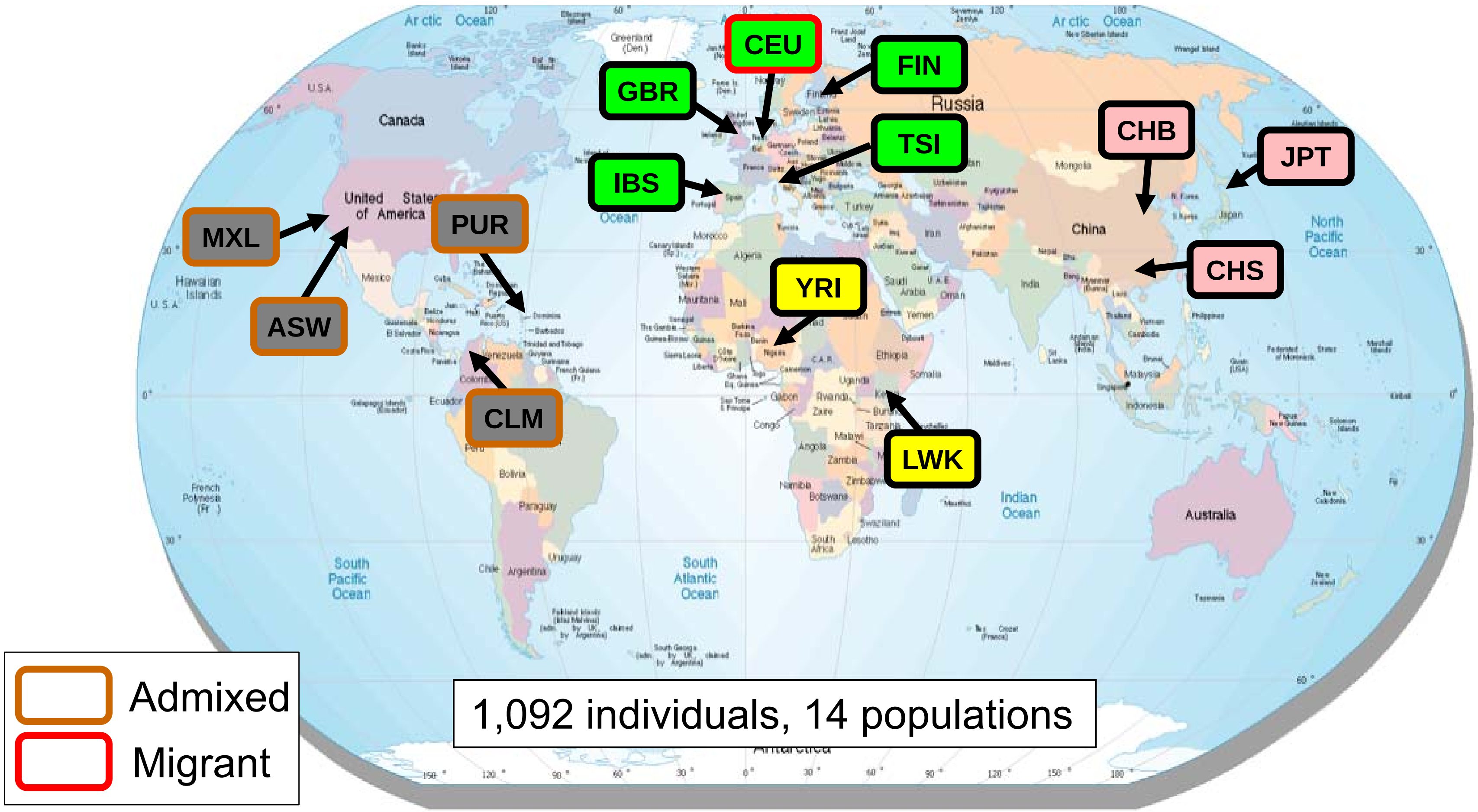


□ Admixed

□ Migrant

179 individuals, 4 populations

Phase 1 samples (2012)

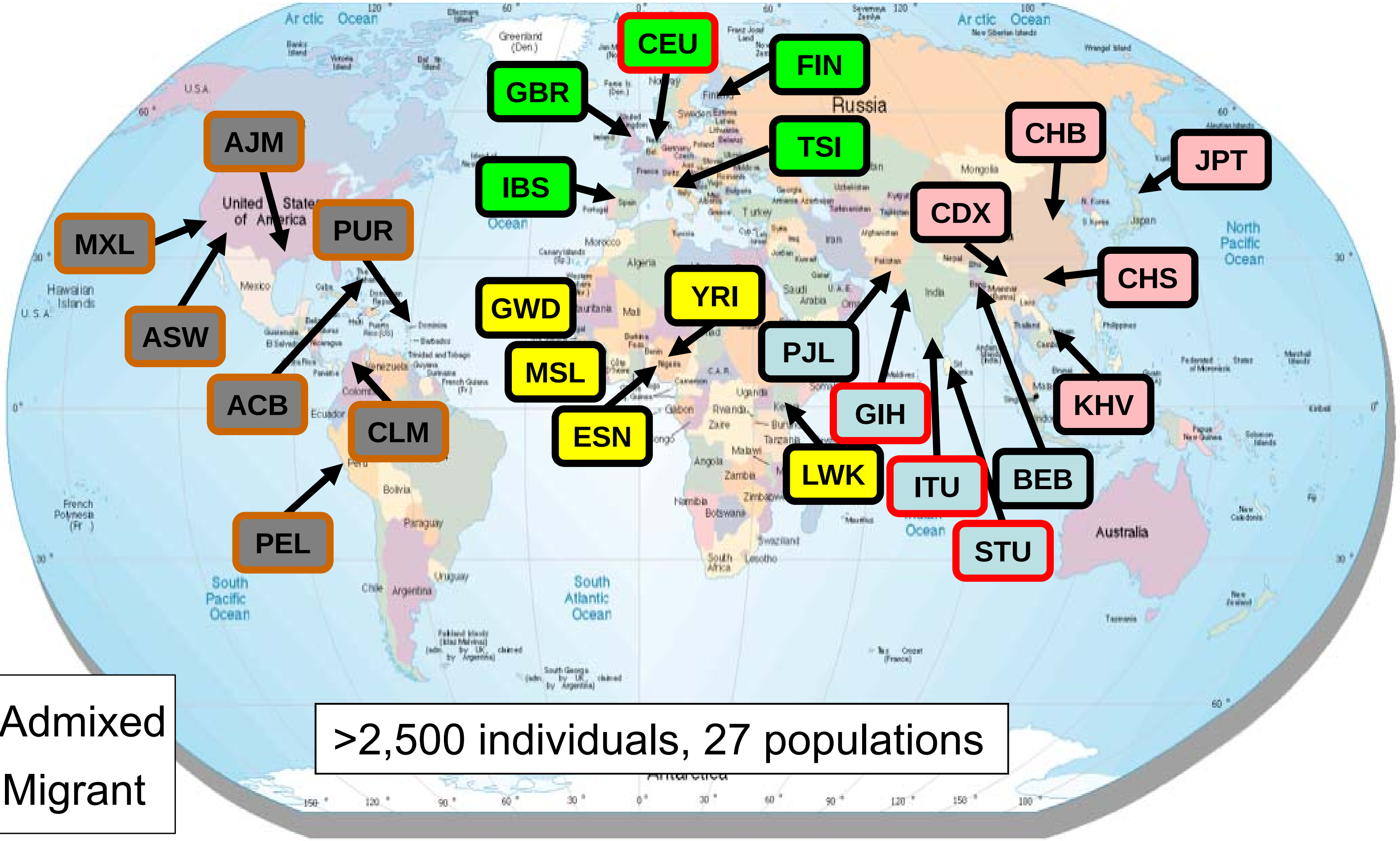


□ Admixed

□ Migrant

1,092 individuals, 14 populations

Full project samples (2014)



