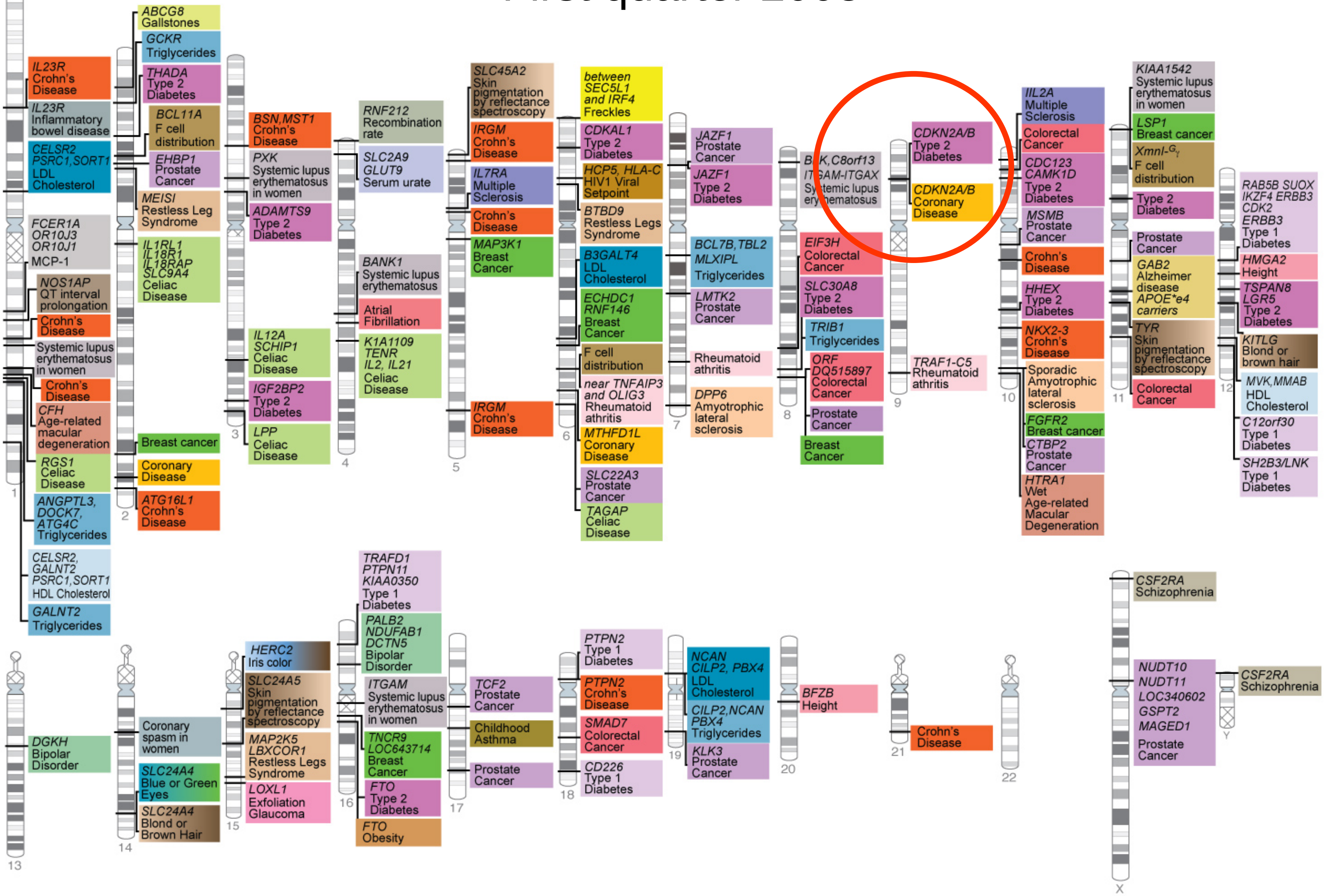
The 1000 Genomes Project:

