CRG

Centre
de Regulació
Genòmica

Standing on the shoulders of giants,

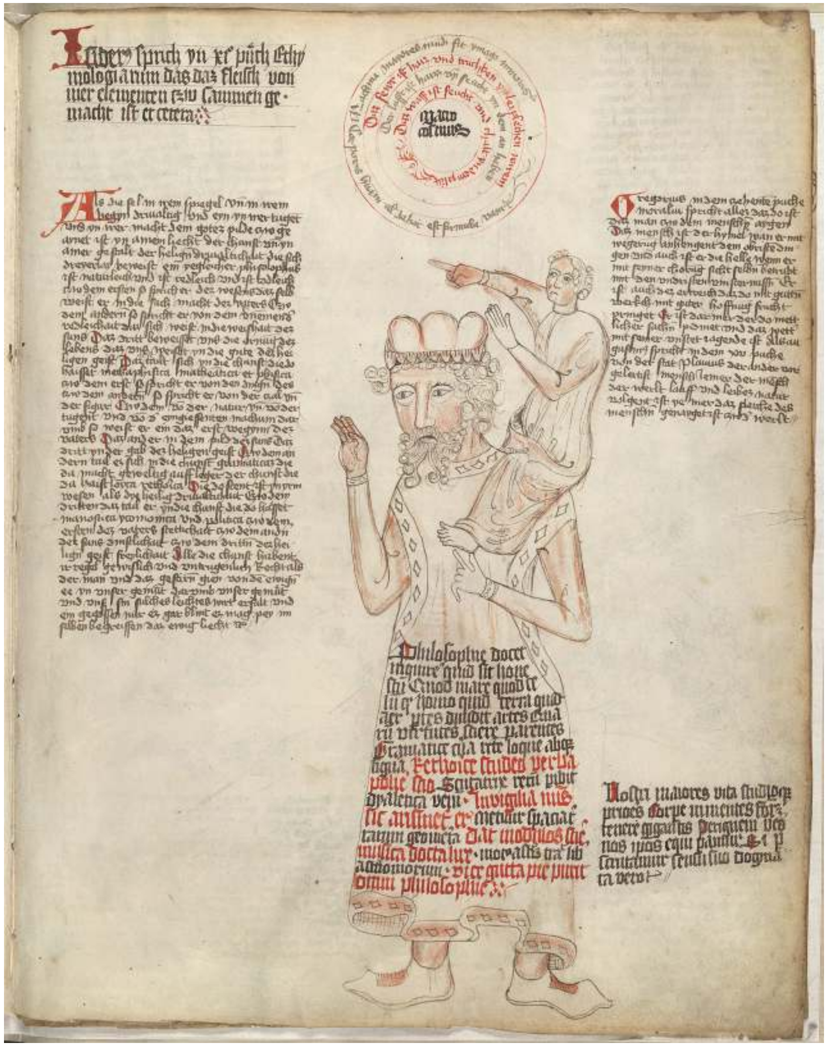
German Demidov,
Bioinformatics
Summer School
2017

EXCELENCIA
SEVERO
OCHOA

ИНСТИТУТ
БИОИНФОРМАТИКИ

Biology and Big Data

- Discovering truth by building on previous discoveries



Why it is useful?

Just one example:

As the result of recent IHEC efforts, a large number of epigenome profiles became publicly available. We have performed the most detailed large-scale data-integration analysis to associate enhancers to their gene targets using all currently available epigenome profiles from four different IHEC data-sets.
We selected cross-cell-types profiles of H3K27ac, H3K4me1, DNA methylation, DNase I hypersensitivity and RNA-Seq to quantify enhancer activity and gene expression. To identify long-range interactions we modelled gene expression and enhancer activity using linear regression, and meta-analyzed individual gene-enhancer models across consortia. To confirm associations, we built a benchmark datasets based on GTEx and published chromatin interactions databases.
We quantified cross-cell-types enhancer activities and gene expressions using up to 177 epigenome profiles. We tested more than 4.3 million gene-enhancer pairs. Approximately 16,000 genes and 60,000 enhancer regions reached significant Bonferroni corrected p-value. Our results confirm previously reported examples of long-range interactions, including the famous FTO-IRX3-IRX5 long-range interactions. As well, it revealed new disease-gene associations, currently not reported in the GWAS Catalog.
This database represents the most detailed regulatory catalog in existence so far. It should empower the future functional interpretations of disease-associated variants by facilitating the precise identification of altered genes.

Using data from consortia

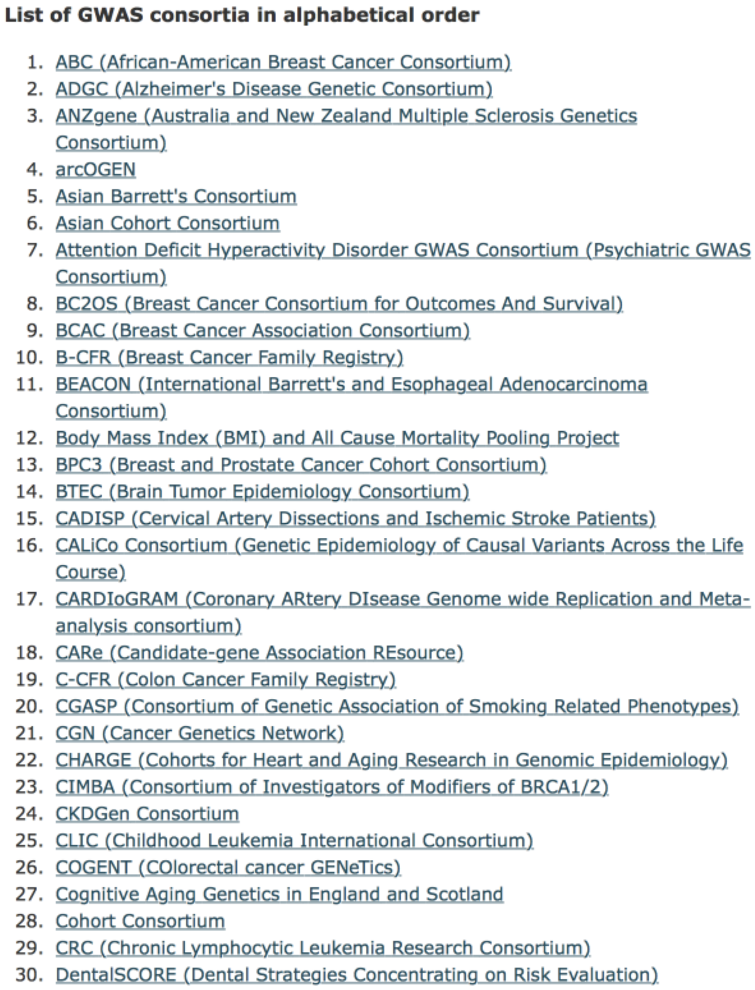
- Which types of data can you obtain from consortia? How to access and download data?
- How to work as a part of consortia? Which problems you may face?

Important Remark

- Workshops “How to use *consortium_name*” usually take ~3 days (ie https://www.encodeproject.org/tutorials/encode-meeting-2016/) , we will try to make an overview in 1 hour
- However, if you want to find more information – google “*consortium_name* workshop”
- There are separate papers (i.e. Ewan Birney, 2012, Nature, about ENCODE)

GWAS Consortia

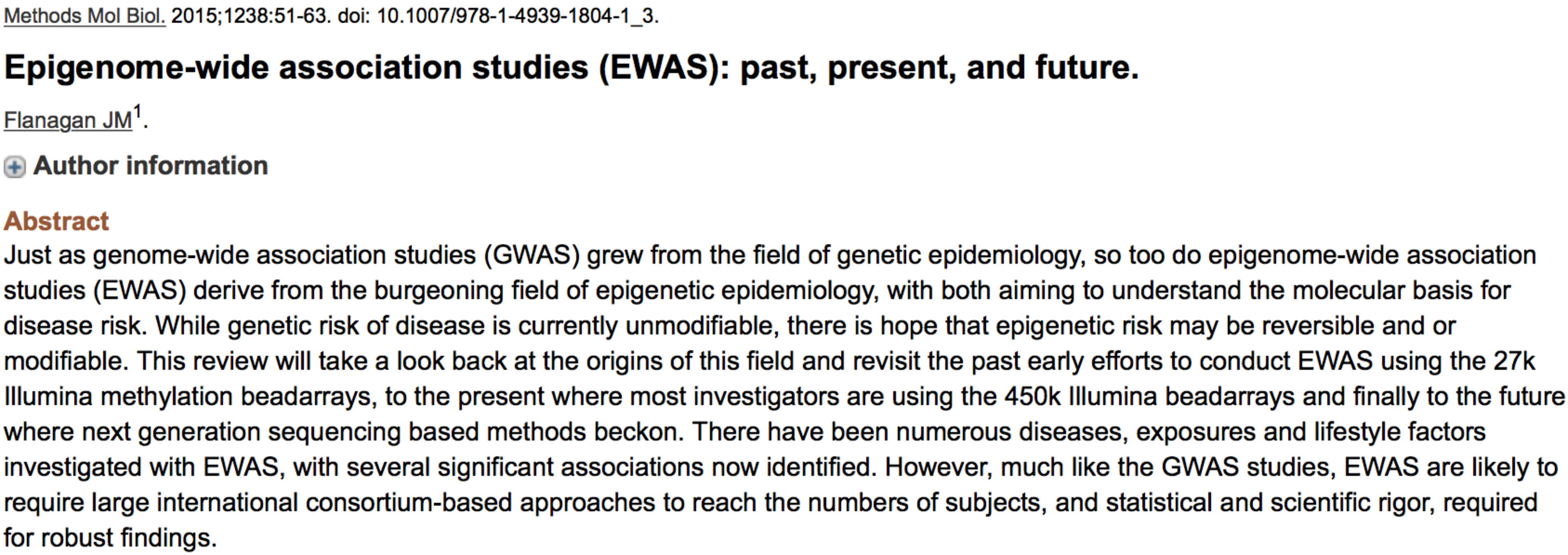
- > http://www.wikigenes.org/e/art/e/185.html
- > 500.000 genotyped people in UK



List of GWAS consortia in alphabetical order

1. ABC (African-American Breast Cancer Consortium)
2. ADGC (Alzheimer's Disease Genetic Consortium)
3. ANZgene (Australia and New Zealand Multiple Sclerosis Genetics Consortium)
4. arcOGEN
5. Asian Barrett's Consortium
6. Asian Cohort Consortium
7. Attention Deficit Hyperactivity Disorder GWAS Consortium (Psychiatric GWAS Consortium)
8. BC2OS (Breast Cancer Consortium for Outcomes And Survival)
9. BCAC (Breast Cancer Association Consortium)
10. B-CFR (Breast Cancer Family Registry)
11. BEACON (International Barrett's and Esophageal Adenocarcinoma Consortium)
12. Body Mass Index (BMI) and All Cause Mortality Pooling Project
13. BPC3 (Breast and Prostate Cancer Cohort Consortium)
14. BTEC (Brain Tumor Epidemiology Consortium)
15. CADISP (Cervical Artery Dissections and Ischemic Stroke Patients)
16. CALiCo Consortium (Genetic Epidemiology of Causal Variants Across the Life Course)
17. CARDIoGRAM (Coronary ARtery DIsease Genome wide Replication and Meta-analysis consortium)
18. CARe (Candidate-gene Association REsource)
19. C-CFR (Colon Cancer Family Registry)
20. CGASP (Consortium of Genetic Association of Smoking Related Phenotypes)
21. CGN (Cancer Genetics Network)
22. CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology)
23. CIMBA (Consortium of Investigators of Modifiers of BRCA1/2)
24. CKDGen Consortium
25. CLIC (Childhood Leukemia International Consortium)
26. COGENT (COlorectal cancer GENeTics)
27. Cognitive Aging Genetics in England and Scotland
28. Cohort Consortium
29. CRC (Chronic Lymphocytic Leukemia Research Consortium)
30. DentalSCORE (Dental Strategies Concentrating on Risk Evaluation)

EWAS Consortia



Methods Mol Biol. 2015;1238:51-63. doi: 10.1007/978-1-4939-1804-1_3.

Epigenome-wide association studies (EWAS): past, present, and future.

Flanagan JM[1].

⊕ Author information

Abstract

Just as genome-wide association studies (GWAS) grew from the field of genetic epidemiology, so too do epigenome-wide association studies (EWAS) derive from the burgeoning field of epigenetic epidemiology, with both aiming to understand the molecular basis for disease risk. While genetic risk of disease is currently unmodifiable, there is hope that epigenetic risk may be reversible and or modifiable. This review will take a look back at the origins of this field and revisit the past early efforts to conduct EWAS using the 27k Illumina methylation beadarrays, to the present where most investigators are using the 450k Illumina beadarrays and finally to the future where next generation sequencing based methods beckon. There have been numerous diseases, exposures and lifestyle factors investigated with EWAS, with several significant associations now identified. However, much like the GWAS studies, EWAS are likely to require large international consortium-based approaches to reach the numbers of subjects, and statistical and scientific rigor, required for robust findings.