Admixed

Migrant

>2,500 individuals, 27 populations

Project design considerations

- In early 2008, still several million $ to sequence a human genome at high coverage (30x)
- But shared genetic variants can be effectively discovered by sequencing at low coverage (2x) and combining information from multiple individuals

Stages of the 1000 Genomes Project

- Pilot, published 2010:
  - 179 genomes at 2x-3x, 2 trios + ~1000 genes at high coverage
- Phase I of main project, published 2012:
  - 1,092 genomes at 3x-4x + high-coverage whole exomes
- Phase 3
  - >2,500 genomes at 4x-6x + exomes, available spring 2013
  - Analysis during 2013-14

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

A THOUSAND GENOMES

Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

HUMAN STEM CELLS

BEYOND THE COURT CASE

Implications for the law, industry and ethics

PAGE 1031

OCEAN PRODUCTIVITY

PHOSPHATE DOWN THE AGES

Key nutrient plentiful after 'snowball' Earth

PAGES 1052 & 1088

AUTUMN BOOKS

THE RECURRING UNIVERSE

Lee Smolin on Roger Penrose's grand idea

PAGE 1034

NATURE.COM/NATURE

28 October 2010 £10

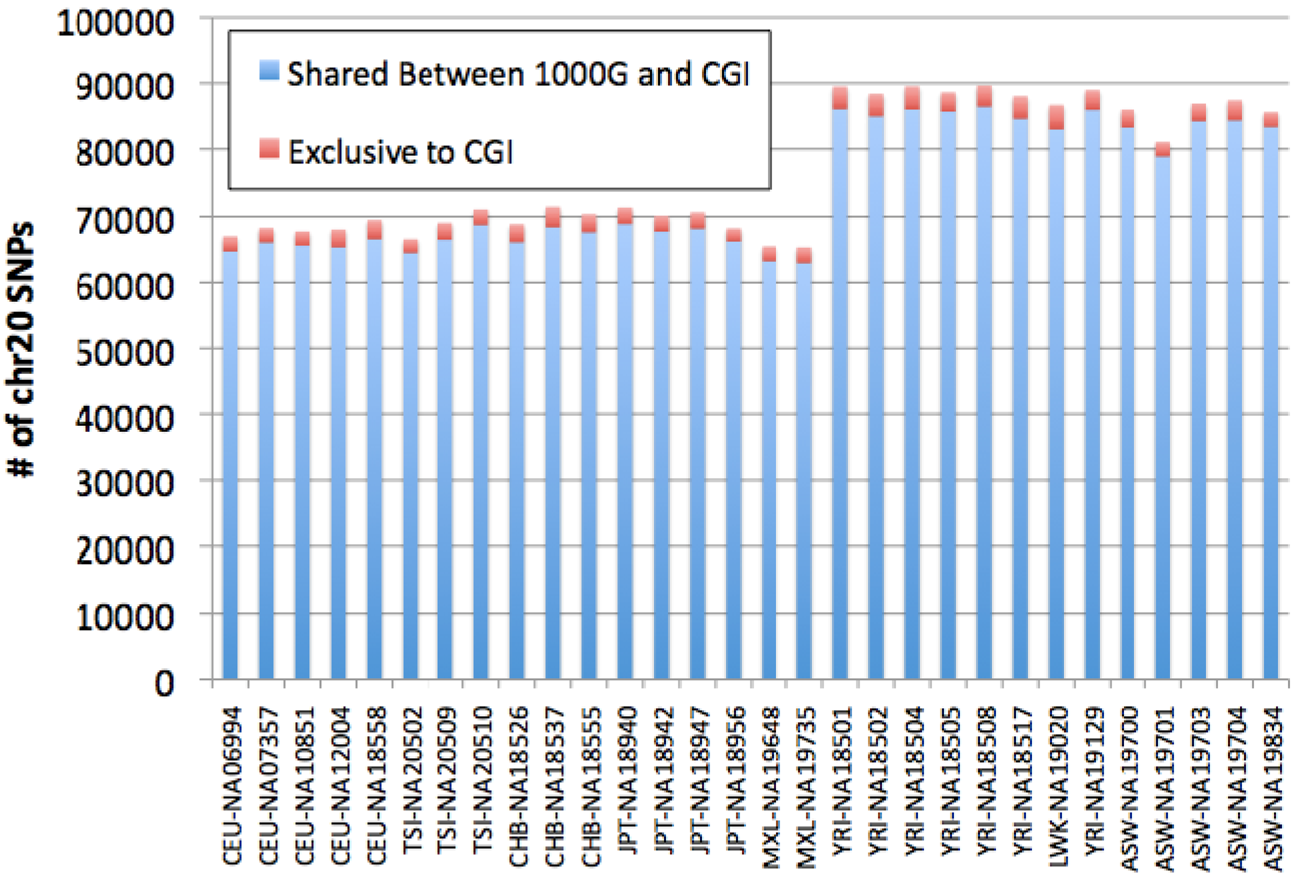
Vol. 467, No. 7319

Many variants discovered

Variant type	Pilot	Phase 1
Total SNPs	15.2 M	37.9 M
Known SNPs	6.8 M	8.2 M
Novel SNPs	8.4 M	29.7 M
Short indels	1.5 M	3.8 M
Large deletions	14 K	14 K

