The 1000 genomes project: A catalogue of human polymorphism created using next generation sequencing

Paul Flicek

Vertebrate Genomics

1000 Genomes

A Deep Catalog of Human Genetic Variation

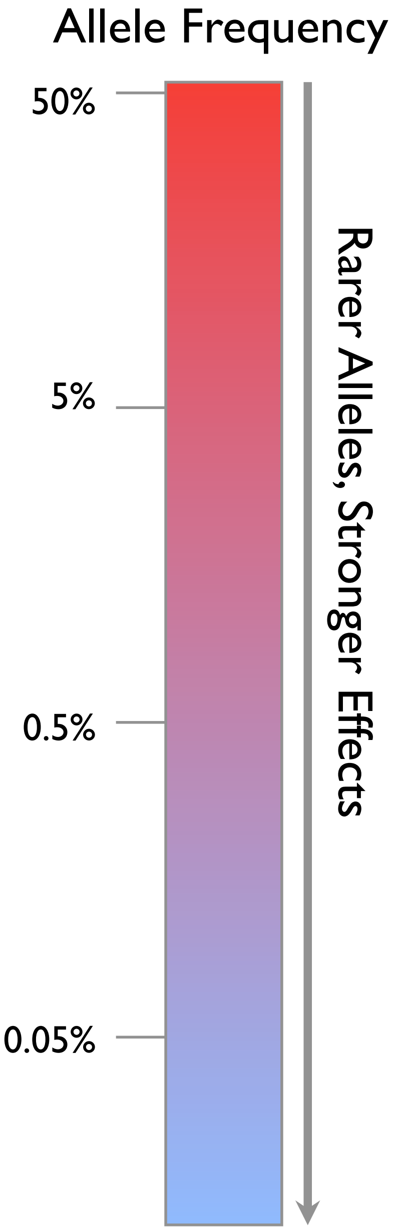
EMBL-EBI

wellcome trust sanger institute

1000 genomes project: motivation

- GWAS shows that systematic association studies can be used to map disease genes
- The first generation of GWAS was well powered only for SNPs with > 5% MAF
- Next generation sequencing now makes it possible to create a complete catalogue of human polymorphism for SNPs and CNVs

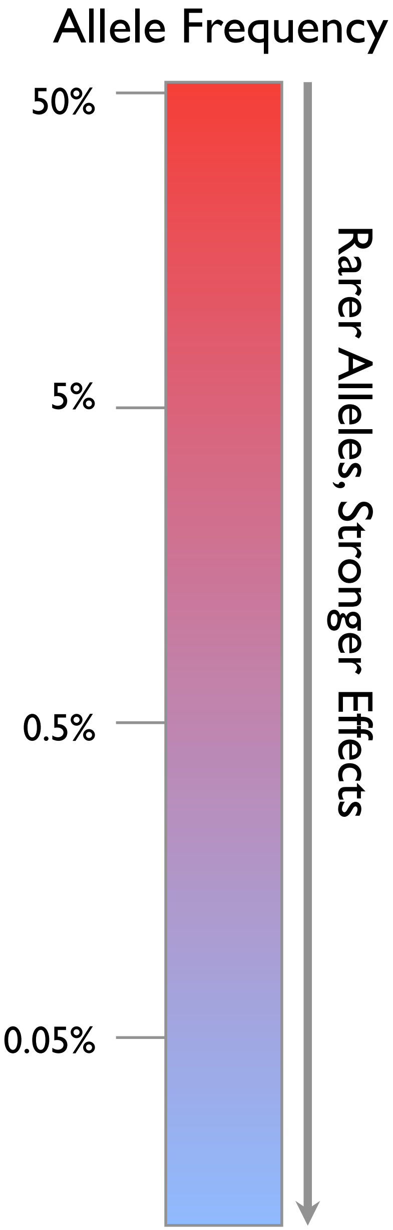
Exploring the full range of genetic variants



High-frequency polymorphisms	*Eg: many now known*	*HapMap* *First generation arrays*
Rare Mutations	*Eg: most mendelian* *MC4R, ABCA1* 1q21.1 in SCZ	*Direct sequencing* *Array-based detection (CNV)*

Slide courtesy David Altshuler

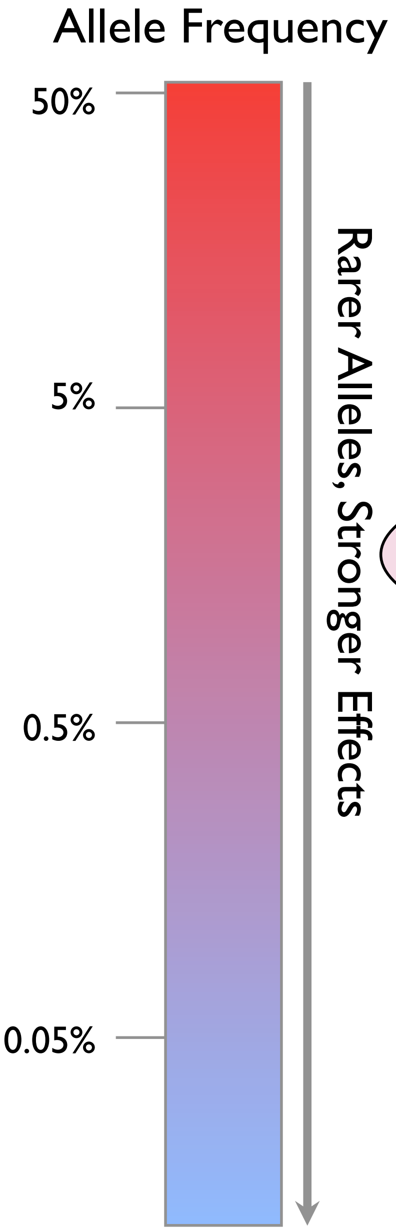
Exploring the full range of genetic variants



High-frequency polymorphisms	*Eg: many now known*	*HapMap* *First generation arrays*
Lower-frequency polymorphisms	*Eg: CFTR delta 508* *PCSK9* C679X	*1000 Genomes Project* *New arrays, imputation*
Rare Mutations	*Eg: most mendelian* *MC4R, ABCA1* 1q21.1 in SCZ	*Direct sequencing* *Array-based detection (CNV)*

Slide courtesy David Altshuler

Exploring the full range of genetic variants



High-frequency polymorphisms	*Eg: many now known*	*HapMap* *First generation arrays*
Lower-frequency polymorphisms	*Eg: CFTR delta 508* *PCSK9 C679X*	*1000 Genomes Project* *New arrays, imputation*
Rare Mutations	*Eg: most mendelian* *MC4R, ABCA1* 1q21.1 in SCZ	*Direct sequencing* *Array-based detection (CNV)*

Slide courtesy David Altshuler

1000 genomes project: primary goals

- A public database of essentially all SNPs and detectable CNVs with allele frequency >1% in each of multiple human population samples

1000 genomes project: primary goals

- A public database of essentially all SNPs and detectable CNVs with allele frequency >1% in each of multiple human population samples
- Pioneer and evaluate methods for:
    - Generating data from next-generation sequencing platforms
    - Exchanging and combining data and analytical methods
    - Discovering and genotyping SNPs and CNVs from nextgen data
    - Imputation with and from next generation sequencing data

1000 genomes project: primary goals

- A public database of essentially all SNPs and detectable CNVs with allele frequency >1% in each of multiple human population samples

- Pioneer and evaluate methods for:
  - Generating data from next-generation sequencing platforms
  - Exchanging and combining data and analytical methods
  - Discovering and genotyping SNPs and CNVs from nextgen data
  - Imputation with and from next generation sequencing data

