Generation Gap: How existing bioinformatics resources are adapting to high-throughput sequencing

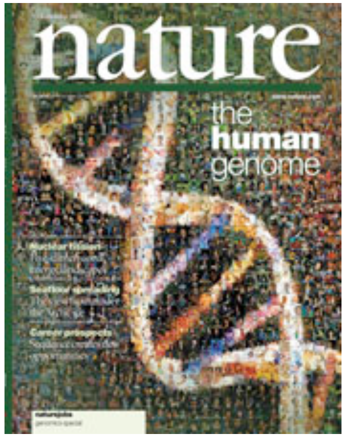
Paul Flicek
Vertebrate Genomics

EMBL-EBI

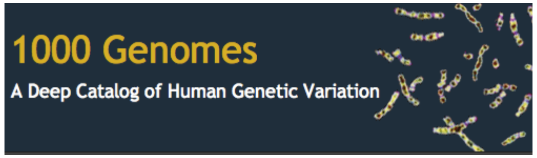
EBI is an Outstation of the European Molecular Biology Laboratory.

Further Evolution of Large-scale Genome Sequencing

- 2000: Human genome working drafts
- Data unit of approximately 10x coverage of human
  - 10 years and cost about $3 billion



- 2008: Major genome centers can sequence the same number of base pairs **every 4 days**
  - 1000 Genome project launched
  - World-wide capacity dramatically increasing



- 2009: **Every 4 hours ($25,000)**
- 2010: **Every 14 minutes ($5,000)**
  - Illumina HiSeq2000 machine produces 200 gigabases per 8 day run (BGI have 128)

illumina

Large-scale genome sequencing

- Today
  - 1000 Genomes, Cancer Genomes, exomes
  - Personal Genomes, Celebrity Genomes, Family Genomes
  - Others
- Soon
  - Thousands of cancer genomes
  - UK 10K
  - Diagnostic laboratories
  - Much, much more
- Results
  - Astronomical amounts of data
  - Catalogs of human variation and mutation

How is next generation sequencing data impacting major bioinformatics resources

- We have always attached a diverse community of users
  - From absolute beginners to ninjas
  - All need support
- Sequencing data is opening up new experiments and driving the transition to human as the model organisms
  - Variation data is the largest component of this change
- Multiple challenges
  - Data access for those who what big and small pieces
  - Annotation and management of the resulting discoveries
  - Your genome is unique and so is everyone else’s genome*

* Identical twins not included

1000 Genomes Project: Primary goals

- Overall: Create a deep catalogue of human variation to provide a better baseline to underpin human genetics

- Discover shared variation (shared = not private to individual) and characterise by allele frequency
    - Aim for effectively all (not just a lot of) common variation
        - For example: any variant down to 1% minor allele frequency in a population in the accessible genome has a 95% chance of being identified
        - The pilot project and simulations will help to determine the precision of this statement
    - Structural variants as well as SNPs
        - Accessible because the project will used paired-end sequencing reads
    - Deeper discovery in gene regions, down to 0.5% to 0.1% MAF

EMBL-EBI

1000 Genomes Project: Outcomes

- 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 data
  - Imputation with and from next generation sequencing data
- Produce an open resource building on HGP, HapMap etc.
  - A control set for sequencing disease samples

- All data publicly available, cell lines available
  - Anonymised samples without phenotypes

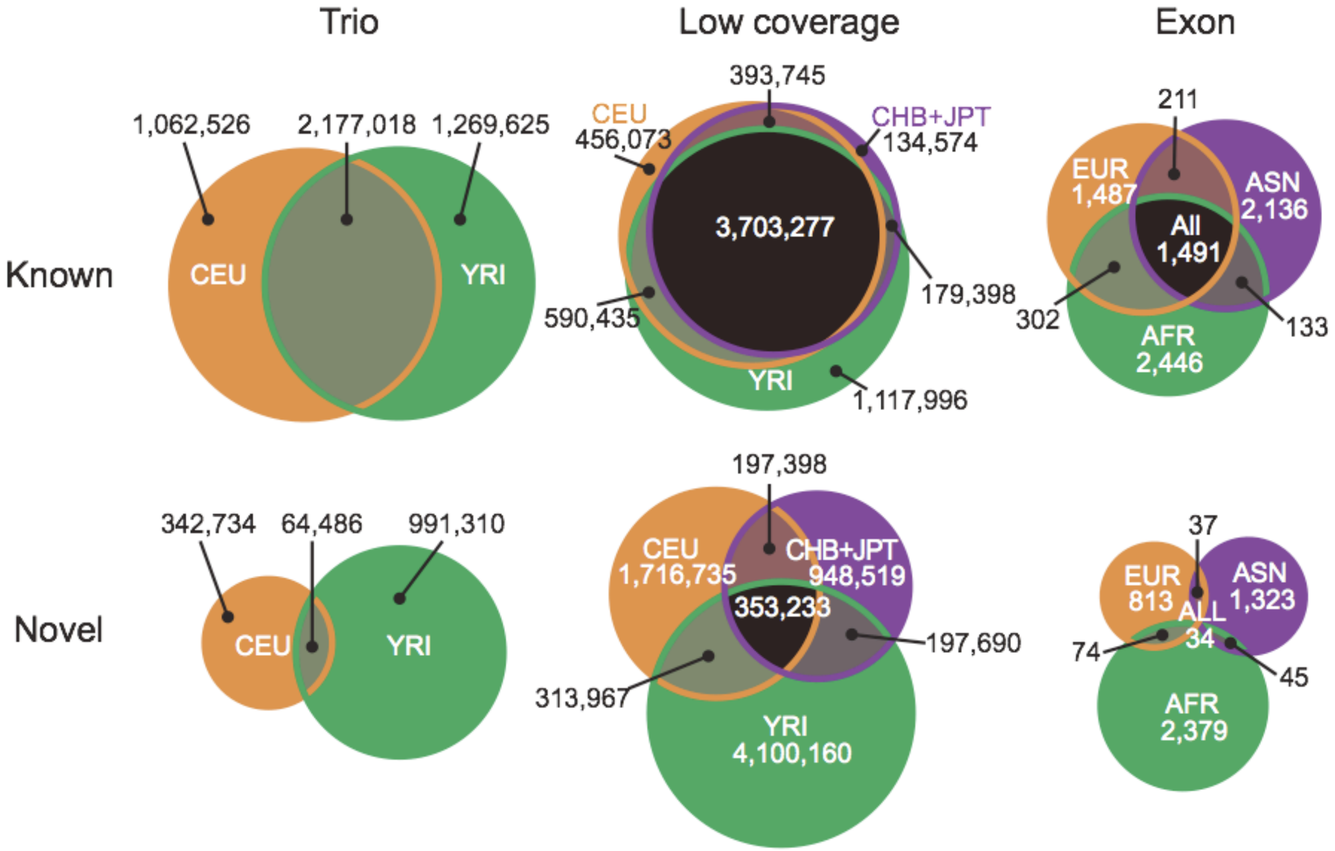
1000 Genomes Project Design and Progress

- Three pilot projects
  - Deep sequence two trios
  - Low coverage (~2X) 60 individuals from each of three populations (180 individuals total)
  - Gene capture for 1000 genes in about 700 individuals

- Pilot data collected in 2008; analysis now finished; paper now submitted to “a major journal”

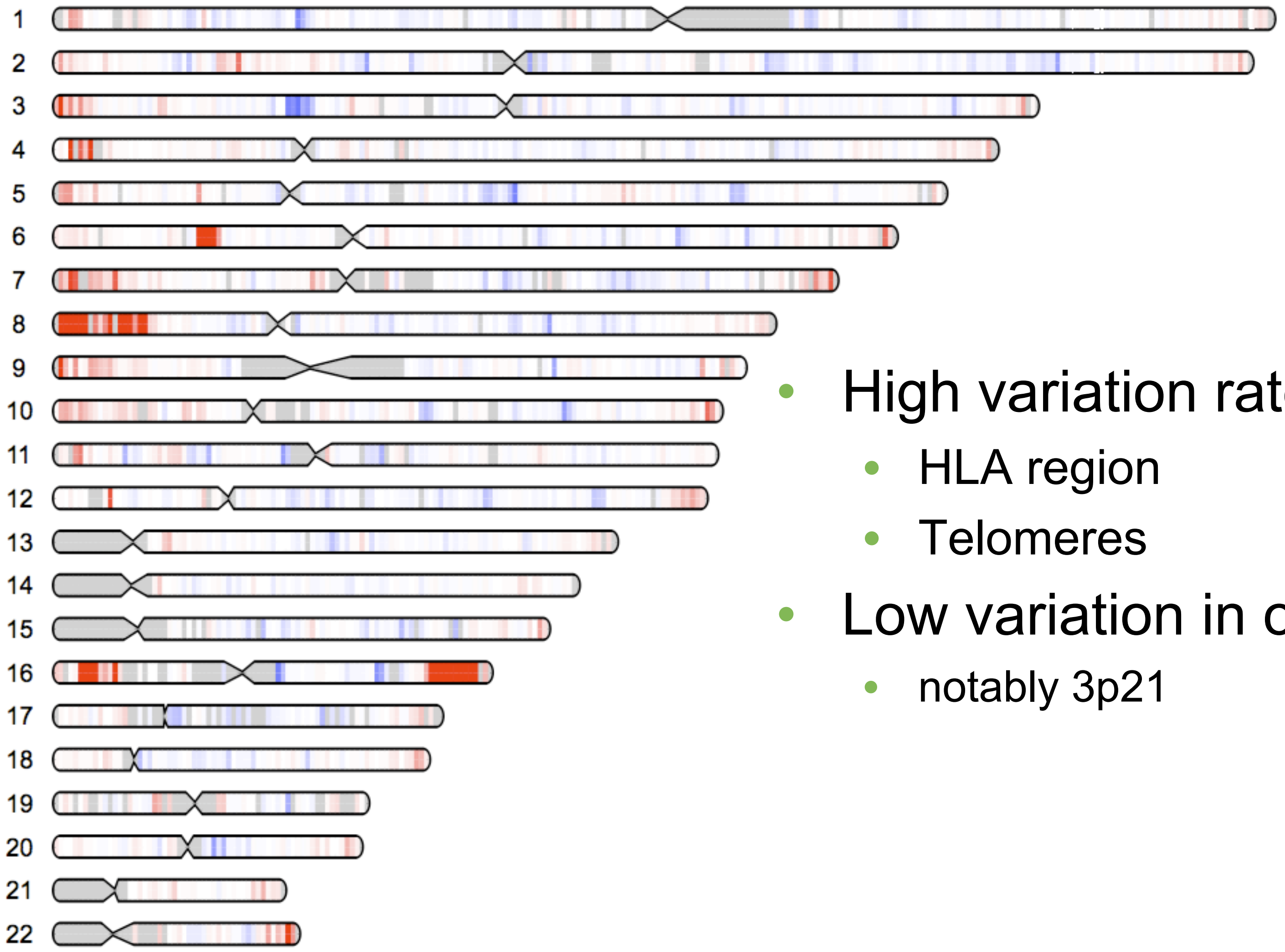
- Full project data collection and analysis underway

Pilot Project SNP Discovery



- 84% of novel SNPs to a single population
  - 4% in all populations
- FDR: <5% for SNPs and <10% for small indels

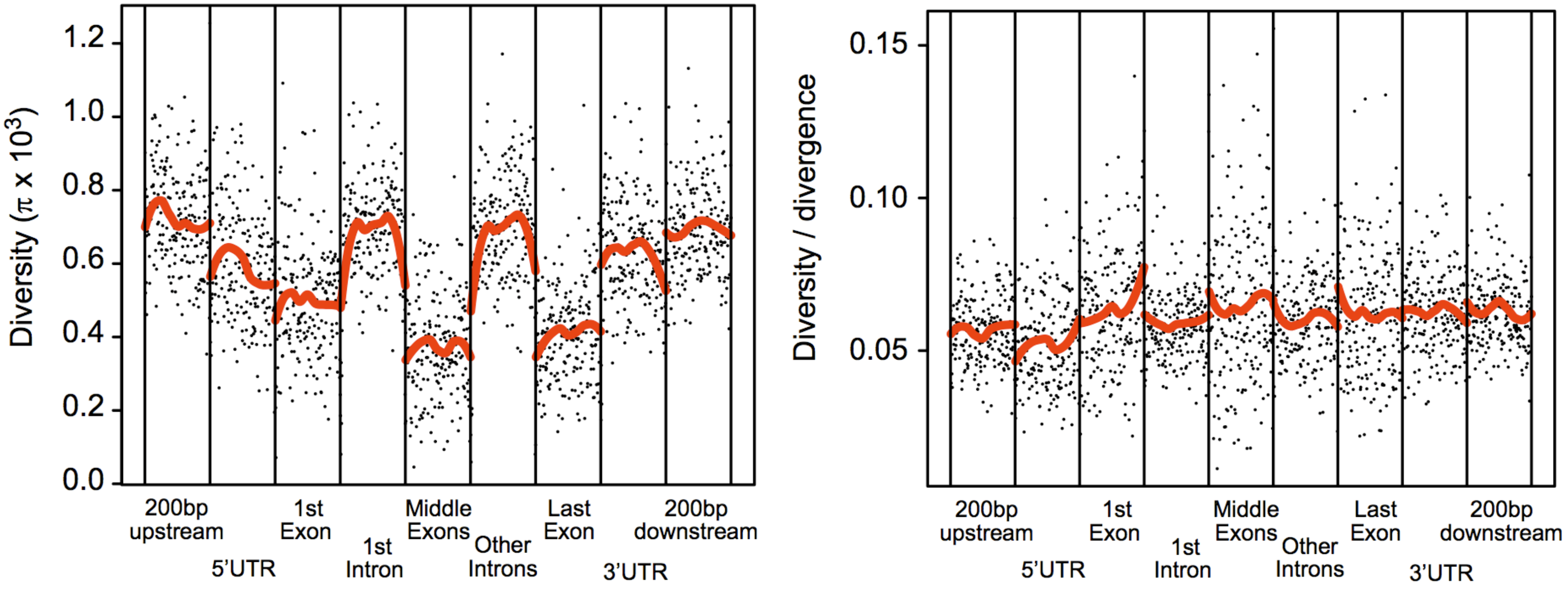
Genome-wide SNP distributions



- High variation rate
  - HLA region
  - Telomeres
- Low variation in other places
  - notably 3p21

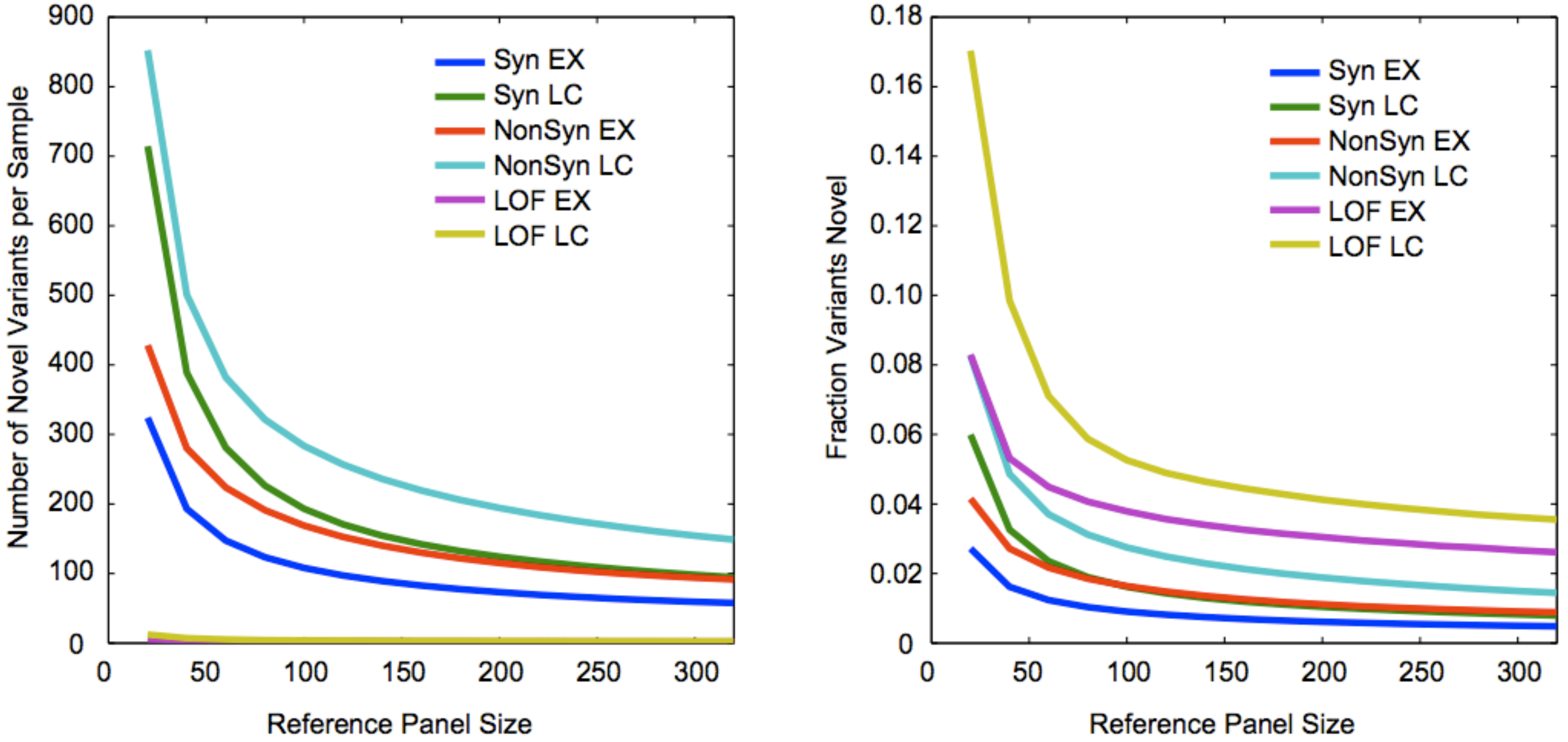
EMBL-EBI

Variation around genes



- Heterozygosity is lowest in middle exons
- Diversity is proportional to divergence
  - Functional constraint is the driving force of gene diversity

Value of additional samples for variant discovery comparing exon and low coverage