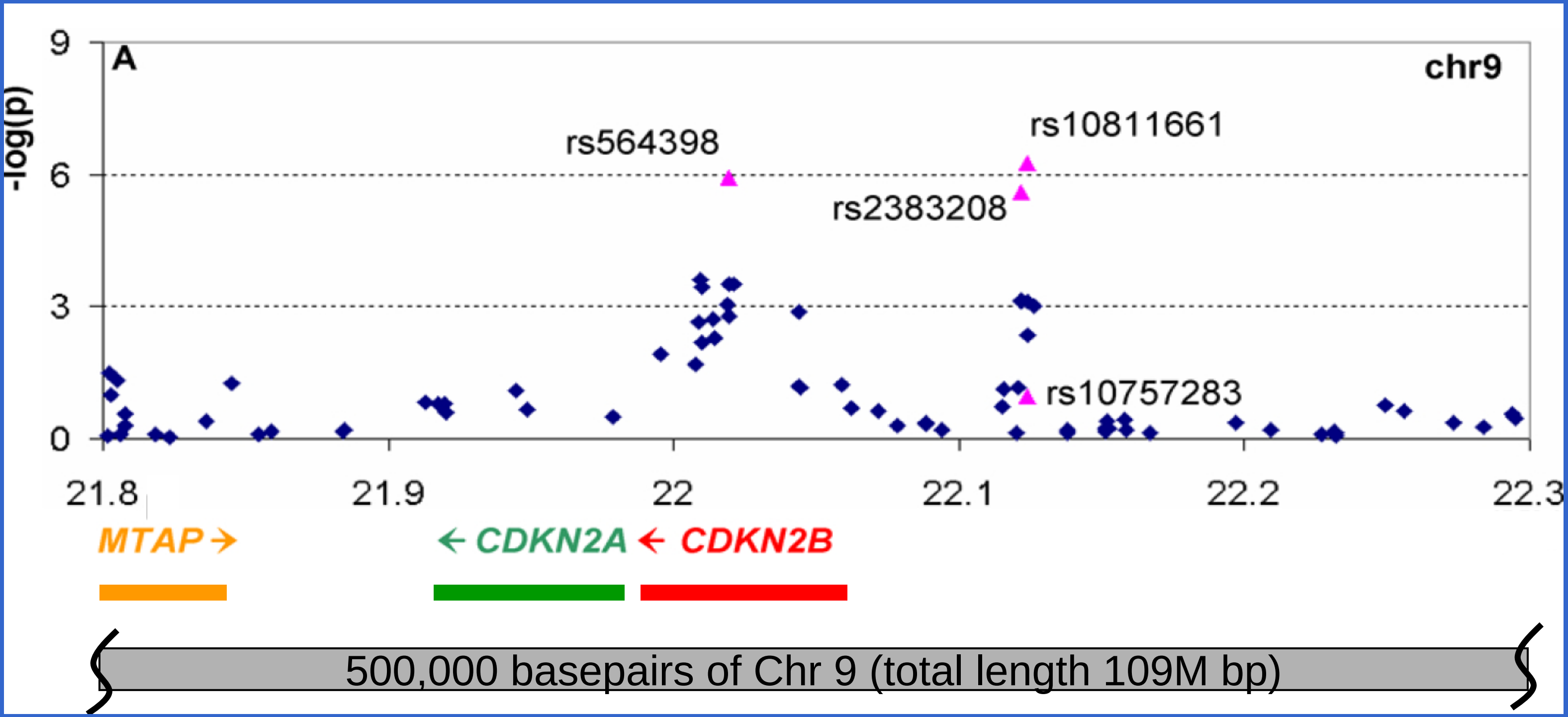
obtaining a deep catalogue of human genetic variation with new sequencing technology