Help establish methodology for rare variants

1000 genomes project: other goals

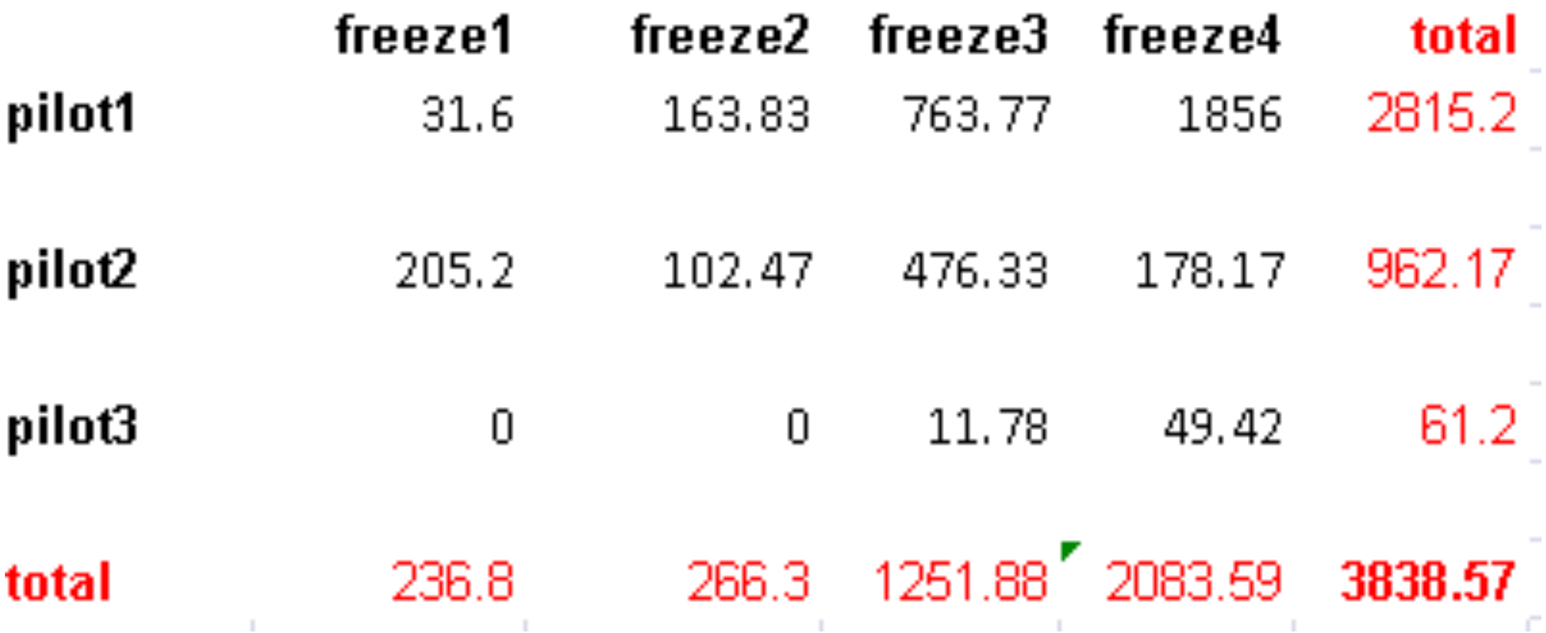
- Enable population genetic studies
  - Identifying regions under selection (now or in the past)
  - Studies of processes of mutation and recombination
  - Population differentiation and history
- Improvement of the human reference sequence
  - Find and fix errors
  - The current reference sequence, and any one individual, is missing sequence present in others
  - Coordinate with the Human Genome Reference Consortium to represent all unique human sequence

Three pilots studies

- Pilot 1: 4x coverage of 180 people
- Pilot 2: 20x coverage of 2 trios
- Pilot 3: targeted sequencing of 1000 genes in 1000 individuals

- Data: 3.8 terabases deposited at the EBI/NCBI to date
  - Illumina/Solexa, 454, and ABI SOLiD platforms
  - Academic genome centers in US, UK, Germany, China and platform companies

Data production (Gb) by pilot and freeze



	freeze1	freeze2	freeze3	freeze4	total
pilot1	31.6	163.83	763.77	1856	2815.2
pilot2	205.2	102.47	476.33	178.17	962.17
pilot3	0	0	11.78	49.42	61.2
total	236.8	266.3	1251.88	2083.59	3838.57

>1,000 x coverage of human genome already in pilots!

Basic Requirements

- Basic File formats
  - New technology sequencing does not produce the same type of raw data as Sanger-style sequencing
  - SRF (sequence read format) stores the raw data for submission
    - srf.sourceforge.net
  - Alignment formats must be efficient if one is mapping half a trillion reads
    - These are being developed now
- Initial analysis tools
  - Most current aligners incorporate the quality scores into the mapping
    - There are now being tested on the 1000 Genomes trio data

Advanced Requirements

- File formats
    - Genome likelihood format
        - Representing an individual genome with appropriate uncertainty
- Advanced analysis tools (mostly under development)
    - SNP calling
        - Trio aware
        - Population based
    - Assembly & Search
        - These can all be theoretically done with de Bruijn graphs, but this is still a research problem for mammalian genomes
    - Enriched fraction analysis for ChIP, MeDIP, DNase, FAIRE
        - Extensive development efforts underway

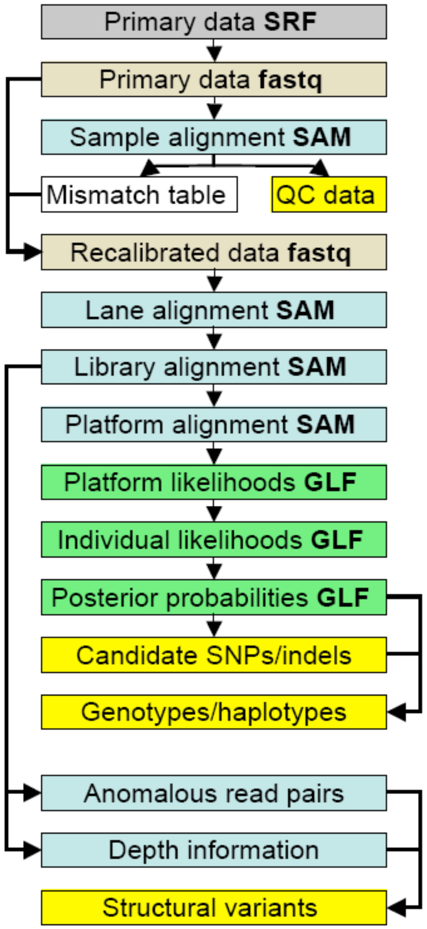
What are we storing?

- Raw data submitted in SRF or SFF (454)
  - Originally
    - Raw and processed intensities for Solexa data
      - NCBI SRA stores intensities in a lossy format
    - Non PF filtered reads (if submitted)
  - Current
    - Raw intensities + base calls + quality + derived files
- SOLiD data arriving in both SRF and native formats
  - Most active development in this area
- Fastq files
  - Most downstream analysis starts here
  - Machine and calibrated quality scores
- Approximately 60 bytes per mapped base originally
  - We need to be at approximately 10 bytes per base for instrumentation
  - Total about 25 bytes per base

Bioinformatics Requirements

- New File formats
  - New technology sequencing does not produce the same type of raw data as Sanger-style sequencing
  - Alignment formats must be efficient if one is mapping half a trillion reads
  - Genome likelihood format
    - Representing an individual genome with appropriate uncertainty
- Initial analysis tools
  - Most current aligners incorporate the quality scores into the mapping
  - Trio aware SNP calling methods

The need for standard data formats



		Unit (one file per …)	Who? (italics if not done yet)
1. Submit	Primary data **SRF**	*lane*	production centres
2. Extract	Primary data **fastq**	*lane*	DCC
3. Map sample	Sample alignment **SAM**	*lane*	Sanger *to DCC*
4. Recalibrate	Mismatch table QC data	*lane*	Sanger *to DCC*
	Recalibrated data **fastq**	*lane*	Sanger *to DCC*
5. Map	Lane alignment **SAM**	*lane*	Data Processing
6. Merge and remove dups	Library alignment **SAM**	*library*	Data Processing
7. Merge	Platform alignment **SAM**	*platform/individual*	Data Processing
8. Calc likelihoods	Platform likelihoods **GLF**	*platform/individual*	Data Processing
9. Combine l'hoods	Individual likelihoods **GLF**	*individual*	Data Processing
10. Apply priors	Posterior probabilities **GLF**	*individual*	Data Processing
11. Call SNPs/indels	Candidate SNPs/indels	*experiment/population*	Data Processing
12. Call genotypes	Genotypes/haplotypes	*individual*	Data Processing
13. Collect read pair info	Anomalous read pairs	*library*	Structural Variation
13. Collect depth info	Depth information	*library*	Structural Variation
14. Call SVs	Structural variants	*experiment and individual*	Structural Variation