Genomics Consortia

- **The Exome Aggregation Consortium**
- **1000 Genomes**
- Human Reference Genome
- International Cancer Genome Consortium
- The Cancer Genome Atlas
- **PanCancer Analysis of Whole Genomes**
- **GTEx**

Epigenomics Consortia

- **ENCODE**
- **Roadmap Epigenomics**
- **BluePrint**
- International Human Epigenome Consortium

ExAC Overivew

- > http://exac.broadinstitute.org/about
- > First thing to do – look and read flagship paper!
- > The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies.

ExAC: Why it is useful

It is used to

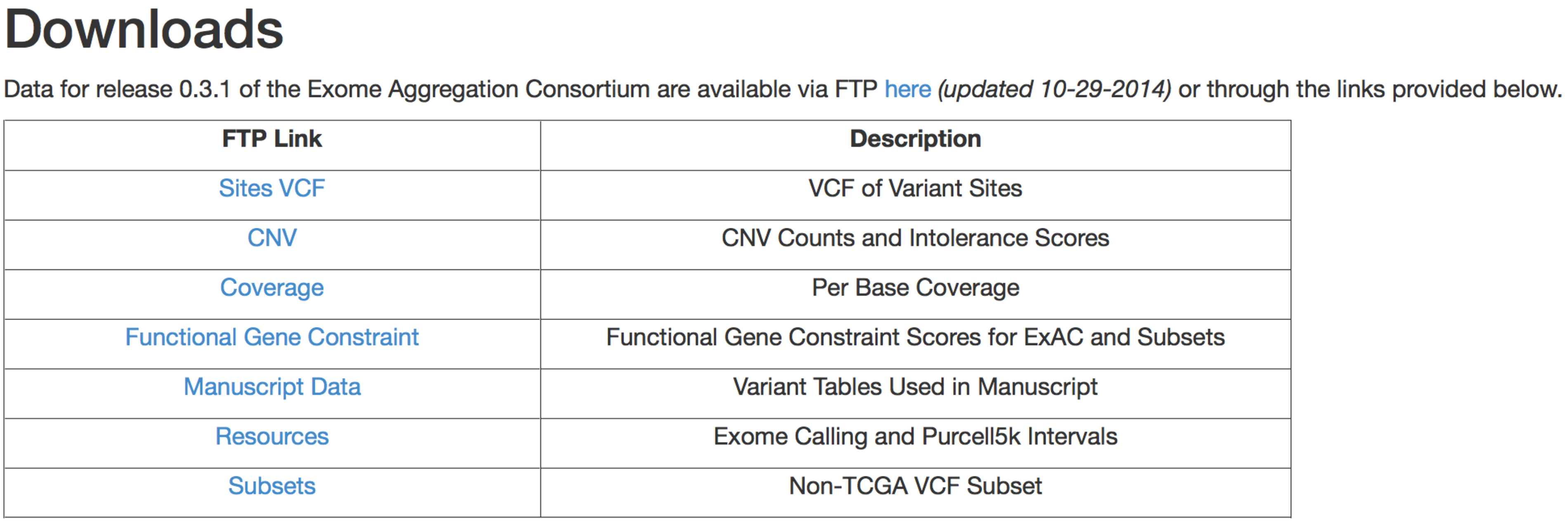
- calculate objective metrics of pathogenicity for sequence variants,
- identify genes subject to strong selection against various classes of mutation; identifying 3,230 genes with near-complete reduction of number of predicted protein-truncating variants, with 72% of these genes having no currently established human disease phenotype,
- efficient filtering of candidate disease-causing variants

ExAC: Results

- ANNOVAR and ATAV were updated using ExAC data
- CADD scores were re-calculated
- Commercial tools such as GoldenHelix and GeneTalk also incorporated ExAC data

ExAC: Download

> Download



Downloads

Data for release 0.3.1 of the Exome Aggregation Consortium are available via FTP here *(updated 10-29-2014)* or through the links provided below.

FTP Link	Description
Sites VCF	VCF of Variant Sites
CNV	CNV Counts and Intolerance Scores
Coverage	Per Base Coverage
Functional Gene Constraint	Functional Gene Constraint Scores for ExAC and Subsets
Manuscript Data	Variant Tables Used in Manuscript
Resources	Exome Calling and Purcell5k Intervals
Subsets	Non-TCGA VCF Subset

ExAC: Methods

- Flagship Paper – Methods – short description with detailed pipelines in Supplementary Information
- 91,796 individual exomes drawn from a wide range of primarily disease-focused consortia

ExAC Quality Assesment

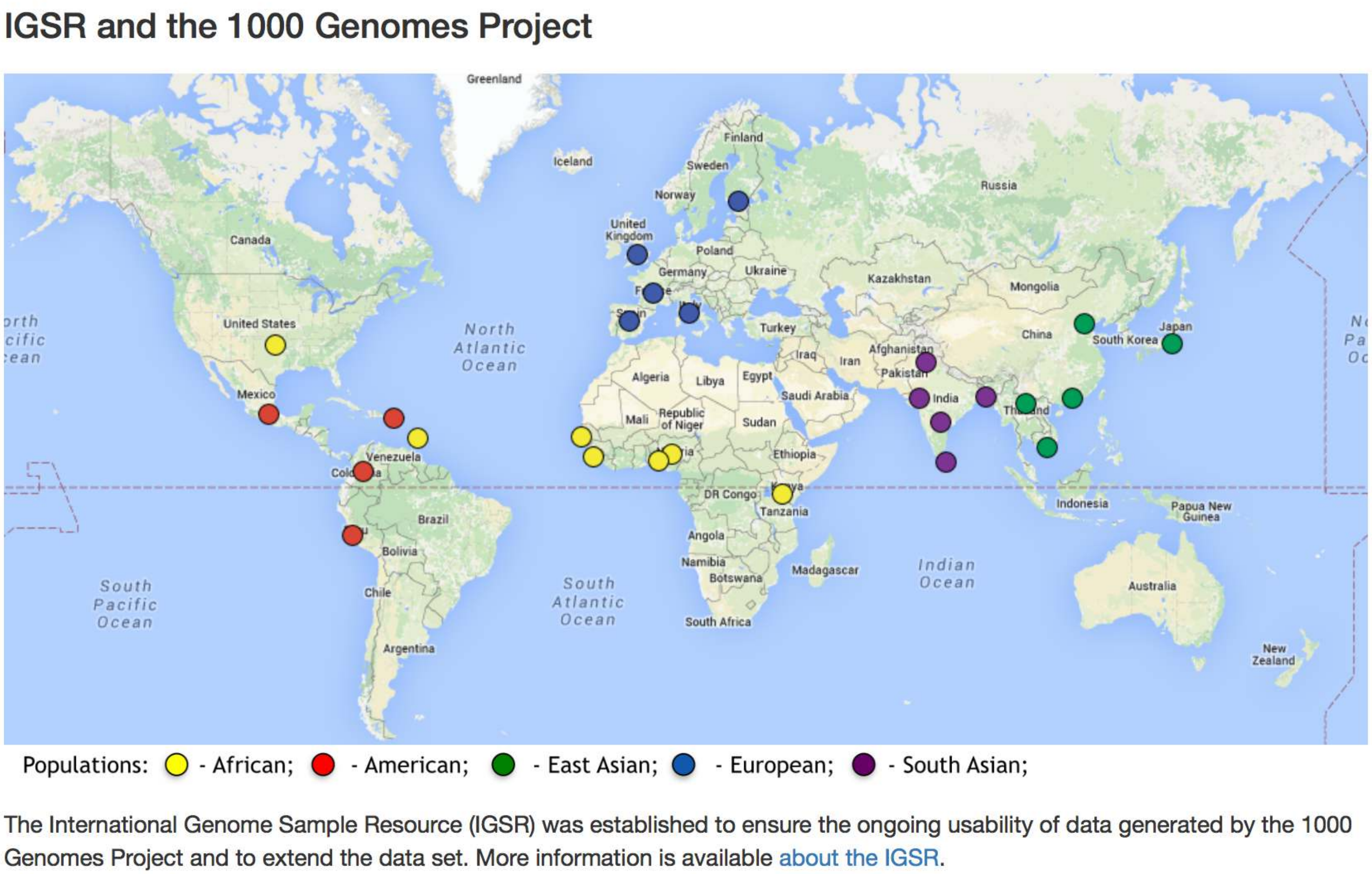
- Comparison within trios: singleton transmission rate of 50.1% (~50%)
- >10.000 samples were checked with SNP Arrays – 97-99% heterozygous concordance
- Platinum standard genome sequenced with 5 different technologies – 99.8% Sensitivity, 0.056% FDR
- Comparison with 13 WGS ~30x, PCR-free
- Indel FDR is higher (4.7%), singleton variants show higher FDR
- FDR is different for different annotation classes (missense, synonymous, protein truncating)

ExAC Sample Filtering

- Only 60.706 samples passed QC out of 91.796
- Set of common SNPs was selected (5.400) and samples with outlier heterozygosity were removed prior to PCA
- Per sample number of variants, transition/ transversion (TiTv) ratio, alternate allele heterozygous/homozygous (Het/Hom) ratio and insertion/deletion (indel) ratio
- Close relatives were removed
- Final coverage: 80% of targeted bases >20x
- 77% were enriched with Agilent Kit (33 MB target)

1000GP

> http://www.internationalgenome.org



IGSR and the 1000 Genomes Project

Populations: 🟡 - African; 🔴 - American; 🟢 - East Asian; 🔵 - European; 🟣 - South Asian;

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

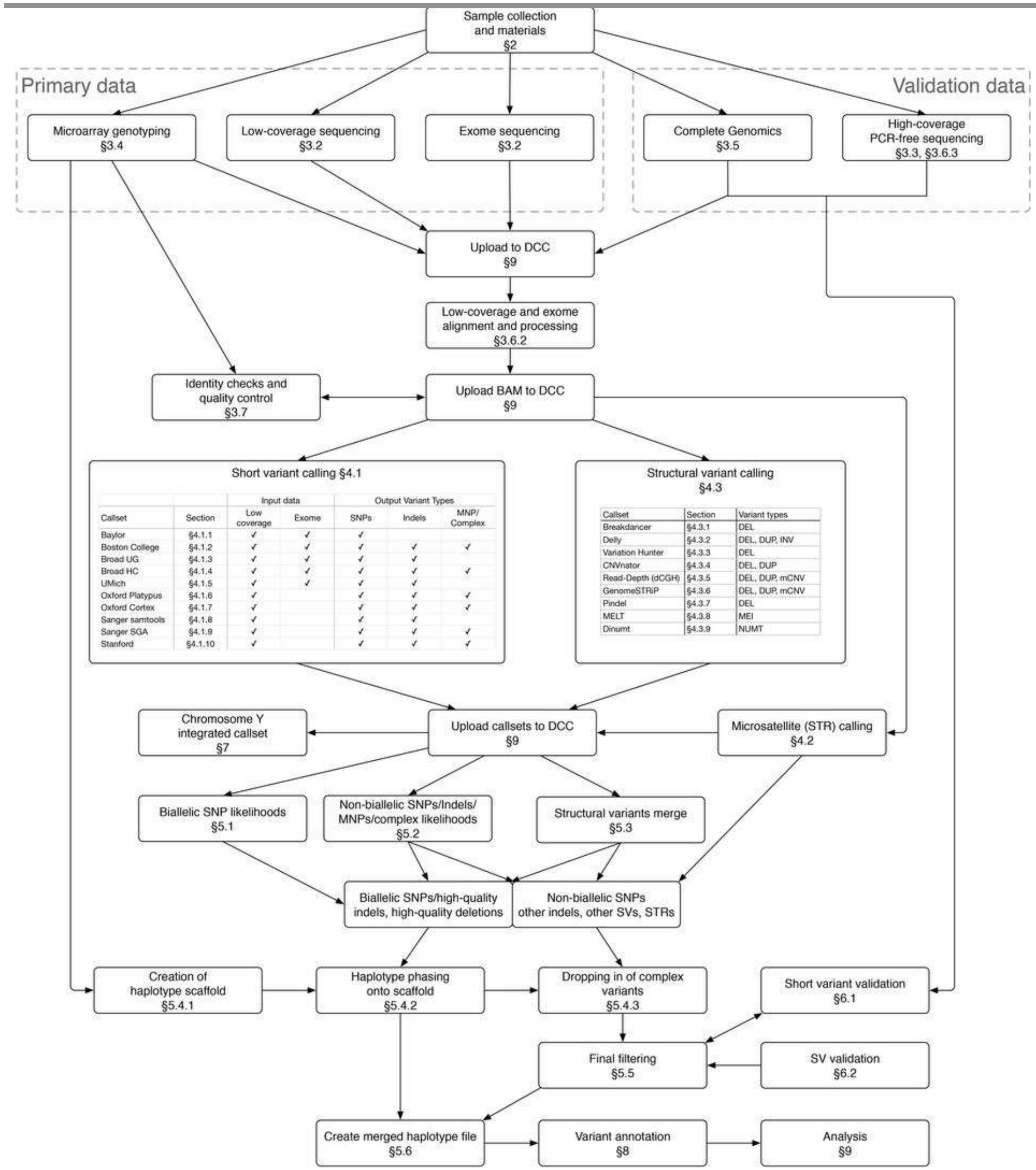
1000GP: Overview, goals

- http://www.internationalgenome.org/data-portal/sample
- Pretty convenient data portal that allows you nice filtering!
- The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied.
- The project planned to sequence each sample to 4x genome coverage; at this depth, sequencing can not discover all variants in each sample, but can allow the detection of most variants with frequencies as low as 1%.