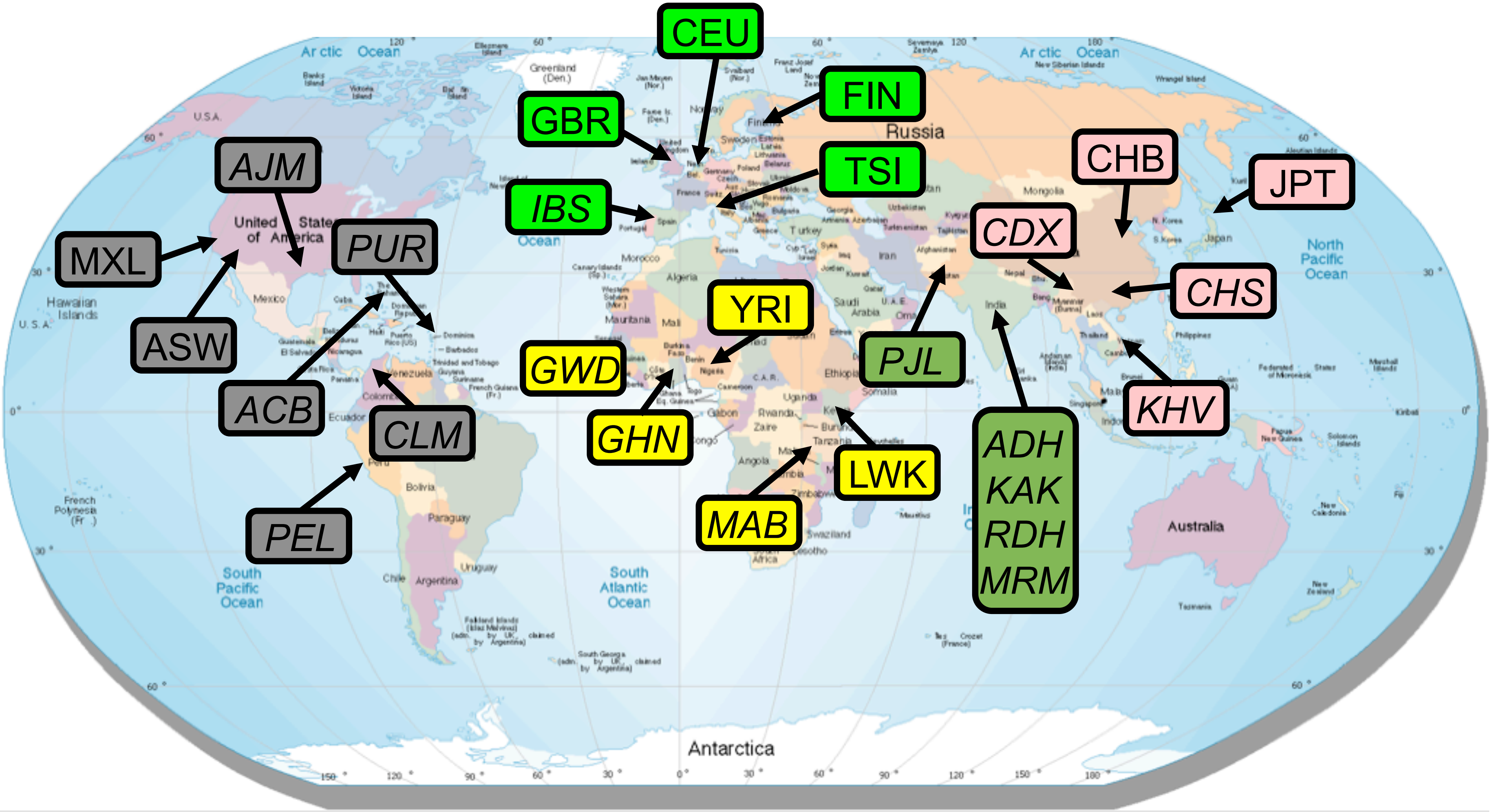
- 220 LC individuals to find 99% of synonymous variants
- 320 LC individuals for 98.5% of non-synonymous

1000 Genomes

Pilot project 180 samples

Extension to 1,100 samples summer 2010

1900 samples end 2010, 2500 samples end 2011



Major population groups comprised of subpopulations of ~100 each

1000 Genomes data by populations

Population	Sequence (gigabases)	Total Coverage
ASW	645	215x
CEU	2368	789x
CHB	1135	378x
CHS	168	56x
GBR	141	47x
JPT	1841	614x
LWK	1087	362x
MXL	216	72x
TSI	1257	419x
YRI	1534	511x

Total number of base pairs as of 11 June – 10.4 TB (12.5 TB including pilot projects)
Approximately 3500x total genome coverage

Putting this scale of data into perspective

- Size of EMBL/Genbank in April 2008 at the start of the 1000 Genomes Project: 235,135,312,328 nucleotides
- The 1000 Genomes project routinely produces the equivalent amount of sequence *every three days*
  - This is only a fraction of world-wide sequence capacity
- Data sizes in biology are now on the same order as those common in physics and astronomy

1KG Data storage infrastructure

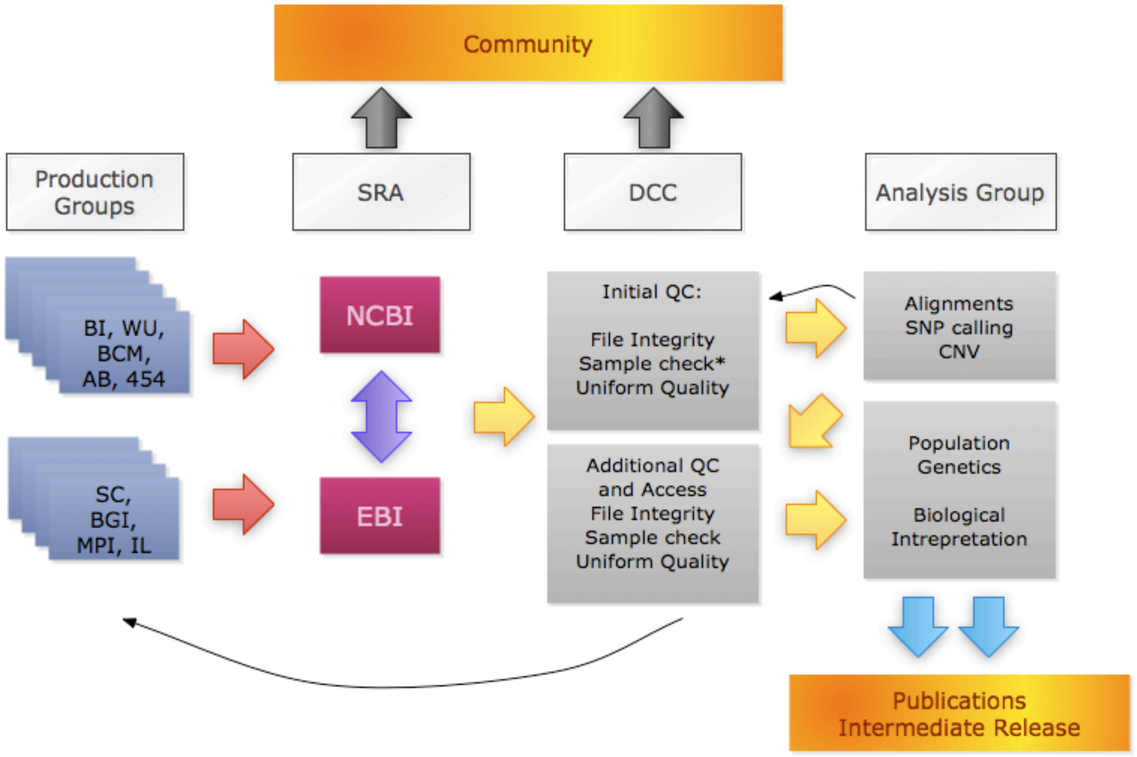
Circa April 2008

Today

Challenges

- Access
  - Data size, storage and transfer
  - Providing access to other researchers that want to use the data
- Annotation of variant data
  - Incorporating published and curated information
  - Integrating data that is collected on the genome index
- Most human research data cannot be openly released
  - How much diagnostic data should be released?
  - From research to clinical practice

1000 Genomes Project: Data Flow

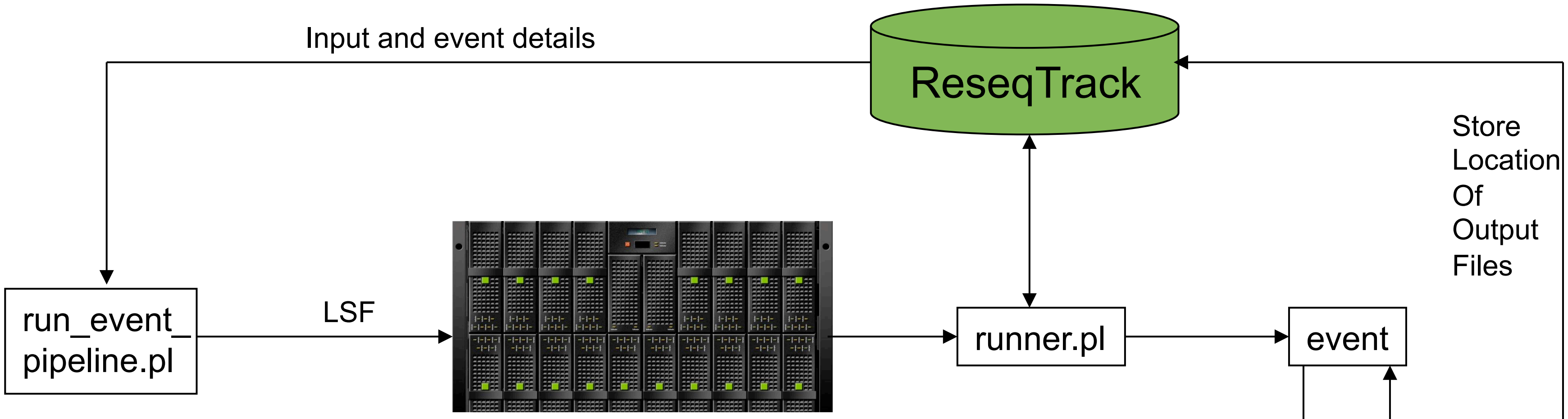


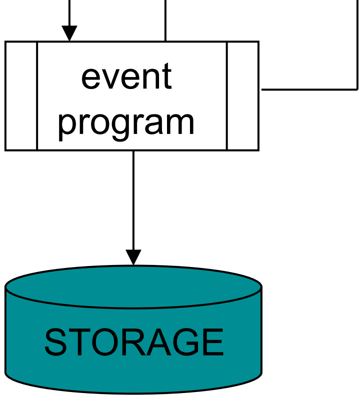
- Developed organically with many loops to relatively smooth system that takes data from sequencing machine to FTP site in about 1 month

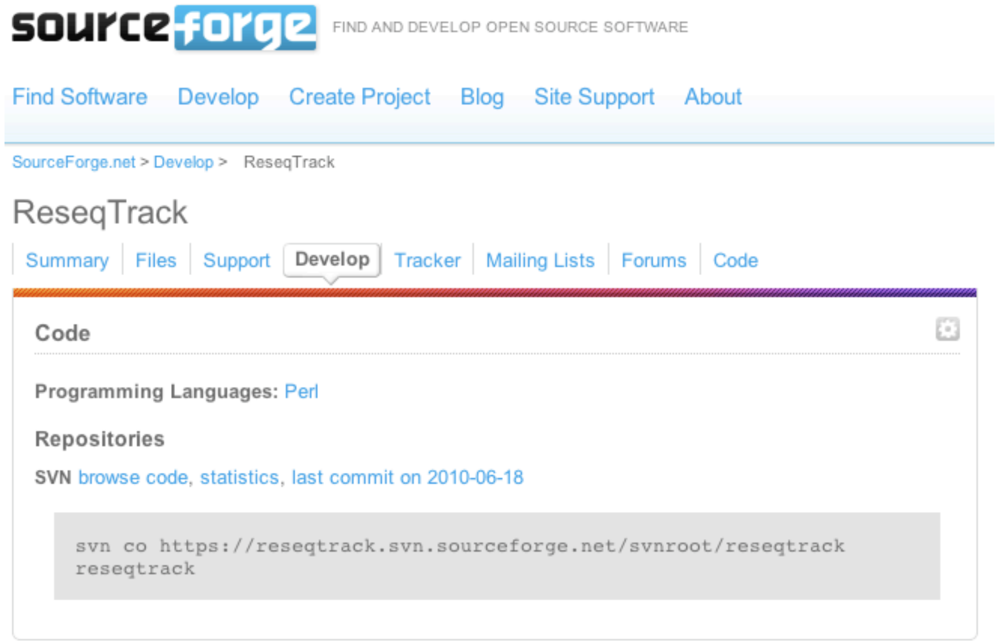
The 1000 Genomes data infrastructure

- Most aspects are running relatively smoothly
  - The pilot project produced about 100,000 sequence and other data files (there are now hundreds of thousands more)
  - Reseqtrack knows where the file is, what has been done to it, potential problems, related result files
- Accurate data transfer and bandwidth remain significant problems
  - File corruption during transfer is still relatively common
  - EBI bandwidth demands have increased about four fold over the course of the project
- The groups using this data are still mostly those within the 1000 Genomes project
  - The demand is growing beyond the project participants

ReseqTrack System: Pipeline overview







ReseqTrack

Code

Programming Languages: Perl

Repositories

SVN browse code, statistics, last commit on 2010-06-18

```
svn co https://reseqtrack.svn.sourceforge.net/svnroot/reseqtrack
reseqtrack
```

L. Clarke

EMBL-EBI

www.1000genomes.org

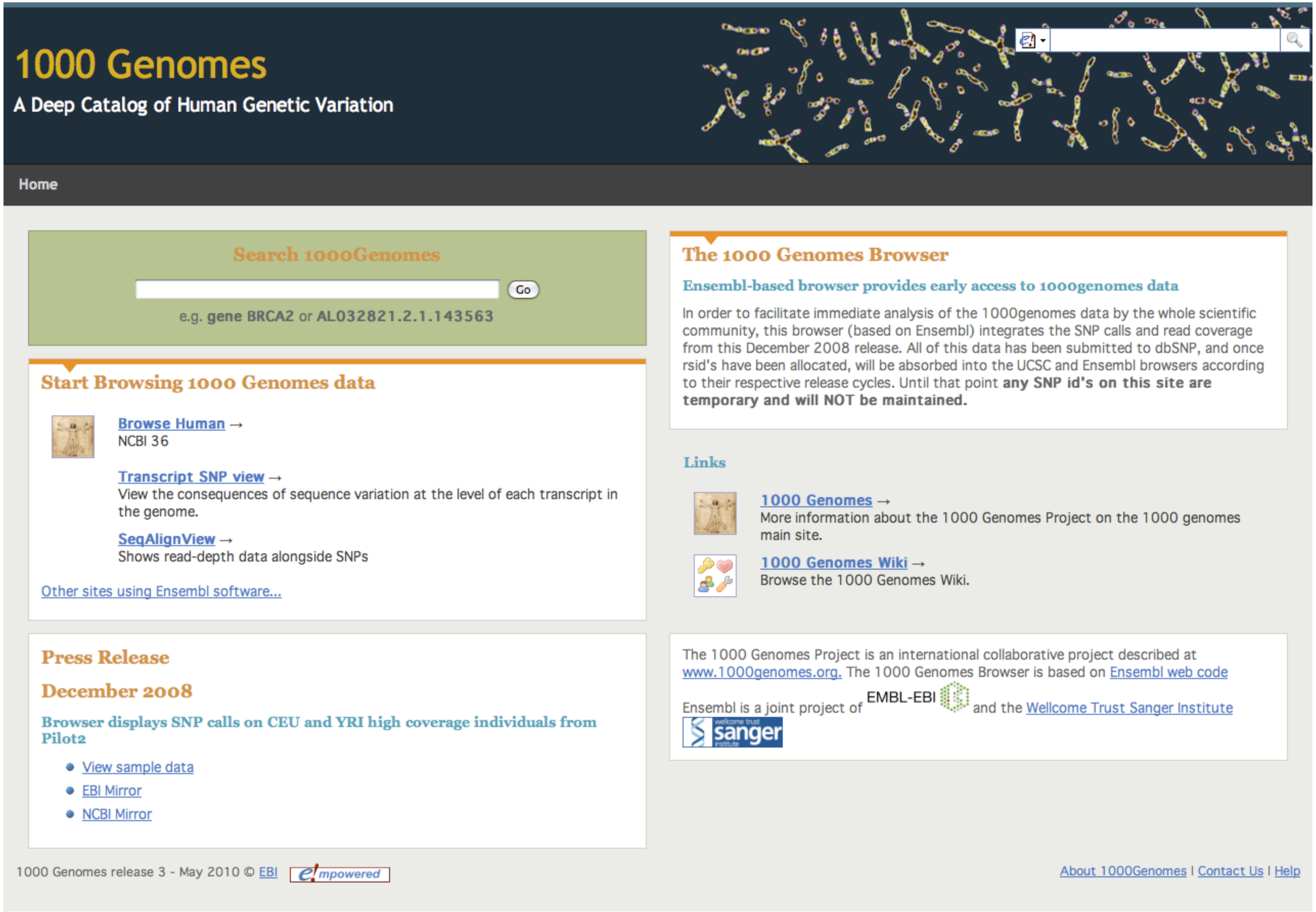
Project information with regular news updates

ftp-trace.ncbi.nih.gov/1000genomes/ftp

ftp.1000genomes.ebi.ac.uk

30-50 terabytes of data

Mostly in data formats that have just been invented
and almost no one has heard of or knows how to use



1000 Genomes

A Deep Catalog of Human Genetic Variation

Search 1000Genomes

e.g. gene BRCA2 or AL032821.2.1.143563

Start Browsing 1000 Genomes data

Browse Human →
NCBI 36

Transcript SNP view →
View the consequences of sequence variation at the level of each transcript in the genome.

SeqAlignView →
Shows read-depth data alongside SNPs

Other sites using Ensembl software...