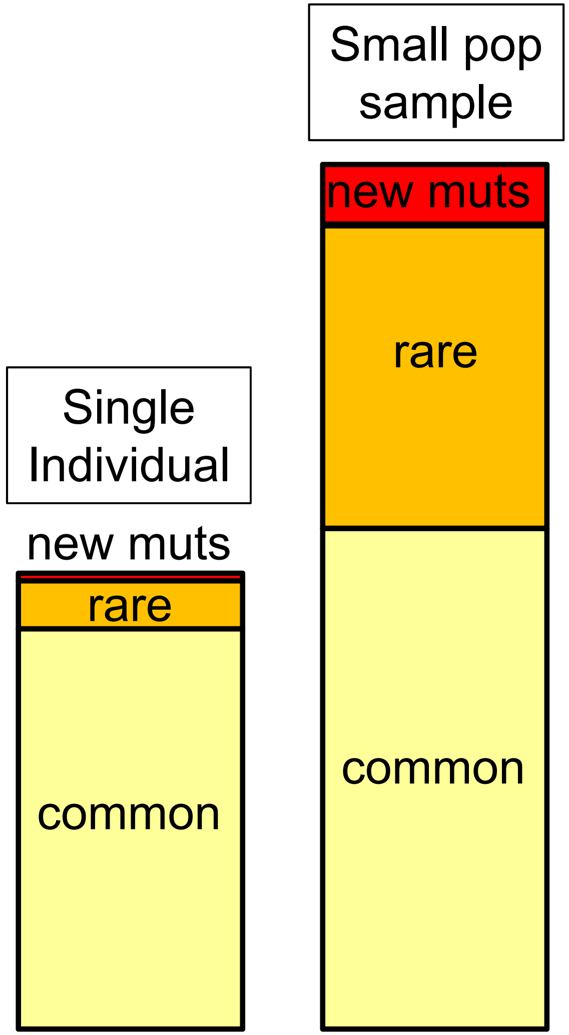
New generation of SNP chips

Discovered most of the variants present in a genome



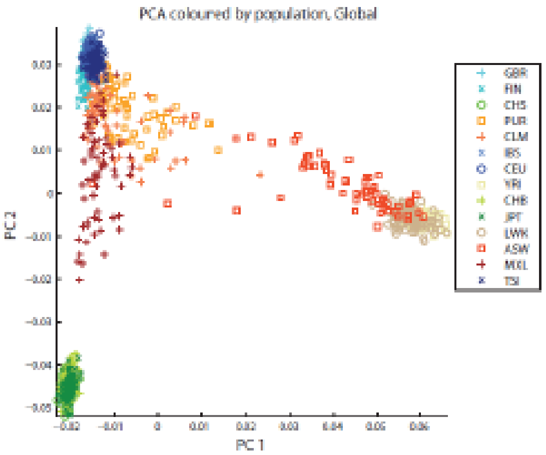
Features of human genetic variation

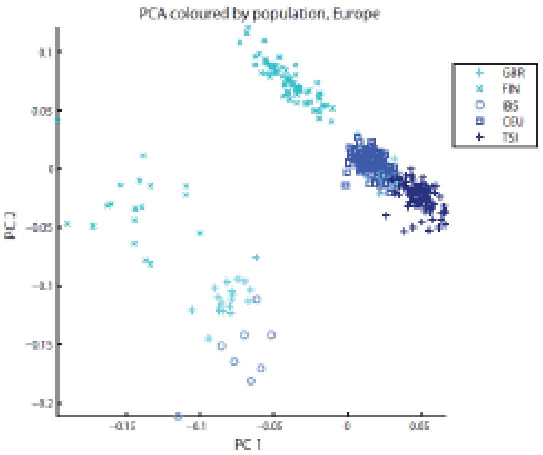
- 3-4 million variants per individual
- Most have no functional consequences
- Some are common in the population, some rare



Insights from common variants (1)

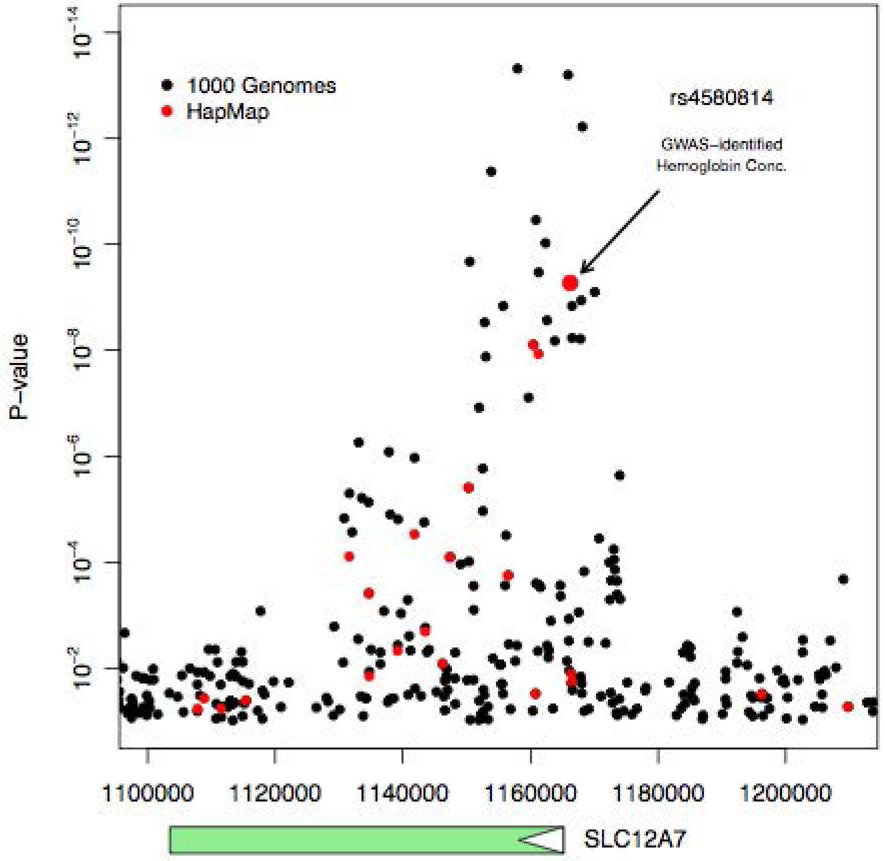
- Geographical origin is predicted by genotype


PCA coloured by population, Global

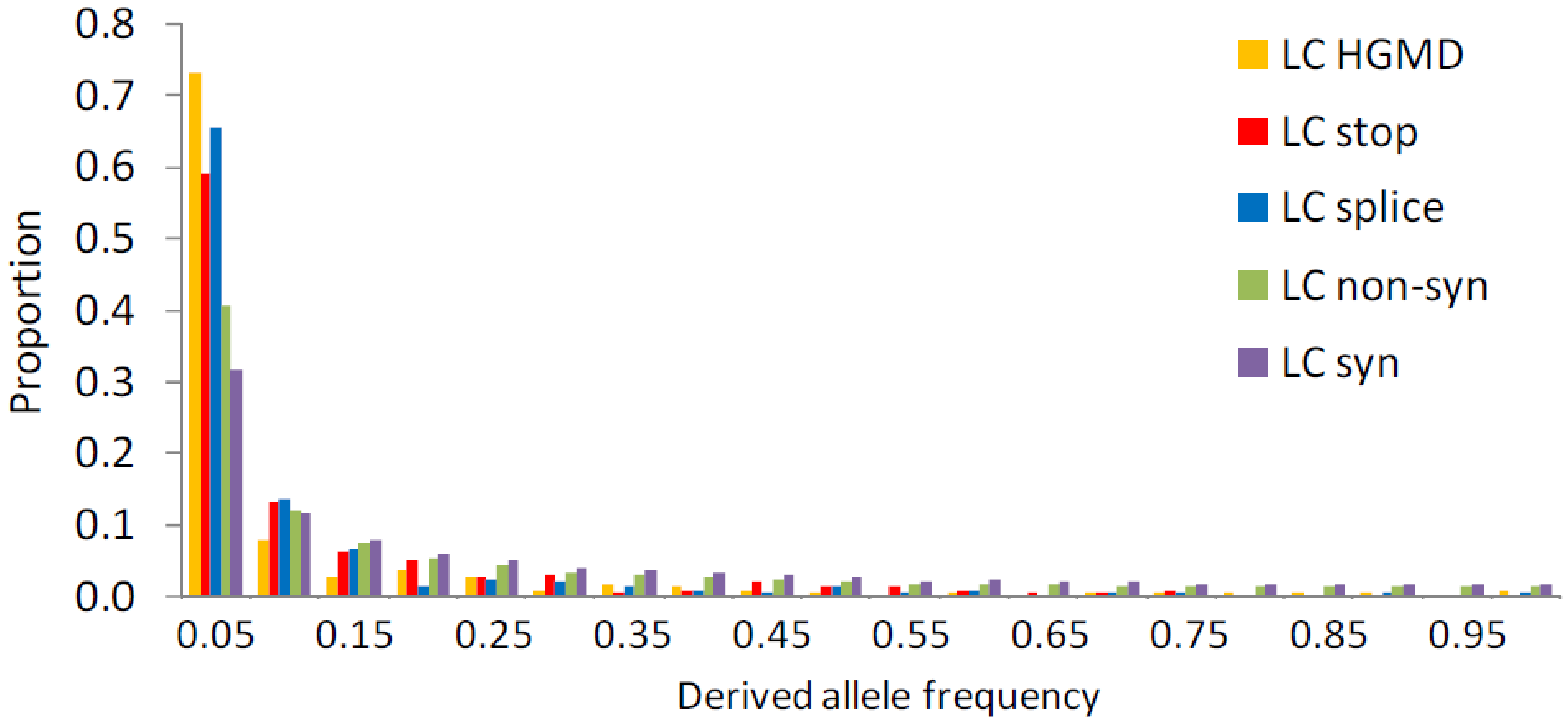

PCA coloured by population, Europe

Insights from common variants (2)