1000GP: Main Publications

- **Pilot:** A map of human genome variation from population-scale sequencing Nature 467, 1061–1073 (28 October 2010)
- **Phase 1:** An integrated map of genetic variation from 1,092 human genomes Nature 491, 56–65 (01 November 2012)
- **Phase 3:** A global reference for human genetic variation Nature 526, 68–74 (01 October 2015)
- An integrated map of structural variation in 2,504 human genomes Nature 526, 75–81 (01 October 2015)

1000GP: Pipeline



Sample collection and materials §2

Primary data

Validation data

Microarray genotyping §3.4

Low-coverage sequencing §3.2

Exome sequencing §3.2

Complete Genomics §3.5

High-coverage PCR-free sequencing §3.3, §3.6.3

Upload to DCC §9

Low-coverage and exome alignment and processing §3.6.2

Identity checks and quality control §3.7

Upload BAM to DCC §9

Short variant calling §4.1

		Input data		Output Variant Types		
Callset	Section	Low coverage	Exome	SNPs	Indels	MNP/ Complex
Baylor	§4.1.1	✓	✓	✓		
Boston College	§4.1.2	✓	✓	✓	✓	✓
Broad UG	§4.1.3	✓	✓	✓	✓	
Broad HC	§4.1.4	✓	✓	✓	✓	✓
UMich	§4.1.5	✓	✓	✓	✓	
Oxford Platypus	§4.1.6	✓		✓	✓	✓
Oxford Cortex	§4.1.7	✓		✓	✓	✓
Sanger samtools	§4.1.8	✓		✓	✓	
Sanger SGA	§4.1.9	✓		✓	✓	✓
Stanford	§4.1.10	✓		✓	✓	✓

Structural variant calling §4.3

Callset	Section	Variant types
Breakdancer	§4.3.1	DEL
Delly	§4.3.2	DEL, DUP, INV
Variation Hunter	§4.3.3	DEL
CNVnator	§4.3.4	DEL, DUP
Read-Depth (dCGH)	§4.3.5	DEL, DUP, mCNV
GenomeSTRiP	§4.3.6	DEL, DUP, mCNV
Pindel	§4.3.7	DEL
MELT	§4.3.8	MEI
Dinumt	§4.3.9	NUMT