Press Release

December 2008

Browser displays SNP calls on CEU and YRI high coverage individuals from Pilot2

- View sample data
- EBI Mirror
- NCBI Mirror

The 1000 Genomes Browser

Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000genomes data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls and read coverage from this December 2008 release. All of this data has been submitted to dbSNP, and once rsid's have been allocated, will be absorbed into the UCSC and Ensembl browsers according to their respective release cycles. Until that point **any SNP id's on this site are temporary and will NOT be maintained.**

Links

1000 Genomes →
More information about the 1000 Genomes Project on the 1000 genomes main site.

1000 Genomes Wiki →
Browse the 1000 Genomes Wiki.

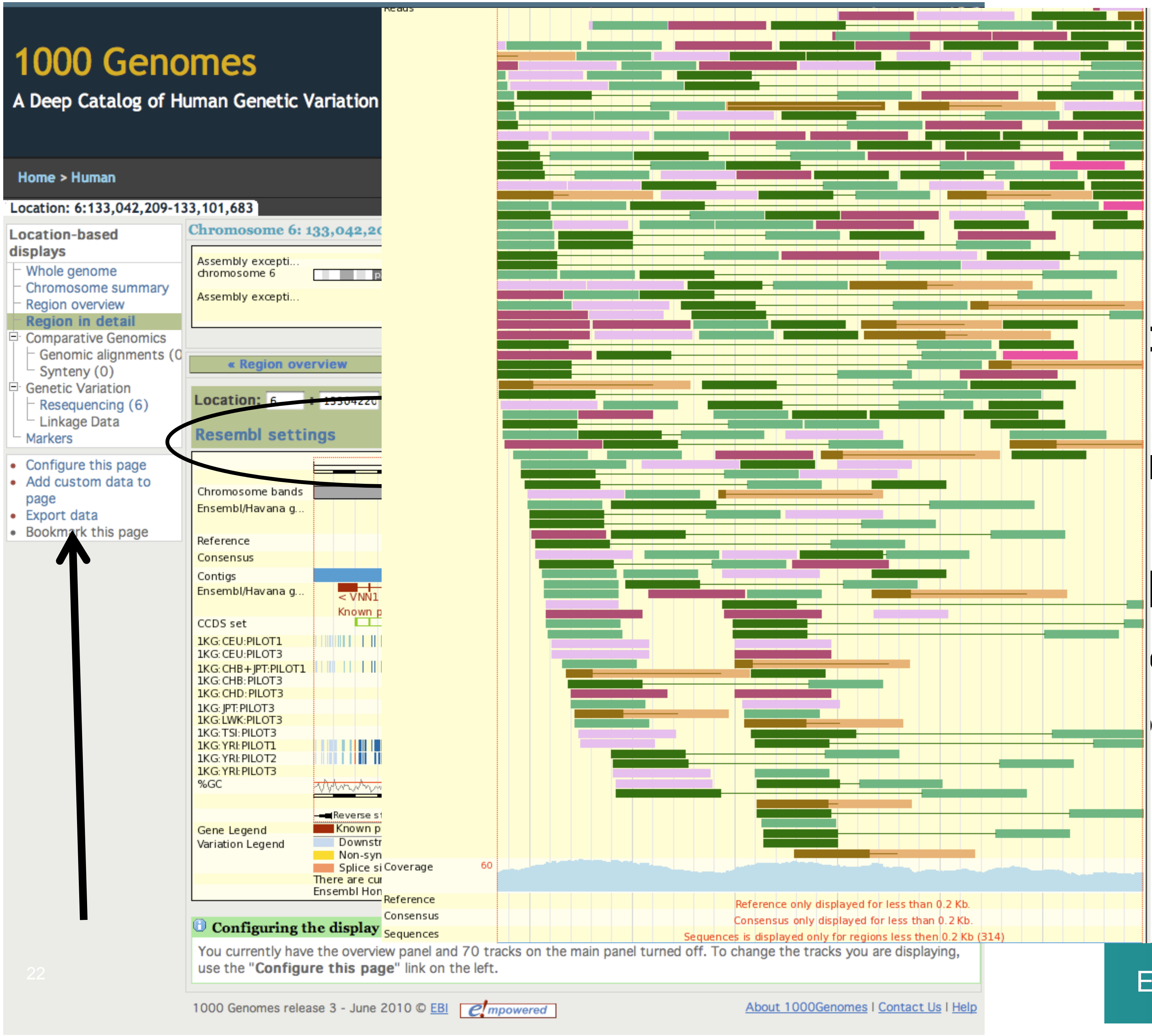
The 1000 Genomes Project is an international collaborative project described at www.1000genomes.org. The 1000 Genomes Browser is based on Ensembl web code

Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute

1000 Genomes release 3 - May 2010 © EBI

About 1000Genomes | Contact Us | Help

1000 Genomes Browser Home Page
http://browser.1000genomes.org



Reference only displayed for less than 0.2 Kb.

Consensus only displayed for less than 0.2 Kb.

Sequences is displayed only for regions less then 0.2 Kb (314)

You currently have the overview panel and 70 tracks on the main panel turned off. To change the tracks you are displaying, use the "Configure this page" link on the left.

1000 Genomes release 3 - June 2010 © EBI e!mpowered

ation

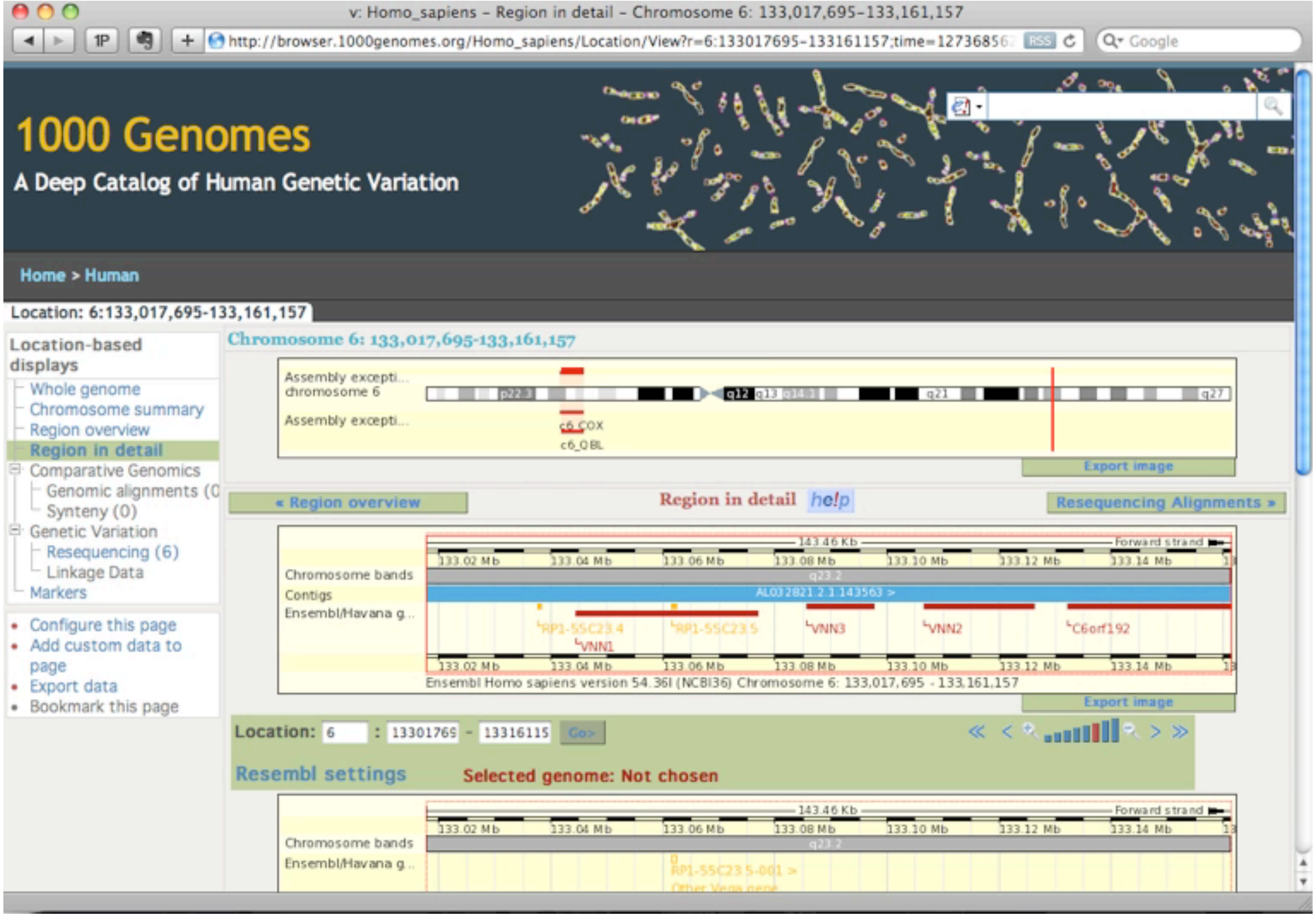
ntext

nes

P tracks

o

pulation



v: Homo_sapiens – Region in detail – Chromosome 6: 133,017,695–133,161,157

http://browser.1000genomes.org/Homo_sapiens/Location/View?r=6:133017695-133161157;time=12736856

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home > Human

Location: 6:133,017,695-133,161,157

Location-based displays
- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Comparative Genomics
  - Genomic alignments (0
  - Synteny (0)
- Genetic Variation
  - Resequencing (6)
  - Linkage Data
- Markers

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Chromosome 6: 133,017,695-133,161,157

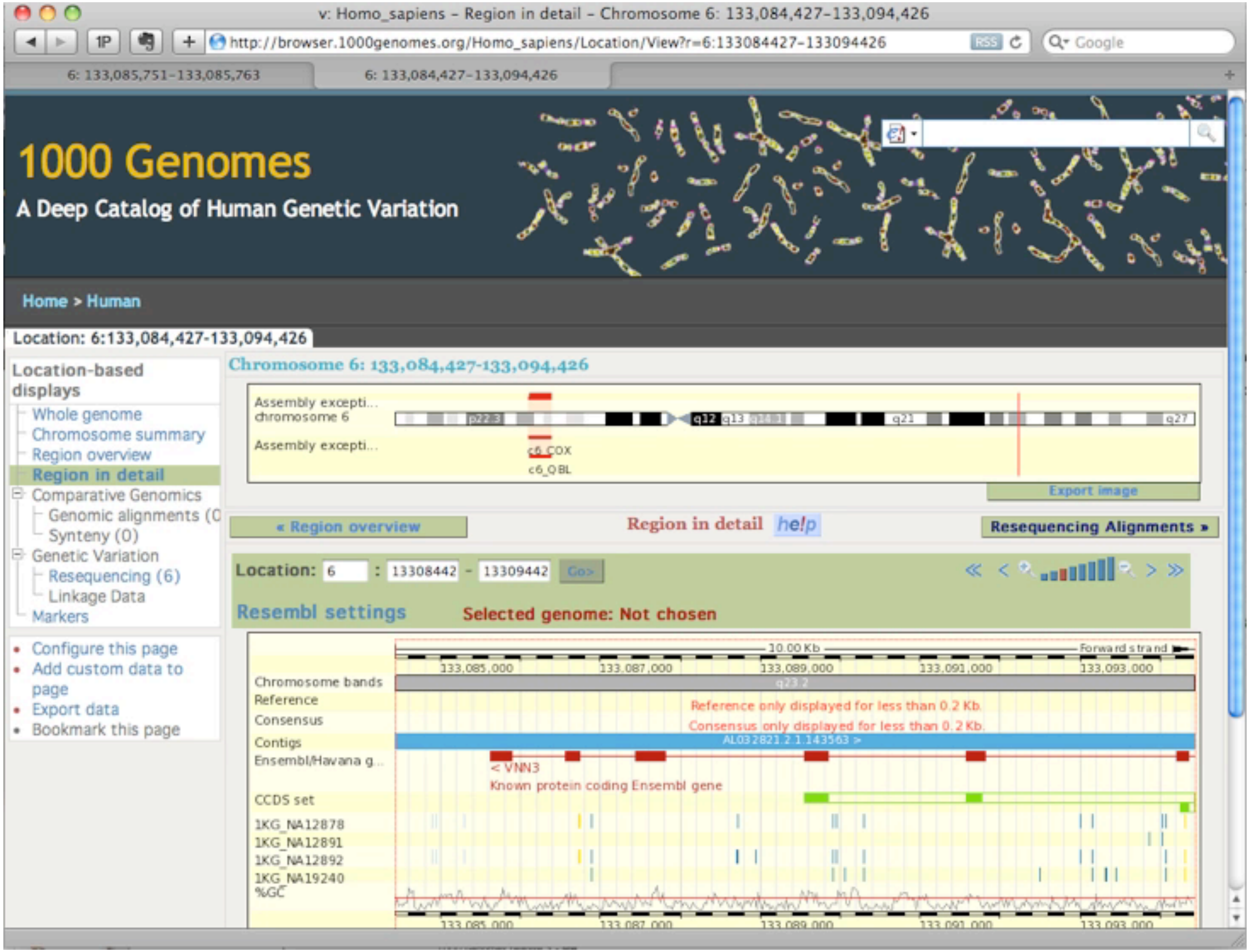
Region in detail

Ensembl Homo sapiens version 54.36l (NCBI36) Chromosome 6: 133,017,695 - 133,161,157

Location: 6 : 13301769 - 13316115 Go>

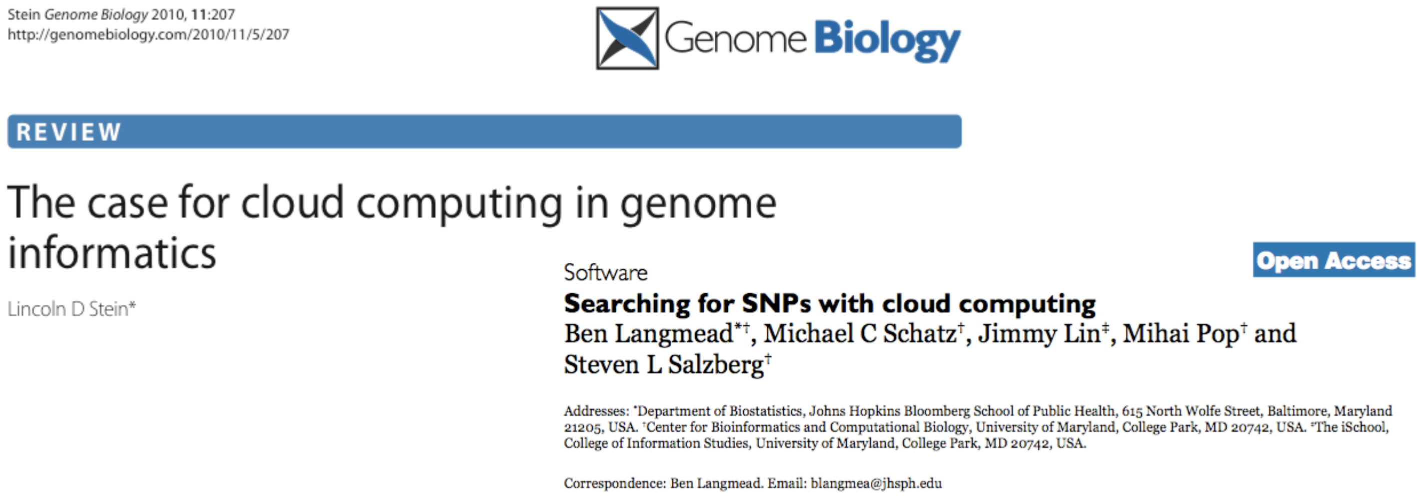
Resembl settings

Selected genome: Not chosen


v: Homo_sapiens - Region in detail - Chromosome 6: 133,084,427-133,094,426
http://browser.1000genomes.org/Homo_sapiens/Location/View?r=6:133084427-133094426
Google
6: 133,085,751-133,085,763
6: 133,084,427-133,094,426
1000 Genomes
A Deep Catalog of Human Genetic Variation
Home > Human
Location: 6:133,084,427-133,094,426
Location-based displays
Whole genome
Chromosome summary
Region overview
Region in detail
Comparative Genomics
Genomic alignments (0
Synteny (0)
Genetic Variation
Resequencing (6)
Linkage Data
Markers
Configure this page
Add custom data to page
Export data
Bookmark this page
Chromosome 6: 133,084,427-133,094,426
Assembly excepti...
chromosome 6
Assembly excepti...
p22.3
q12
q13
q14.1
q21
q27
c6_COX
c6_QBL
Export image
« Region overview
Region in detail
help
Resequencing Alignments »
Location: 6 : 13308442 - 13309442 Go>
Resembl settings
Selected genome: Not chosen
10.00 Kb
Forward strand
133,085,000
133,087,000
133,089,000
133,091,000
133,093,000
Chromosome bands
q23.2
Reference
Reference only displayed for less than 0.2 Kb.
Consensus
Consensus only displayed for less than 0.2 Kb.
Contigs
AL032821.2.1.143563 >
Ensembl/Havana g...
< VNN3
Known protein coding Ensembl gene
CCDS set
1KG_NA12878
1KG_NA12891
1KG_NA12892
1KG_NA19240
%GC
133,085,000
133,087,000
133,089,000
133,091,000
133,093,000

Expanding data availability with the cloud

- Amazon Web Services
  - The final 1000 Genomes Pilot alignment files (BAMs) are now loaded into the Amazon EC2 cloud and have been formally announced last week
  - Some files had to be split to accommodate the 5 Gb max file size of the S3 storage
- Anyone can use the data with standard AWS costs per computer hour (as low as 8.5¢ per CPU hour)
- We will be developing publicly accessible applications within the cloud environment and expect that others will as well



Stein *Genome Biology* 2010, 11:207
http://genomebiology.com/2010/11/5/207

Genome **Biology**

REVIEW

The case for cloud computing in genome informatics

Lincoln D Stein*

Software

Open Access

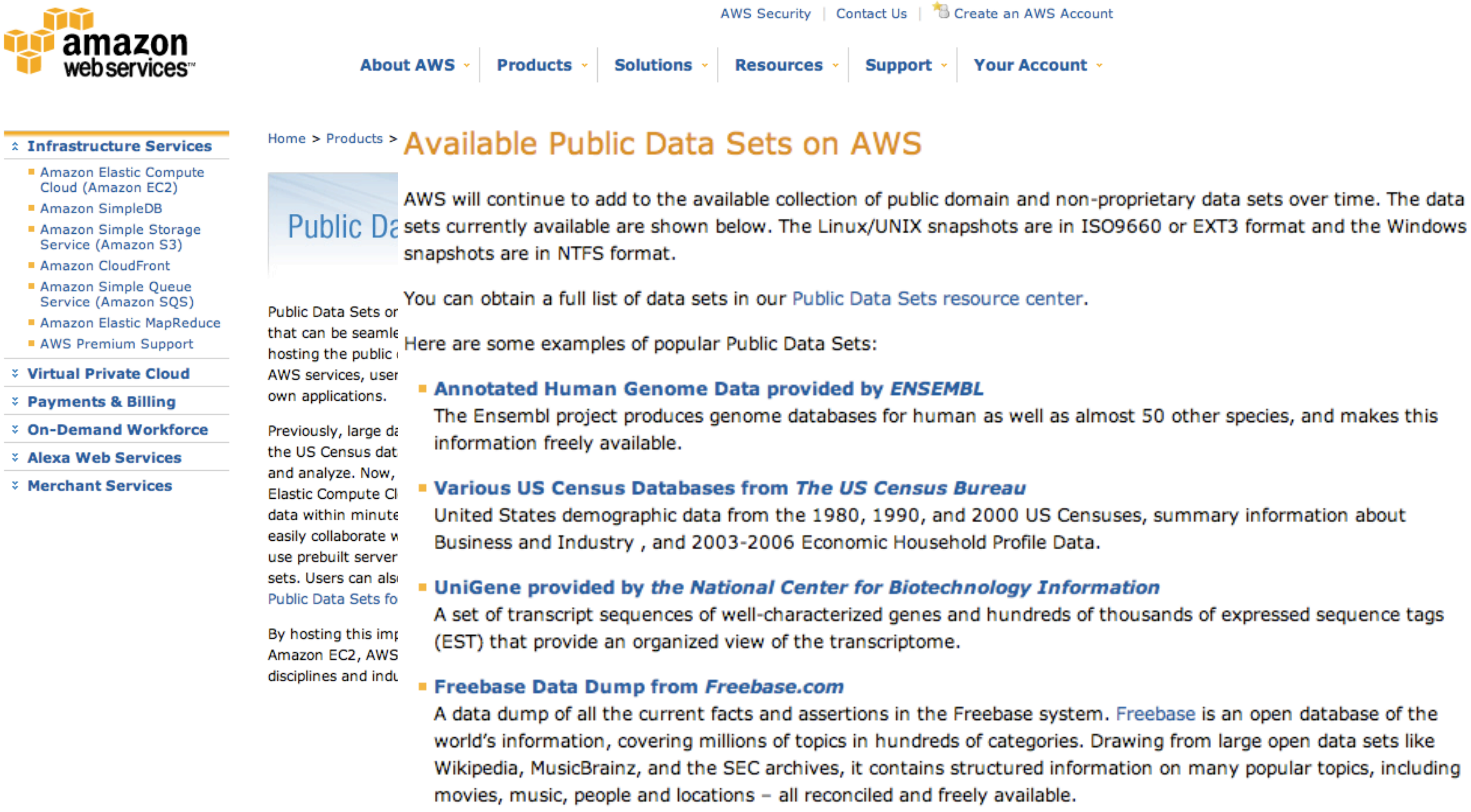
Searching for SNPs with cloud computing

Ben Langmead*†, Michael C Schatz†, Jimmy Lin‡, Mihai Pop† and Steven L Salzberg†

Addresses: *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA. †Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA. ‡The iSchool, College of Information Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: blangmea@jhsph.edu

AWS Public Data Sets



Public Data Sets or that can be seamle hosting the public AWS services, user own applications.