- GWASs are usually carried out using SNP chips
- The best hit on the chip is often not the causal variant
- Better candidates for the causal variant can be obtained by imputation using sequence data

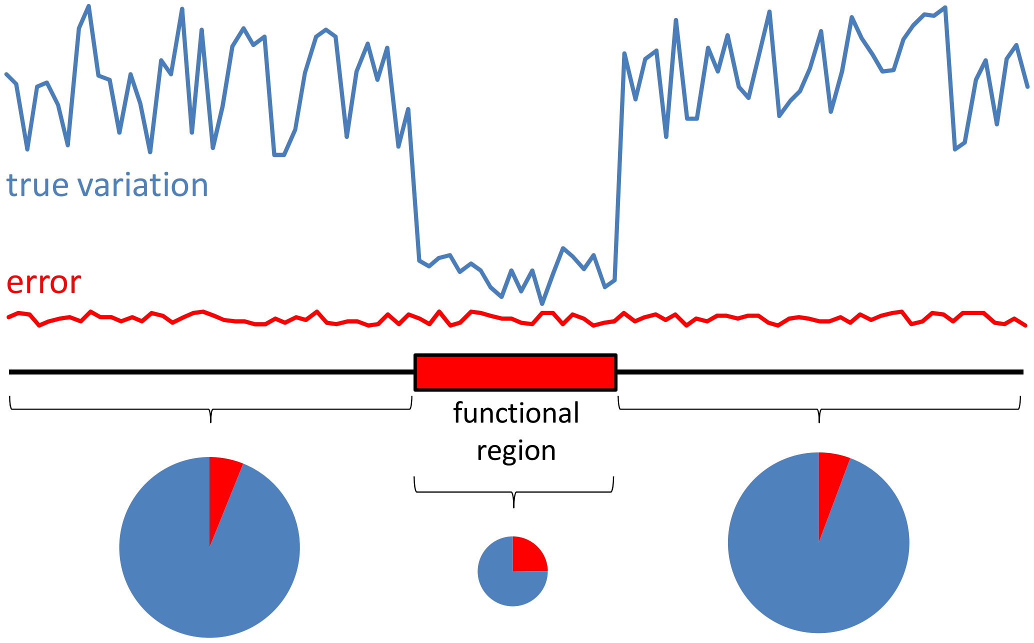


Strong functional variants are mostly rare in the population



Insights into functional variants (1)

- Enriched in errors of many kinds, extra QC and validation required



MacArthur and Tyler-Smith (2010)
Hum. Mol. Genet. **19(R2),** R131-136

Insights into functional variants (2)

- Each individual carries ~100 genes in an inactive form, ~20 with both copies inactive

2,951 raw > 1,269 filtered

Category	Filtered number/individual (CEU)	
	All	Homozygous
nonsense SNP	26.2	5.2
splice SNP	11.2	1.9
frameshift indel	38.2	9.2
large LoF deletion	28.3	6.2
total	103.9	22.3

MacArthur *et al*. (2012) *Science* **335**, 823-828

Insights into functional variants (3)

- Each individual carries ~2 (0-7) known disease-causing variants, and these are expected to impact health in ~10% of carriers

578 raw > 45 filtered

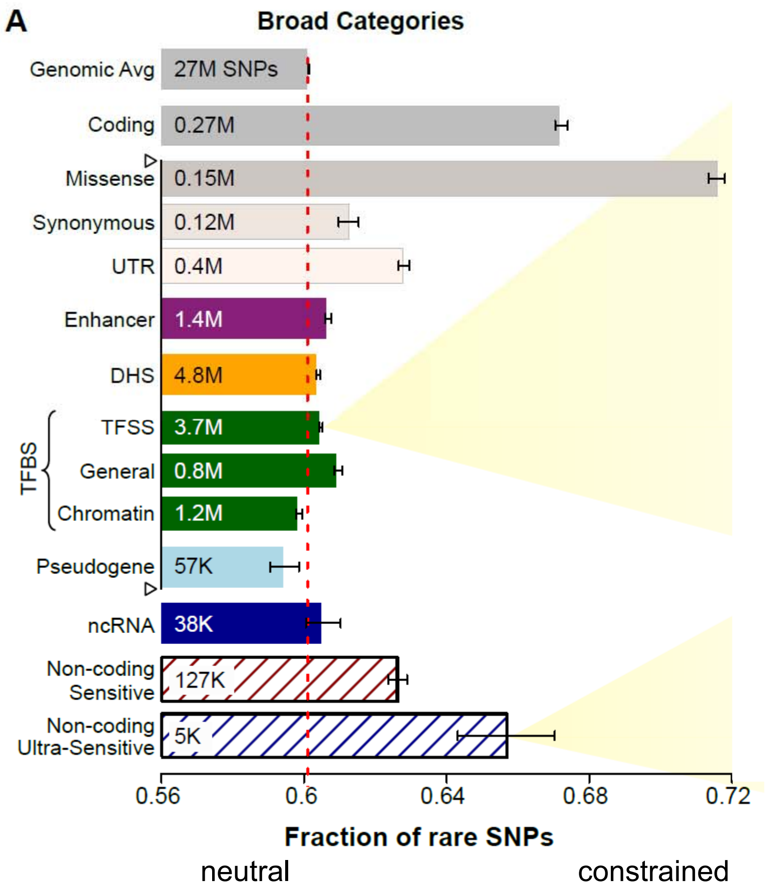
Xue *et al*. (2012) *Am. J. Hum. Genet*. **91**, 1022-1032

Insights into functional variants (4)

- Population genetics can help to discover new *non-coding* functional variants

Khurana *et al*. (2013) *Science* **342**, 84

Insights into functional variants (4)