Chromosome Y integrated callset §7

Upload callsets to DCC §9

Microsatellite (STR) calling §4.2

Biallelic SNP likelihoods §5.1

Non-biallelic SNPs/Indels/ MNPs/complex likelihoods §5.2

Structural variants merge §5.3

Biallelic SNPs/high-quality indels, high-quality deletions

Non-biallelic SNPs other indels, other SVs, STRs

Creation of haplotype scaffold §5.4.1

Haplotype phasing onto scaffold §5.4.2

Dropping in of complex variants §5.4.3

Short variant validation §6.1

Final filtering §5.5

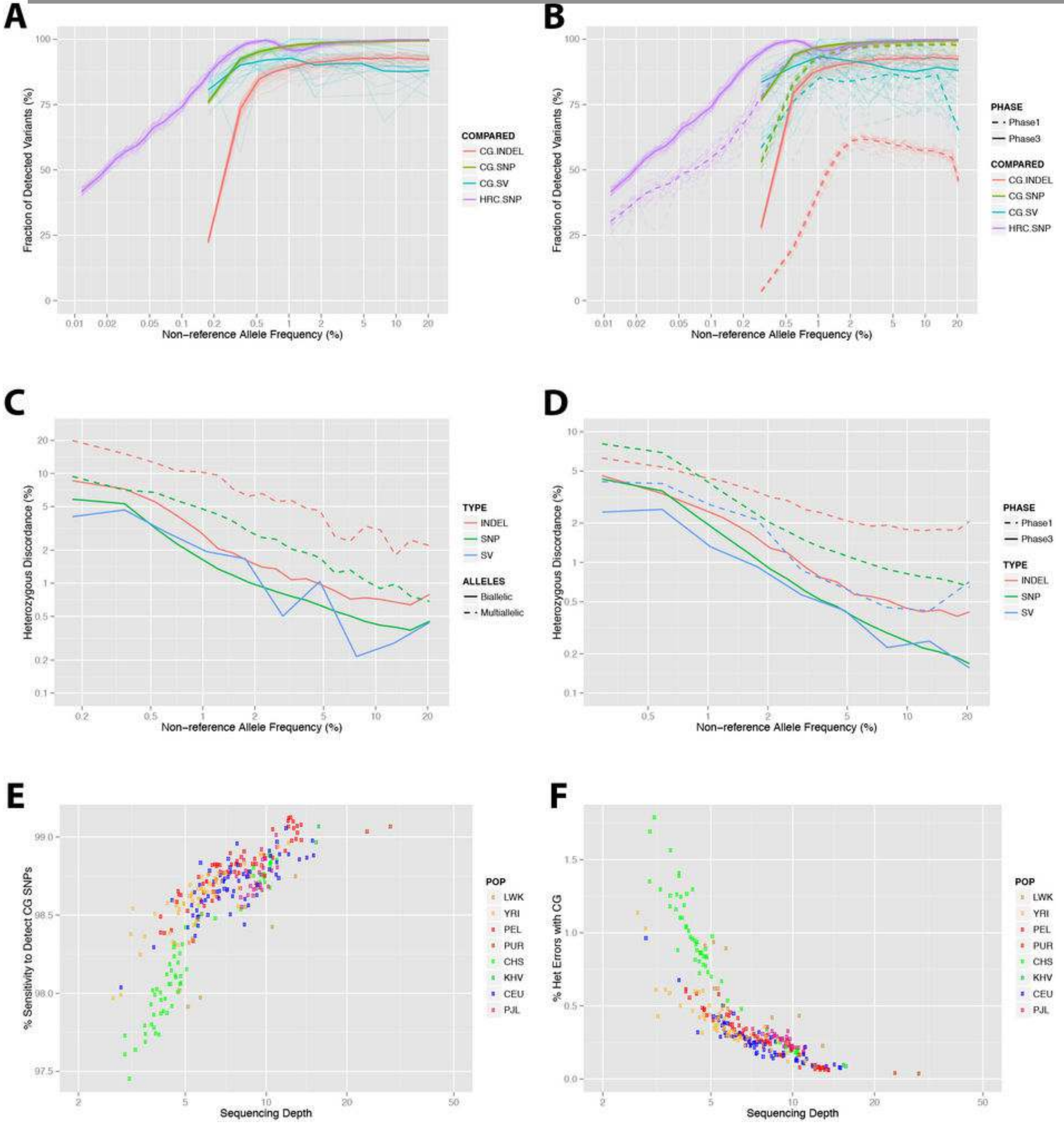
SV validation §6.2

Create merged haplotype file §5.6

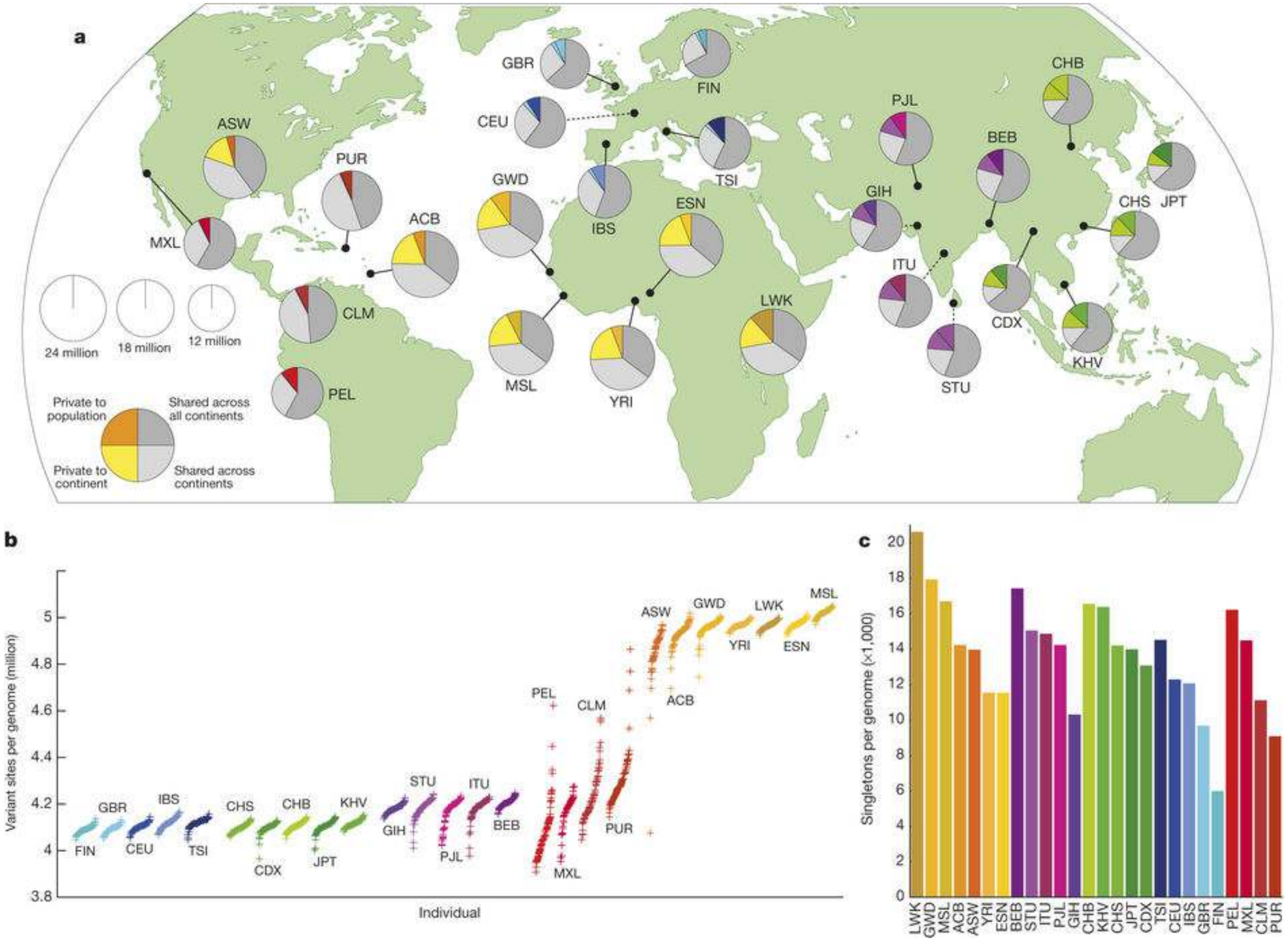
Variant annotation §8

Analysis §9

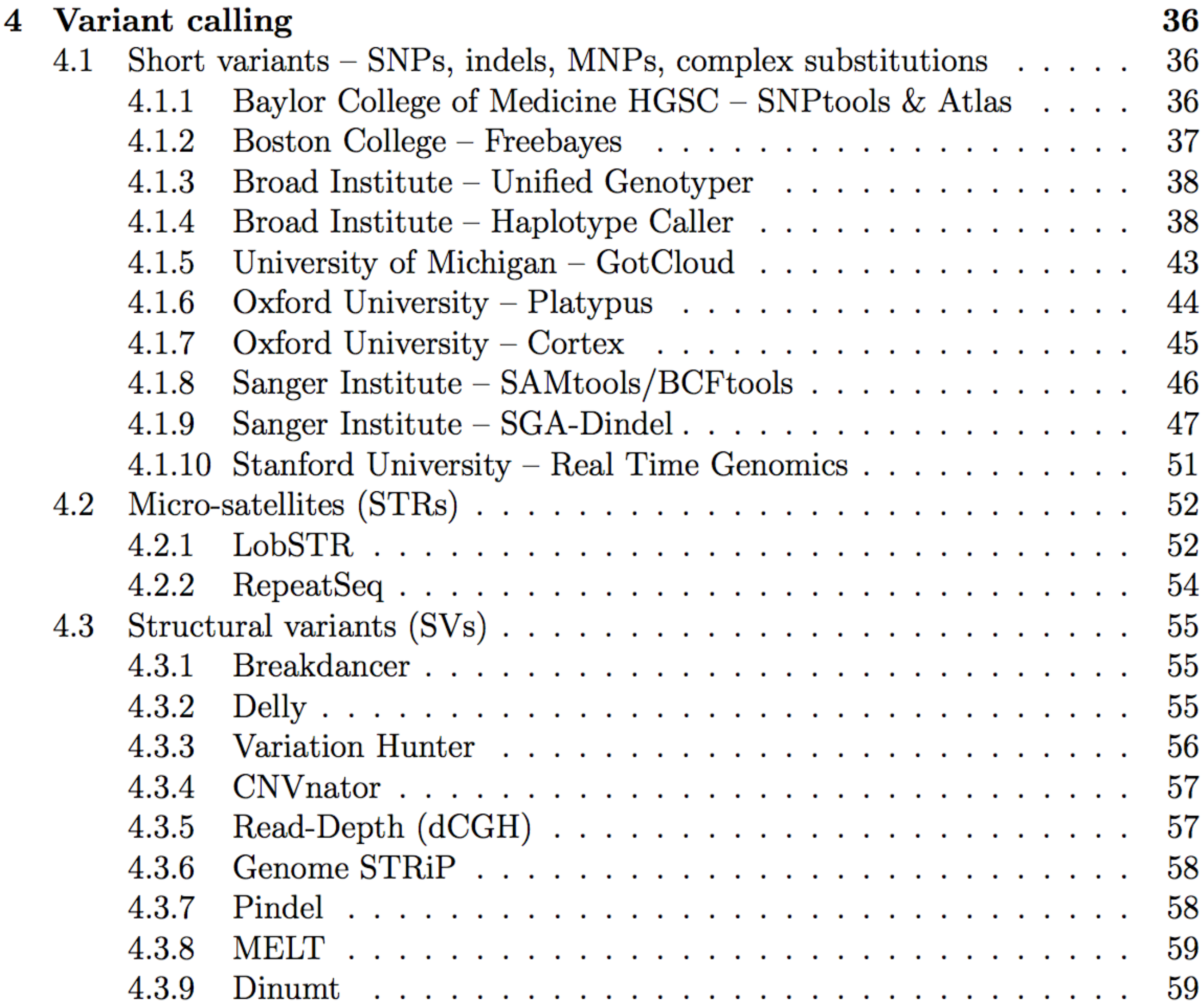
1000GP: Power of Detection, Heterozygous Discordance, Sequencing Depth



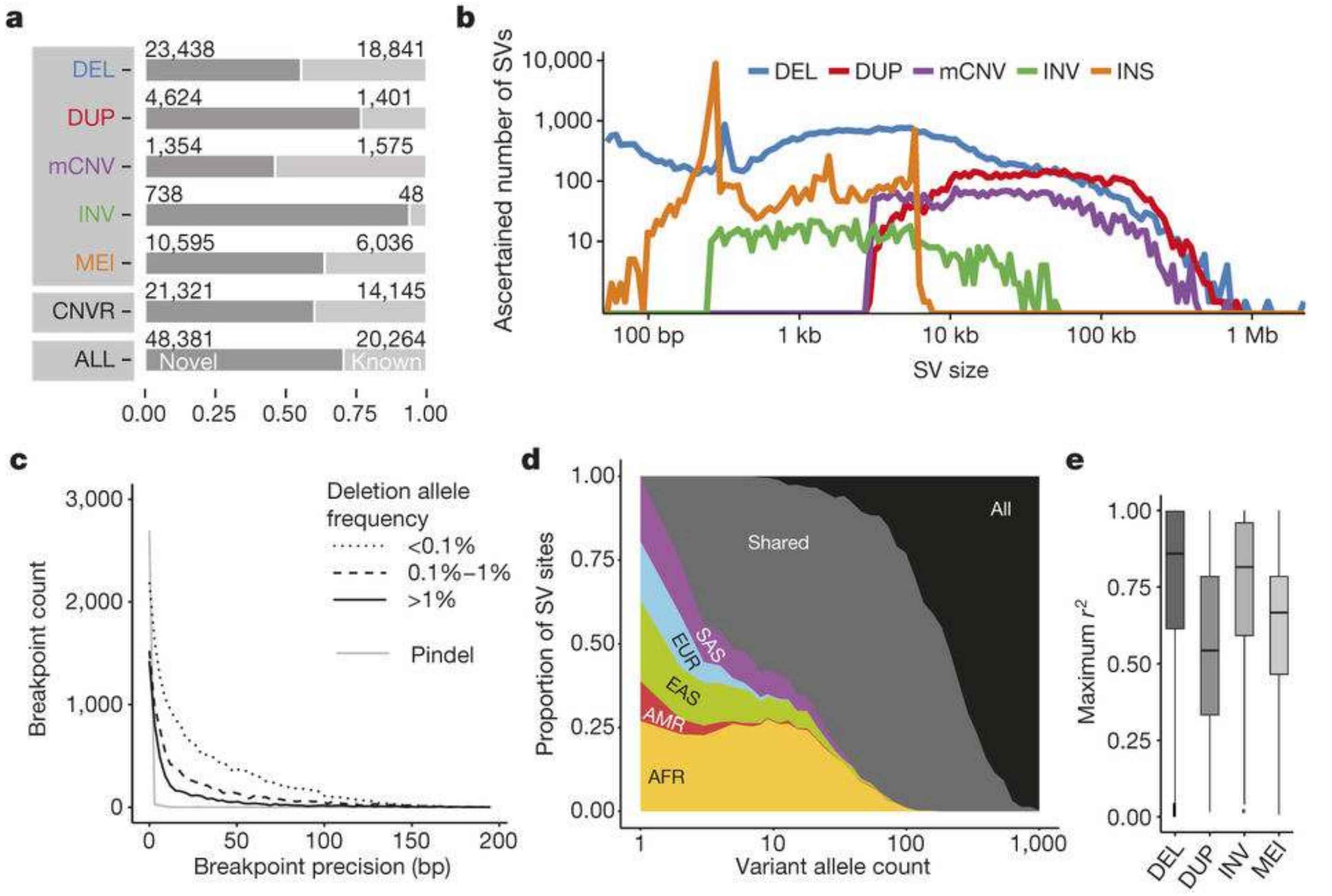
1000GP: Results



1000GP: Variant Calling



1000GP: CNVs


a
DEL
23,438
18,841
DUP
4,624
1,401
mCNV
1,354
1,575
INV
738
48
MEI
10,595
6,036
CNVR
21,321
14,145
ALL
48,381
20,264
Novel
Known
0.00
0.25
0.50
0.75
1.00
b
Ascertained number of SVs
10,000
1,000
100
10
DEL
DUP
mCNV
INV
INS
100 bp
1 kb
10 kb
100 kb
1 Mb
SV size
c
Breakpoint count
3,000
2,000
1,000
0
Deletion allele frequency
<0.1%
0.1%–1%
>1%
Pindel
0
50
100
150
200
Breakpoint precision (bp)
d
Proportion of SV sites
1.00
0.75
0.50
0.25
0.00
All
Shared
SAS
EUR
EAS
AMR
AFR
1
10
100
1,000
Variant allele count
e
Maximum r^2
1.00
0.75
0.50
0.25
0.00
DEL
DUP
INV
MEI

1000GP: CNVs concordance

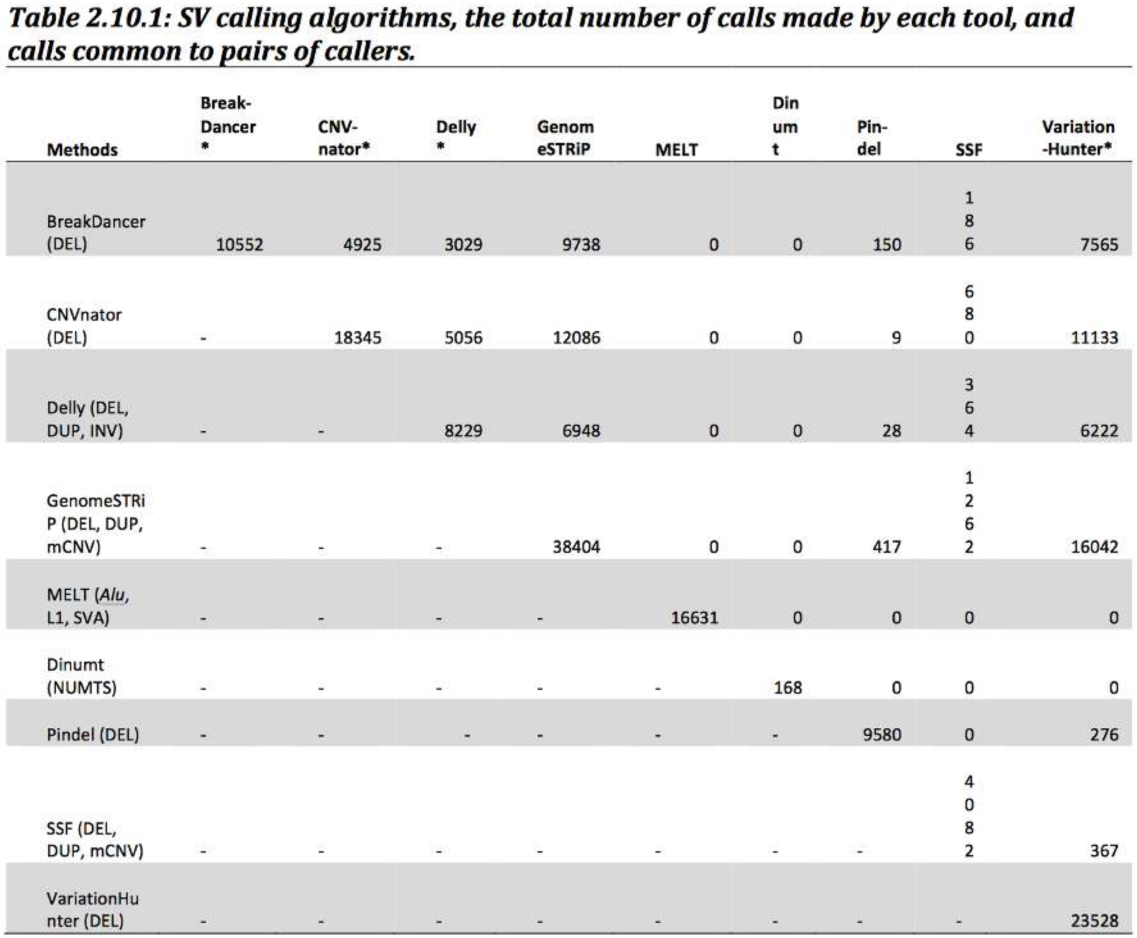
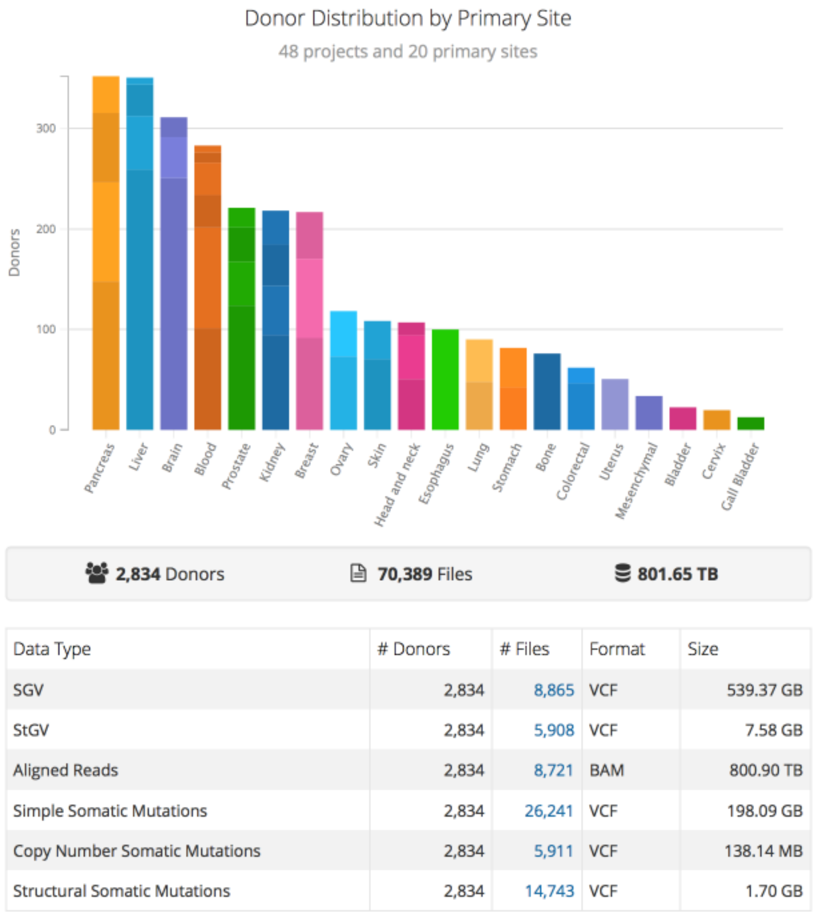


Table 2.10.1: SV calling algorithms, the total number of calls made by each tool, and calls common to pairs of callers.