First quarter 2008


IL23R Crohn's Disease
IL23R Inflammatory bowel disease
CELSR2 PSRC1,SORT1 LDL Cholesterol
FCER1A OR10J3 OR10J1 MCP-1
NOS1AP QT interval prolongation
Crohn's Disease
Systemic lupus erythematosus in women
Crohn's Disease
CFH Age-related macular degeneration
RGS1 Celiac Disease
ANGPTL3, DOCK7, ATG4C Triglycerides
CELSR2, GALNT2 PSRC1,SORT1 HDL Cholesterol
GALNT2 Triglycerides
1
ABCG8 Gallstones
GCKR Triglycerides
THADA Type 2 Diabetes
BCL11A F cell distribution
EHBP1 Prostate Cancer
MEISI Restless Leg Syndrome
IL1RL1 IL18R1 IL18RAP SLC9A4 Celiac Disease
Breast cancer
Coronary Disease
ATG16L1 Crohn's Disease
2
BSN,MST1 Crohn's Disease
PXK Systemic lupus erythematosus in women
ADAMTS9 Type 2 Diabetes
IL12A SCHIP1 Celiac Disease
IGF2BP2 Type 2 Diabetes
LPP Celiac Disease
3
RNF212 Recombination rate
SLC2A9 GLUT9 Serum urate
BANK1 Systemic lupus erythematosus
Atrial Fibrillation
K1A1109 TENR IL2, IL21 Celiac Disease
4
SLC45A2 Skin pigmentation by reflectance spectroscopy
IRGM Crohn's Disease
IL7RA Multiple Sclerosis
Crohn's Disease
MAP3K1 Breast Cancer
IRGM Crohn's Disease
5
between SEC5L1 and IRF4 Freckles
CDKAL1 Type 2 Diabetes
HCP5, HLA-C HIV1 Viral Setpoint
BTBD9 Restless Legs Syndrome
B3GALT4 LDL Cholesterol
ECHDC1 RNF146 Breast Cancer
F cell distribution
near TNFAIP3 and OLIG3 Rheumatoid athritis
MTHFD1L Coronary Disease
SLC22A3 Prostate Cancer
TAGAP Celiac Disease
6
JAZF1 Prostate Cancer
JAZF1 Type 2 Diabetes
BCL7B,TBL2 MLXIPL Triglycerides
LMTK2 Prostate Cancer
Rheumatoid athritis
DPP6 Amyotrophic lateral sclerosis
7
BLK,C8orf13 ITGAM-ITGAX Systemic lupus erythematosus
EIF3H Colorectal Cancer
SLC30A8 Type 2 Diabetes
TRIB1 Triglycerides
ORF DQ515897 Colorectal Cancer
Prostate Cancer
Breast Cancer
8
CDKN2A/B Type 2 Diabetes
CDKN2A/B Coronary Disease
TRAF1-C5 Rheumatoid athritis
9
IL2A Multiple Sclerosis
Colorectal Cancer
CDC123 CAMK1D Type 2 Diabetes
MSMB Prostate Cancer
Crohn's Disease
HHEX Type 2 Diabetes
NKX2-3 Crohn's Disease
Sporadic Amyotrophic lateral sclerosis
FGFR2 Breast cancer
CTBP2 Prostate Cancer
HTRA1 Wet Age-related Macular Degeneration
10
KIAA1542 Systemic lupus erythematosus in women
LSP1 Breast cancer
XmnI-G_γ F cell distribution
Type 2 Diabetes
Prostate Cancer
GAB2 Alzheimer disease APOE*e4 carriers
TYR Skin pigmentation by reflectance spectroscopy
Colorectal Cancer
11
RAB5B SUOX IKZF4 ERBB3 CDK2 ERBB3 Type 1 Diabetes
HMGA2 Height
TSPAN8 LGR5 Type 2 Diabetes
KITLG Blond or brown hair
MVK,MMAB HDL Cholesterol
C12orf30 Type 1 Diabetes
SH2B3/LNK Type 1 Diabetes
12
DGKH Bipolar Disorder
13
Coronary spasm in women
SLC24A4 Blue or Green Eyes
SLC24A4 Blond or Brown Hair
14
HERC2 Iris color
SLC24A5 Skin pigmentation by reflectance spectroscopy
MAP2K5 LBXCOR1 Restless Legs Syndrome
LOXL1 Exfoliation Glaucoma
15
TRAFD1 PTPN11 KIAA0350 Type 1 Diabetes
PALB2 NDUFAB1 DCTN5 Bipolar Disorder
ITGAM Systemic lupus erythematosus in women
TNCR9 LOC643714 Breast Cancer
FTO Type 2 Diabetes
FTO Obesity
16
TCF2 Prostate Cancer
Childhood Asthma
Prostate Cancer
17
PTPN2 Type 1 Diabetes
PTPN2 Crohn's Disease
SMAD7 Colorectal Cancer
CD226 Type 1 Diabetes
18
NCAN CILP2, PBX4 LDL Cholesterol
CILP2,NCAN PBX4 Triglycerides
KLK3 Prostate Cancer
19
BFZB Height
20
Crohn's Disease
21
22
CSF2RA Schizophrenia
NUDT10 NUDT11 LOC340602 GSPT2 MAGED1 Prostate Cancer
X
CSF2RA Schizophrenia
Y

Manolio, Brooks, Collins, J. Clin. Invest., May 2008

Chromosome 9p21: diabetes, coronary heart disease. Three genes, multiple SNPs



Zeggini et al, *Science* 2007; 316:1336-1341.