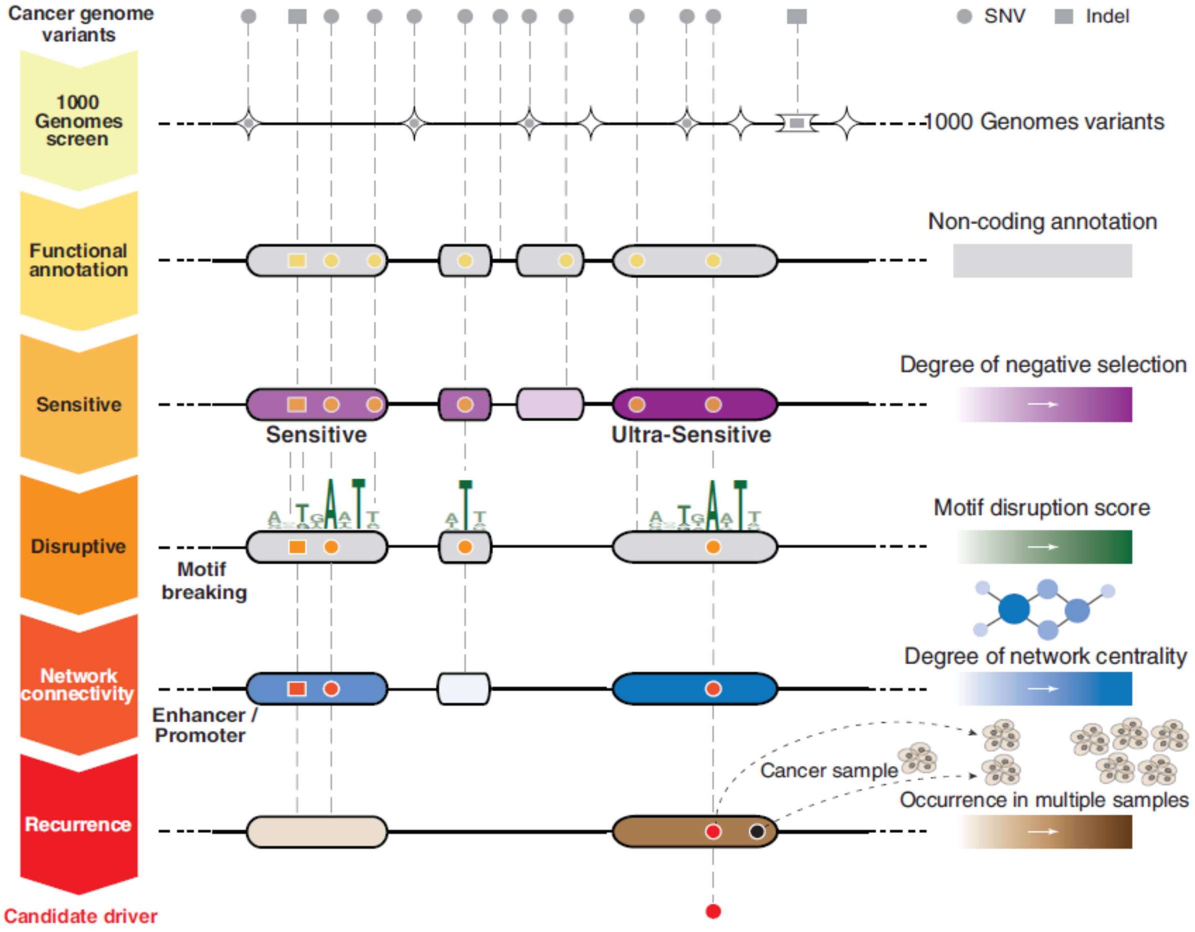


Khurana *et al*. (2013) *Science* **342**, 84

SME Bioinformatics, Hinxton

6th March 2014

Insights into functional variants (4)



FunSeq

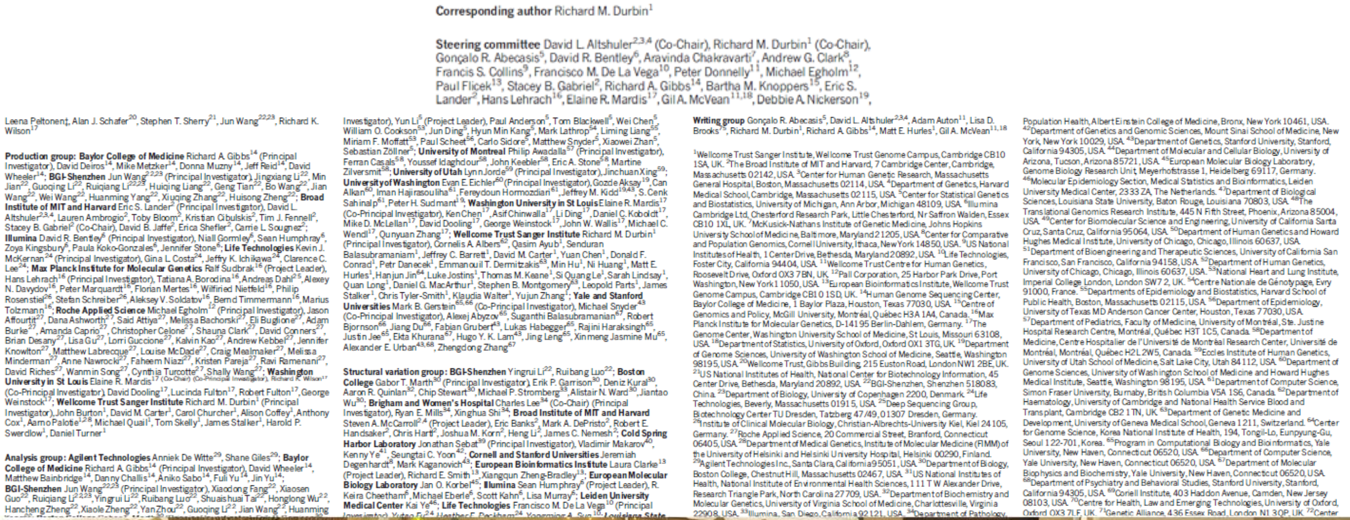
New Non-coding cancer driver mutations

Khurana *et al*. (2013) *Science* **342**, 84

Final thoughts

- One of the very few large-scale sources of open access genomic data
- Samples (cell lines) available
- No phenotypes
- A standard and lasting resource for the human genomics field
- A legacy model for additional populations

Acknowledgements



Corresponding author Richard M. Durbin[1]

Steering committee David L. Altshuler[2,3,4] (Co-Chair), Richard M. Durbin[1] (Co-Chair), Gonçalo R. Abecasis[5], David R. Bentley[6], Aravinda Chakravarti[7], Andrew G. Clark[8], Francis S. Collins[9], Francisco M. De La Vega[10], Peter Donnelly[11], Michael Egholm[12], Paul Flicek[13], Stacey B. Gabriel[2], Richard A. Gibbs[14], Bartha M. Knoppers[15], Eric S. Lander[2], Hans Lehrach[16], Elaine R. Mardis[17], Gil A. McVean[11,18], Debbie A. Nickerson[19], Leena Peltonen†, Alan J. Schafer[20], Stephen T. Sherry[21], Jun Wang[22,23], Richard K. Wilson[17]