Slide courtesy Richard Durbin

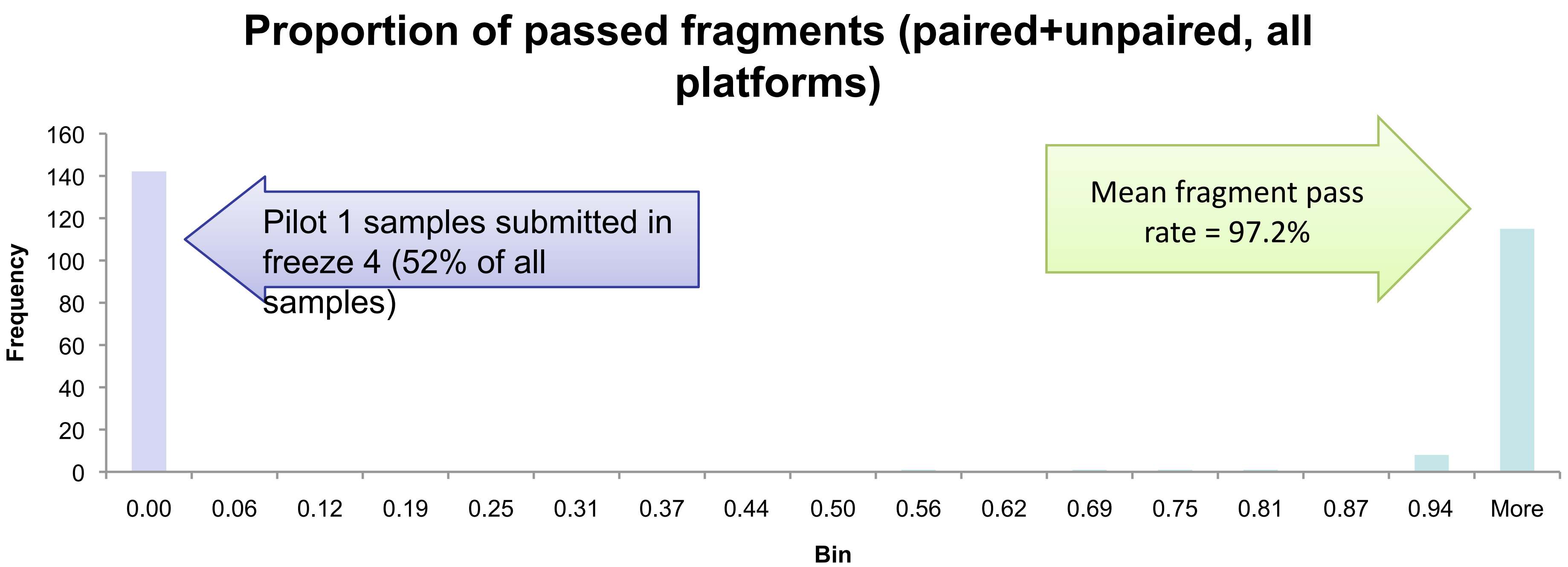
Needing unprecedented data quality

- Variation in human DNA is ≈0.1-0.2%
- However, 90% of this is already in dbSNP, so event rate for new variants ≈1:10,000
- Per base error rates must be <1:100,000
- Must account for error properties of raw data

Analysis in process for Freeze 1,2,3

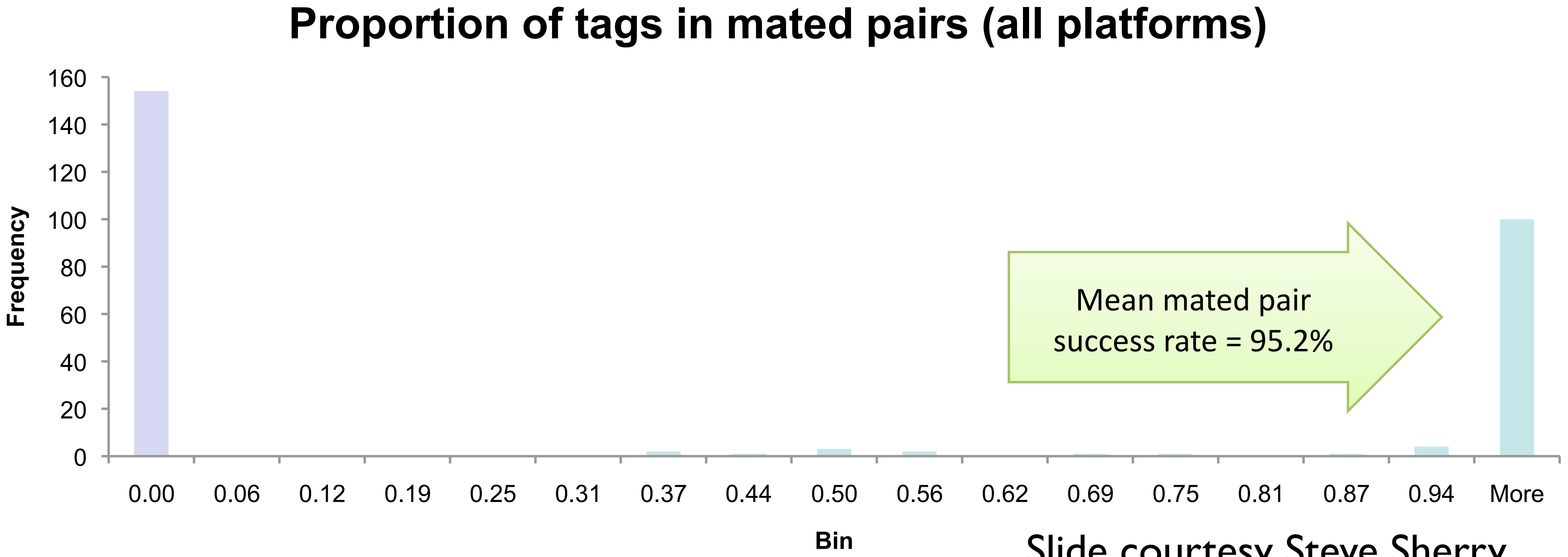
QC filter criteria	Preliminary measure
Read flagged as failed by center	<0.5% overall
'N' calls in first 25 bases	3.22%
Quality score <3 in first 25 bases	0.86%
Average fragment pass rate	97.2%
Average pairing success rate	95.2%

Slide courtesy Steve Sherry



Pilot 1 samples submitted in freeze 4 (52% of all samples)