Methods	BreakDancer*	CNVnator*	Delly*	GenomeSTRiP	MELT	Dinumt	Pindel	SSF	VariationHunter*
BreakDancer (DEL)	10552	4925	3029	9738	0	0	150	186	7565
CNVnator (DEL)	-	18345	5056	12086	0	0	9	680	11133
Delly (DEL, DUP, INV)	-	-	8229	6948	0	0	28	364	6222
GenomeSTRiP (DEL, DUP, mCNV)	-	-	-	38404	0	0	417	1262	16042
MELT (*Alu*, L1, SVA)	-	-	-	-	16631	0	0	0	0
Dinumt (NUMTS)	-	-	-	-	-	168	0	0	0
Pindel (DEL)	-	-	-	-	-	-	9580	0	276
SSF (DEL, DUP, mCNV)	-	-	-	-	-	-	-	4082	367
VariationHunter (DEL)	-	-	-	-	-	-	-	-	23528

Pan Cancer Analysis Of WG

> https://dcc.icgc.org/pcawg



2,834 Donors **70,389** Files **801.65 TB**

Data Type	# Donors	# Files	Format	Size
SGV	2,834	8,865	VCF	539.37 GB
StGV	2,834	5,908	VCF	7.58 GB
Aligned Reads	2,834	8,721	BAM	800.90 TB
Simple Somatic Mutations	2,834	26,241	VCF	198.09 GB
Copy Number Somatic Mutations	2,834	5,911	VCF	138.14 MB
Structural Somatic Mutations	2,834	14,743	VCF	1.70 GB

Pan Cancer Analysis Of WG

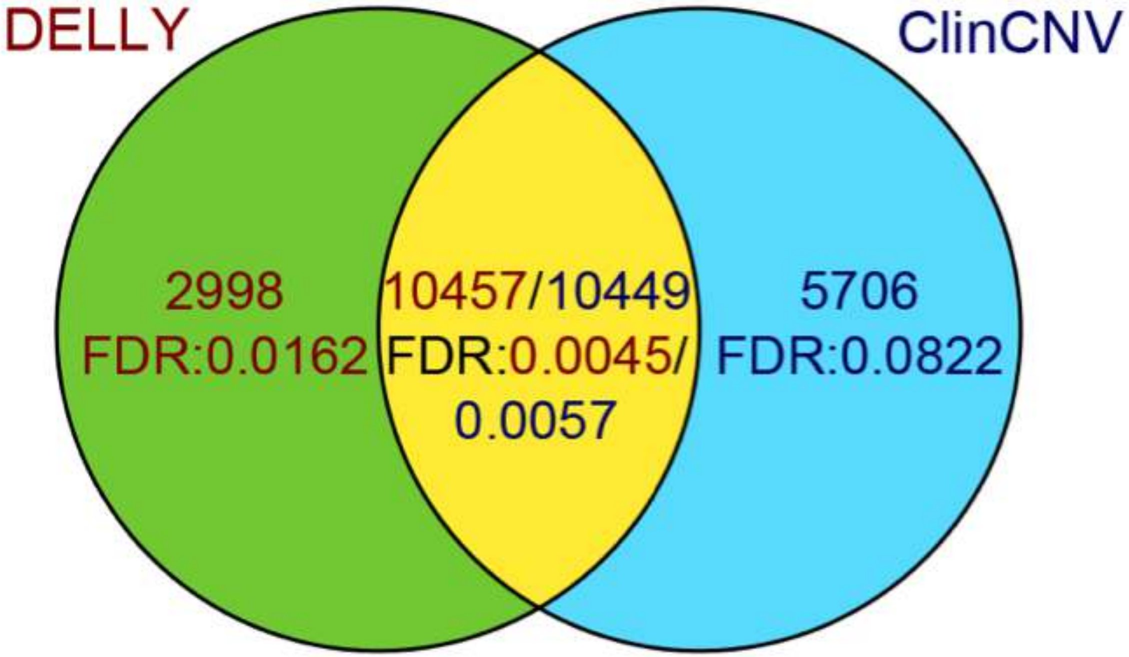
1. Novel somatic mutation calling methods
2. Analysis of mutations in regulatory regions
3. Integration of the transcriptome and genome
4. Integration of the epigenome and genome
5. Consequences of somatic mutations on pathway and network activity
6. Patterns of structural variations, signatures, genomic correlations, retrotransposons and mobile elements
7. Mutation signatures and processes
8. *Germline cancer genome*
9. Inferring driver mutations and identifying cancer genes and pathways
10. Translating cancer genomes to the clinic
11. Evolution and heterogeneity
12. Portals, visualization and software infrastructure
13. Molecular subtypes and classification
14. Analysis of mutations in non-coding RNA
15. Mitochondrial
16. Pathogens

PCAWG, WG8: Validation

- High-coverage validation
- 3 main callers: Broad Institute – Haplotype Caller, Annai-RTG (private company), Freebayes (EMBL-DKFZ)
- 50 samples, 5000 sites per sample sequenced with ~1000 depth
- ~2300 SNVs, ~2700 indels
- SNP Recall/PPV/concordance ~0.995
- Indels: 0.94 Recall, 0.91 PPV, concordance 0.88

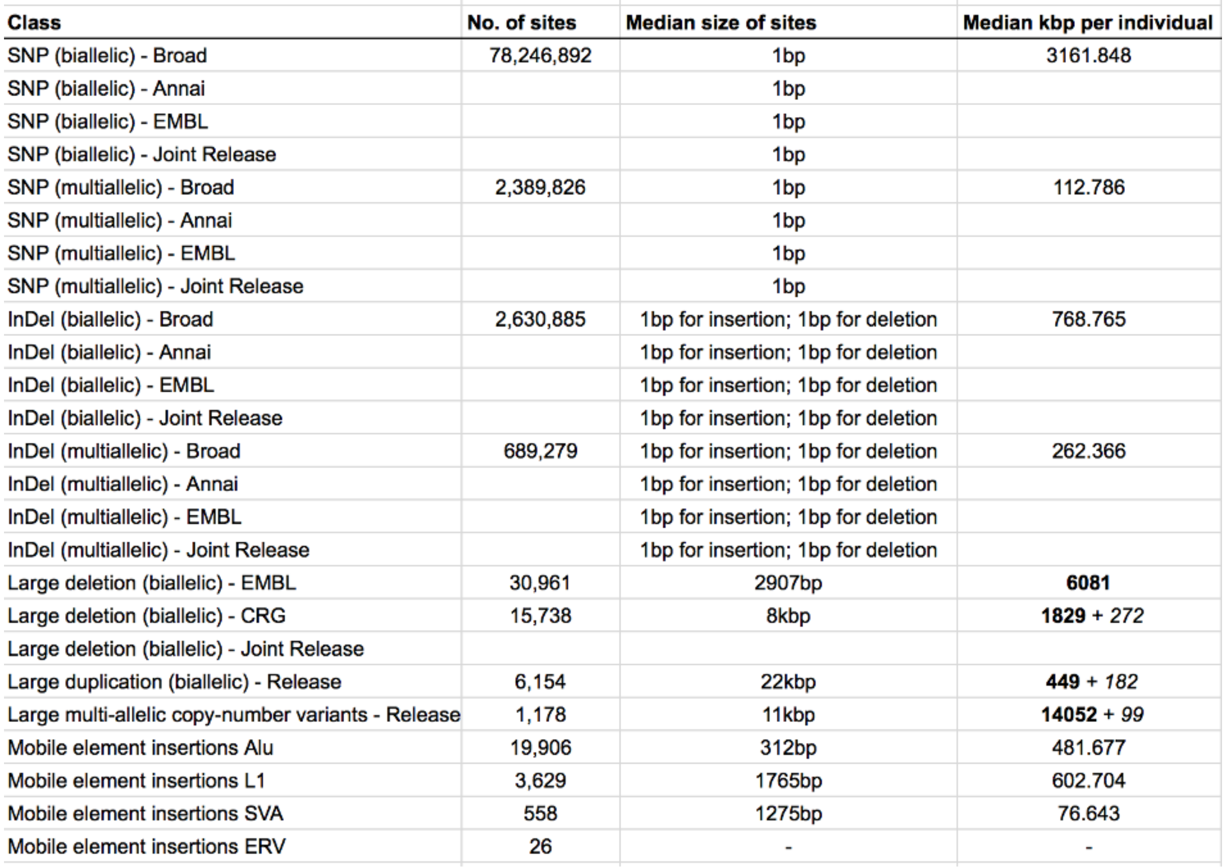
PCAWG WG8, CNVs

> CNVs



PCAWG WG8: Results

> Sensitivity, deletions only ~60%, duplications ~40%!