Previously, large da the US Census dat and analyze. Now, Elastic Compute Cl data within minute easily collaborate w use prebuilt server sets. Users can als Public Data Sets fo

By hosting this imp Amazon EC2, AWS disciplines and indu

Available Public Data Sets on AWS

AWS will continue to add to the available collection of public domain and non-proprietary data sets over time. The data sets currently available are shown below. The Linux/UNIX snapshots are in ISO9660 or EXT3 format and the Windows snapshots are in NTFS format.

You can obtain a full list of data sets in our Public Data Sets resource center.

Here are some examples of popular Public Data Sets:

- **Annotated Human Genome Data provided by *ENSEMBL***
  The Ensembl project produces genome databases for human as well as almost 50 other species, and makes this information freely available.
- **Various US Census Databases from *The US Census Bureau***
  United States demographic data from the 1980, 1990, and 2000 US Censuses, summary information about Business and Industry , and 2003-2006 Economic Household Profile Data.
- **UniGene provided by *the National Center for Biotechnology Information***
  A set of transcript sequences of well-characterized genes and hundreds of thousands of expressed sequence tags (EST) that provide an organized view of the transcriptome.
- **Freebase Data Dump from *Freebase.com***
  A data dump of all the current facts and assertions in the Freebase system. Freebase is an open database of the world's information, covering millions of topics in hundreds of categories. Drawing from large open data sets like Wikipedia, MusicBrainz, and the SEC archives, it contains structured information on many popular topics, including movies, music, people and locations – all reconciled and freely available.

We have just launched a complete Ensembl genome browser mirror within EC2 (http://useast.ensembl.org).

Challenges

- Access
  - Data size, storage and transfer
  - Providing access to other researchers that want to use the data
- Annotation of variant data
  - Incorporating published and curated information
  - Integrating data that is collected on the genome index
- Most human research data cannot be openly released
  - How much diagnostic data should be released?
  - From research to clinical practice

Ensembl

- Ensembl’s mission is to enable genomic science by providing high-quality, integrated annotation on vertebrate genomes within a consistent and accessible infrastructure.

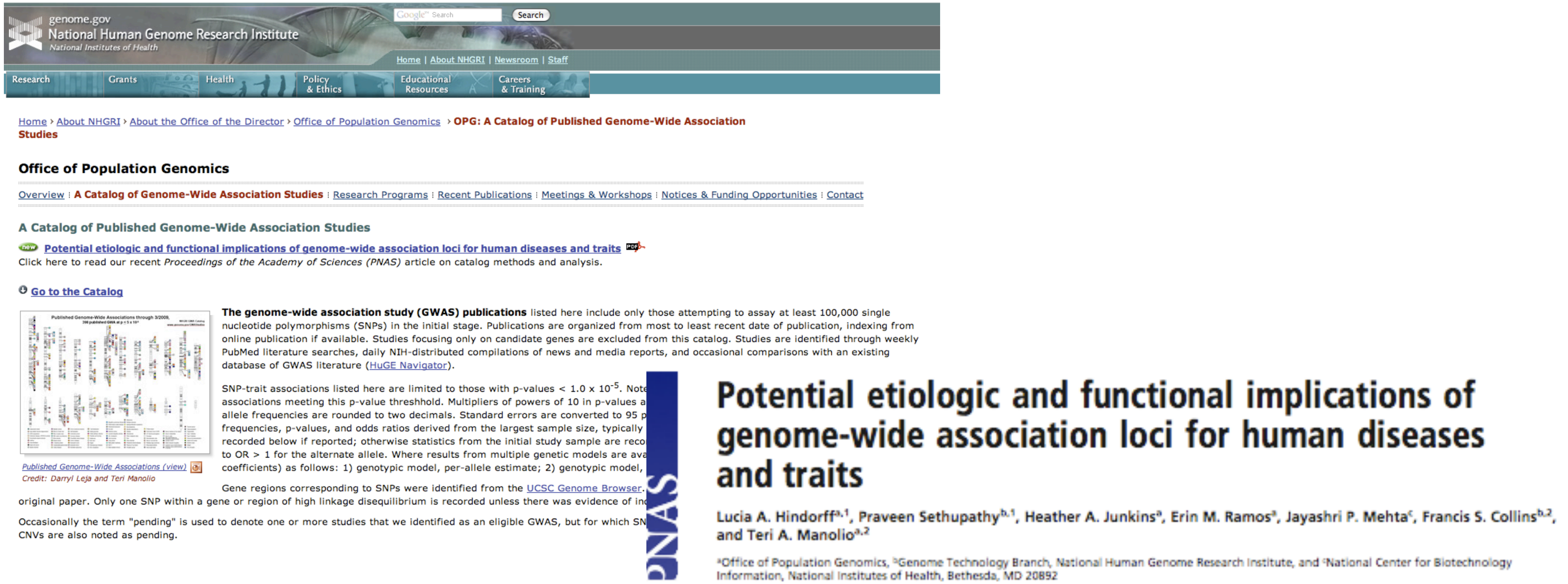
- Creating and providing core value-added data sets
  - High-quality evidence-based gene sets
  - Multiple alignments
  - Gene homology and paralogy relationships
  - Genome variation including SNPs, genotypes and CNV/SV data
  - Integrative analysis of genome regulation

- Roadmap includes extensive support for data on multiple individuals
  - Human cell lines, mouse strains
  - Favouring integrated information

EMBL-EBI

Phenotype annotation - Genomic

- Genome wide association study data on 672 phenotypes
- Currently over 60,000 phenotype annotations
  - All high-quality, curated and publication based
- Data is growing with every Ensembl release



genome.gov
National Human Genome Research Institute
National Institutes of Health

Home › About NHGRI › About the Office of the Director › Office of Population Genomics › **OPG: A Catalog of Published Genome-Wide Association Studies**

Office of Population Genomics

Overview : **A Catalog of Genome-Wide Association Studies** : Research Programs : Recent Publications : Meetings & Workshops : Notices & Funding Opportunities : Contact

A Catalog of Published Genome-Wide Association Studies

new **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits** PDF
Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

Go to the Catalog

Published Genome-Wide Associations through 3/2009, 398 published GWA at p ≤ 5 x 10⁻⁸

Published Genome-Wide Associations (view)
Credit: Darryl Leja and Teri Manolio

The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database of GWAS literature (HuGE Navigator).

SNP-trait associations listed here are limited to those with p-values < 1.0×10^{-5}. Note
associations meeting this p-value threshhold. Multipliers of powers of 10 in p-values a
allele frequencies are rounded to two decimals. Standard errors are converted to 95 p
frequencies, p-values, and odds ratios derived from the largest sample size, typically
recorded below if reported; otherwise statistics from the initial study sample are reco
to OR > 1 for the alternate allele. Where results from multiple genetic models are ava
coefficients) as follows: 1) genotypic model, per-allele estimate; 2) genotypic model,

Gene regions corresponding to SNPs were identified from the UCSC Genome Browser.
original paper. Only one SNP within a gene or region of high linkage disequilibrium is recorded unless there was evidence of in

Occasionally the term "pending" is used to denote one or more studies that we identified as an eligible GWAS, but for which SN
CNVs are also noted as pending.

PNAS

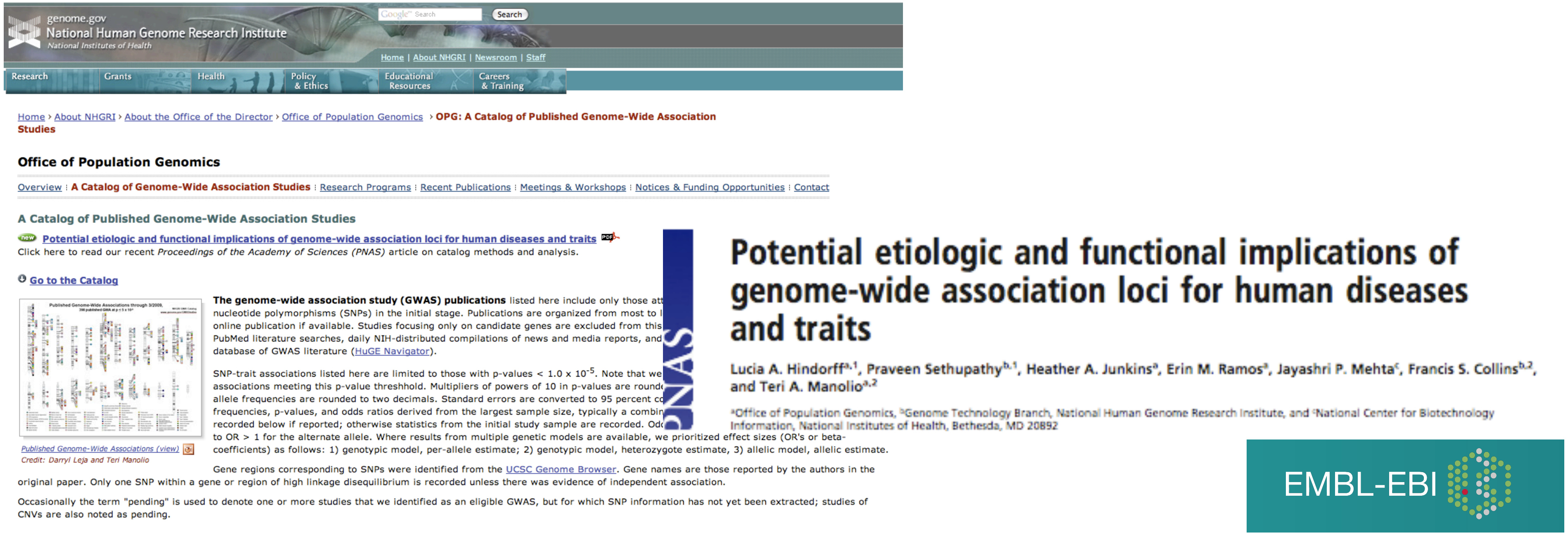
Potential etiologic and functional implications of genome-wide association loci for human diseases and traits

Lucia A. Hindorff[a,1], Praveen Sethupathy[b,1], Heather A. Junkins[a], Erin M. Ramos[a], Jayashri P. Mehta[c], Francis S. Collins[b,2], and Teri A. Manolio[a,2]

[a]Office of Population Genomics, [b]Genome Technology Branch, National Human Genome Research Institute, and [c]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20892

Annotation of the variation catalog in Ensembl

- Incorporated variation annotations representing 134 distinct phenotypes
- Reference 186 publications
- Currently 1120 variations with annotated information
  - All high-quality, curated and publication based
- Data is growing with every Ensembl release


genome.gov
National Human Genome Research Institute
National Institutes of Health
Google Search
Search
Home | About NHGRI | Newsroom | Staff
Research
Grants
Health
Policy & Ethics
Educational Resources
Careers & Training

Home › About NHGRI › About the Office of the Director › Office of Population Genomics › OPG: A Catalog of Published Genome-Wide Association Studies

Office of Population Genomics

Overview ¦ **A Catalog of Genome-Wide Association Studies** ¦ Research Programs ¦ Recent Publications ¦ Meetings & Workshops ¦ Notices & Funding Opportunities ¦ Contact

A Catalog of Published Genome-Wide Association Studies

new **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits** PDF

Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

Go to the Catalog

Published Genome-Wide Associations (view)

Credit: Darryl Leja and Teri Manolio

The genome-wide association study (GWAS) publications listed here include only those att nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to l online publication if available. Studies focusing only on candidate genes are excluded from this PubMed literature searches, daily NIH-distributed compilations of news and media reports, and database of GWAS literature (HuGE Navigator).

SNP-trait associations listed here are limited to those with p-values < 1.0 x 10^{-5}. Note that we associations meeting this p-value threshhold. Multipliers of powers of 10 in p-values are rounde allele frequencies are rounded to two decimals. Standard errors are converted to 95 percent co frequencies, p-values, and odds ratios derived from the largest sample size, typically a combin recorded below if reported; otherwise statistics from the initial study sample are recorded. Odd to OR > 1 for the alternate allele. Where results from multiple genetic models are available, we prioritized effect sizes (OR's or beta-coefficients) as follows: 1) genotypic model, per-allele estimate; 2) genotypic model, heterozygote estimate, 3) allelic model, allelic estimate.

Gene regions corresponding to SNPs were identified from the UCSC Genome Browser. Gene names are those reported by the authors in the original paper. Only one SNP within a gene or region of high linkage disequilibrium is recorded unless there was evidence of independent association.

Occasionally the term "pending" is used to denote one or more studies that we identified as an eligible GWAS, but for which SNP information has not yet been extracted; studies of CNVs are also noted as pending.

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits

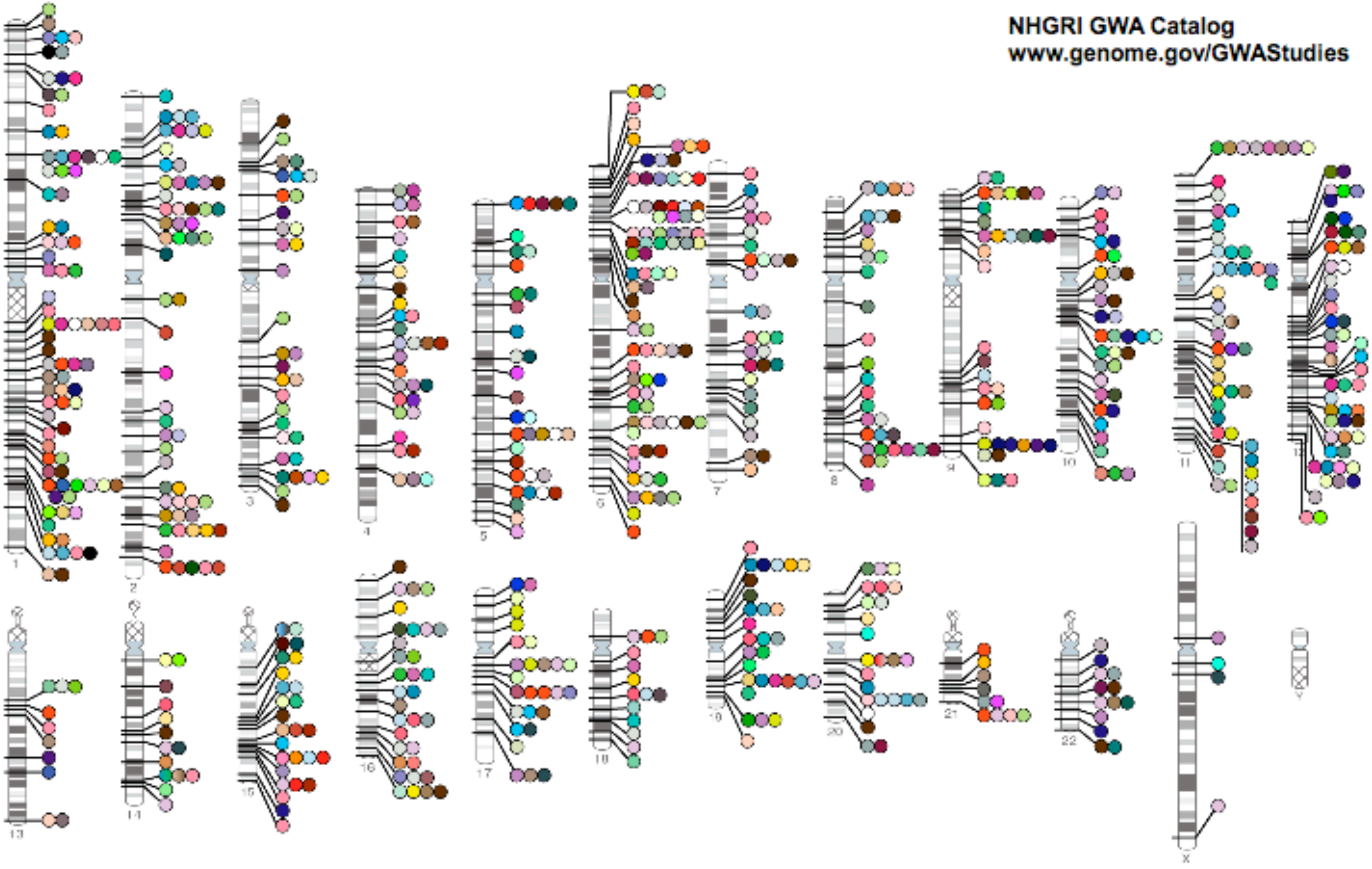
Lucia A. Hindorff[a,1], Praveen Sethupathy[b,1], Heather A. Junkins[a], Erin M. Ramos[a], Jayashri P. Mehta[c], Francis S. Collins[b,2], and Teri A. Manolio[a,2]