After GWAS “hit”, what next?

(remember, these are associations, not causes)

One region (~Mb), multiple genes, or sometimes no genes (!), multiple SNPs to sort through

Which is the right gene? What is the “causal” variant?

The current SNP catalog is not complete – may not have the causal variant

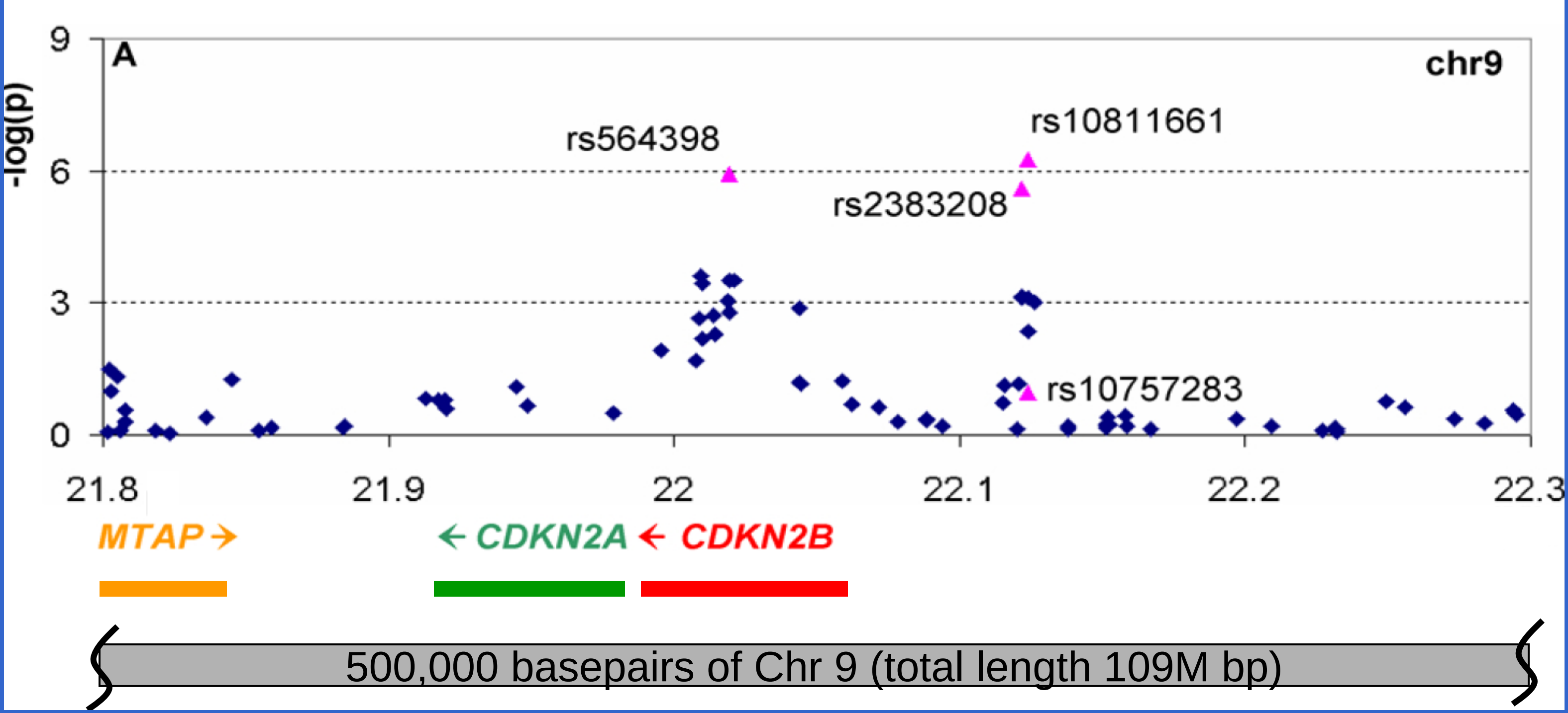
After a GWAS “hit”, what next?