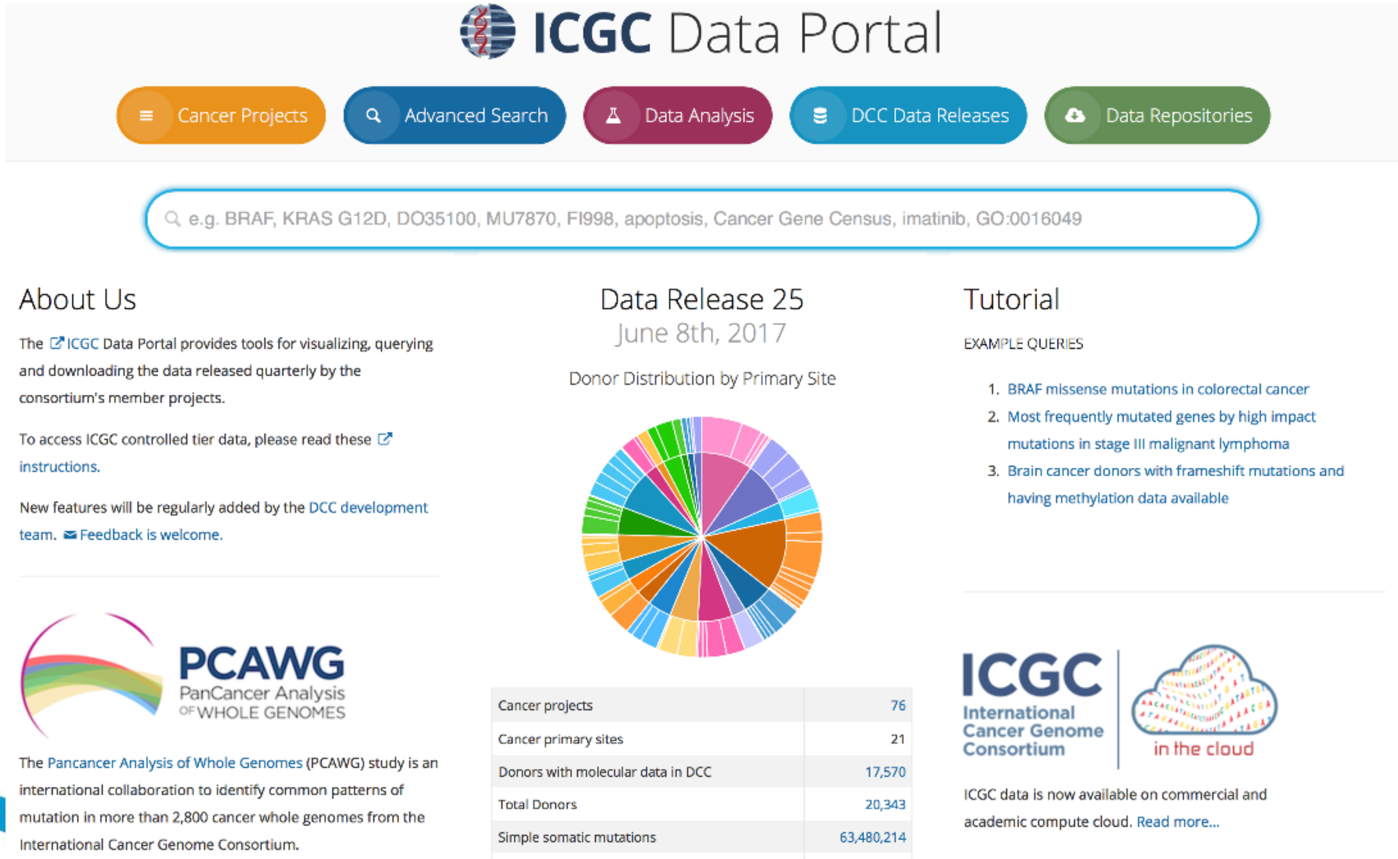


Class	No. of sites	Median size of sites	Median kbp per individual
SNP (biallelic) - Broad	78,246,892	1bp	3161.848
SNP (biallelic) - Annai		1bp	
SNP (biallelic) - EMBL		1bp	
SNP (biallelic) - Joint Release		1bp	
SNP (multiallelic) - Broad	2,389,826	1bp	112.786
SNP (multiallelic) - Annai		1bp	
SNP (multiallelic) - EMBL		1bp	
SNP (multiallelic) - Joint Release		1bp	
InDel (biallelic) - Broad	2,630,885	1bp for insertion; 1bp for deletion	768.765
InDel (biallelic) - Annai		1bp for insertion; 1bp for deletion	
InDel (biallelic) - EMBL		1bp for insertion; 1bp for deletion	
InDel (biallelic) - Joint Release		1bp for insertion; 1bp for deletion	
InDel (multiallelic) - Broad	689,279	1bp for insertion; 1bp for deletion	262.366
InDel (multiallelic) - Annai		1bp for insertion; 1bp for deletion	
InDel (multiallelic) - EMBL		1bp for insertion; 1bp for deletion	
InDel (multiallelic) - Joint Release		1bp for insertion; 1bp for deletion	
Large deletion (biallelic) - EMBL	30,961	2907bp	**6081**
Large deletion (biallelic) - CRG	15,738	8kbp	**1829** + *272*
Large deletion (biallelic) - Joint Release			
Large duplication (biallelic) - Release	6,154	22kbp	**449** + *182*
Large multi-allelic copy-number variants - Release	1,178	11kbp	**14052** + *99*
Mobile element insertions Alu	19,906	312bp	481.677
Mobile element insertions L1	3,629	1765bp	602.704
Mobile element insertions SVA	558	1275bp	76.643
Mobile element insertions ERV	26	-	-

Further Information

> Flagship paper is not informative :/

> 16 papers are released in bioRxiv



ICGC Data Portal
≡ Cancer Projects
Advanced Search
Data Analysis
DCC Data Releases
Data Repositories
e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

To access ICGC controlled tier data, please read these instructions.

New features will be regularly added by the DCC development team. Feedback is welcome.

PCAWG PanCancer Analysis of Whole Genomes

The Pancancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium.

Data Release 25

June 8th, 2017

Donor Distribution by Primary Site

Cancer projects	76
Cancer primary sites	21
Donors with molecular data in DCC	17,570
Total Donors	20,343
Simple somatic mutations	63,480,214

Tutorial

EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

ICGC International Cancer Genome Consortium in the cloud

ICGC data is now available on commercial and academic compute cloud. Read more...

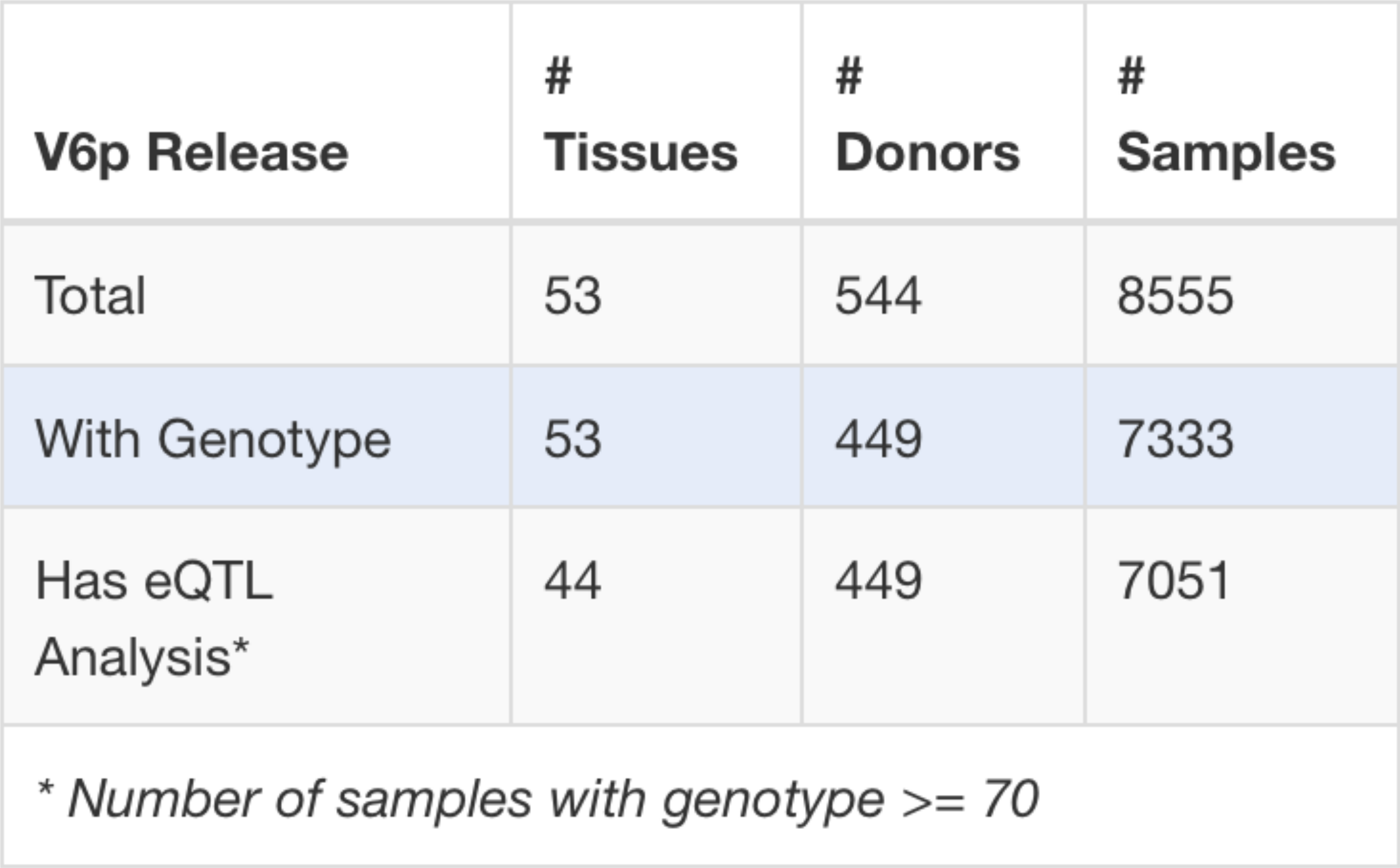
GTEx

- The Genotype-Tissue Expression project aims to provide to the scientific community a resource with which to study human gene expression and regulation and its relationship to genetic variation
- Variations in gene expression that are highly correlated with genetic variation can be identified as expression quantitative trait loci, or eQTLs

GTEx

- A lot of genetic changes associated with common human diseases, such as heart disease, cancer, diabetes, asthma, and stroke, lies outside of the protein-coding regions of genes
- The comprehensive identification of human eQTLs will greatly help to identify genes whose expression is affected by genetic variation

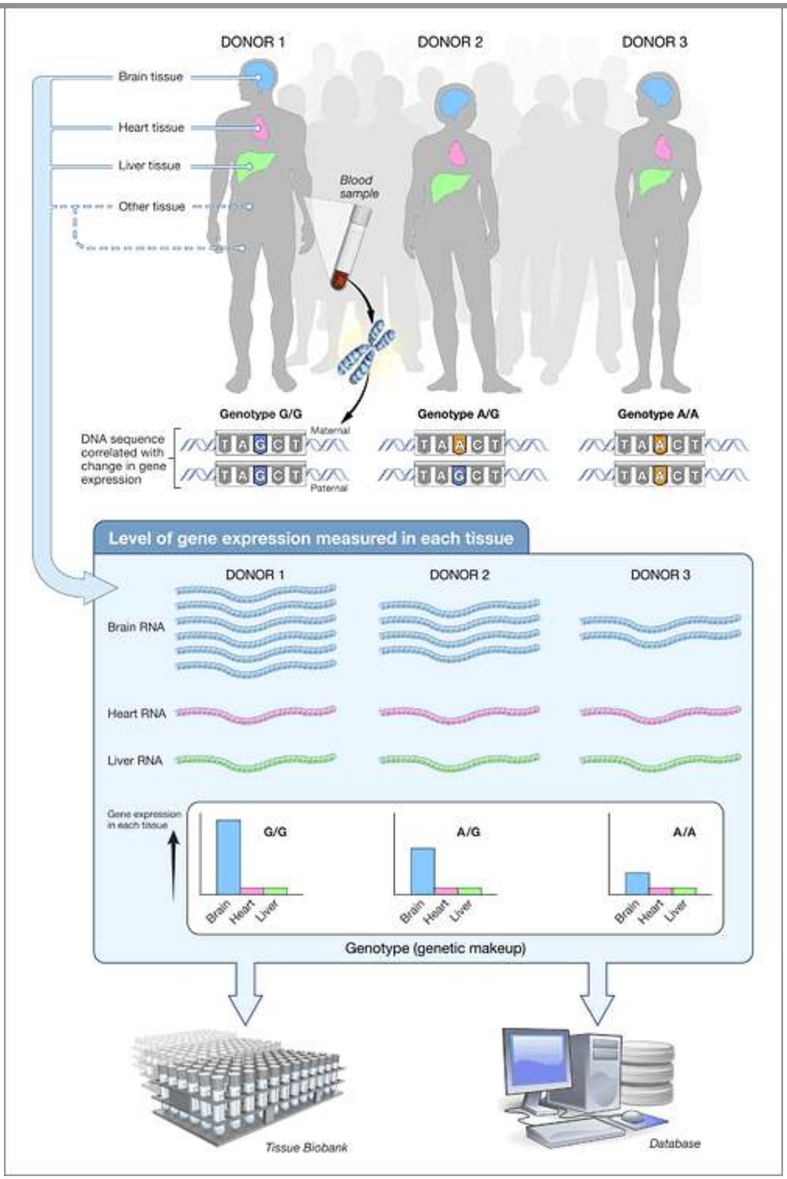
GTEx Data Overview



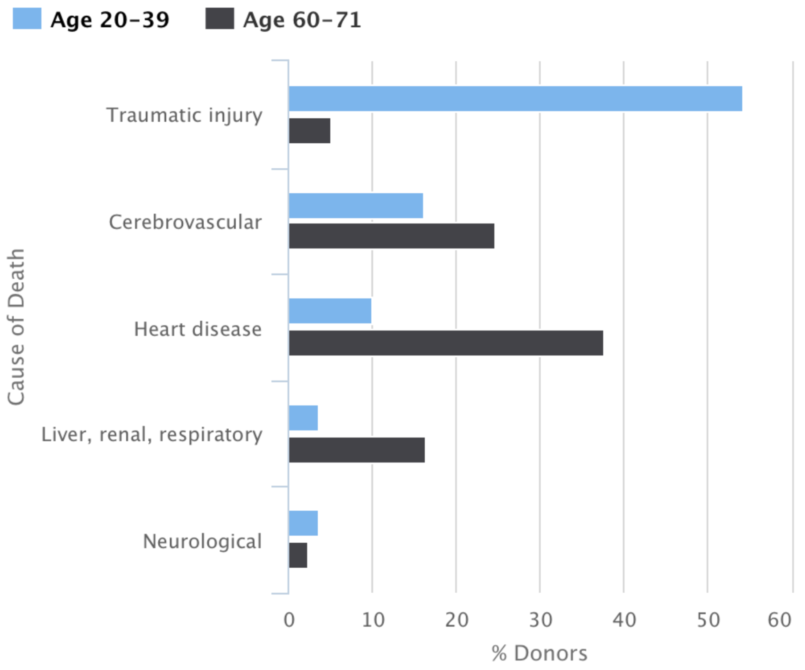
V6p Release	# Tissues	# Donors	# Samples
Total	53	544	8555
With Genotype	53	449	7333
Has eQTL Analysis*	44	449	7051