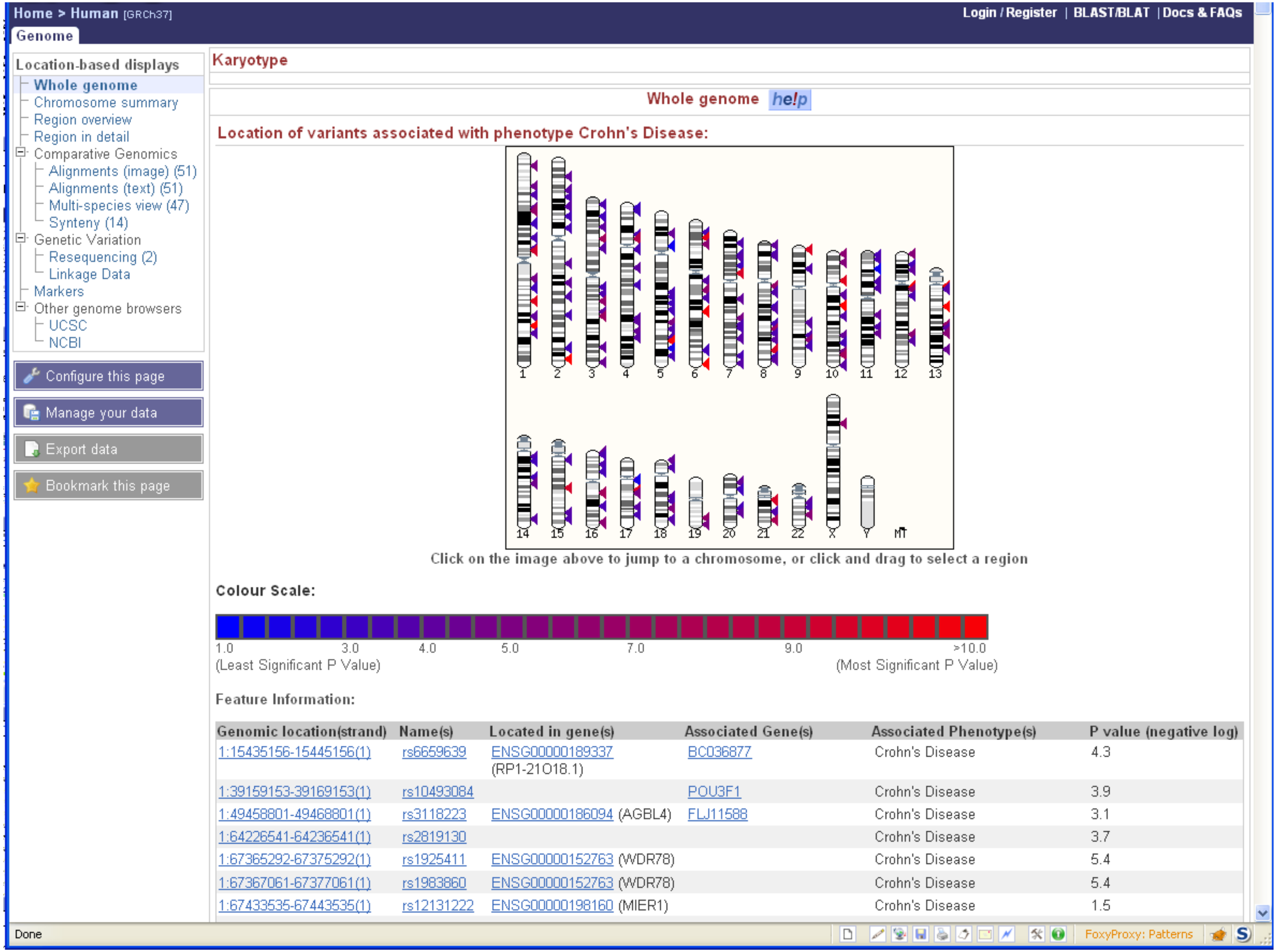
[a]Office of Population Genomics, [b]Genome Technology Branch, National Human Genome Research Institute, and [c]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20892

EMBL-EBI

Published Genome-Wide Associations through 3/2010, 779 published GWA at $p \leq 5 \times 10^{-8}$ for 148 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies





Location of variants associated with phenotype Crohn's Disease:

Click on the image above to jump to a chromosome, or click and drag to select a region

Colour Scale:

Feature Information:

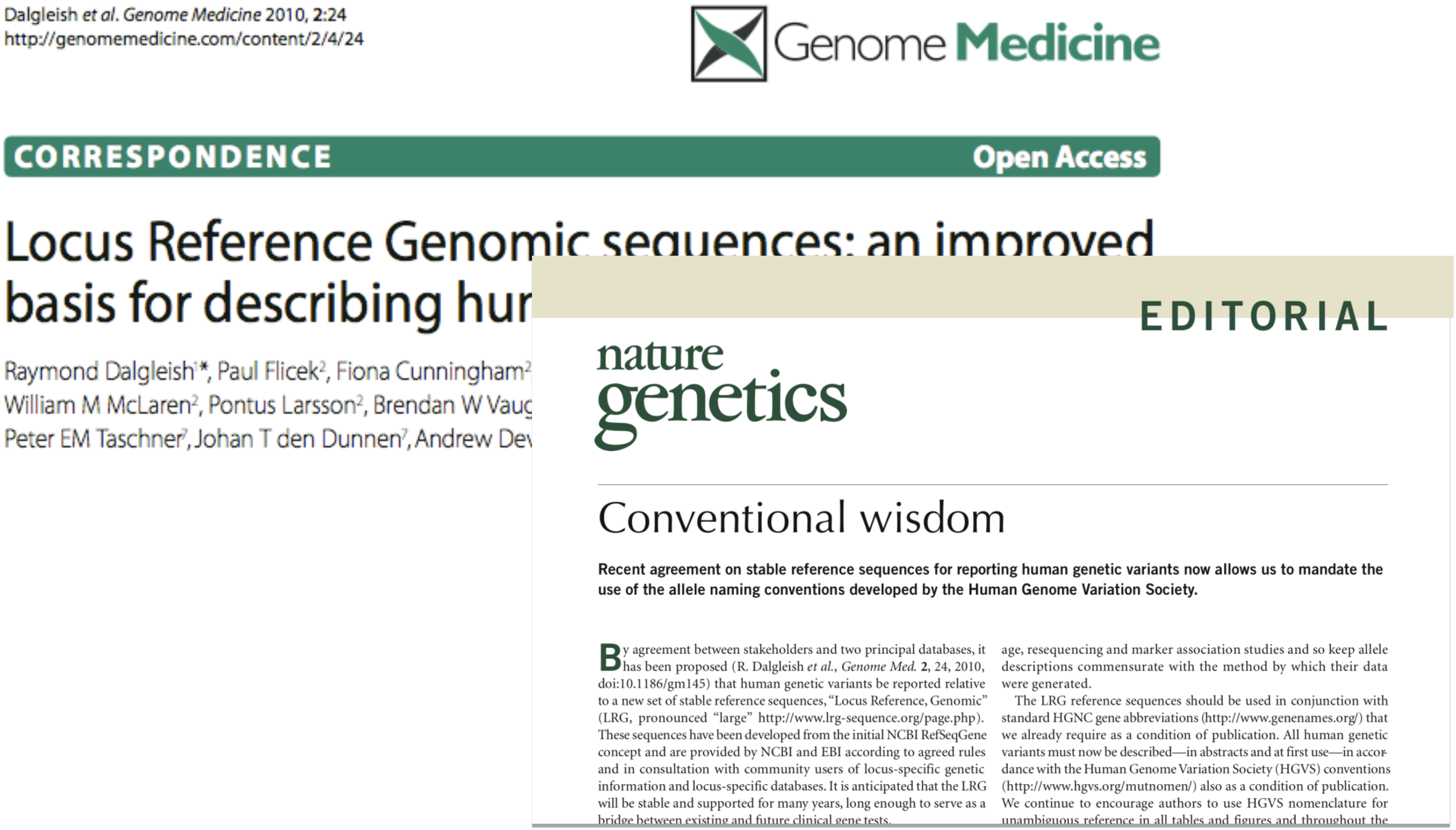
Genomic location(strand)	Name(s)	Located in gene(s)	Associated Gene(s)	Associated Phenotype(s)	P value (negative log)
1:15435156-15445156(1)	rs6659639	ENSG00000189337 (RP1-21O18.1)	BC036877	Crohn's Disease	4.3
1:39159153-39169153(1)	rs10493084		POU3F1	Crohn's Disease	3.9
1:49458801-49468801(1)	rs3118223	ENSG00000186094 (AGBL4)	FLJ11588	Crohn's Disease	3.1
1:64226541-64236541(1)	rs2819130			Crohn's Disease	3.7
1:67365292-67375292(1)	rs1925411	ENSG00000152763 (WDR78)		Crohn's Disease	5.4
1:67367061-67377061(1)	rs1983860	ENSG00000152763 (WDR78)		Crohn's Disease	5.4
1:67433535-67443535(1)	rs12131222	ENSG00000198160 (MIER1)		Crohn's Disease	1.5

Phenotype annotation - Gene-based

- Over 1400 LSDBs on the Human Genome Variation Society website
  - Data integration with central resources has been challenging

- Locus Reference Genomic Sequences
  - An informatics solution
    - Stable, community-determined sequence
    - Collaboration with the NCBI & Gen2phen
  - Extension and generalisation of NCBI’s RefSeqGene project
  - http://www.lrg-sequence.org

Dalgleish, et al. *Genome Medicine* 2010, 2:24



Dalgleish *et al. Genome Medicine* 2010, **2**:24
http://genomemedicine.com/content/2/4/24

Genome **Medicine**

CORRESPONDENCE **Open Access**

Locus Reference Genomic sequences: an improved basis for describing hur

Raymond Dalgleish[1]*, Paul Flicek[2], Fiona Cunningham[2]
William M McLaren[2], Pontus Larsson[2], Brendan W Vaug
Peter EM Taschner[7], Johan T den Dunnen[7], Andrew Dev

EDITORIAL

nature genetics

Conventional wisdom

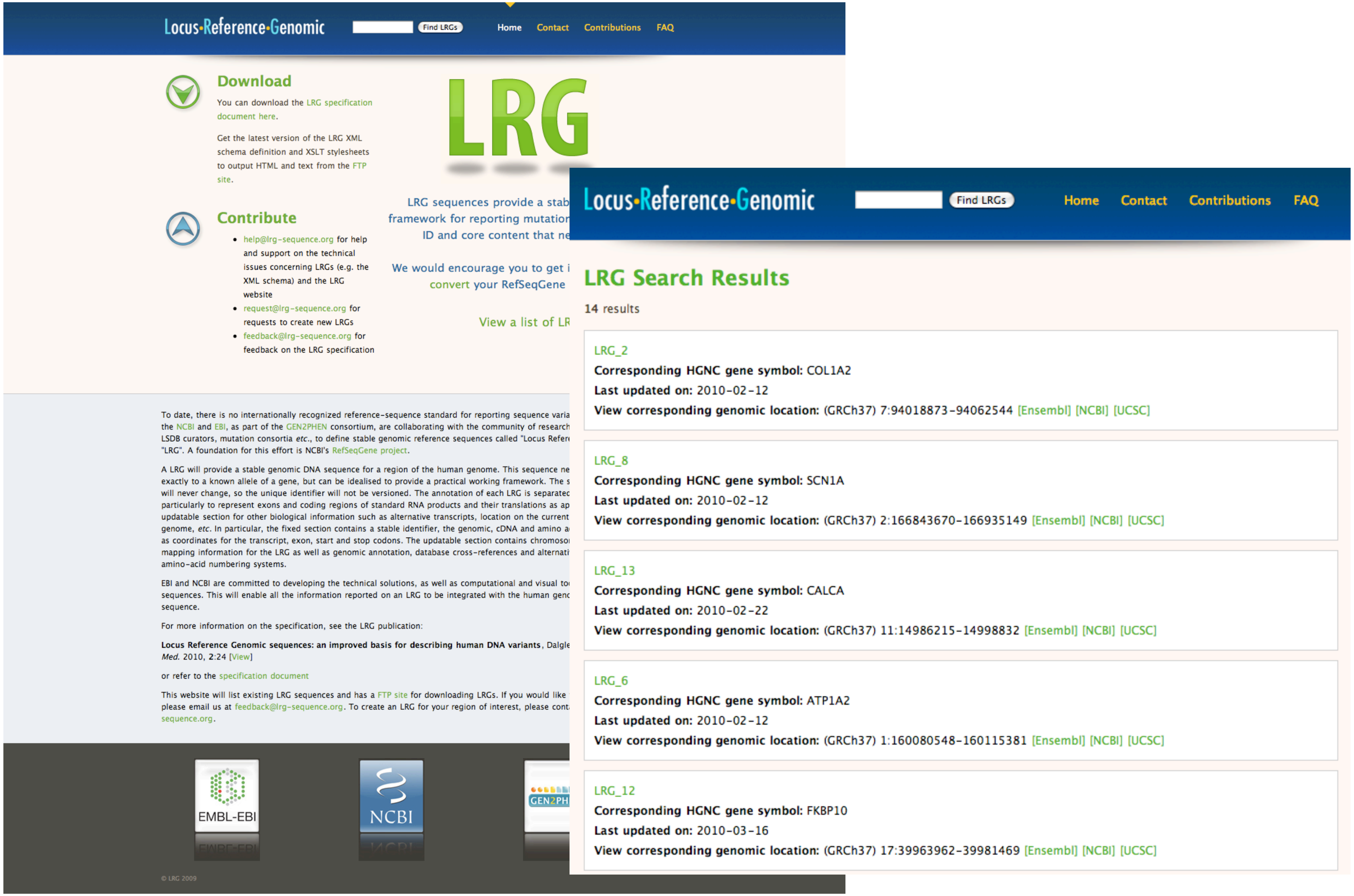
Recent agreement on stable reference sequences for reporting human genetic variants now allows us to mandate the use of the allele naming conventions developed by the Human Genome Variation Society.

By agreement between stakeholders and two principal databases, it has been proposed (R. Dalgleish *et al.*, *Genome Med.* **2**, 24, 2010, doi:10.1186/gm145) that human genetic variants be reported relative to a new set of stable reference sequences, “Locus Reference, Genomic” (LRG, pronounced “large” http://www.lrg-sequence.org/page.php). These sequences have been developed from the initial NCBI RefSeqGene concept and are provided by NCBI and EBI according to agreed rules and in consultation with community users of locus-specific genetic information and locus-specific databases. It is anticipated that the LRG will be stable and supported for many years, long enough to serve as a bridge between existing and future clinical gene tests.

age, resequencing and marker association studies and so keep allele descriptions commensurate with the method by which their data were generated.

The LRG reference sequences should be used in conjunction with standard HGNC gene abbreviations (http://www.genenames.org/) that we already require as a condition of publication. All human genetic variants must now be described—in abstracts and at first use—in accordance with the Human Genome Variation Society (HGVS) conventions (http://www.hgvs.org/mutnomen/) also as a condition of publication. We continue to encourage authors to use HGVS nomenclature for unambiguous reference in all tables and figures and throughout the

EMBL-EBI



Download

You can download the LRG specification document here.

Get the latest version of the LRG XML schema definition and XSLT stylesheets to output HTML and text from the FTP site.

Contribute

- help@lrg-sequence.org for help and support on the technical issues concerning LRGs (e.g. the XML schema) and the LRG website
- request@lrg-sequence.org for requests to create new LRGs
- feedback@lrg-sequence.org for feedback on the LRG specification

LRG sequences provide a stab
framework for reporting mutatior
ID and core content that ne

We would encourage you to get i
convert your RefSeqGene

View a list of LR

To date, there is no internationally recognized reference-sequence standard for reporting sequence varia
the NCBI and EBI, as part of the GEN2PHEN consortium, are collaborating with the community of research
LSDB curators, mutation consortia *etc.*, to define stable genomic reference sequences called "Locus Refere
"LRG". A foundation for this effort is NCBI's RefSeqGene project.

A LRG will provide a stable genomic DNA sequence for a region of the human genome. This sequence ne
exactly to a known allele of a gene, but can be idealised to provide a practical working framework. The s
will never change, so the unique identifier will not be versioned. The annotation of each LRG is separatec
particularly to represent exons and coding regions of standard RNA products and their translations as ap
updatable section for other biological information such as alternative transcripts, location on the current
genome, *etc*. In particular, the fixed section contains a stable identifier, the genomic, cDNA and amino a
as coordinates for the transcript, exon, start and stop codons. The updatable section contains chromosor
mapping information for the LRG as well as genomic annotation, database cross-references and alternati
amino-acid numbering systems.

EBI and NCBI are committed to developing the technical solutions, as well as computational and visual toc
sequences. This will enable all the information reported on an LRG to be integrated with the human genc
sequence.

For more information on the specification, see the LRG publication:

Locus Reference Genomic sequences: an improved basis for describing human DNA variants, Dalgle
Med. 2010, **2**:24 [View]

or refer to the specification document

This website will list existing LRG sequences and has a FTP site for downloading LRGs. If you would like
please email us at feedback@lrg-sequence.org. To create an LRG for your region of interest, please conta
sequence.org.

© LRG 2009

LRG Search Results

14 results

LRG_2
Corresponding HGNC gene symbol: COL1A2
Last updated on: 2010-02-12
View corresponding genomic location: (GRCh37) 7:94018873-94062544 [Ensembl] [NCBI] [UCSC]

LRG_8
Corresponding HGNC gene symbol: SCN1A
Last updated on: 2010-02-12
View corresponding genomic location: (GRCh37) 2:166843670-166935149 [Ensembl] [NCBI] [UCSC]

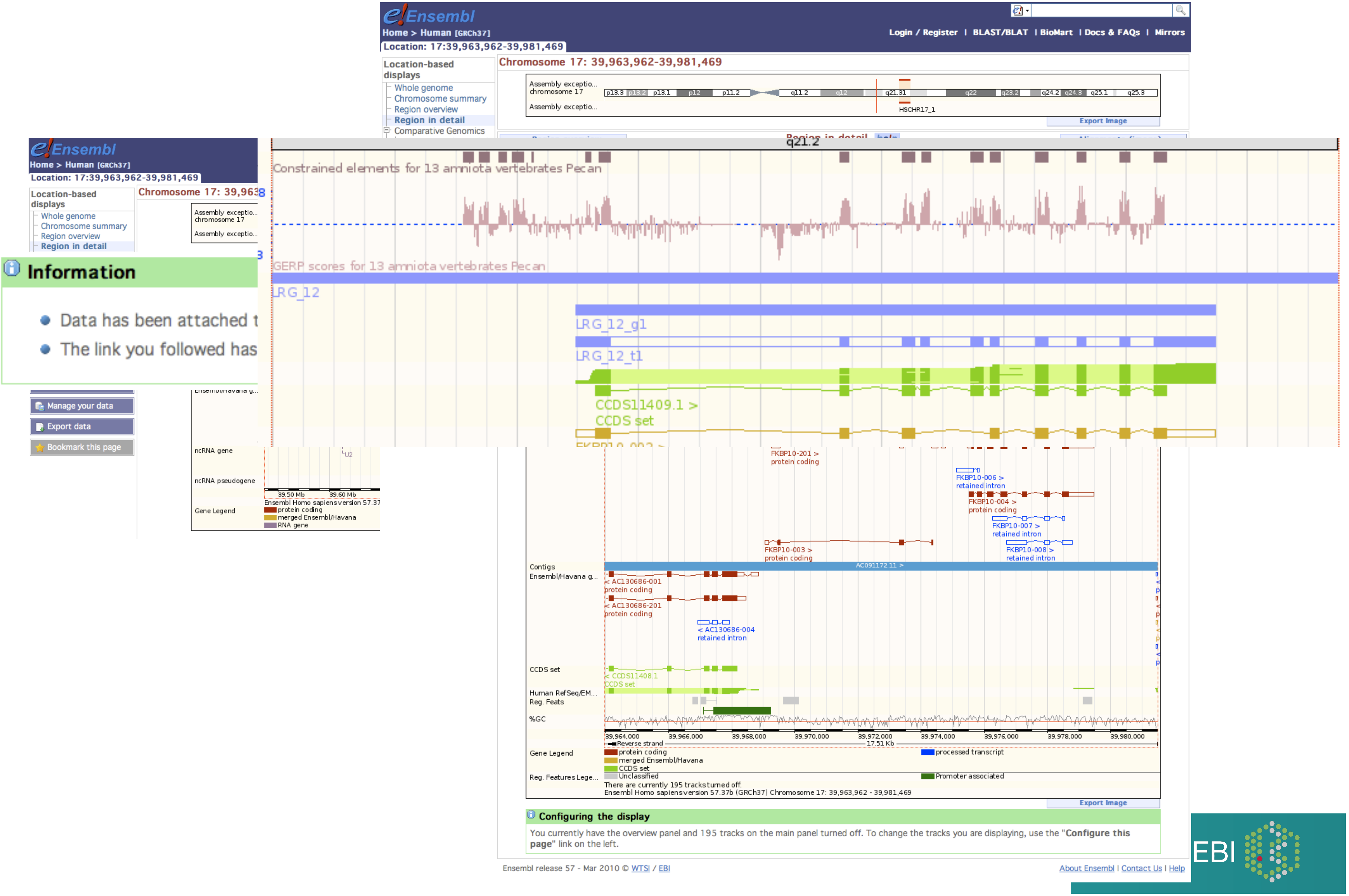
LRG_13
Corresponding HGNC gene symbol: CALCA
Last updated on: 2010-02-22
View corresponding genomic location: (GRCh37) 11:14986215-14998832 [Ensembl] [NCBI] [UCSC]

LRG_6
Corresponding HGNC gene symbol: ATP1A2
Last updated on: 2010-02-12
View corresponding genomic location: (GRCh37) 1:160080548-160115381 [Ensembl] [NCBI] [UCSC]

LRG_12
Corresponding HGNC gene symbol: FKBP10
Last updated on: 2010-03-16
View corresponding genomic location: (GRCh37) 17:39963962-39981469 [Ensembl] [NCBI] [UCSC]

EMBL-EBI

LRGs in Ensembl



Information

- Data has been attached t
- The link you followed has

There are currently 195 tracks turned off.