Production group: Baylor College of Medicine Richard A. Gibbs[14] (Principal Investigator), David Deiros[14], Mike Metzker[14], Donna Muzny[14], Jeff Reid[14], David Wheeler[14]; **BGI-Shenzhen** Jun Wang[22,23] (Principal Investigator), Jingxiang Li[22], Min Jian[22], Guoqing Li[22], Ruiqiang Li[22,23], Huiqing Liang[22], Geng Tian[22], Bo Wang[22], Jian Wang[22], Wei Wang[22], Huanming Yang[22], Xiuqing Zhang[22], Huisong Zheng[22]; **Broad Institute of MIT and Harvard** Eric S. Lander[2] (Principal Investigator), David L. Altshuler[2,3,4], Lauren Ambrogio[2], Toby Bloom[2], Kristian Cibulskis[2], Tim J. Fennell[2], Stacey B. Gabriel[2] (Co-Chair), David B. Jaffe[2], Erica Shefler[2], Carrie L. Sougnez[2]; **Illumina** David R. Bentley[6] (Principal Investigator), Niall Gormley[6], Sean Humphray[6], Zoya Kingsbury[6], Paula Koko-Gonzales[6], Jennifer Stone[6]; **Life Technologies** Kevin J. McKernan[24] (Principal Investigator), Gina L. Costa[24], Jeffry K. Ichikawa[24], Clarence C. Lee[24]; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak[16] (Project Leader), Hans Lehrach[16] (Principal Investigator), Tatiana A. Borodina[16], Andreas Dahl[25], Alexey N. Davydov[16], Peter Marquardt[16], Florian Mertes[16], Wilfiried Nietfeld[16], Philip Rosenstiel[26], Stefan Schreiber[26], Aleksey V. Soldatov[16], Bernd Timmermann[16], Marius Tolzmann[16]; **Roche Applied Science** Michael Egholm[12] (Principal Investigator), Jason Affourtit[27], Dana Ashworth[27], Said Attiya[27], Melissa Bachorski[27], Eli Buglione[27], Adam Burke[27], Amanda Caprio[27], Christopher Celone[27], Shauna Clark[27], David Conners[27], Brian Desany[27], Lisa Gu[27], Lorri Guccione[27], Kalvin Kao[27], Andrew Kebbel[27], Jennifer Knowlton[27], Matthew Labrecque[27], Louise McDade[27], Craig Mealmaker[27], Melissa Minderman[27], Anne Nawrocki[27], Faheem Niazi[27], Kristen Pareja[27], Ravi Ramenani[27], David Riches[27], Wanmin Song[27], Cynthia Turcotte[27], Shally Wang[27]; **Washington University in St Louis** Elaine R. Mardis[17] (Co-Chair) (Co-Principal Investigator), Richard K. Wilson[17] (Co-Principal Investigator), David Dooling[17], Lucinda Fulton[17], Robert Fulton[17], George Weinstock[17]; **Wellcome Trust Sanger Institute** Richard M. Durbin[1] (Principal Investigator), John Burton[1], David M. Carter[1], Carol Churcher[1], Alison Coffey[1], Anthony Cox[1], Aarno Palotie[1,28], Michael Quail[1], Tom Skelly[1], James Stalker[1], Harold P. Swerdlow[1], Daniel Turner[1]

Analysis group: Agilent Technologies Anniek De Witte[29], Shane Giles[29]; **Baylor College of Medicine** Richard A. Gibbs[14] (Principal Investigator), David Wheeler[14], Matthew Bainbridge[14], Danny Challis[14], Aniko Sabo[14], Fuli Yu[14], Jin Yu[14]; **BGI-Shenzhen** Jun Wang[22,23] (Principal Investigator), Xiaodong Fang[22], Xiaosen Guo[22], Ruiqiang Li[22,23], Yingrui Li[22], Ruibang Luo[22], Shuaishuai Tai[22], Honglong Wu[22], Hancheng Zheng[22], Xiaole Zheng[22], Yan Zhou[22], Guoqing Li[22], Jian Wang[22], Huanming

Afidalina Tumian[18]; **Roche Applied Science** Brian Desany[27] (Project Leader), James Knight[27], Roger Winer[27]; **The Translational Genomics Research Institute** David W. Craig[48] (Principal Investigator), Steve M. Beckstrom-Sternberg[48], Alexis Christoforides[48], Ahmet A. Kurdoglu[48], John V. Pearson[48], Shripad A. Sinari[48], Waibhav D. Tembe[48]; **University of California, Santa Cruz** David Haussler[49] (Principal Investigator), Angie S. Hinrichs[49], Sol J. Katzman[49], Andrew Kern[49], Robert M. Kuhn[49]; **University of Chicago** Molly Przeworski[50] (Co-Chair, Population Genetics) (Principal Investigator), Ryan D. Hernandez[51], Bryan Howie[52], Joanna L. Kelley[52], S. Cord Melton[52]; **University of Michigan** Gonçalo R. Abecasis[5] (Co-Chair) (Principal Investigator), Yun Li[5] (Project Leader), Paul Anderson[5], Tom Blackwell[5], Wei Chen[5], William O. Cookson[53], Jun Ding[5], Hyun Min Kang[5], Mark Lathrop[54], Liming Liang[55], Miriam F. Moffatt[53], Paul Scheet[56], Carlo Sidore[5], Matthew Snyder[5], Xiaowei Zhan[5], Sebastian Zöllner[5]; **University of Montreal** Philip Awadalla[57] (Principal Investigator), Ferran Casals[58], Youssef Idaghdour[58], John Keebler[58], Eric A. Stone[58], Martine Zilversmit[58]; **University of Utah** Lynn Jorde[59] (Principal Investigator), Jinchuan Xing[59]; **University of Washington** Evan E. Eichler[60] (Principal Investigator), Gozde Aksay[19], Can Alkan[60], Iman Hajirasouliha[61], Fereydoun Hormozdiari[61], Jeffrey M. Kidd[19,43], S. Cenk Sahinalp[61], Peter H. Sudmant[19]; **Washington University in St Louis** Elaine R. Mardis[17] (Co-Principal Investigator), Ken Chen[17], Asif Chinwalla[17], Li Ding[17], Daniel C. Koboldt[17], Mike D. McLellan[17], David Dooling[17], George Weinstock[17], John W. Wallis[17], Michael C. Wendl[17], Qunyuan Zhang[17]; **Wellcome Trust Sanger Institute** Richard M. Durbin[1] (Principal Investigator), Cornelis A. Albers[62], Qasim Ayub[1], Senduran Balasubramaniam[1], Jeffrey C. Barrett[1], David M. Carter[1], Yuan Chen[1], Donald F. Conrad[1], Petr Danecek[1], Emmanouil T. Dermitzakis[63], Min Hu[1], Ni Huang[1], Matt E. Hurles[1], Hanjun Jin[64], Luke Jostins[1], Thomas M. Keane[1], Si Quang Le[1], Sarah Lindsay[1], Quan Long[1], Daniel G. MacArthur[1], Stephen B. Montgomery[63], Leopold Parts[1], James Stalker[1], Chris Tyler-Smith[1], Klaudia Walter[1], Yujun Zhang[1]; **Yale and Stanford Universities** Mark B. Gerstein[65,66] (Co-Principal Investigator), Michael Snyder[43] (Co-Principal Investigator), Alexej Abyzov[65], Suganthi Balasubramanian[67], Robert Bjornson[66], Jiang Du[66], Fabian Grubert[43], Lukas Habegger[65], Rajini Haraksingh[65], Justin Jee[65], Ekta Khurana[67], Hugo Y. K. Lam[43], Jing Leng[65], Xinmeng Jasmine Mu[65], Alexander E. Urban[43,68], Zhengdong Zhang[67]

Structural variation group: BGI-Shenzhen Yingrui Li[22], Ruibang Luo[22]; **Boston College** Gabor T. Marth[30] (Principal Investigator), Erik P. Garrison[30], Deniz Kural[30], Aaron R. Quinlan[32], Chip Stewart[30], Michael P. Stromberg[33], Alistair N. Ward[30], Jiantao Wu[30]; **Brigham and Women's Hospital** Charles Lee[34] (Co-Chair) (Principal Investigator), Ryan E. Mills[34], Xinghua Shi[34]; **Broad Institute of MIT and Harvard** Steven A. McCarroll[2,4] (Project Leader), Eric Banks[2], Mark A. DePristo[2], Robert E. Handsaker[2], Chris Hartl[2], Joshua M. Korn[2], Heng Li[2], James C. Nemesh[2]; **Cold Spring Harbor Laboratory** Jonathan Sebat[39] (Principal Investigator), Vladimir Makarov[40], Kenny Ye[41], Seungtai C. Yoon[42]; **Cornell and Stanford Universities** Jeremiah Degenhardt[8], Mark Kaganovich[43]; **European Bioinformatics Institute** Laura Clarke[13] (Project Leader), Richard E. Smith[13], Xiangqun Zheng-Bradley[13]; **European Molecular Biology Laboratory** Jan O. Korbel[45]; **Illumina** Sean Humphray[6] (Project Leader), R. Keira Cheetham[6], Michael Eberle[6], Scott Kahn[6], Lisa Murray[6]; **Leiden University Medical Center** Kai Ye[46]; **Life Technologies** Francisco M. De La Vega[10] (Principal Investigator), Yutao Fu[24], Heather E. Peckham[24], Yongming A. Sun[10]; **Louisiana State**