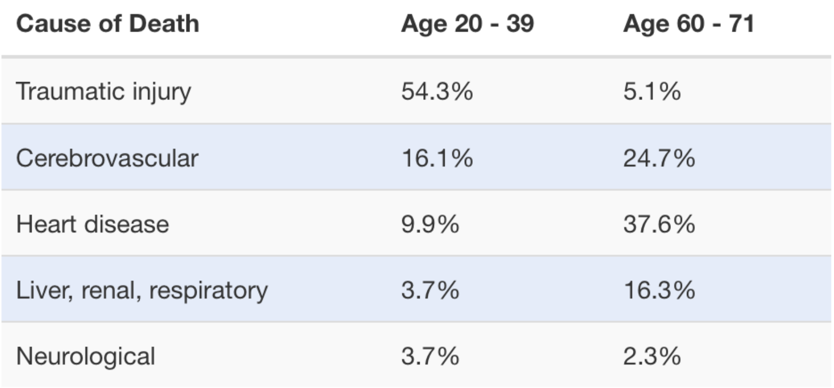
** Number of samples with genotype >= 70*

GTEx Scheme


DONOR 1
DONOR 2
DONOR 3
Brain tissue
Heart tissue
Liver tissue
Other tissue
Blood sample
Genotype G/G
Genotype A/G
Genotype A/A
DNA sequence correlated with change in gene expression
Maternal
Paternal
T A G C T
T A G C T
T A A C T
T A G C T
T A A C T
T A A C T
Level of gene expression measured in each tissue
DONOR 1
DONOR 2
DONOR 3
Brain RNA
Heart RNA
Liver RNA
Gene expression in each tissue
G/G
A/G
A/A
Brain
Heart
Liver
Genotype (genetic makeup)
Tissue Biobank
Database

GTEx: Causes of Death



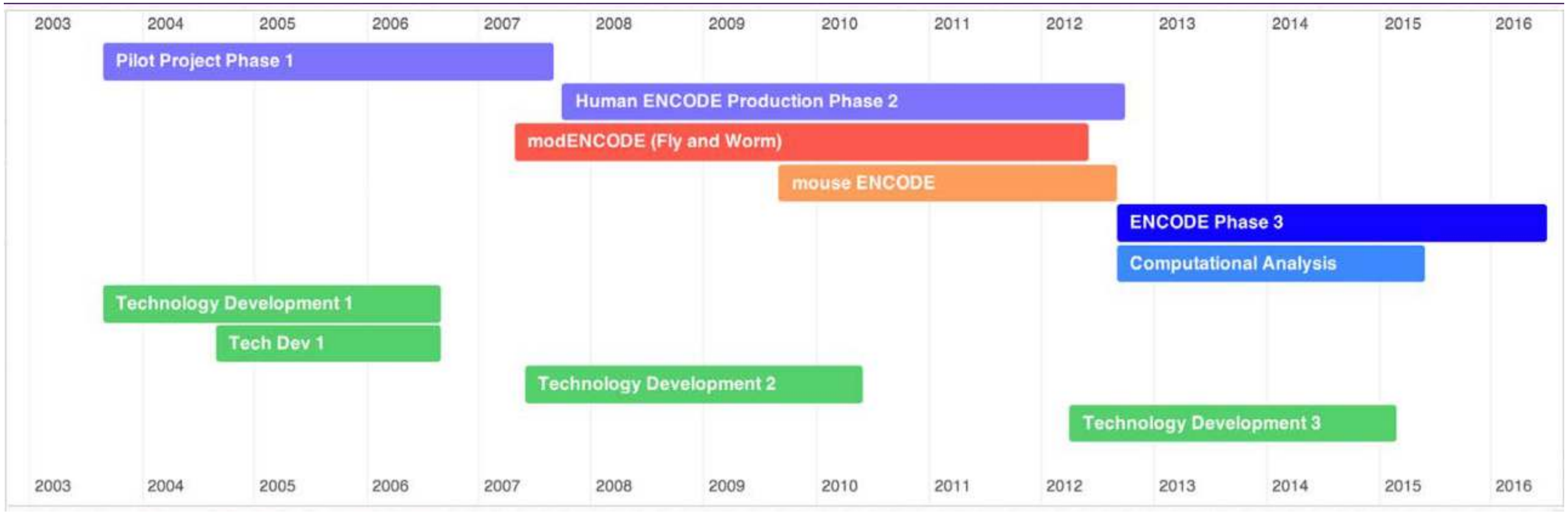


Cause of Death	Age 20 - 39	Age 60 - 71
Traumatic injury	54.3%	5.1%
Cerebrovascular	16.1%	24.7%
Heart disease	9.9%	37.6%
Liver, renal, respiratory	3.7%	16.3%
Neurological	3.7%	2.3%

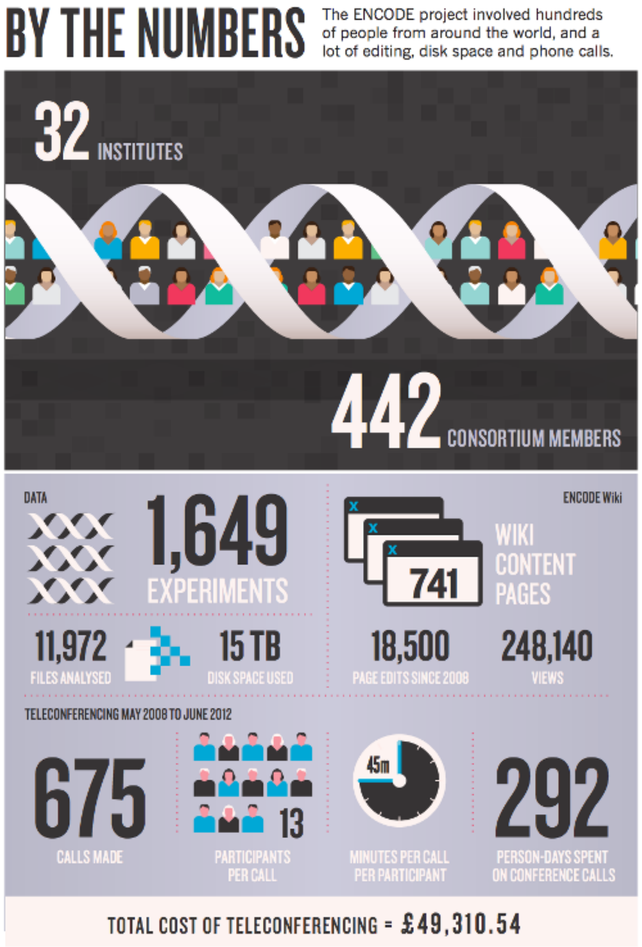
ENCODE: Overview

- https://www.encodeproject.org
- Encyclopedia of DNA elements
- The goal of ENCODE is to build a comprehensive parts list of functional elements in the human (mouse/fly/worm) genome

ENCODE Timeline



ENCODE as for 2012



BY THE NUMBERS The ENCODE project involved hundreds of people from around the world, and a lot of editing, disk space and phone calls.

32 INSTITUTES

442 CONSORTIUM MEMBERS

DATA

1,649 EXPERIMENTS

11,972 FILES ANALYSED

15 TB DISK SPACE USED

ENCODE Wiki

741 WIKI CONTENT PAGES

18,500 PAGE EDITS SINCE 2008

248,140 VIEWS

TELECONFERENCING MAY 2008 TO JUNE 2012

675 CALLS MADE

13 PARTICIPANTS PER CALL

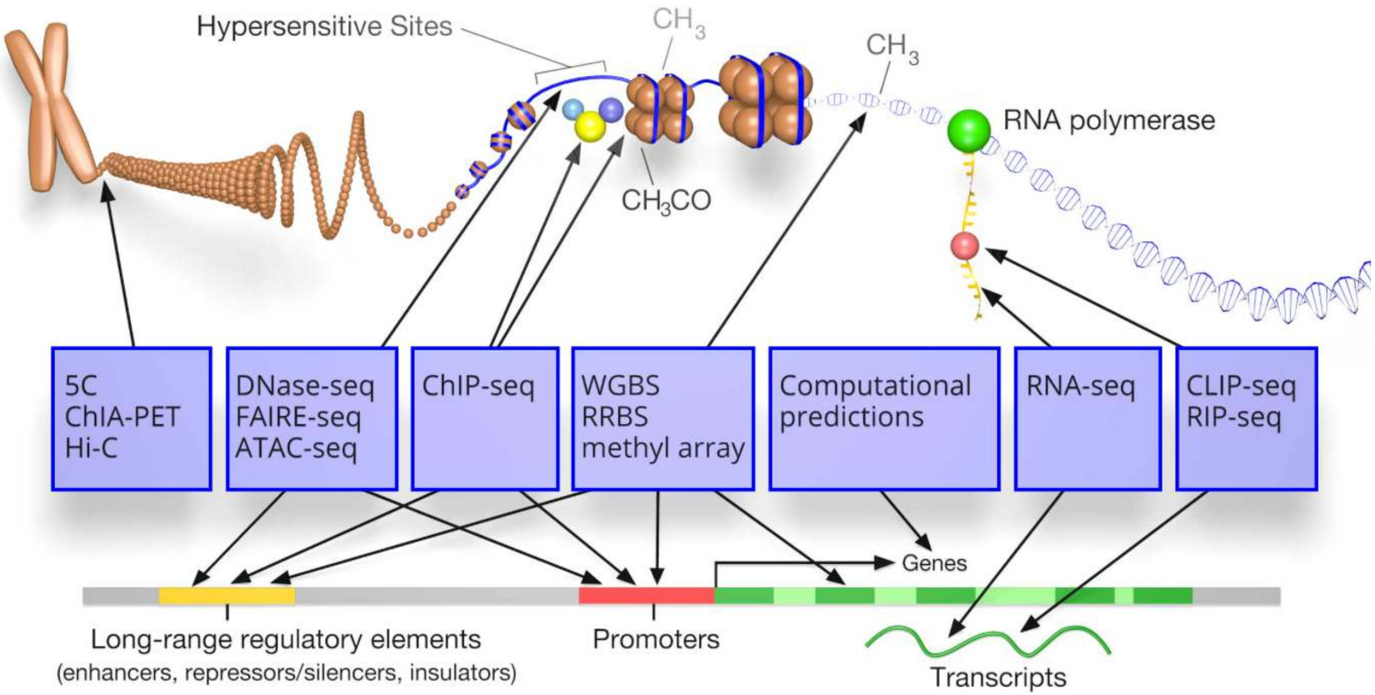
45m MINUTES PER CALL PER PARTICIPANT

292 PERSON-DAYS SPENT ON CONFERENCE CALLS

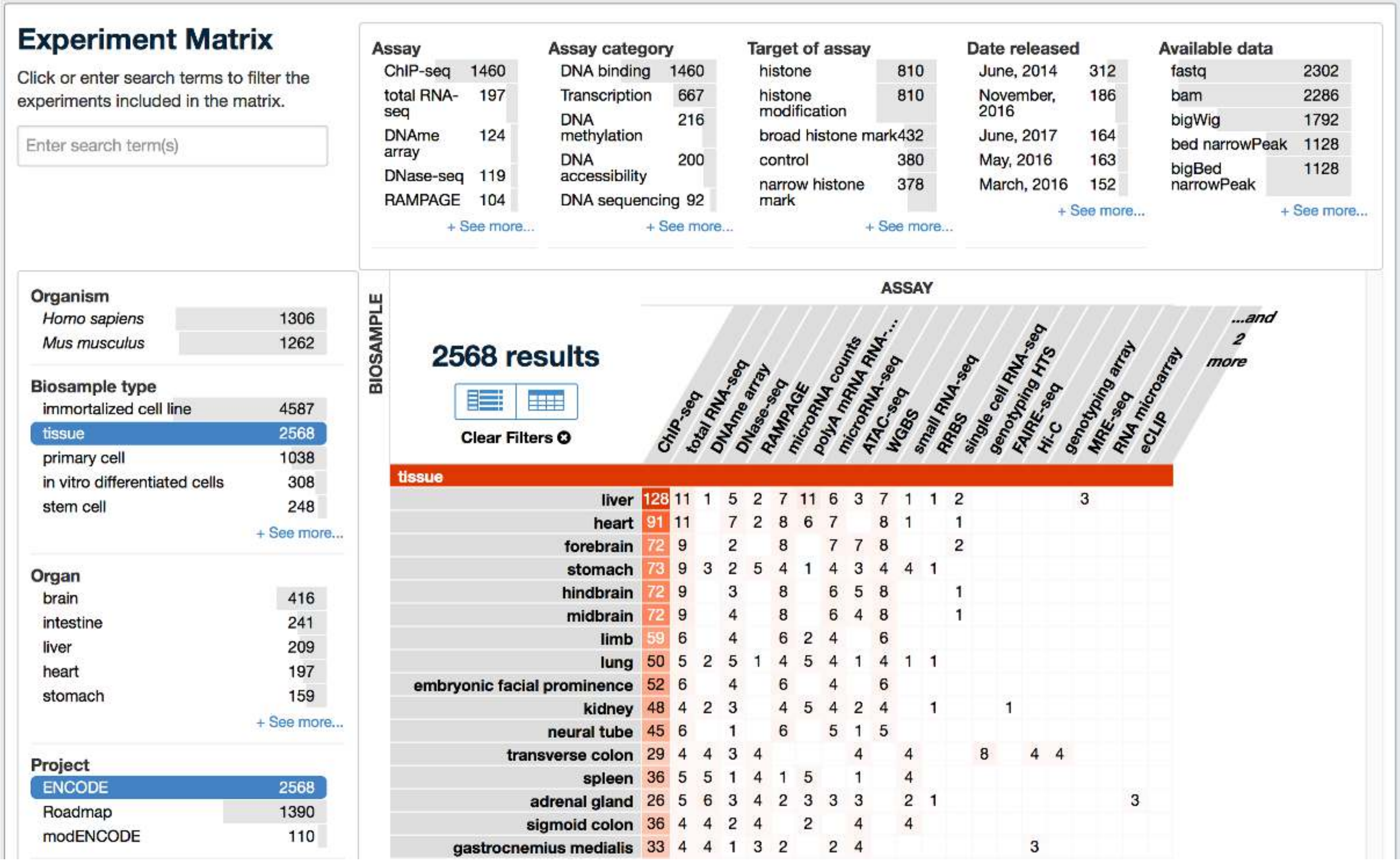
TOTAL COST OF TELECONFERENCING = £49,310.54

ENCODE: Types of Data

> https://www.encodeproject.org



ENCODE: Data Matrix



Experiment Matrix

Click or enter search terms to filter the experiments included in the matrix.