Ensembl Homo sapiens version 57.37b (GRCh37) Chromosome 17: 39,963,962 - 39,981,469

Configuring the display

You currently have the overview panel and 195 tracks on the main panel turned off. To change the tracks you are displaying, use the "**Configure this page**" link on the left.

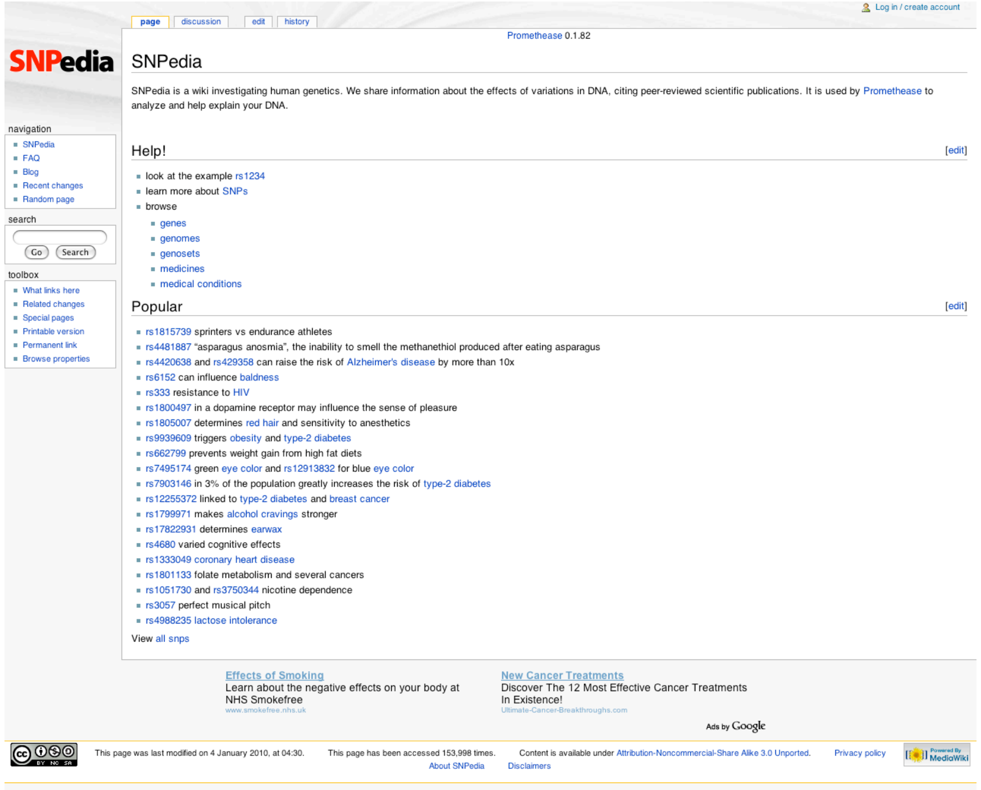
Ensembl release 57 - Mar 2010 © WTSI / EBI

About Ensembl | Contact Us | Help

EBI

Integrating live external data sources

- SNPedia
  - Wiki-based system for editing information about SNP annotations
    - Current data on 12418 SNPs
  - Licensed under a Creative Commons Attribution-Noncommercial-Share Alike license
  - www.snpedia.com
- Realtime updates in Ensembl



Promethease 0.1.82

SNPedia

SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by Promethease to analyze and help explain your DNA.

Help! [edit]

- look at the example rs1234
- learn more about SNPs
- browse
  - genes
  - genomes
  - genosets
  - medicines
  - medical conditions

Popular [edit]

- rs1815739 sprinters vs endurance athletes
- rs4481887 “asparagus anosmia”, the inability to smell the methanethiol produced after eating asparagus
- rs4420638 and rs429358 can raise the risk of Alzheimer's disease by more than 10x
- rs6152 can influence baldness
- rs333 resistance to HIV
- rs1800497 in a dopamine receptor may influence the sense of pleasure
- rs1805007 determines red hair and sensitivity to anesthetics
- rs9939609 triggers obesity and type-2 diabetes
- rs662799 prevents weight gain from high fat diets
- rs7495174 green eye color and rs12913832 for blue eye color
- rs7903146 in 3% of the population greatly increases the risk of type-2 diabetes
- rs12255372 linked to type-2 diabetes and breast cancer
- rs1799971 makes alcohol cravings stronger
- rs17822931 determines earwax
- rs4680 varied cognitive effects
- rs1333049 coronary heart disease
- rs1801133 folate metabolism and several cancers
- rs1051730 and rs3750344 nicotine dependence
- rs3057 perfect musical pitch
- rs4988235 lactose intolerance

View all snps

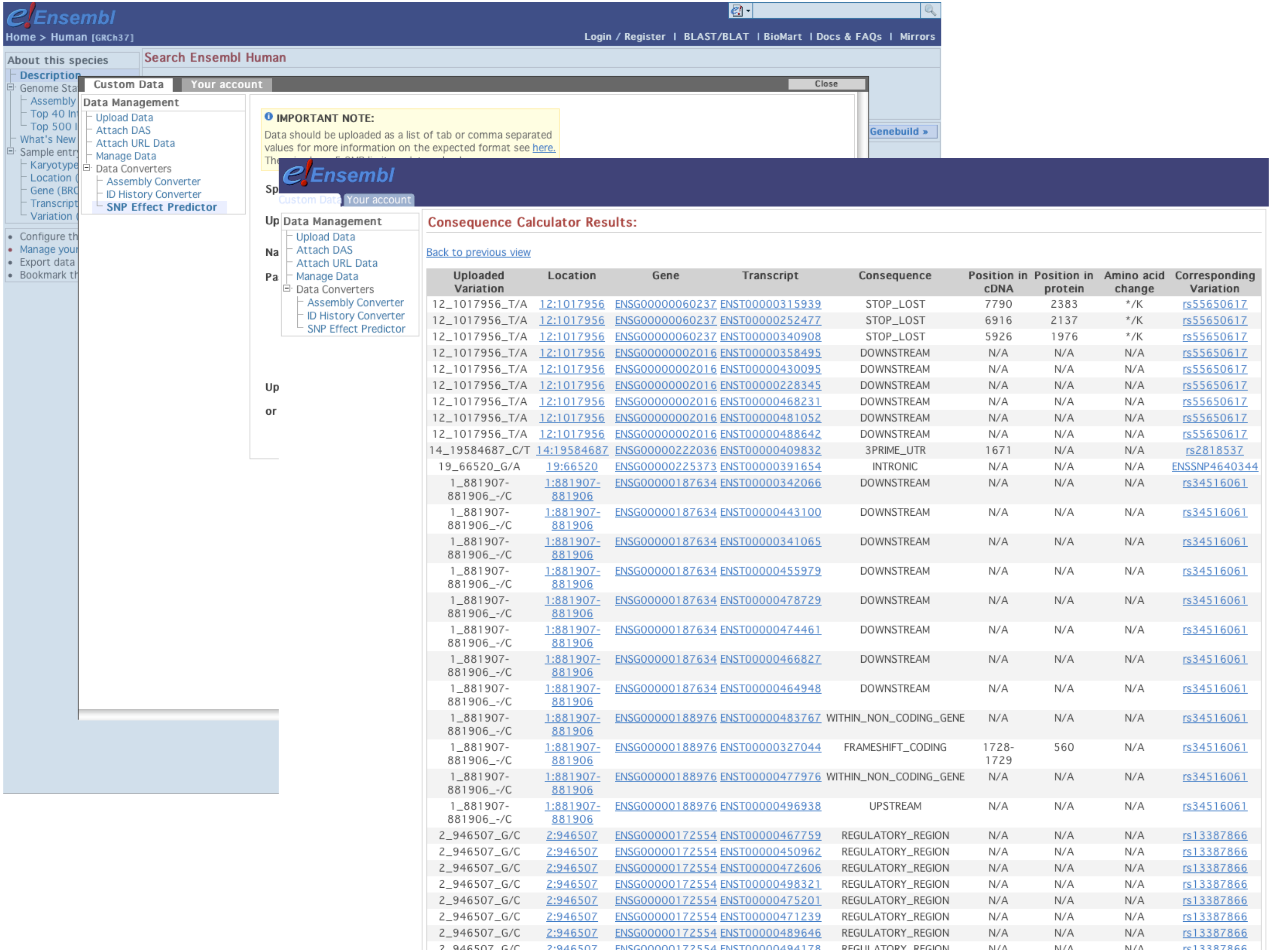
This page was last modified on 4 January 2010, at 04:30. This page has been accessed 153,998 times. Content is available under Attribution-Noncommercial-Share Alike 3.0 Unported. Privacy policy About SNPedia Disclaimers

SNP Effect Prediction tool

- Calculates the effect of SNPs in the context of Ensembl genes and regulatory features
  - Web and API interface
  - Code back-ported to support NCBI36 assembly
  - Programmatic support for tab-delimited and VCF files

- Previously SNP effects were pre-computed for all Ensembl species with variation databases and known SNPs
  - Supports all species and arbitrary SNPs
  - Easily integrated into analysis pipelines

McLaren, et al. *Bioinformatics.* 2010



Home > Human [GRCh37]

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs | Mirrors

About this species

Search Ensembl Human

Custom Data | Your account | Close

Data Management

- Upload Data
- Attach DAS
- Attach URL Data
- Manage Data
- Data Converters
  - Assembly Converter
  - ID History Converter
  - SNP Effect Predictor

IMPORTANT NOTE:

Data should be uploaded as a list of tab or comma separated values for more information on the expected format see here.

Consequence Calculator Results:

Back to previous view

Uploaded Variation	Location	Gene	Transcript	Consequence	Position in cDNA	Position in protein	Amino acid change	Corresponding Variation
12_1017956_T/A	12:1017956	ENSG00000060237	ENST00000315939	STOP_LOST	7790	2383	*/K	rs55650617
12_1017956_T/A	12:1017956	ENSG00000060237	ENST00000252477	STOP_LOST	6916	2137	*/K	rs55650617
12_1017956_T/A	12:1017956	ENSG00000060237	ENST00000340908	STOP_LOST	5926	1976	*/K	rs55650617
12_1017956_T/A	12:1017956	ENSG00000002016	ENST00000358495	DOWNSTREAM	N/A	N/A	N/A	rs55650617
12_1017956_T/A	12:1017956	ENSG00000002016	ENST00000430095	DOWNSTREAM	N/A	N/A	N/A	rs55650617
12_1017956_T/A	12:1017956	ENSG00000002016	ENST00000228345	DOWNSTREAM	N/A	N/A	N/A	rs55650617
12_1017956_T/A	12:1017956	ENSG00000002016	ENST00000468231	DOWNSTREAM	N/A	N/A	N/A	rs55650617
12_1017956_T/A	12:1017956	ENSG00000002016	ENST00000481052	DOWNSTREAM	N/A	N/A	N/A	rs55650617
12_1017956_T/A	12:1017956	ENSG00000002016	ENST00000488642	DOWNSTREAM	N/A	N/A	N/A	rs55650617
14_19584687_C/T	14:19584687	ENSG00000222036	ENST00000409832	3PRIME_UTR	1671	N/A	N/A	rs2818537
19_66520_G/A	19:66520	ENSG00000225373	ENST00000391654	INTRONIC	N/A	N/A	N/A	ENSSNP4640344
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000342066	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000443100	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000341065	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000455979	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000478729	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000474461	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000466827	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000187634	ENST00000464948	DOWNSTREAM	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000188976	ENST00000483767	WITHIN_NON_CODING_GENE	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000188976	ENST00000327044	FRAMESHIFT_CODING	1728-1729	560	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000188976	ENST00000477976	WITHIN_NON_CODING_GENE	N/A	N/A	N/A	rs34516061
1_881907-881906_-/C	1:881907-881906	ENSG00000188976	ENST00000496938	UPSTREAM	N/A	N/A	N/A	rs34516061
2_946507_G/C	2:946507	ENSG00000172554	ENST00000467759	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000450962	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000472606	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000498321	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000475201	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000471239	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000489646	REGULATORY_REGION	N/A	N/A	N/A	rs13387866
2_946507_G/C	2:946507	ENSG00000172554	ENST00000494178	REGULATORY_REGION	N/A	N/A	N/A	rs13387866

McLaren, et al. *Bioinformatics.* 2010

EMBL-EBI

Challenges

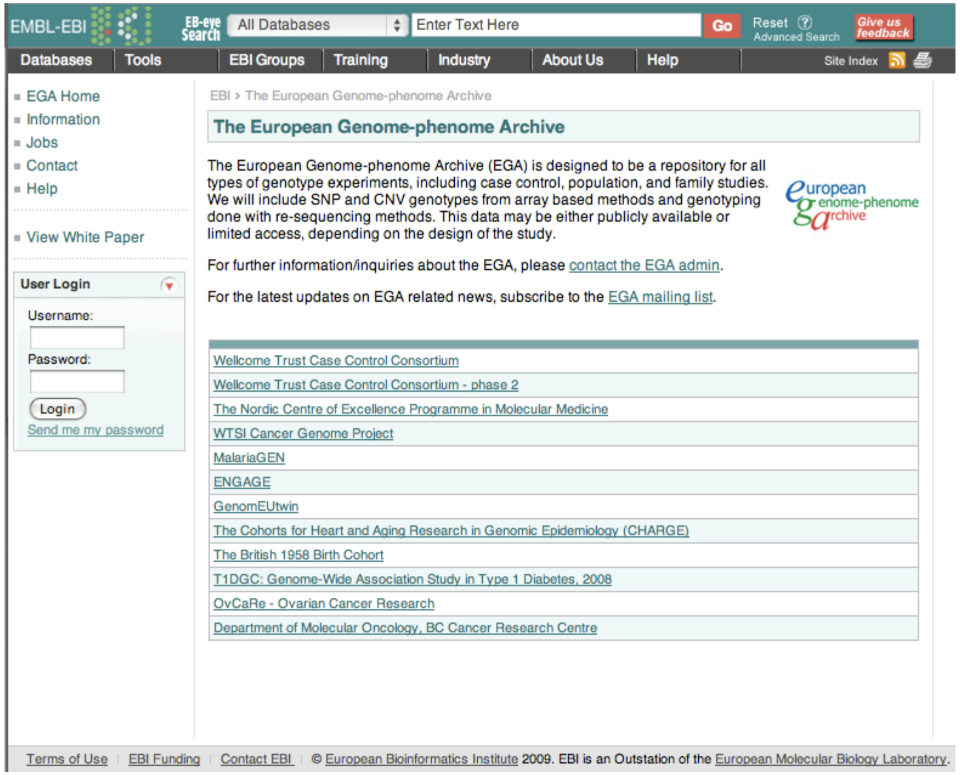
- Access
  - Data size, storage and transfer
  - Providing access to other researchers that want to use the data

- Annotation of variant data
  - Incorporating published and curated information
  - Integrating data that is collected on the genome index

- Most human research data cannot be openly released
  - How much diagnostic data should be released?
  - From research to clinical practice

The European Genome-phenome Archive

- Secure storage and authorised access to all types of data sets that might be generated in the context of research into molecular medicine
  - Sequence; Genotypes
  - Transcriptomics; Proteomics
  - Phenotype data
- Enable the collection of larger cohorts and maximisation of resource use
  - Sequencing capacity is increasing dramatically
  - Analysis capacity is increasing more slowly



EBI > The European Genome-phenome Archive

The European Genome-phenome Archive

The European Genome-phenome Archive (EGA) is designed to be a repository for all types of genotype experiments, including case control, population, and family studies. We will include SNP and CNV genotypes from array based methods and genotyping done with re-sequencing methods. This data may be either publicly available or limited access, depending on the design of the study.

For further information/inquiries about the EGA, please contact the EGA admin.

For the latest updates on EGA related news, subscribe to the EGA mailing list.

- Wellcome Trust Case Control Consortium
- Wellcome Trust Case Control Consortium - phase 2
- The Nordic Centre of Excellence Programme in Molecular Medicine
- WTSI Cancer Genome Project
- MalariaGEN
- ENGAGE
- GenomEUtwin
- The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)
- The British 1958 Birth Cohort
- T1DGC: Genome-Wide Association Study in Type 1 Diabetes, 2008
- OvCaRe - Ovarian Cancer Research
- Department of Molecular Oncology, BC Cancer Research Centre

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2009. EBI is an Outstation of the European Molecular Biology Laboratory.