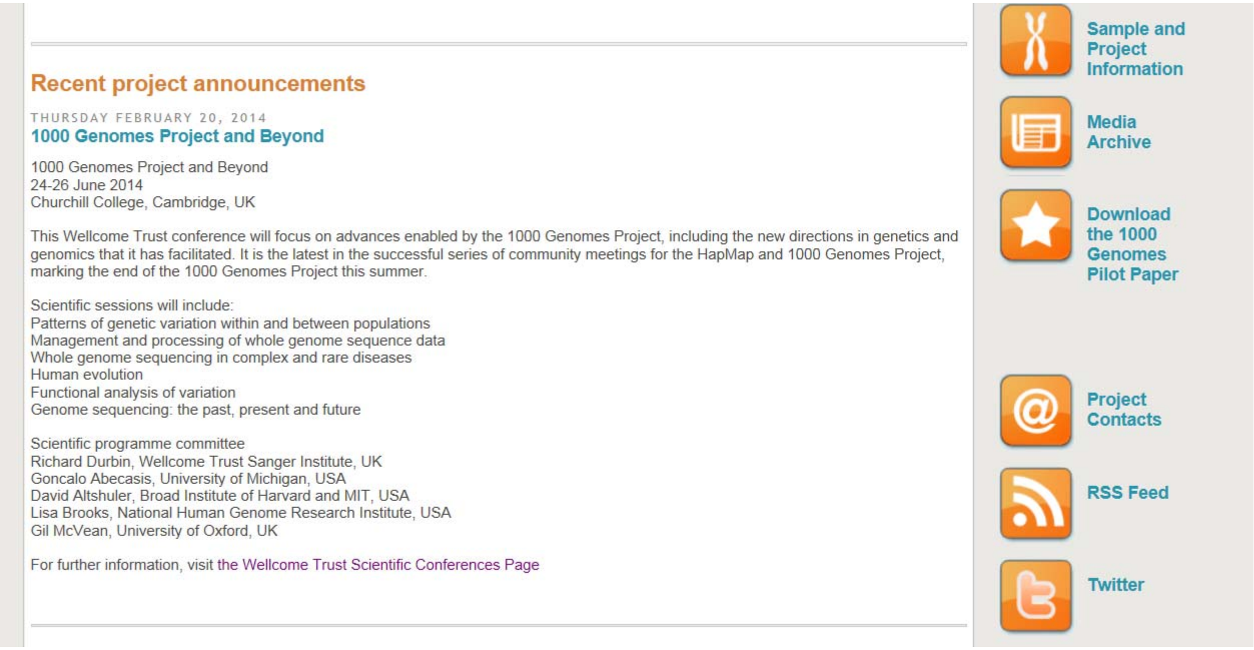
(Co-Chair), Leena Peltonen† (Co-Chair), Gonçalo R. Abecasis[5], Carlos D. Bustamante, Neda Gharani[69], Richard A. Gibbs[14], Lynn Jorde[59], Jane S. Kaye[70], Alastair Kent[71], Taosha Li[22], Amy L. McGuire[72], Gil A. McVean[11,18], Pilar N. Ossorio[73], Charles N. Rotimi[74], Yeyang Su[22], Lorraine H. Toji[69], Chris Tyler-Smith[1]

Scientific management Lisa D. Brooks[75], Adam L. Felsenfeld[75], Jean E. McEwen[75], Assya Abdallah[76], Christopher R. Juenger[77], Nicholas C. Clemm[75], Francis S. Collins[9], Audrey Duncanson[20], Eric D. Green[78], Mark S. Guyer[75], Jane L. Peterson[75], Alan J. Schafer[20]

Writing group Gonçalo R. Abecasis[5], David L. Altshuler[2,3,4], Adam Auton[11], Lisa D. Brooks[75], Richard M. Durbin[1], Richard A. Gibbs[14], Matt E. Hurles[1], Gil A. McVean[11,18]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. [2]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115, USA. [5]Center for Statistical Genetics and Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [6]Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [7]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [8]Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA. [9]US National Institutes of Health, 1 Center Drive, Bethesda, Maryland 20892, USA. [10]Life Technologies, Foster City, California 94404, USA. [11]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [12]Pall Corporation, 25 Harbor Park Drive, Port Washington, New York 11050, USA. [13]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK. [14]Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA. [15]Centre of Genomics and Policy, McGill University, Montréal, Québec H3A 1A4, Canada. [16]Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany. [17]The Genome Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [18]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [19]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [20]Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [21]US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA. [22]BGI-Shenzhen, Shenzhen 518083, China. [23]Department of Biology, University of Copenhagen 2200, Denmark. [24]Life Technologies, Beverly, Massachusetts 01915, USA. [25]Deep Sequencing Group, Biotechnology Center TU Dresden, Tatzberg 47/49, 01307 Dresden, Germany. [26]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel 24105, Germany. [27]Roche Applied Science, 20 Commercial Street, Branford, Connecticut 06405, USA. [28]Department of Medical Genetics, Institute of Molecular Medicine (FIMM) of the University of Helsinki and Helsinki University Hospital, Helsinki 00290, Finland. [29]Agilent Technologies Inc., Santa Clara, California 95051, USA. [30]Department of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA. [31]US National Institutes of Health, National Institute of Environmental Health Sciences, 111 T W Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [32]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA. [33]Illumina, San Diego, California 92121, USA. [34]Department of Pathology,

Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [42]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA. [43]Department of Genetics, Stanford University, Stanford, California 94305, USA. [44]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA. [45]European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstrasse 1, Heidelberg 69117, Germany. [46]Molecular Epidemiology Section, Medical Statistics and Bioinformatics, Leiden University Medical Center, 2333 ZA, The Netherlands. [47]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA. [48]The Translational Genomics Research Institute, 445 N Fifth Street, Phoenix, Arizona 85004, USA. [49]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [50]Department of Human Genetics and Howard Hughes Medical Institute, University of Chicago, Chicago, Illinois 60637, USA. [51]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, USA. [52]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [53]National Heart and Lung Institute, Imperial College London, London SW7 2, UK. [54]Centre Nationale de Génotypage, Evry 91000, France. [55]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [56]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [57]Department of Pediatrics, Faculty of Medicine, University of Montréal, Ste. Justine Hospital Research Centre, Montréal, Québec H3T 1C5, Canada. [58]Department of Medicine, Centre Hospitalier de l'Université de Montréal Research Center, Université de Montréal, Montréal, Québec H2L 2W5, Canada. [59]Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. [60]Department of Genome Sciences, University of Washington School of Medicine and Howard Hughes Medical Institute, Seattle, Washington 98195, USA. [61]Department of Computer Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. [62]Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 1TN, UK. [63]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva 1211, Switzerland. [64]Center for Genome Science, Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul 122-701, Korea. [65]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. [66]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA. [67]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. [68]Department of Psychiatry and Behavioral Studies, Stanford University, Stanford, California 94305, USA. [69]Coriell Institute, 403 Haddon Avenue, Camden, New Jersey 08103, USA. [70]Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK. [71]Genetic Alliance, 436 Essex Road, London N1 3QP, UK. [72]Center



Recent project announcements

THURSDAY FEBRUARY 20, 2014

1000 Genomes Project and Beyond

1000 Genomes Project and Beyond
24-26 June 2014
Churchill College, Cambridge, UK

This Wellcome Trust conference will focus on advances enabled by the 1000 Genomes Project, including the new directions in genetics and genomics that it has facilitated. It is the latest in the successful series of community meetings for the HapMap and 1000 Genomes Project, marking the end of the 1000 Genomes Project this summer.