- One could get lucky (gene is a likely candidate based on previously known function*; a known associated SNP is a variant that prevents any gene function)
- Gene expression correlates with believed function (e.g. tissue specific, disease specific)
- Conservation of sequence between genomes of many mammals

- Get a complete list of variants in the region, and one of them will be right. Need to sequence the associated region in many people.

*CDKN: evidence for a role in islet cell growth. Also a tumor suppressor.

Chromosome 9p21: diabetes, coronary heart disease. Three genes, multiple SNPs



Good bet on the gene, but what is the cause?

1000 Genomes Project: A resource for aiding human genetics studies

- An essentially complete list of all variants in human populations

- To provide a catalog of almost all variants in regions of all possible GWAS hits (i.e., the whole genome) ahead of time, so studies do not need to sequence their samples

(Gives the complete list of candidates, but still have to follow up on all candidate variants!)

Other potential benefits for Whole Genome Association studies

- The new variants will be associated by LD context with all existing variants, increasing the power of GWAS
- Better design of future assays for variation
- Access to lower frequency variants than current designs, e.g. down to 1%. (At what frequency do disease-causing variations occur in the population?)
- Can find alternate alleles in region of interest (disease could be caused by more than one variant in a single gene)

1000 Genomes primary goals: how many more variants?

“Essentially all” (not just a lot of) common variation genome-wide: any variant occurring in the population down to 1% allele frequency.

Deeper in gene regions (0.5%-0.1%)

All variant types (SNPs, insertions/deletions, and structural variants)

Place variants in their haplotype context (what other variants are they associated with?)

Do this in multiple populations—enough people at random that “all” (medically relevant) variation will be represented

How to do?

- Sequence in three populations to start: European, Africa, East Asian*; 500 individuals each

- Need to understand exactly how much sequence needed from each individual to build haplotype information

- A one-year pilot phase to test theory and technology:
  - What will it take for the new platforms to produce data that are useful for this?
  - How much sequence from each individual is needed?
  - Do we have enough from each population?
  - Build analytical infrastructure

- Two year main project

*Samples are mostly those already collected for HapMap under appropriate consent for fully anonymous release of genomic data. Some new anonymous samples will be needed.

1000 Genomes Pilots

Started Feb 2008, ~ 300 Gb data already

– Pilot 1: 180 samples @ less sequence each:
CEU (European) 4x, YRI (African) 2x, CHB/JPT (East Asian) 2x

~10 people done

– Pilot 2: CEU and YRI families (two parents, one child) @ high levels of sequence (20X)

CEU trio mostly complete
YRI trio in progress

– Pilot 3: 1000 genes in 1000 people

Starting

– Test multiple platforms/protocols

– Develop and evaluate methods for data collection and analysis

Simulations of trios,
1000 people at 2x,
plus samples at 4x, 8x

Additional goals

Not just SNPs: structural variation (2bp to >1M bp)

Population genetic studies

- Identifying regions under selection (now or in the past)
- Studies of processes of mutation and recombination
- Population differentiation and history

Improvement of the human reference sequence

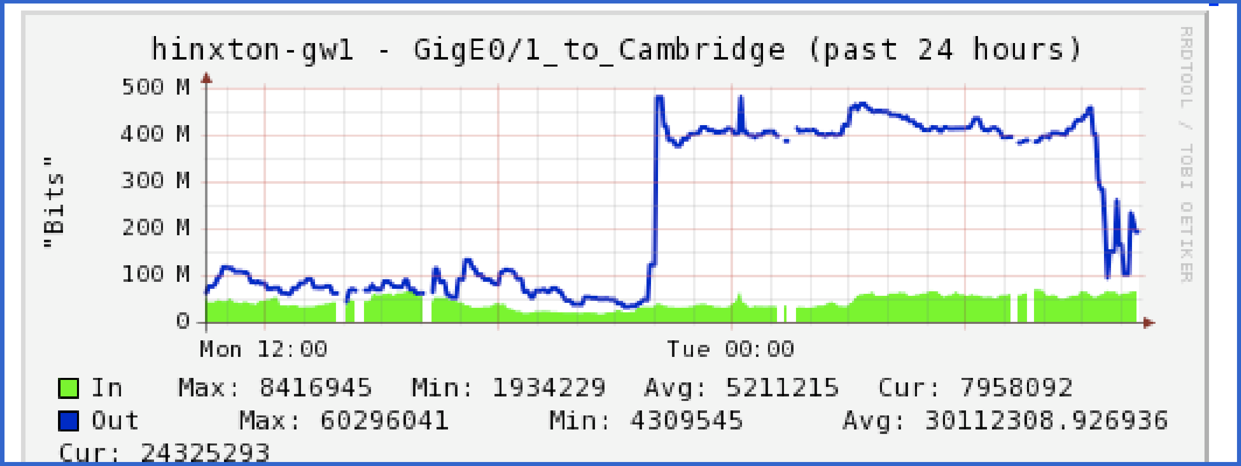
- Find and fix errors
- The current reference sequence, and any one individual, is missing sequence present in others
- Coordinate with the Human Genome Reference Consortium to represent all unique human sequence