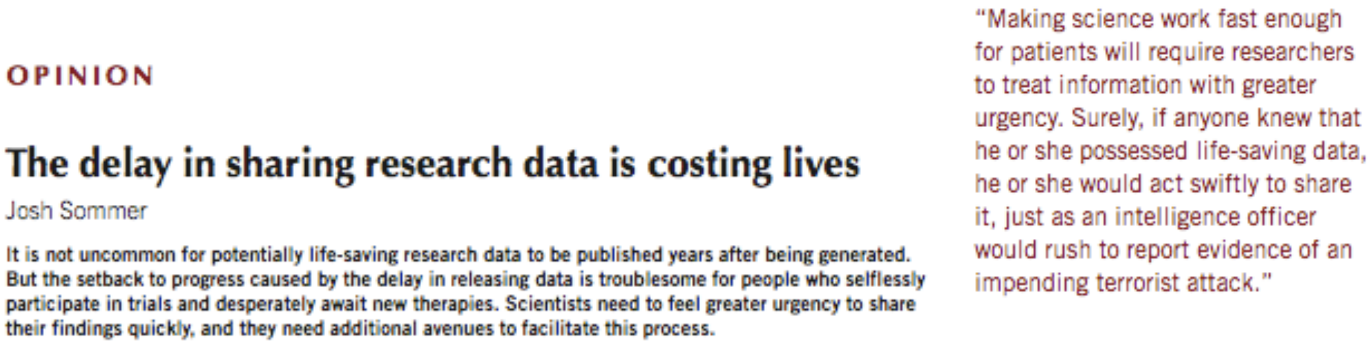
EGA Data Acceptance and Access

european genome-phenome archive

- Access decisions will remain with the data generating body
  - Distributed model
  - Transparency to the data generators
  - EGA manages the access granted
  - Users can also be restricted to particular collections within a study
- EGA is the European peer database to dbGAP (NCBI)
  - dbGAP has adopted a more centralised model of data access decisions
  - We plan data exchange of meta data and more extensive discussions are on going to increase data discoverability
  - Working toward a common application for both databases to lower administrative burden

Community Benefits of the EGA

- Data subject to access controls is a burden and it limits the number of researchers that will reuse the resources
  - This may slow the pace of science and prevent serendipitous discovery
- However…
  - Five years ago accessing this type of data was impossible
  - Now it is just incredibly difficult
  - This is real progress
  - Complicated or overly onerous data access agreements are more likely to be ignored



OPINION

The delay in sharing research data is costing lives

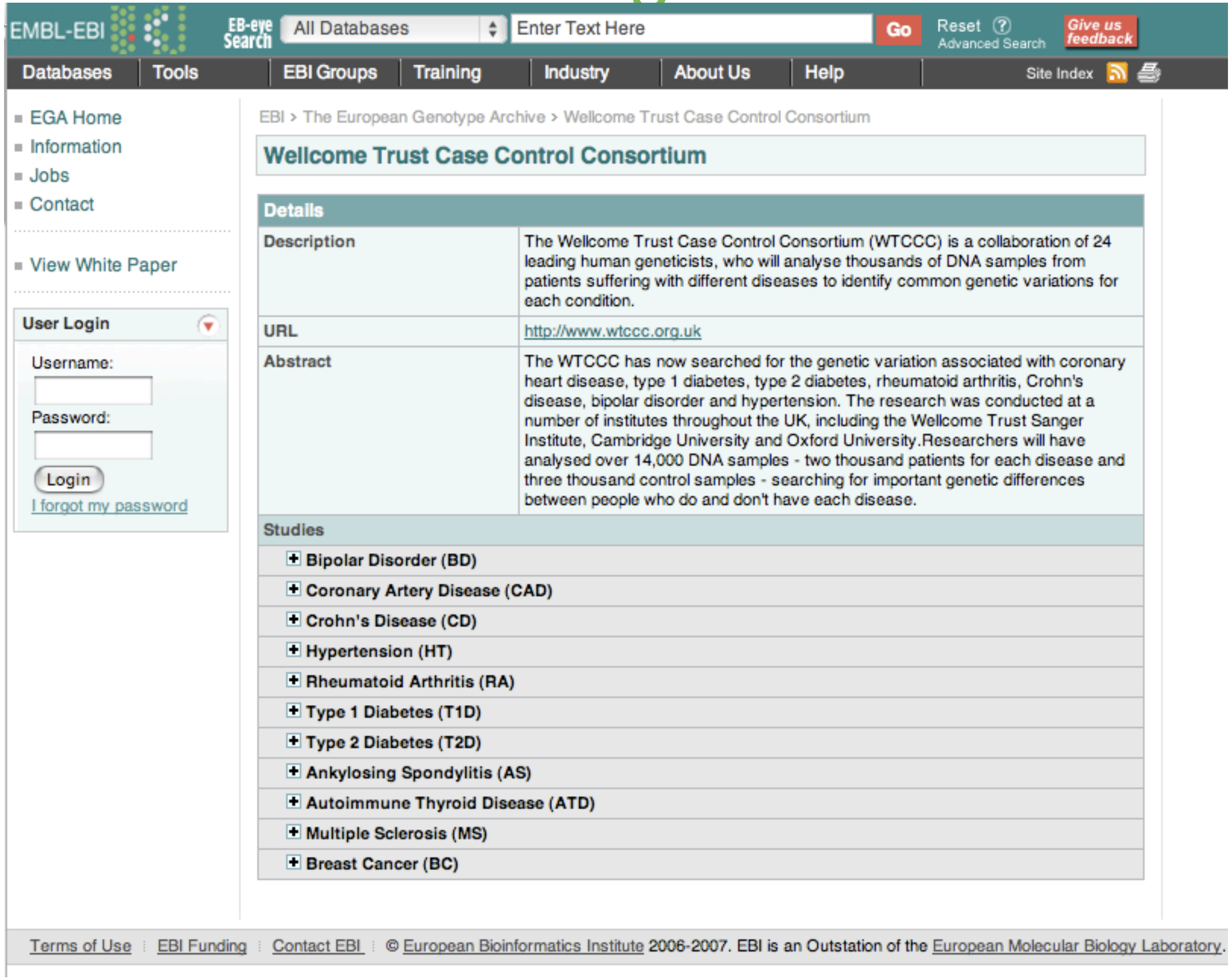
Josh Sommer

It is not uncommon for potentially life-saving research data to be published years after being generated. But the setback to progress caused by the delay in releasing data is troublesome for people who selflessly participate in trials and desperately await new therapies. Scientists need to feel greater urgency to share their findings quickly, and they need additional avenues to facilitate this process.

"Making science work fast enough for patients will require researchers to treat information with greater urgency. Surely, if anyone knew that he or she possessed life-saving data, he or she would act swiftly to share it, just as an intelligence officer would rush to report evidence of an impending terrorist attack."

Nature Medicine, July 2010

EGA Consortium Page for WTCCC



EBI > The European Genotype Archive > Wellcome Trust Case Control Consortium

Wellcome Trust Case Control Consortium

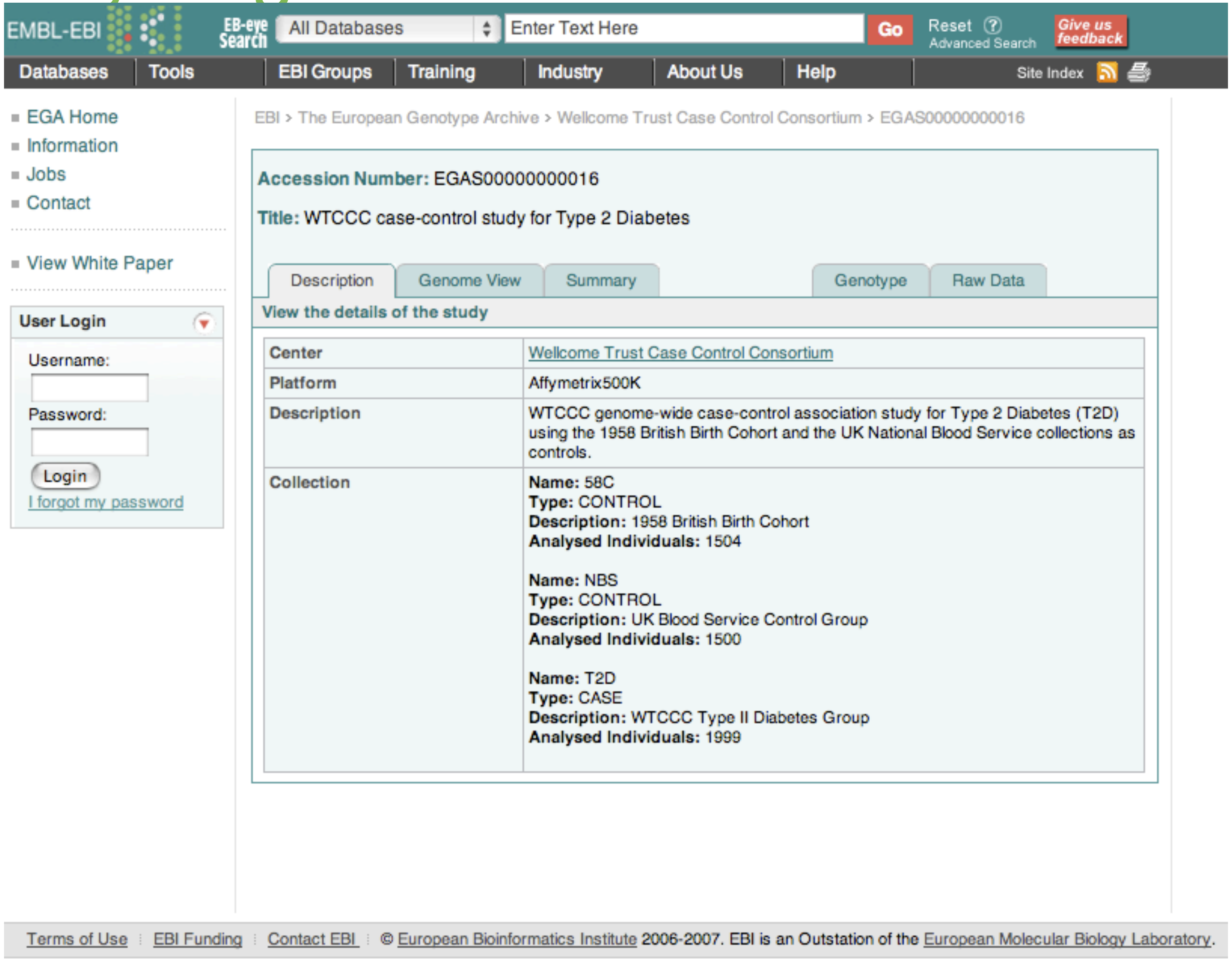
Details	
Description	The Wellcome Trust Case Control Consortium (WTCCC) is a collaboration of 24 leading human geneticists, who will analyse thousands of DNA samples from patients suffering with different diseases to identify common genetic variations for each condition.
URL	http://www.wtccc.org.uk
Abstract	The WTCCC has now searched for the genetic variation associated with coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. The research was conducted at a number of institutes throughout the UK, including the Wellcome Trust Sanger Institute, Cambridge University and Oxford University.Researchers will have analysed over 14,000 DNA samples - two thousand patients for each disease and three thousand control samples - searching for important genetic differences between people who do and don't have each disease.

Studies

- + Bipolar Disorder (BD)
- + Coronary Artery Disease (CAD)
- + Crohn's Disease (CD)
- + Hypertension (HT)
- + Rheumatoid Arthritis (RA)
- + Type 1 Diabetes (T1D)
- + Type 2 Diabetes (T2D)
- + Ankylosing Spondylitis (AS)
- + Autoimmune Thyroid Disease (ATD)
- + Multiple Sclerosis (MS)
- + Breast Cancer (BC)

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2006-2007. EBI is an Outstation of the European Molecular Biology Laboratory.

Study Page for WTCCC T2D



EBI > The European Genotype Archive > Wellcome Trust Case Control Consortium > EGAS00000000016

Accession Number: EGAS00000000016

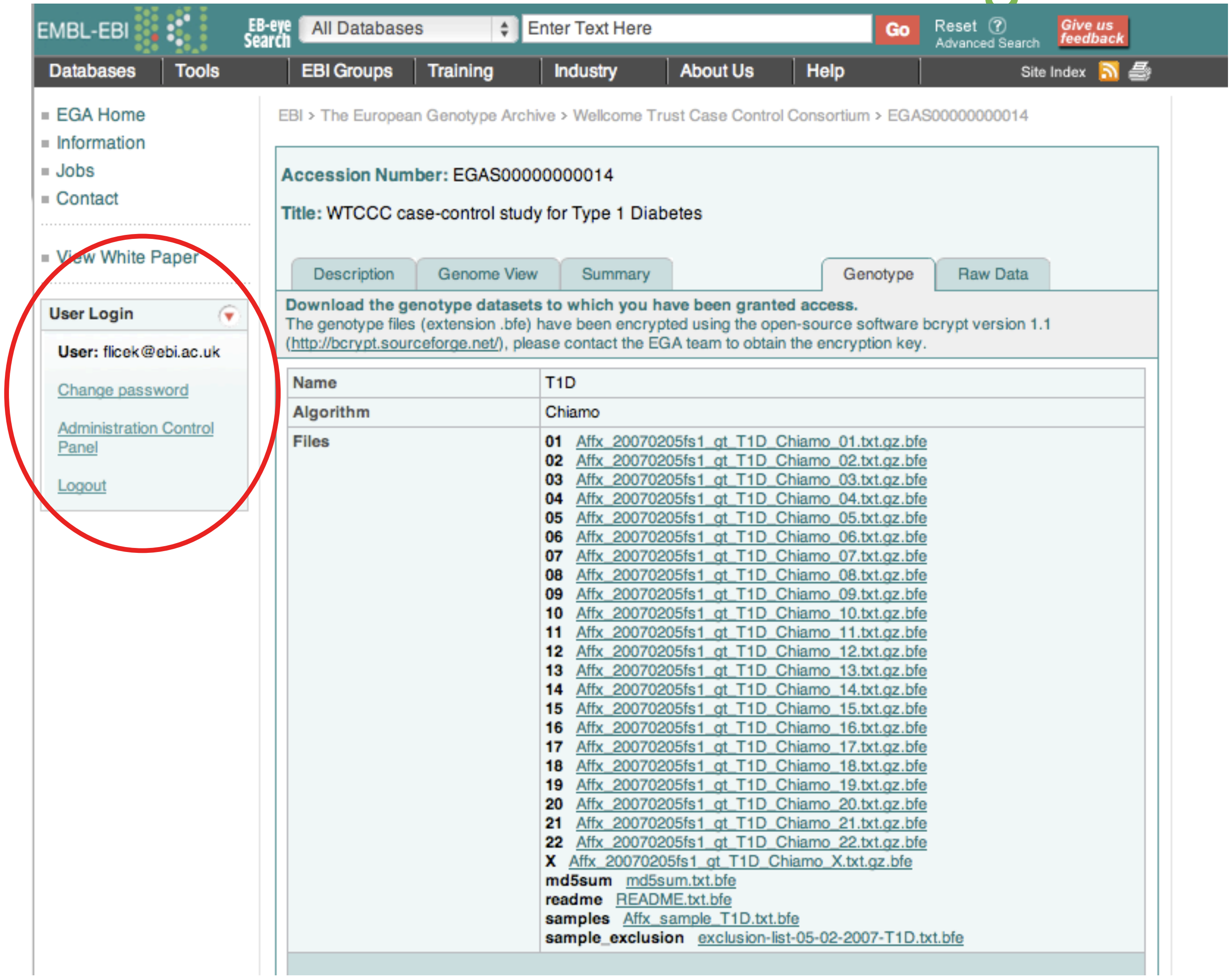
Title: WTCCC case-control study for Type 2 Diabetes

View the details of the study

Center	Wellcome Trust Case Control Consortium
Platform	Affymetrix500K
Description	WTCCC genome-wide case-control association study for Type 2 Diabetes (T2D) using the 1958 British Birth Cohort and the UK National Blood Service collections as controls.
Collection	**Name:** 58C **Type:** CONTROL **Description:** 1958 British Birth Cohort **Analysed Individuals:** 1504 **Name:** NBS **Type:** CONTROL **Description:** UK Blood Service Control Group **Analysed Individuals:** 1500 **Name:** T2D **Type:** CASE **Description:** WTCCC Type II Diabetes Group **Analysed Individuals:** 1999

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2006-2007. EBI is an Outstation of the European Molecular Biology Laboratory.

WTCCC T1D Data Access Page



User: flicek@ebi.ac.uk

EBI > The European Genotype Archive > Wellcome Trust Case Control Consortium > EGAS00000000014

Accession Number: EGAS00000000014

Title: WTCCC case-control study for Type 1 Diabetes

Download the genotype datasets to which you have been granted access.
The genotype files (extension .bfe) have been encrypted using the open-source software bcrypt version 1.1 (http://bcrypt.sourceforge.net/), please contact the EGA team to obtain the encryption key.

Name	T1D
Algorithm	Chiamo
Files	**01** Affx_20070205fs1_gt_T1D_Chiamo_01.txt.gz.bfe **02** Affx_20070205fs1_gt_T1D_Chiamo_02.txt.gz.bfe **03** Affx_20070205fs1_gt_T1D_Chiamo_03.txt.gz.bfe **04** Affx_20070205fs1_gt_T1D_Chiamo_04.txt.gz.bfe **05** Affx_20070205fs1_gt_T1D_Chiamo_05.txt.gz.bfe **06** Affx_20070205fs1_gt_T1D_Chiamo_06.txt.gz.bfe **07** Affx_20070205fs1_gt_T1D_Chiamo_07.txt.gz.bfe **08** Affx_20070205fs1_gt_T1D_Chiamo_08.txt.gz.bfe **09** Affx_20070205fs1_gt_T1D_Chiamo_09.txt.gz.bfe **10** Affx_20070205fs1_gt_T1D_Chiamo_10.txt.gz.bfe **11** Affx_20070205fs1_gt_T1D_Chiamo_11.txt.gz.bfe **12** Affx_20070205fs1_gt_T1D_Chiamo_12.txt.gz.bfe **13** Affx_20070205fs1_gt_T1D_Chiamo_13.txt.gz.bfe **14** Affx_20070205fs1_gt_T1D_Chiamo_14.txt.gz.bfe **15** Affx_20070205fs1_gt_T1D_Chiamo_15.txt.gz.bfe **16** Affx_20070205fs1_gt_T1D_Chiamo_16.txt.gz.bfe **17** Affx_20070205fs1_gt_T1D_Chiamo_17.txt.gz.bfe **18** Affx_20070205fs1_gt_T1D_Chiamo_18.txt.gz.bfe **19** Affx_20070205fs1_gt_T1D_Chiamo_19.txt.gz.bfe **20** Affx_20070205fs1_gt_T1D_Chiamo_20.txt.gz.bfe **21** Affx_20070205fs1_gt_T1D_Chiamo_21.txt.gz.bfe **22** Affx_20070205fs1_gt_T1D_Chiamo_22.txt.gz.bfe **X** Affx_20070205fs1_gt_T1D_Chiamo_X.txt.gz.bfe **md5sum** md5sum.txt.bfe **readme** README.txt.bfe **samples** Affx_sample_T1D.txt.bfe **sample_exclusion** exclusion-list-05-02-2007-T1D.txt.bfe