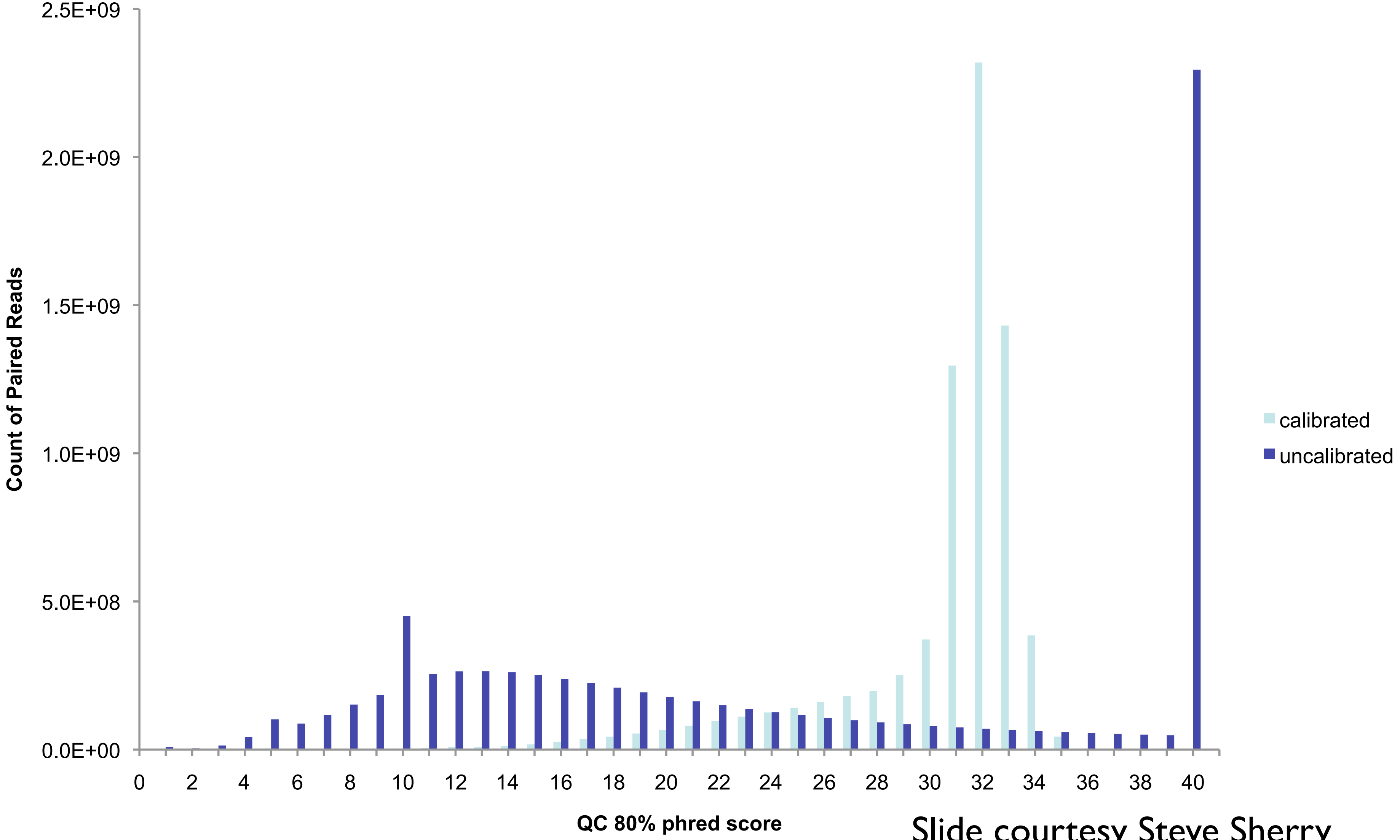
Mean fragment pass rate = 97.2%



Mean mated pair success rate = 95.2%

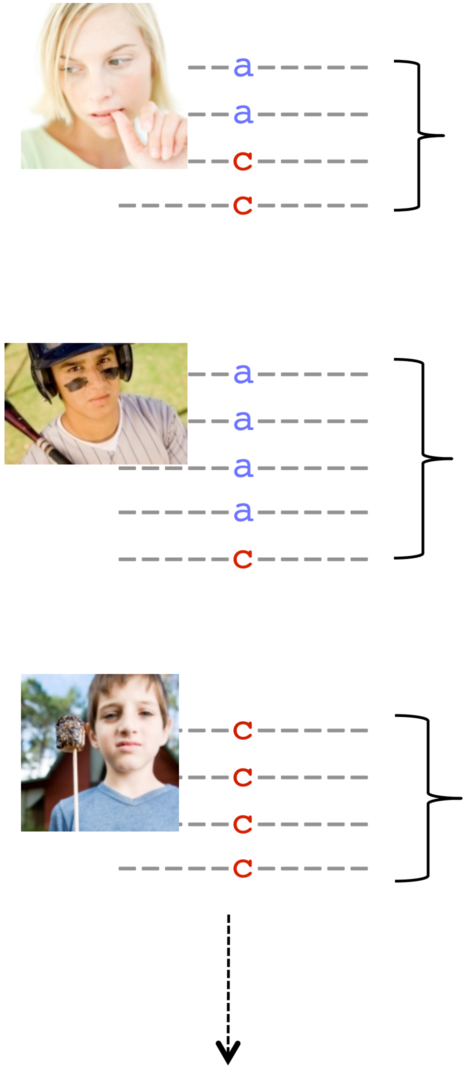
Slide courtesy Steve Sherry

Sanger Center calibrated vs. un-calibrated Illumina phred score Q -- 20th percentile (80% bases with score > Q)



Slide courtesy Steve Sherry

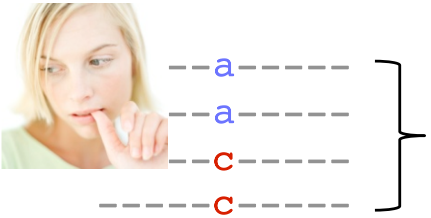
Genotype likelihood format: GLF



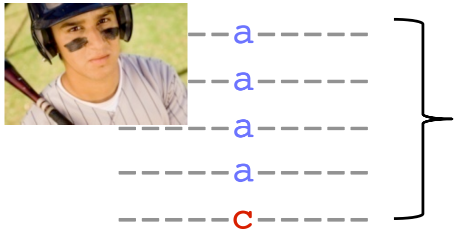
P(SNP)

Slide courtesy Gabor Marth

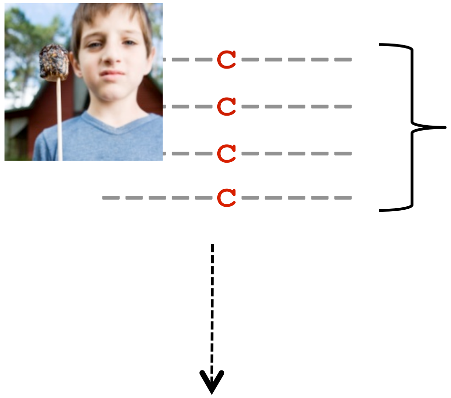
Genotype likelihood format: GLF



$P(B_1=aacc|G_1=aa)$
$P(B_1=aacc|G_1=cc)$
$P(B_1=aacc|G_1=ac)$



$P(B_i=aaaacc|G_i=aa)$
$P(B_i=aaaacc|G_i=cc)$
$P(B_i=aaaacc|G_i=ac)$



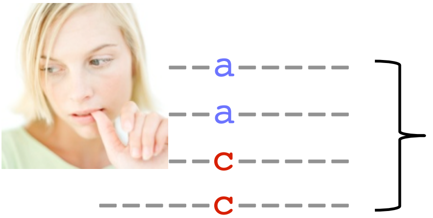
$P(B_n=cccc|G_n=aa)$
$P(B_n=cccc|G_n=cc)$
$P(B_n=cccc|G_n=ac)$

P(SNP)

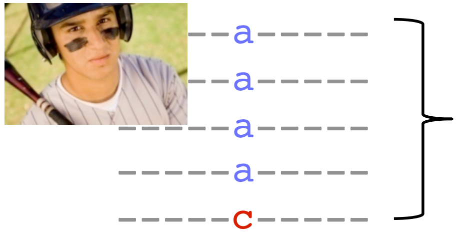
“genotype likelihoods”

Slide courtesy Gabor Marth

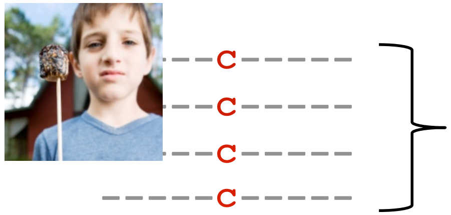
Genotype likelihood format: GLF



$P(B_1=aacc|G_1=aa)$
$P(B_1=aacc|G_1=cc)$
$P(B_1=aacc|G_1=ac)$



$P(B_i=aaaacc|G_i=aa)$
$P(B_i=aaaacc|G_i=cc)$
$P(B_i=aaaacc|G_i=ac)$



$P(B_n=cccc|G_n=aa)$
$P(B_n=cccc|G_n=cc)$
$P(B_n=cccc|G_n=ac)$

P(SNP)

“genotype likelihoods”

$\mathrm{Prior}(G_1,\ldots,G_i,\ldots,G_n)$

$P(G_1=aa|B_1=aacc; B_i=aaaacc; B_n= cccc)$
$P(G_1=cc|B_1=aacc; B_i=aaaacc; B_n= cccc)$
$P(G_1=ac|B_1=aacc; Bi=aaaacc; B_n= cccc)$

$P(G_i=aa|B_1=aacc; B_i=aaaacc; B_n= cccc)$
$P(G_i=cc|B_1=aacc; B_i=aaaacc; B_n= cccc)$
$P(G_i=ac|B_1=aacc; Bi=aaaacc; B_n= cccc)$