Impractical without new sequencing technologies

- Project requires ~18,000 Gb
- “Old” tech (2006): >$1B
- New tech (2008): ~$50M

Challenges in “drinking from the firehose”

- Data handling, informatics resources: a LOT of data—the *initial* deposition increased the total sequence data available in the public domain by 10%, overnight



- Analysis, analysis, analysis…

- Samples, with appropriate consent for use in genomic studies and ***data release***

1000 Genomes Consortium

Production: Sanger Institute, Beijing Genomics Institute, Baylor College of Medicine, Broad Institute, Washington University of St Louis

Analysis: many statistical and population geneticists

Data Coordination: European Bioinformatics Institute, National Center for Biotechnology Information

Samples/ELSI: expertise in ethics and population sampling

Funding: Wellcome Trust, Beijing Genomics Institute, National Institutes of Health/NHGRI

A Public Resource

- Data publicly available shortly after it is produced

  -raw sequence data in the Short Read Archive

  -SNPs and other variant data in dbSNP

- Cell lines available

1000 Genomes Project steering co-chairs:

Richard Durbin	Wellcome Trust Sanger Institute
David Altshuler	Broad Institute

NHGRI Staff:

Lisa Brooks
Jean McEwen
Adam Felsenfeld

1000 Genomes

A Deep Catalog of Human Genetic Variation

Samples and ELSI Group

Leena Peltonen (co-chair) Sanger Institute
Bartha Knoppers (co-chair) University of Montreal
Aravinda Chakravarti (co-chair) Johns Hopkins
Gonçalo Abecasis University of Michigan
Richard Gibbs Baylor College of Medicine
Lynn Jorde University of Utah
Eric Juengst Case Western Reserve University
Jane Kaye Oxford University
Alastair Kent Genetic Interest Group
Rick Kittles University of Chicago
Jim Mullikin National Human Genome Research Institute
Mike Province Washington University in St. Louis
Charles Rotimi Howard University
Yeyang Su Beijing Genomics Institute
Chris Tyler-Smith Sanger Institute
Ling Yang Beijing Genomics Institute

Production Group

Elaine Mardis (co-chair) Washington University in St. Louis
Stacey Gabriel (co-chair) Broad Institute
Richard Durbin Sanger Institute
Richard Gibbs Baylor College of Medicine
David Jaffe Broad Institute
Ruiqiang Li Beijing Genomics Institute
Donna Muzny Baylor College of Medicine
Chad Nusbaum Broad Institute
Aarno Palotie Sanger Institute
Dan Turner Sanger Institute
Jun Wang Beijing Genomics Institute
Wei Wang Beijing Genomics Institute
Rick Wilson Washington University in St. Louis

Data Flow Group (being formed)

Paul Flicek (co-chair) European Bioinforma
Stephen Sherry (co-chair) National Center
Ewan Birney European Bioinformatics Instit
Clive Brown Sanger Institute
David Dooling Washington University in St.
Richard Gibbs Baylor College of Medicine
Sol Katzman University of California, Santa
Hoda Khouri National Center for Biotechnology Information
Martin Shumway National Center for Biotechnology Information
Jun Wang Beijing Genomics Institute
George Weinstock Baylor College of Medicine
(Broad representative)