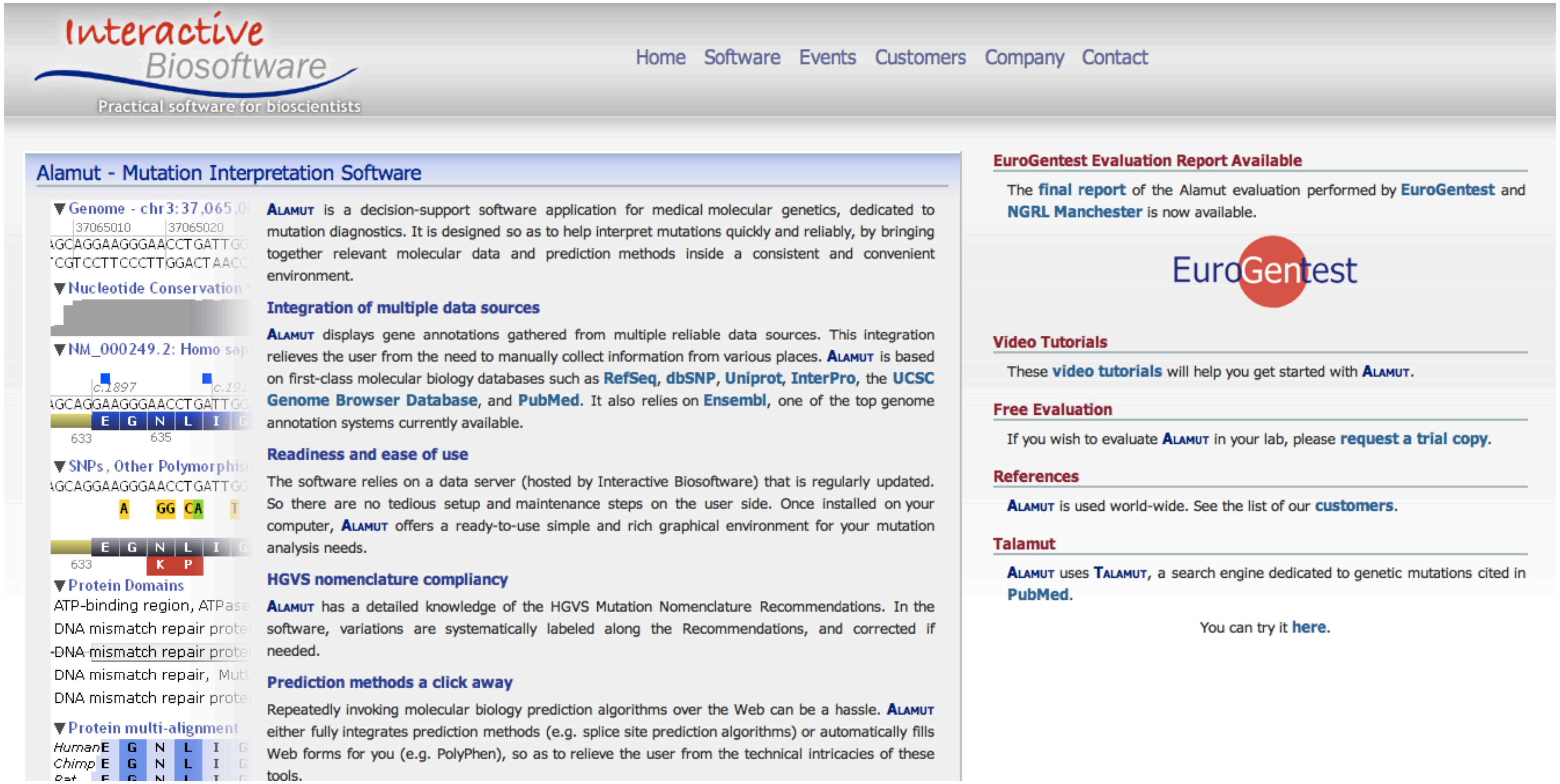
Beyond research toward medical practice

- Needs:
  - Consistent, traceable data generation and analysis routines
  - Robust annotation based on public information sources such as those at the EBI
  - Reporting into medical records

- Data storage:
  - Probably not necessary for primary data as costs drop
  - Individual variant catalogs are already much smaller than MRI data
  - May prevent some liability issues

Enabling clinical services

- Multiple commercial clinical services built on annotation/Ensembl
  - Alamut - Mutation Interpretation Software - http://www.interactive-biosoftware.com/



Alamut - Mutation Interpretation Software

ALAMUT is a decision-support software application for medical molecular genetics, dedicated to mutation diagnostics. It is designed so as to help interpret mutations quickly and reliably, by bringing together relevant molecular data and prediction methods inside a consistent and convenient environment.

Integration of multiple data sources

ALAMUT displays gene annotations gathered from multiple reliable data sources. This integration relieves the user from the need to manually collect information from various places. ALAMUT is based on first-class molecular biology databases such as **RefSeq**, **dbSNP**, **Uniprot**, **InterPro**, the **UCSC Genome Browser Database**, and **PubMed**. It also relies on **Ensembl**, one of the top genome annotation systems currently available.

Readiness and ease of use

The software relies on a data server (hosted by Interactive Biosoftware) that is regularly updated. So there are no tedious setup and maintenance steps on the user side. Once installed on your computer, ALAMUT offers a ready-to-use simple and rich graphical environment for your mutation analysis needs.

HGVS nomenclature compliancy

ALAMUT has a detailed knowledge of the HGVS Mutation Nomenclature Recommendations. In the software, variations are systematically labeled along the Recommendations, and corrected if needed.

Prediction methods a click away

Repeatedly invoking molecular biology prediction algorithms over the Web can be a hassle. ALAMUT either fully integrates prediction methods (e.g. splice site prediction algorithms) or automatically fills Web forms for you (e.g. PolyPhen), so as to relieve the user from the technical intricacies of these tools.

EuroGentest Evaluation Report Available

The **final report** of the Alamut evaluation performed by **EuroGentest** and **NGRL Manchester** is now available.

EuroGentest

Video Tutorials

These **video tutorials** will help you get started with ALAMUT.

Free Evaluation

If you wish to evaluate ALAMUT in your lab, please **request a trial copy**.

References

ALAMUT is used world-wide. See the list of our **customers**.

Talamut

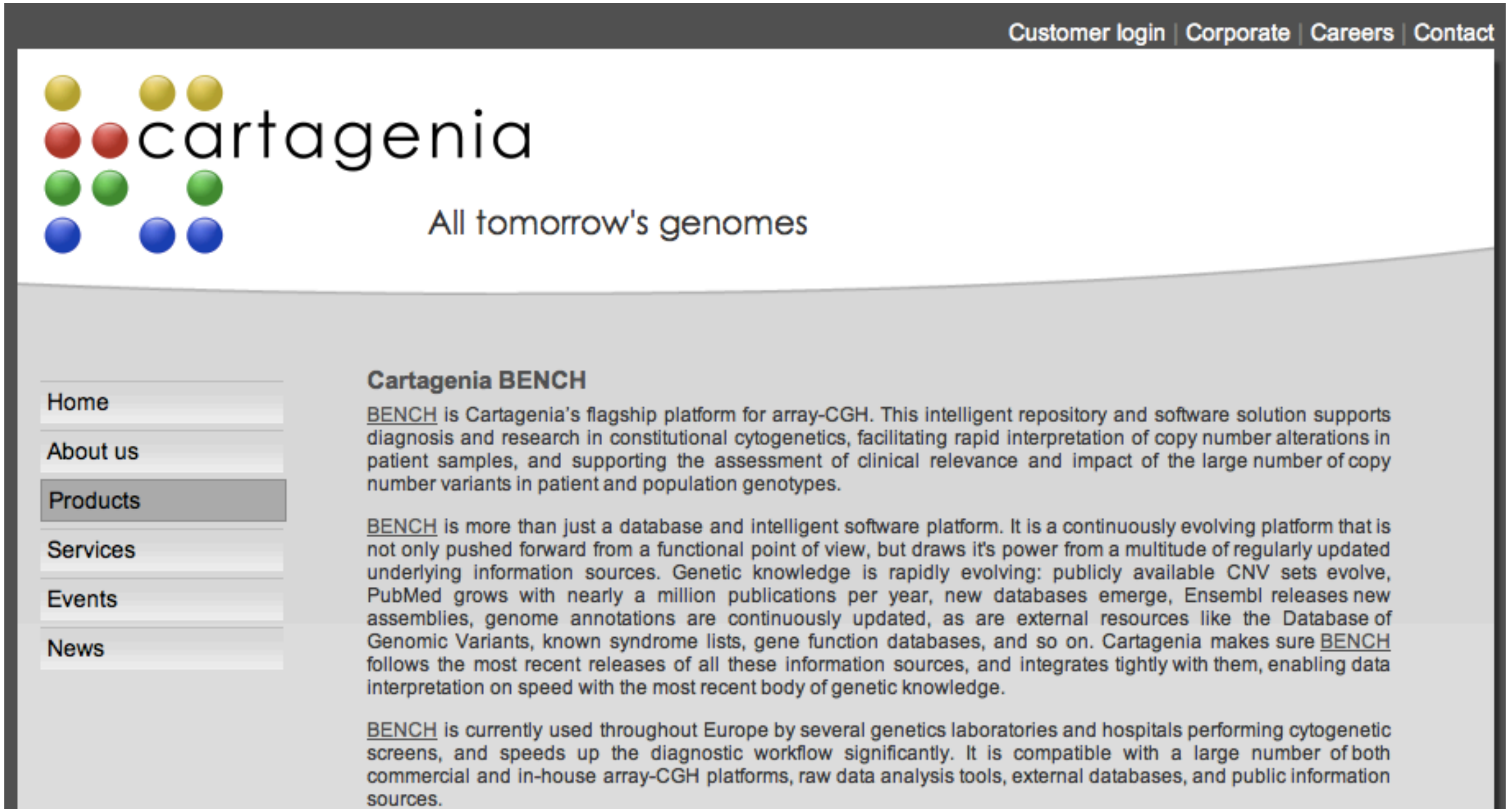
ALAMUT uses TALAMUT, a search engine dedicated to genetic mutations cited in **PubMed**.

You can try it **here**.

EMBL-EBI

Enabling clinical services

- Multiple commercial clinical services built on annotation/Ensembl
  - BENCH - array CGH platform - http://www.cartagenia.com/



Cartagenia BENCH

BENCH is Cartagenia's flagship platform for array-CGH. This intelligent repository and software solution supports diagnosis and research in constitutional cytogenetics, facilitating rapid interpretation of copy number alterations in patient samples, and supporting the assessment of clinical relevance and impact of the large number of copy number variants in patient and population genotypes.

BENCH is more than just a database and intelligent software platform. It is a continuously evolving platform that is not only pushed forward from a functional point of view, but draws it's power from a multitude of regularly updated underlying information sources. Genetic knowledge is rapidly evolving: publicly available CNV sets evolve, PubMed grows with nearly a million publications per year, new databases emerge, Ensembl releases new assemblies, genome annotations are continuously updated, as are external resources like the Database of Genomic Variants, known syndrome lists, gene function databases, and so on. Cartagenia makes sure BENCH follows the most recent releases of all these information sources, and integrates tightly with them, enabling data interpretation on speed with the most recent body of genetic knowledge.

BENCH is currently used throughout Europe by several genetics laboratories and hospitals performing cytogenetic screens, and speeds up the diagnostic workflow significantly. It is compatible with a large number of both commercial and in-house array-CGH platforms, raw data analysis tools, external databases, and public information sources.

LRGs are already part of HL7

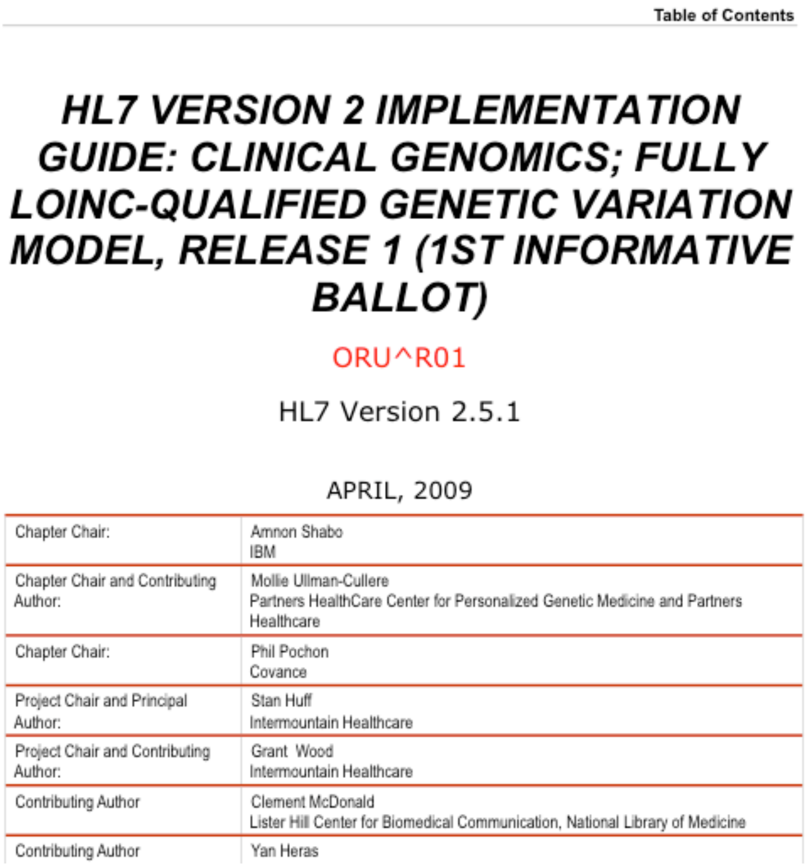


Table of Contents

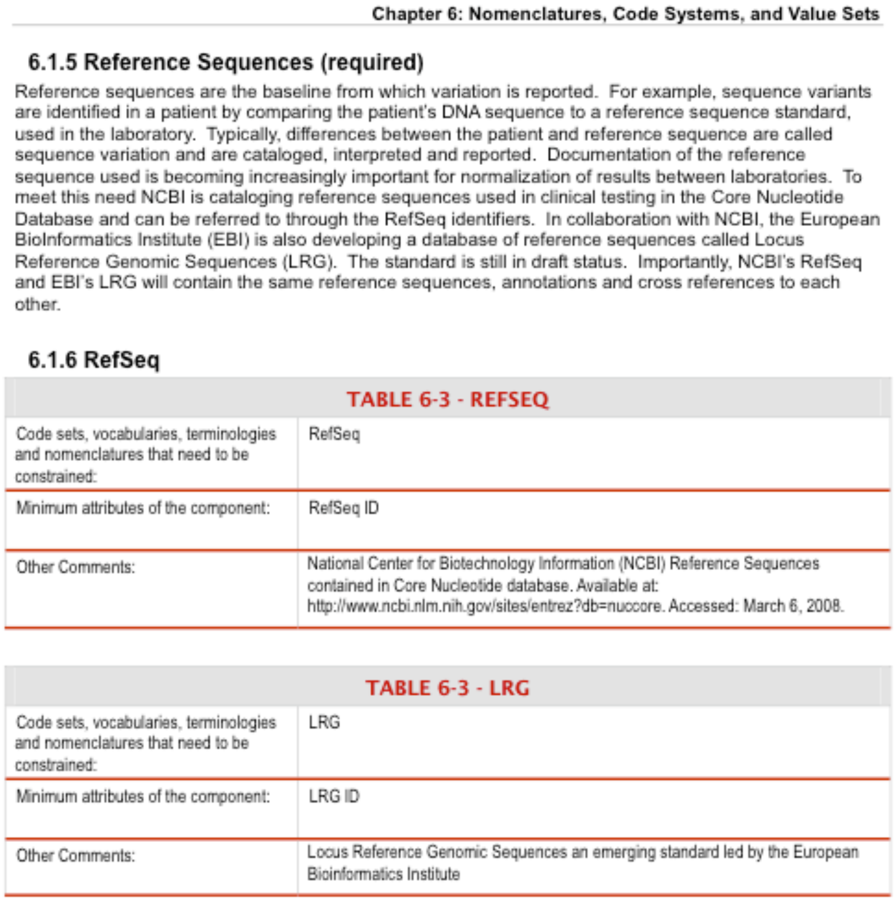
HL7 VERSION 2 IMPLEMENTATION GUIDE: CLINICAL GENOMICS; FULLY LOINC-QUALIFIED GENETIC VARIATION MODEL, RELEASE 1 (1ST INFORMATIVE BALLOT)

ORU^R01

HL7 Version 2.5.1

APRIL, 2009

Chapter Chair:	Amnon Shabo IBM
Chapter Chair and Contributing Author:	Mollie Ullman-Cullere Partners HealthCare Center for Personalized Genetic Medicine and Partners Healthcare
Chapter Chair:	Phil Pochon Covance
Project Chair and Principal Author:	Stan Huff Intermountain Healthcare
Project Chair and Contributing Author:	Grant Wood Intermountain Healthcare
Contributing Author	Clement McDonald Lister Hill Center for Biomedical Communication, National Library of Medicine
Contributing Author	Yan Heras



Chapter 6: Nomenclatures, Code Systems, and Value Sets

6.1.5 Reference Sequences (required)

Reference sequences are the baseline from which variation is reported. For example, sequence variants are identified in a patient by comparing the patient's DNA sequence to a reference sequence standard, used in the laboratory. Typically, differences between the patient and reference sequence are called sequence variation and are cataloged, interpreted and reported. Documentation of the reference sequence used is becoming increasingly important for normalization of results between laboratories. To meet this need NCBI is cataloging reference sequences used in clinical testing in the Core Nucleotide Database and can be referred to through the RefSeq identifiers. In collaboration with NCBI, the European BioInformatics Institute (EBI) is also developing a database of reference sequences called Locus Reference Genomic Sequences (LRG). The standard is still in draft status. Importantly, NCBI's RefSeq and EBI's LRG will contain the same reference sequences, annotations and cross references to each other.

6.1.6 RefSeq

TABLE 6-3 - REFSEQ

Code sets, vocabularies, terminologies and nomenclatures that need to be constrained:	RefSeq
Minimum attributes of the component:	RefSeq ID
Other Comments:	National Center for Biotechnology Information (NCBI) Reference Sequences contained in Core Nucleotide database. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore. Accessed: March 6, 2008.

TABLE 6-3 - LRG

Code sets, vocabularies, terminologies and nomenclatures that need to be constrained:	LRG
Minimum attributes of the component:	LRG ID
Other Comments:	Locus Reference Genomic Sequences an emerging standard led by the European Bioinformatics Institute

Annotating the variation catalogue created by the 1000 Genomes projects (and other similar projects) will be one of the major future challenges in human genomics

The results of this annotation will change the way that medicine is practised. And will impact society.

Acknowledgements

- Vertebrate Genomics Team
  - Ilkka Lappalainen, Jonathan Hinton, Vasudev Kumanduri, Jeff Almeida-King, Michael Maguire (European Genome-phenome Archive)
  - Laura Clarke, Holly Zheng-Bradley, Rick Smith (1000 Genomes DCC)
  - Fiona Cunningham, Graham Ritchie, Will McLaren, Pontus Larsson (Ensembl Variation)
  - Glenn Proctor, Eugene Kulesha, Stephen Keenan (Ensembl Software)

- NCBI DCC Team: Steve Sherry, Martin Shumway, Justin Paschall, Chunlin Xiao
- European Nucleotide Archive / European Read Archive Team
- 1000 Genomes and WTCCC Projects
- The Protein and Nucleotide Database Group

- Funding: Wellcome Trust, EMBL, UK MRC, European Union

1000 Genomes
A Deep Catalog of Human Genetic Variation

http://www.1000genomes.org

EMBL-EBI