$P(G_n=aa|B_1=aacc; B_i=aaaacc; B_n= cccc)$
$P(G_n=cc|B_1=aacc; B_i=aaaacc; B_n= cccc)$
$P(G_n=ac|B_1=aacc; Bi=aaaacc; B_n= cccc)$

“genotype probabilities”

Slide courtesy Gabor Marth

Initial experience: SNP calling

- Deep coverage (20x) parent-offspring trio
  - 4,047,762 single base polymorphisms
    - 88% were already present in dbSNP

Analyses by Goncalo Abecasis, Richard Durbin, Stacey Gabriel

Initial experience: SNP calling

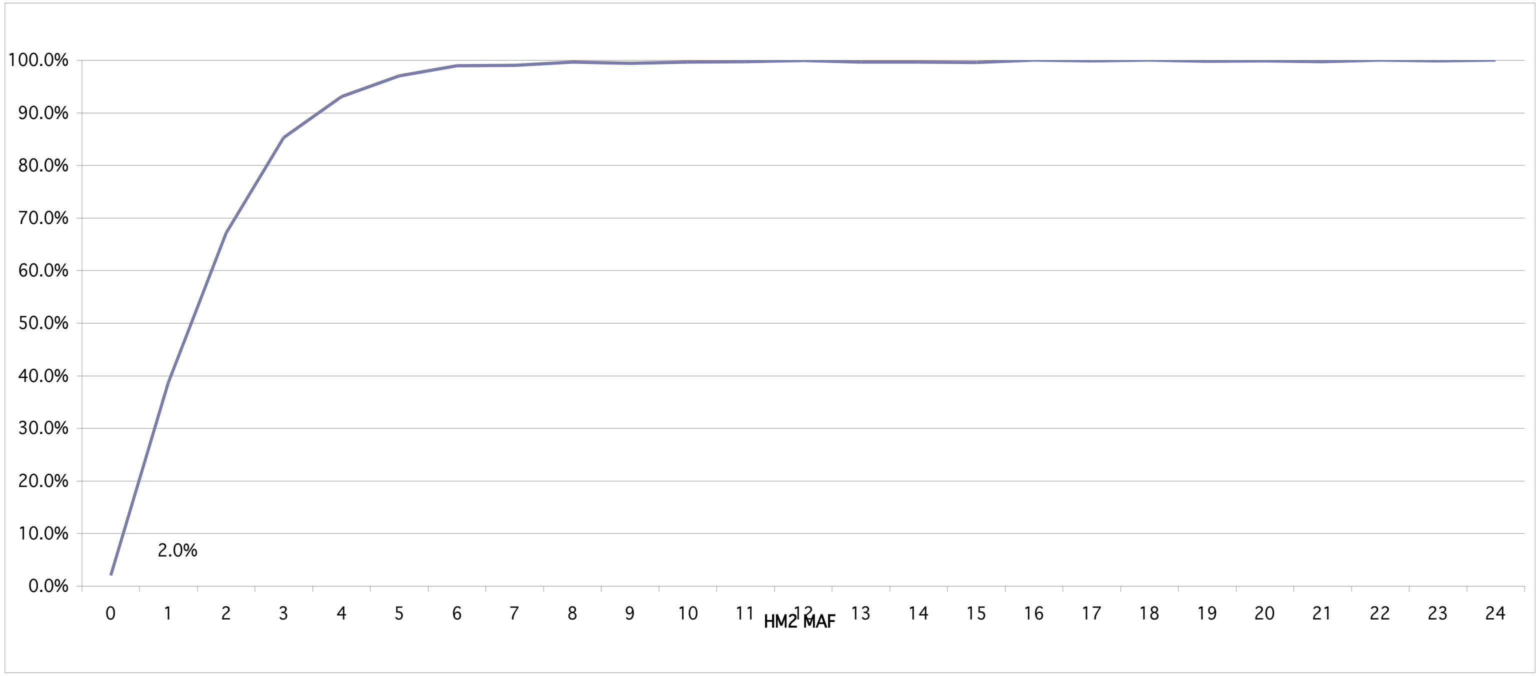
- Deep coverage (20x) parent-offspring trio
  - 4,047,762 single base polymorphisms
    - 88% were already present in dbSNP

- Validation testing of SNPs not in dbSNP
  - 1,200 tested using sequenom
  - 1,068 successful assays
  - 95% validated as true positive, in HWE, etc

Analyses by Goncalo Abecasis, Richard Durbin, Stacey Gabriel

Initial experience: SNP calling from 4x coverage in 36 unrelateds

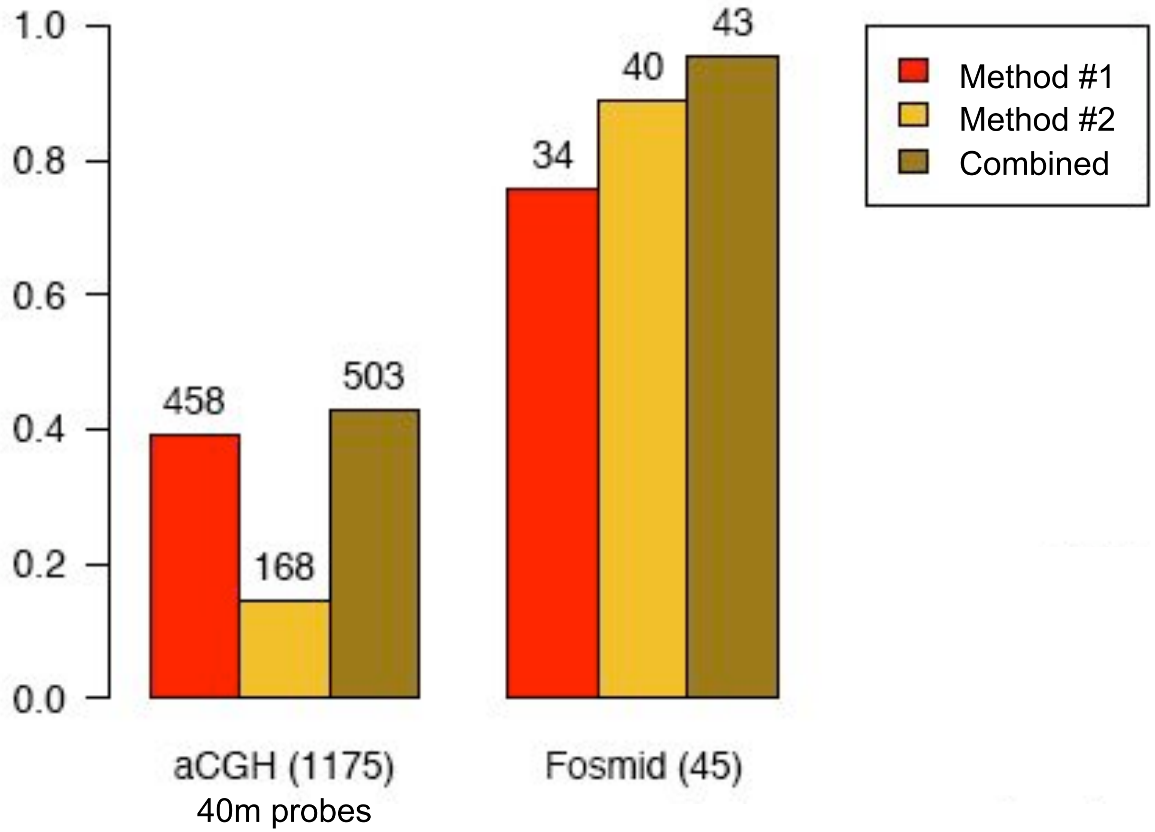
- Haplotype-informed SNP calling (knows tree at each site)
- 93% detection for alleles seen 4x, 97% for alleles seen 5x
- >50% novel (compared to ≈10% for trio sample)