Scientific sessions will include:
Patterns of genetic variation within and between populations
Management and processing of whole genome sequence data
Whole genome sequencing in complex and rare diseases
Human evolution
Functional analysis of variation
Genome sequencing: the past, present and future

Scientific programme committee
Richard Durbin, Wellcome Trust Sanger Institute, UK
Goncalo Abecasis, University of Michigan, USA
David Altshuler, Broad Institute of Harvard and MIT, USA
Lisa Brooks, National Human Genome Research Institute, USA
Gil McVean, University of Oxford, UK

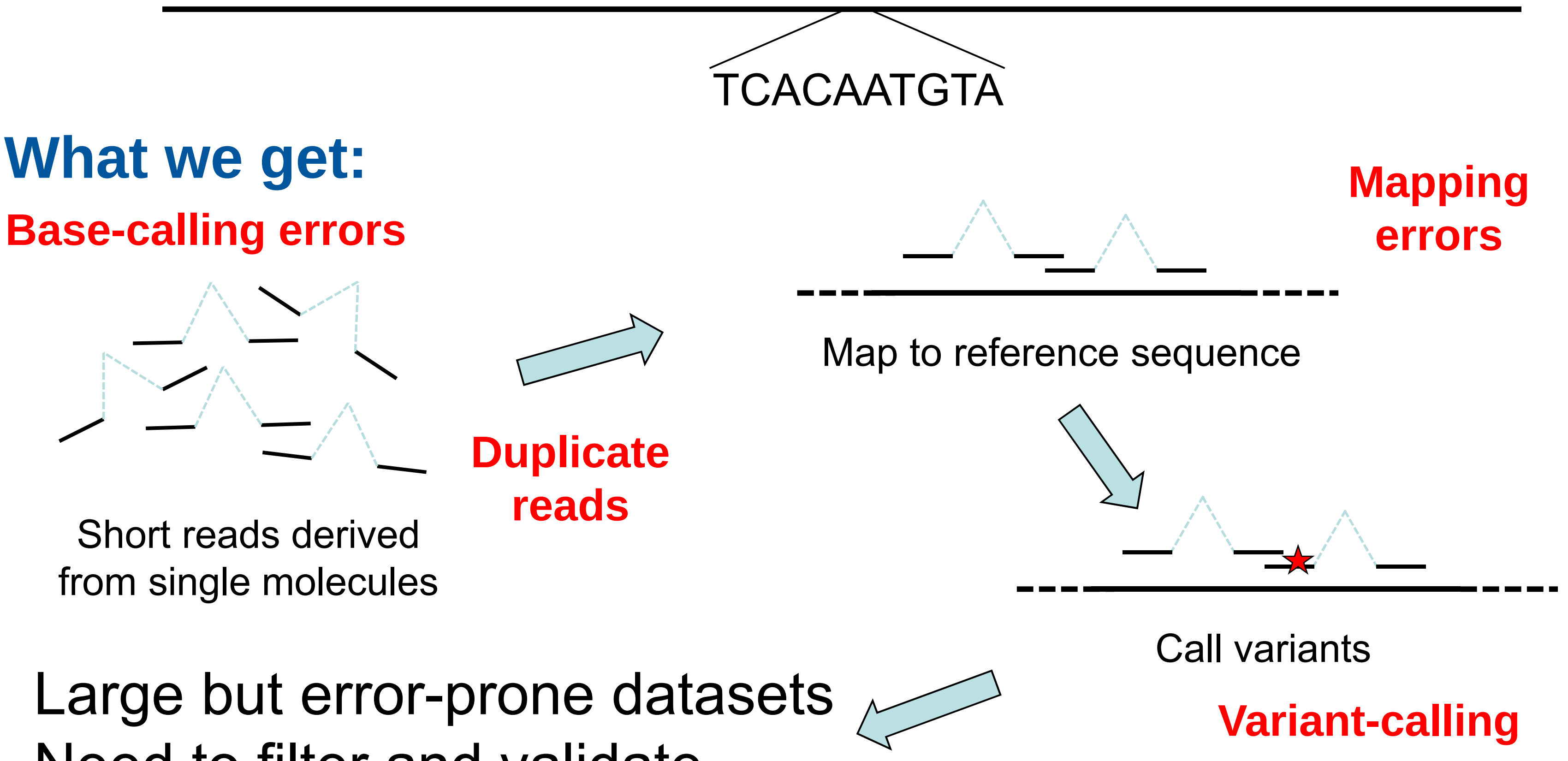
For further information, visit the Wellcome Trust Scientific Conferences Page

Sample and Project Information
Media Archive
Download the 1000 Genomes Pilot Paper
Project Contacts
RSS Feed
Twitter

http://www.1000genomes.org/

Sequencing technology

What we want:



TCACAATGTA

What we get:

Base-calling errors

Short reads derived from single molecules

Duplicate reads

Map to reference sequence

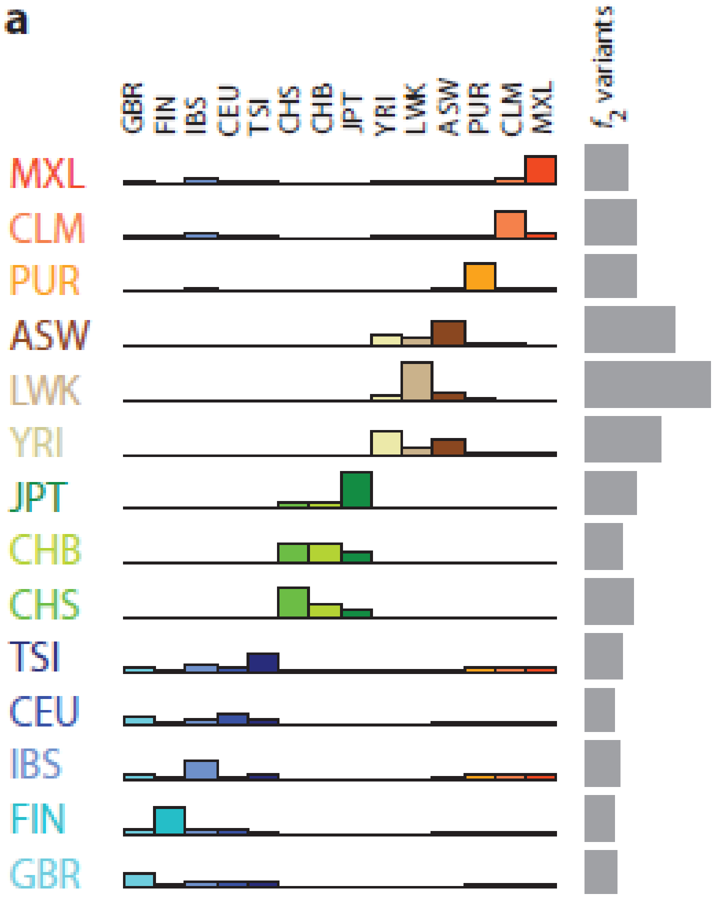
Mapping errors

Call variants

Variant-calling errors

Large but error-prone datasets
Need to filter and validate

Rare variants tend to be population-specific



Phase 1