Steering Committee

Richard Durbin (co-chair) Sanger Institute
David Altshuler (co-chair) Broad / MGH / Harvard
Gonçalo Abecasis University of Michigan
Aravinda Chakravarti Johns Hopkins
Andrew Clark Cornell University
Francis Collins National Human Genome Research Institute
Peter Donnelly Oxford University
Paul Flicek European Bioinformatics Institute
Stacey Gabriel Broad Institute
Richard Gibbs Baylor College of Medicine
Bartha Knoppers University of Montreal
Eric Lander Broad Institute
Elaine Mardis Washington University in St. Louis
Gil McVean Oxford University
Debbie Nickerson University of Washington
Leena Peltonen Sanger Institute
Stephen Sherry National Center for Biotechnology Information
Rick Wilson Washington University in St. Louis
Huanming (Henry) Yang Beijing Genomics Institute

Funders

Alan Schafer Wellcome Trust
Francis Collins National Human Genome Research Institute
Lisa Brooks National Human Genome Research Institute
Audrey Duncanson Wellcome Trust
Adam Felsenfeld National Human Genome Research Institute
Mark Guyer National Human Genome Research Institute
Ruth Jamieson Wellcome Trust
Li Kang Chinese Academy of Science
Yingli [illegible] Shenzhen Local Municipal Government
Jean McEwen National Human Genome Research Institute
Jane Peterson National Human Genome Research Institute
Anne Pierson National Human Genome Research Institute
Zhiwu Ren National Planning and Development Committee
Jian Wang Beijing Genomics Institute

Analysis Group

Gil McVean (co-chair) Oxford University
Gonçalo Abecasis (co-chair) University of Michigan
David Altshuler Broad / MGH / Harvard
Paul de Bakker Broad / BWH / Harvard
Brian Browning University of Auckland
Sharon Browning University of Auckland
Carlos Bustamante Cornell University
David Carter Sanger Institute
Aravinda Chakravarti Johns Hopkins
Andrew Clark Cornell University
Don Conrad Sanger Institute
Mark Daly Broad / MGH / Harvard
Manolis Dermitzakis Sanger Institute
Peter Donnelly Oxford University
Richard Durbin Sanger Institute
Evan Eichler University of Washington
Paul Flicek European Bioinformatics Institute
Bryan Howie Oxford University
Matt Hurles Sanger Institute
David Jaffe Broad Institute
Lynn Jorde University of Utah
Hoda Khouri National Center for Biotechnology Information
Eric Lander Broad Institute
Charles Lee Brigham and Women's Hospital
Guoqing Li Beijing Genomics Institute
Heng Li Sanger Institute
Ruiqiang Li Beijing Genomics Institute
Yingrui Li Beijing Genomics Institute
Yun Li University of Michigan
Jonathan Marchini Oxford University
Gabor Marth Boston College
Steve McCarroll Broad Institute
Jim Mullikin National Human Genome Research Institute
Simon Myers Oxford University
Rasmus Nielsen University of California, Berkeley
Alkes Price Broad / Harvard
Jonathan Pritchard University of Chicago
Mike Province Washington University in St Louis
Molly Przeworski University of Chicago
Shaun Purcell Broad / MGH / Harvard
Noah Rosenberg University of Michigan
Pardis Sabeti Broad / Harvard
Paul Scheet University of Michigan
Steven Schaffner Broad Institute
Jonathan Sebat Cold Spring Harbor Laboratory
Stephen Sherry National Center for Biotechnology Information
Matthew Stephens University of Chicago
Simon Tavaré University of Southern California
Chris Tyler-Smith Sanger Institute
Jun Wang Beijing Genomics Institute
David Wheeler Baylor College of Medicine
Hongkun Zheng Beijing Genomics Institute

www.1000genomes.org

Medical Sequencing

- Finding sequence variants that underlie disease
- Ideal: Sequence whole genomes of patients vs. healthy people, identify differences
- Reality: Too expensive now
- Challenge: Too many variants to sort through
- Solution: Pick candidate regions (e.g., GWAS; by function; by other previous findings); or "exomes" (practical very soon).

Medical Sequencing

Example: Autism

- Choose candidates based on function e.g., in neuronal synapses
- Sequence those genes in multiple affected and unaffected individuals
- Follow-up all differences (will find many differences, so this step needs to be relatively easy)