Slide courtesy Richard Durbin

Example: structural variants from 1000G data

Proportion of validated SVs identified



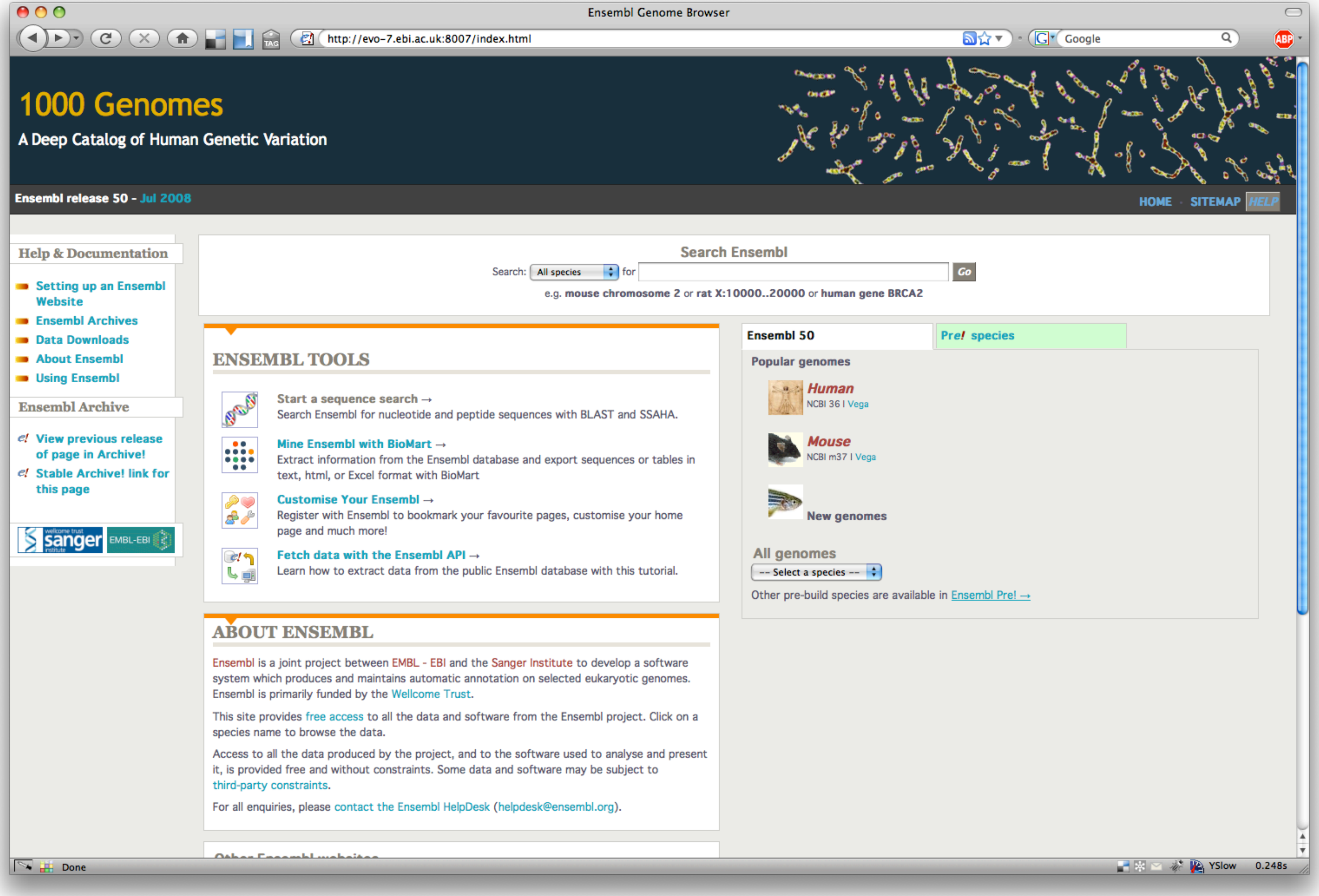
Slide courtesy Matt Hurles for 1000G SV group

Raw and summary data distribution

- Continue to be available from SRA/ERA with more extensive discoverability within these resources and supported on 1000genomes.org
- 1000 Genomes specific data that is not appropriate for archives, such as simulation data, will continue to be provided on the EBI/NCBI dedicated FTP sites

1000 Genomes Browser

- Based on Ensembl and potentially including the Resembl plugin developed by Illumina
- A separate installation managed and updated at the EBI and available within the 1000genomes.org domain
- SNPs, GLF and coverage data for all individuals
- “Full data” for the trios and other high coverage individuals such as NA18507 using Resembl if available
- Built on current version of Ensembl web code (with project specific “skinning”)
    - Expected update to new Ensembl interface in late Q1 or early Q2 2009



Ensembl Genome Browser

http://evo-7.ebi.ac.uk:8007/index.html

1000 Genomes

A Deep Catalog of Human Genetic Variation

Ensembl release 50 - Jul 2008

HOME · SITEMAP · HELP

Help & Documentation

- Setting up an Ensembl Website
- Ensembl Archives
- Data Downloads
- About Ensembl
- Using Ensembl

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page

Search Ensembl

Search: All species for [ ] Go

e.g. **mouse chromosome 2** or **rat X:10000..20000** or **human gene BRCA2**

ENSEMBL TOOLS

Start a sequence search →
Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.

Mine Ensembl with BioMart →
Extract information from the Ensembl database and export sequences or tables in text, html, or Excel format with BioMart

Customise Your Ensembl →
Register with Ensembl to bookmark your favourite pages, customise your home page and much more!

Fetch data with the Ensembl API →
Learn how to extract data from the public Ensembl database with this tutorial.

ABOUT ENSEMBL

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other Ensembl websites

Ensembl 50 | Pre! species

Popular genomes

Human
NCBI 36 | Vega

Mouse
NCBI m37 | Vega

New genomes

All genomes

-- Select a species --

Other pre-build species are available in Ensembl Pre! →

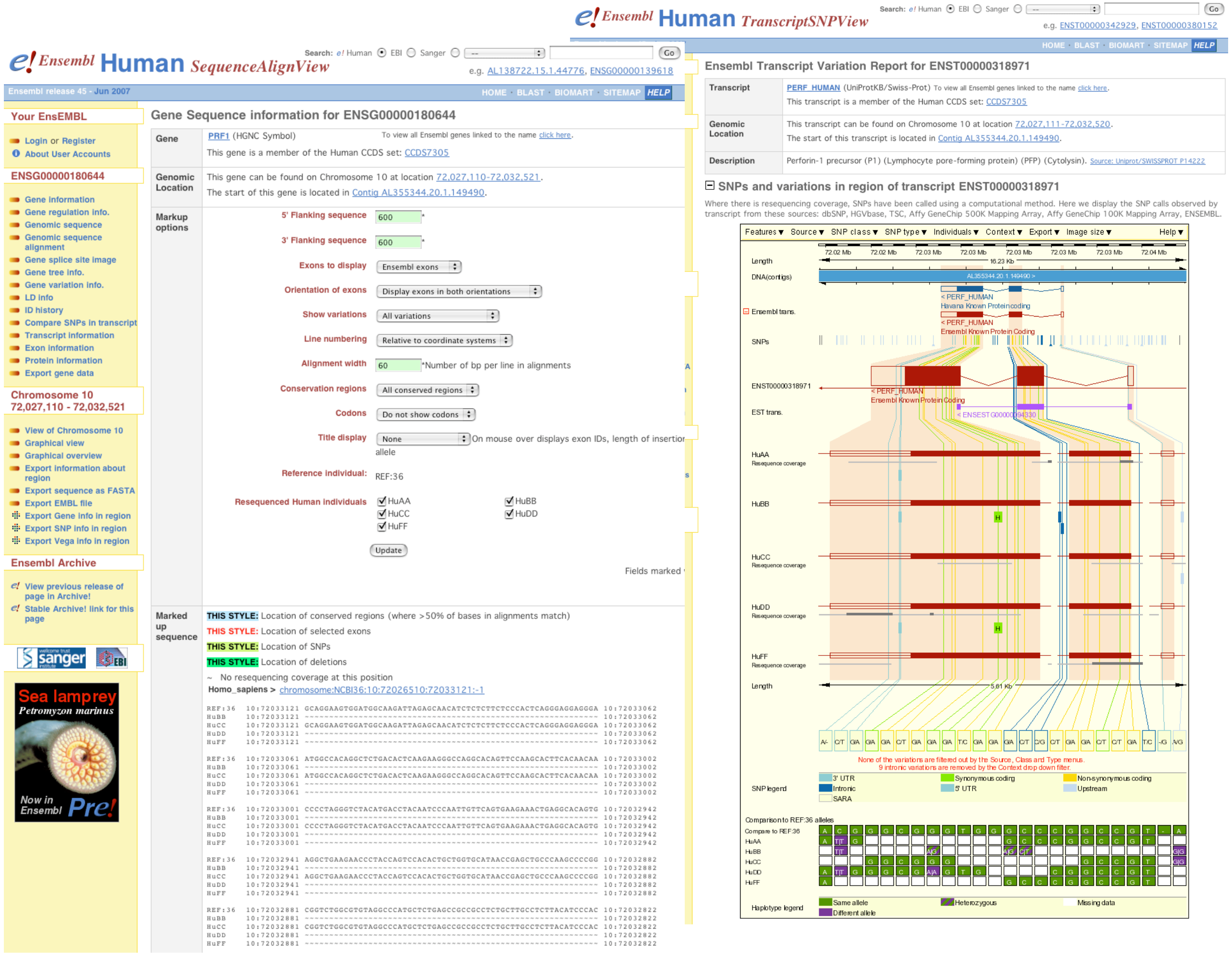
Done

YSlow 0.248s

Interaction with dbSNP and Ensembl and UCSC Browsers

- Data will be loading into each browser once it has been “released” by the project
- These SNPs will be deposited in dbSNP and from there make their way to the major browsers
- Support from Ensembl and UCSC for data beyond SNPs and CNVs will likely be more limited and less up to date than what is available at the project portal

Ensembl/Resembl Displays



e! Ensembl Human *SequenceAlignView*

Search: e! Human ● EBI ○ Sanger ○ [--] [ ] (Go)

e.g. AL138722.15.1.44776, ENSG00000139618

Ensembl release 45 - Jun 2007

HOME · BLAST · BIOMART · SITEMAP · HELP

Your EnsEMBL

- Login or Register
- About User Accounts

ENSG00000180644

- Gene information
- Gene regulation info.
- Genomic sequence
- Genomic sequence alignment
- Gene splice site image
- Gene tree info.
- Gene variation info.
- LD info
- ID history
- Compare SNPs in transcript
- Transcript information
- Exon information
- Protein information
- Export gene data

Chromosome 10 72,027,110 - 72,032,521

- View of Chromosome 10
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page

Sea lamprey
Petromyzon marinus
Now in Ensembl Pre!

Gene Sequence information for ENSG00000180644

Gene PRF1 (HGNC Symbol) To view all Ensembl genes linked to the name click here.
This gene is a member of the Human CCDS set: CCDS7305

Genomic Location This gene can be found on Chromosome 10 at location 72,027,110-72,032,521.
The start of this gene is located in Contig AL355344.20.1.149490.

Markup options

5' Flanking sequence	600 *
3' Flanking sequence	600 *
Exons to display	Ensembl exons
Orientation of exons	Display exons in both orientations
Show variations	All variations
Line numbering	Relative to coordinate systems
Alignment width	60 *Number of bp per line in alignments
Conservation regions	All conserved regions
Codons	Do not show codons
Title display	None On mouse over displays exon IDs, length of insertion allele
Reference individual:	REF:36
Resequenced Human individuals	☑ HuAA ☑ HuBB ☑ HuCC ☑ HuDD ☑ HuFF

(Update)

Fields marked

Marked up sequence

THIS STYLE: Location of conserved regions (where >50% of bases in alignments match)
THIS STYLE: Location of selected exons
THIS STYLE: Location of SNPs
THIS STYLE: Location of deletions
~ No resequencing coverage at this position
Homo_sapiens > chromosome:NCBI36:10:72026510:72033121:-1

```
REF:36   10:72033121 GCAGGAAGTGGATGGCAAGATTAGAGCAACATCTCTCTTCTCCCACTCAGGGAGGAGGGA 10:72033062
HuBB     10:72033121 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72033062
HuCC     10:72033121 GCAGGAAGTGGATGGCAAGATTAGAGCAACATCTCTCTTCTCCCACTCAGGGAGGAGGGA 10:72033062
HuDD     10:72033121 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72033062
HuFF     10:72033121 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72033062

REF:36   10:72033061 ATGGCCACAGGCTCTGACACTCAAGAAGGGCCAGGCACAGTTCCAAGCACTTCACAACAA 10:72033002
HuBB     10:72033061 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72033002
HuCC     10:72033061 ATGGCCACAGGCTCTGACACTCAAGAAGGGCCAGGCACAGTTCCAAGCACTTCACAACAA 10:72033002
HuDD     10:72033061 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72033002
HuFF     10:72033061 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72033002

REF:36   10:72033001 CCCCTAGGGTCTACATGACCTACAATCCCAATTGTTCAGTGAAGAAACTGAGGCACAGTG 10:72032942
HuBB     10:72033001 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032942
HuCC     10:72033001 CCCCTAGGGTCTACATGACCTACAATCCCAATTGTTCAGTGAAGAAACTGAGGCACAGTG 10:72032942
HuDD     10:72033001 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032942
HuFF     10:72033001 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032942

REF:36   10:72032941 AGGCTGAAGAACCCTACCAGTCCACACTGCTGGTGCATAACCGAGCTGCCCAAGCCCCGG 10:72032882
HuBB     10:72032941 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032882
HuCC     10:72032941 AGGCTGAAGAACCCTACCAGTCCACACTGCTGGTGCATAACCGAGCTGCCCAAGCCCCGG 10:72032882
HuDD     10:72032941 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032882
HuFF     10:72032941 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032882

REF:36   10:72032881 CGGTCTGGCGTGTAGGCCCATGCTCTGAGCCGCCGCCTCTGCTTGCCTCTTACATCCCAC 10:72032822
HuBB     10:72032881 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032822
HuCC     10:72032881 CGGTCTGGCGTGTAGGCCCATGCTCTGAGCCGCCGCCTCTGCTTGCCTCTTACATCCCAC 10:72032822
HuDD     10:72032881 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032822
HuFF     10:72032881 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 10:72032822
```

e! Ensembl Human *TranscriptSNPView*

Search: e! Human ● EBI ○ Sanger ○ [--] [ ] (Go)

e.g. ENST00000342929, ENST00000380152

HOME · BLAST · BIOMART · SITEMAP · HELP

Ensembl Transcript Variation Report for ENST00000318971

Transcript PERF_HUMAN (UniProtKB/Swiss-Prot) To view all Ensembl genes linked to the name click here.
This transcript is a member of the Human CCDS set: CCDS7305

Genomic Location This transcript can be found on Chromosome 10 at location 72,027,111-72,032,520.
The start of this transcript is located in Contig AL355344.20.1.149490.

Description Perforin-1 precursor (P1) (Lymphocyte pore-forming protein) (PFP) (Cytolysin). Source: Uniprot/SWISSPROT P14222

⊟ SNPs and variations in region of transcript ENST00000318971

Where there is resequencing coverage, SNPs have been called using a computational method. Here we display the SNP calls observed by transcript from these sources: dbSNP, HGVbase, TSC, Affy GeneChip 500K Mapping Array, Affy GeneChip 100K Mapping Array, ENSEMBL.

None of the variations are filtered out by the Source, Class and Type menus.
9 intronic variations are removed by the Context drop down filter.

SNP legend

3' UTR	Synonymous coding	Non-synonymous coding
Intronic	5' UTR	Upstream
SARA		

Comparison to REF:36 alleles

Compare to REF:36	A	C	G	G	G	C	G	G	G	T	G	G	G	C	C	C	G	G	C	C	G	T	-	A
HuAA	A	T|T	G										G	C	C	C	G	G	C	C	G	T		
HuBB		T|T						A|G					A|G	C|T										G|G
HuCC				G	G	C	G	G	G									G	C	C	G	T		G|G
HuDD	A	T|T	G	G	G	C	G	A|A	G	T	G					C	G	G	C	C	G	T		
HuFF	A												G	C	C	C	G	G	C	C	G	T		

Haplotype legend

Same allele	Heterozygous	Missing data
Different allele		

Putting this scale of data into perspective

- Current size of EMBL/Genbank: 235,135,312,328 nucleotides

- During September and October the 1000 Genomes project produced the equivalent of EMBL/GenBank *every week*

- Raw data is freely available now
  - ftp://ftp.1000genomes.ebi.ac.uk
  - ftp://ftp-trace.ncbi.nih.gov/1000genomes/

1000 genomes project: plans

- Pilots show high quality data collected at scale, and that variants can be called reliably

- Project has now set as its target:
  - 1,200 people sequenced each to 4 x coverage
  - Data collection completed by winter 2009

- Quarterly data releases starting Jan 2009

Success Measures

1. The DCC is providing data as fast or faster than the analysis group can handle it
2. The Production Group is creating data as fast or faster than the DCC can handle it
3. The manufactures are expanding machine capacity as fast or faster than the production centers can handle it

What will the 1000 genomes project provide to human genetics community?

- Essentially all SNPs (MAF >1%) in each sample
  - Will find many, but not all, variants 0.2-1% MAF
- Highly complete catalogue of CNVs
- Information required for imputation of lower frequency alleles into existing GWAS samples
- Content for next generation more poweful arrays
- A set of validated methods for use of next generation sequencing in disease samples

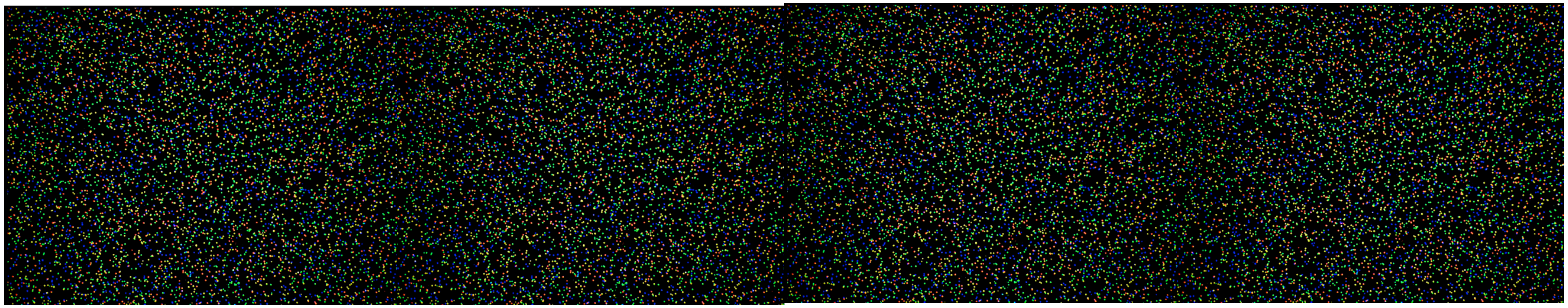
Sequencing data production is now just an order of magnitude behind CERN

- The Large Hadron Collider produces only 15 petabytes per year from a single point source
- The LHC grid is 140 computer centres in 33 countries
- Tier 0 (CERN) can write data to the ten Tier 1 centers at 1.3 GB/sec sustained and have tested long transfers at more than 3 GB/sec

Sequencing data production is now just an order of magnitude behind CERN

- Sequencing is producing data in hundreds of centers in dozens of countries with 9 production centers and two Tier 0 sites

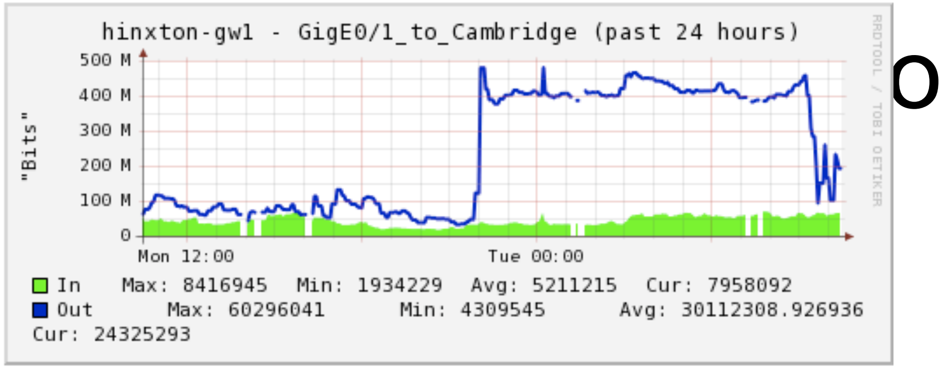
- The 1000 Genomes grid is, umm…



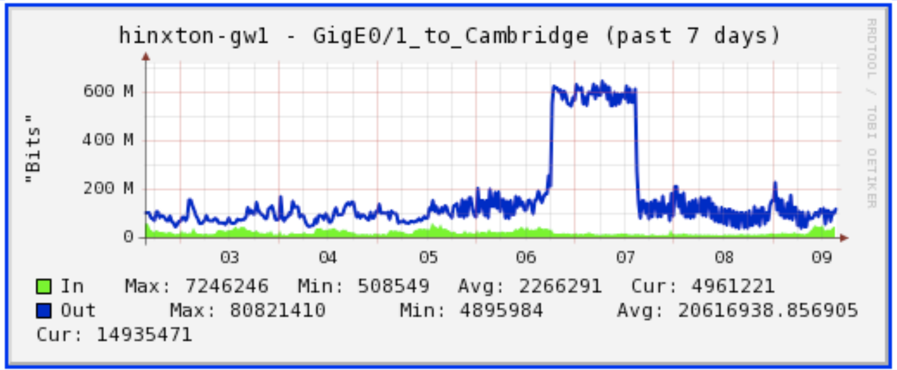
Data Transfer Infrastructure

- FTP does not work well for terabytes of data
- “Old fashioned” solutions
  - Copy the data onto a hard drive and mail the hard drive around the world
  - (Significant personnel costs)
- Infrastructure solutions
  - Create/buy dedicated lines for point to point transfer or direct connection to faster points on the backbone
  - Expensive to do collaborative analysis, but will probably be part of the solution
- Advanced technology solutions
  - Asperasoft
  - Uses udp to transfer files to avoid tcp
  - Can quickly saturate connections

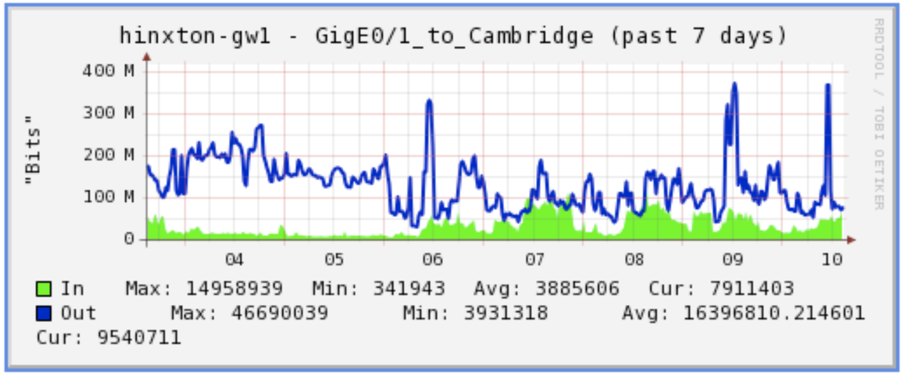
1000 Genomes Data moving through
outer



April Data Push



June Data Transfer



The last week

Acknowledgements

- EBI: Laura Clarke, Zam Iqbal, Guy Cochrane, Rasko Leinonen, Eugene Kulesha, Stephen Keenen
- NCBI: Martin Shumway, Hoda Khouri, Justin Paschall, Bob Sanders
- 1000 Genomes Data Flow group members

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

Data Flow Group (being formed)

Paul Flicek (co-chair) European Bioinforma
Stephen Sherry (co-chair) National Center
Ewan Birney European Bioinformatics Institu
Clive Brown Sanger Institute
David Dooling Washington University in St.
Richard Gibbs Baylor College of Medicine
Sol Katzman University of California, Sa
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)

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

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
[illegible] Kang Chinese Academy of Sciences
Yong Liu Shenzhen Local Municipal Government
Jean McEwen National Human Genome Research Institute
Jane Peterson National Human Genome Research Institute
[illegible]ine 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