Enter search term(s)

Assay	
ChIP-seq	1460
total RNA-seq	197
DNAme array	124
DNase-seq	119
RAMPAGE	104

+ See more...

Assay category	
DNA binding	1460
Transcription	667
DNA methylation	216
DNA accessibility	200
DNA sequencing	92

+ See more...

Target of assay	
histone	810
histone modification	810
broad histone mark	432
control	380
narrow histone mark	378

+ See more...

Date released	
June, 2014	312
November, 2016	186
June, 2017	164
May, 2016	163
March, 2016	152

+ See more...

Available data	
fastq	2302
bam	2286
bigWig	1792
bed narrowPeak	1128
bigBed narrowPeak	1128

+ See more...

Organism	
Homo sapiens	1306
Mus musculus	1262

Biosample type	
immortalized cell line	4587
tissue	2568
primary cell	1038
in vitro differentiated cells	308
stem cell	248

+ See more...

Organ	
brain	416
intestine	241
liver	209
heart	197
stomach	159

+ See more...

Project	
ENCODE	2568
Roadmap	1390
modENCODE	110

BIOSAMPLE

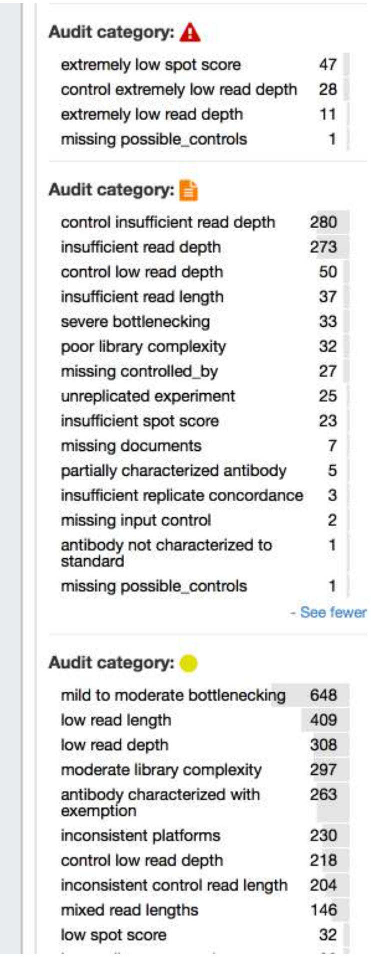
ASSAY

2568 results

Clear Filters ⊗

tissue	ChIP-seq	total RNA-seq	DNAme array	DNase-seq	RAMPAGE	microRNA counts	polyA mRNA RNA-...	microRNA-seq	ATAC-seq	WGBS	small RNA-seq	RRBS	single cell RNA-seq	genotyping HTS	FAIRE-seq	Hi-C	genotyping array	MRE-seq	RNA microarray	eCLIP	...and 2 more
liver	128	11	1	5	2	7	11	6	3	7	1	1	2					3			
heart	91	11		7	2	8	6	7		8	1		1								
forebrain	72	9		2		8		7	7	8			2								
stomach	73	9	3	2	5	4	1	4	3	4	4	1									
hindbrain	72	9		3		8		6	5	8			1								
midbrain	72	9		4		8		6	4	8			1								
limb	59	6		4		6	2	4		6											
lung	50	5	2	5	1	4	5	4	1	4	1	1									
embryonic facial prominence	52	6		4		6		4		6											
kidney	48	4	2	3		4	5	4	2	4		1			1						
neural tube	45	6		1		6		5	1	5											
transverse colon	29	4	4	3	4				4		4			8		4	4				
spleen	36	5	5	1	4	1	5		1		4										
adrenal gland	26	5	6	3	4	2	3	3	3		2	1								3	
sigmoid colon	36	4	4	2	4		2		4		4										
gastrocnemius medialis	33	4	4	1	3	2		2	4							3					

ENCODE: Audit Category



Audit category: ⚠

extremely low spot score	47
control extremely low read depth	28
extremely low read depth	11
missing possible_controls	1

Audit category: 📄

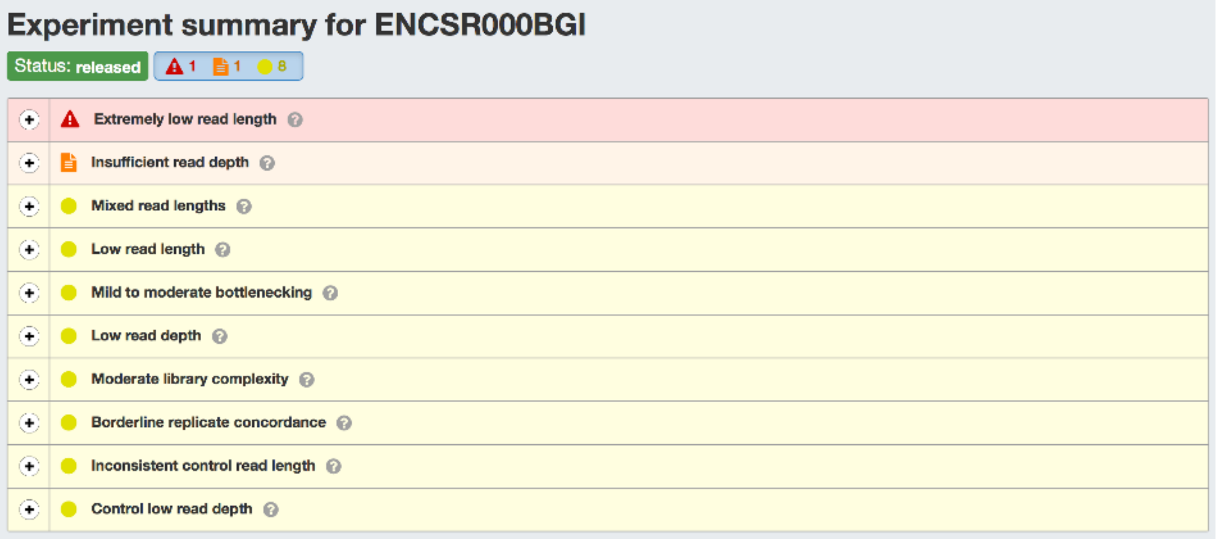
control insufficient read depth	280
insufficient read depth	273
control low read depth	50
insufficient read length	37
severe bottlenecking	33
poor library complexity	32
missing controlled_by	27
unreplicated experiment	25
insufficient spot score	23
missing documents	7
partially characterized antibody	5
insufficient replicate concordance	3
missing input control	2
antibody not characterized to standard	1
missing possible_controls	1

- See fewer

Audit category: ●

mild to moderate bottlenecking	648
low read length	409
low read depth	308
moderate library complexity	297
antibody characterized with exemption	263
inconsistent platforms	230
control low read depth	218
inconsistent control read length	204
mixed read lengths	146
low spot score	32

Each sample can have multiple QC issues and can still
Be available for downloading!

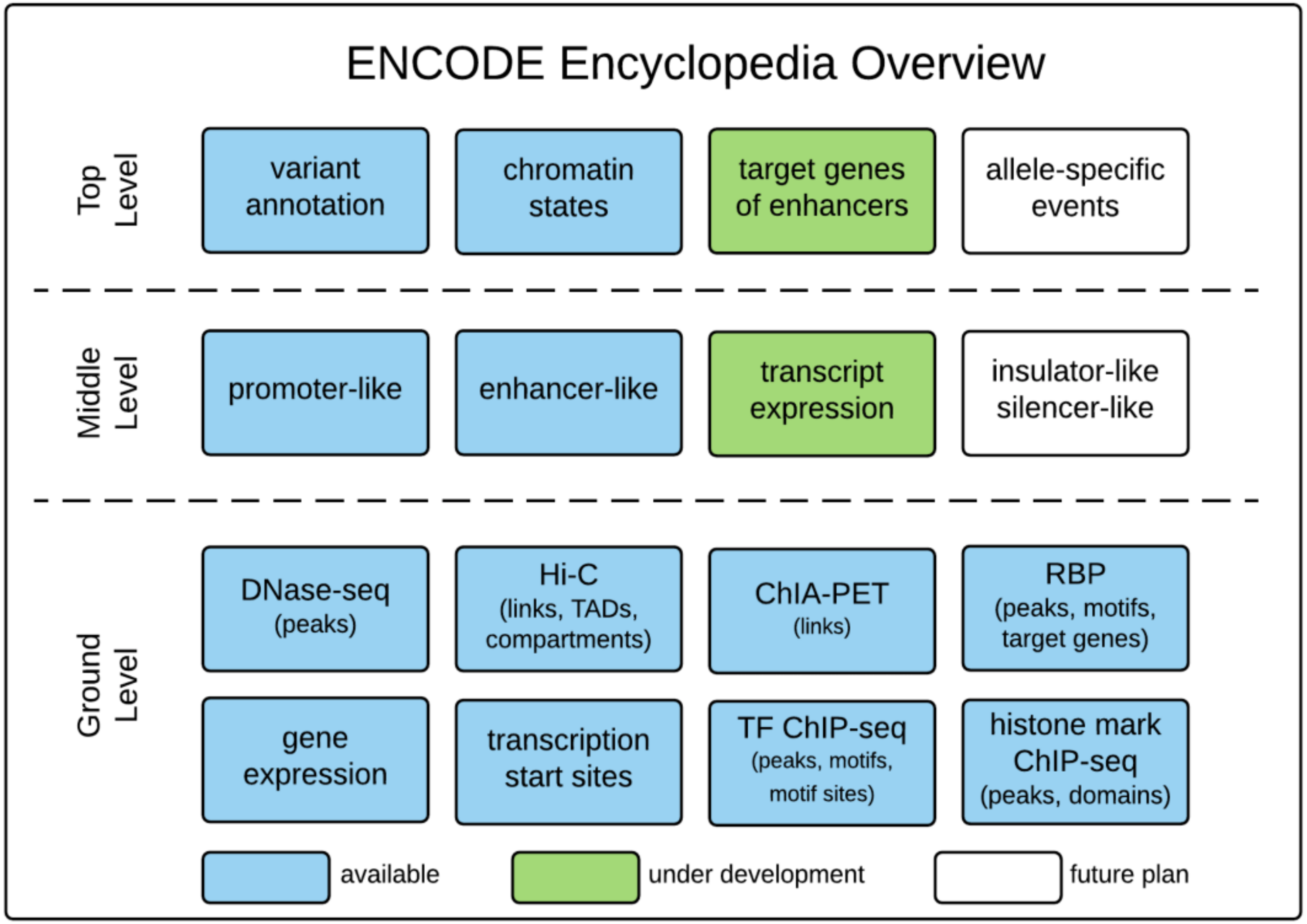


Experiment summary for ENCSR000BGI

Status: released ⚠ 1 📄 1 ● 8

- ⚠ Extremely low read length
- 📄 Insufficient read depth
- ● Mixed read lengths
- ● Low read length
- ● Mild to moderate bottlenecking
- ● Low read depth
- ● Moderate library complexity
- ● Borderline replicate concordance
- ● Inconsistent control read length
- ● Control low read depth

ENCODE: Result of Analysis

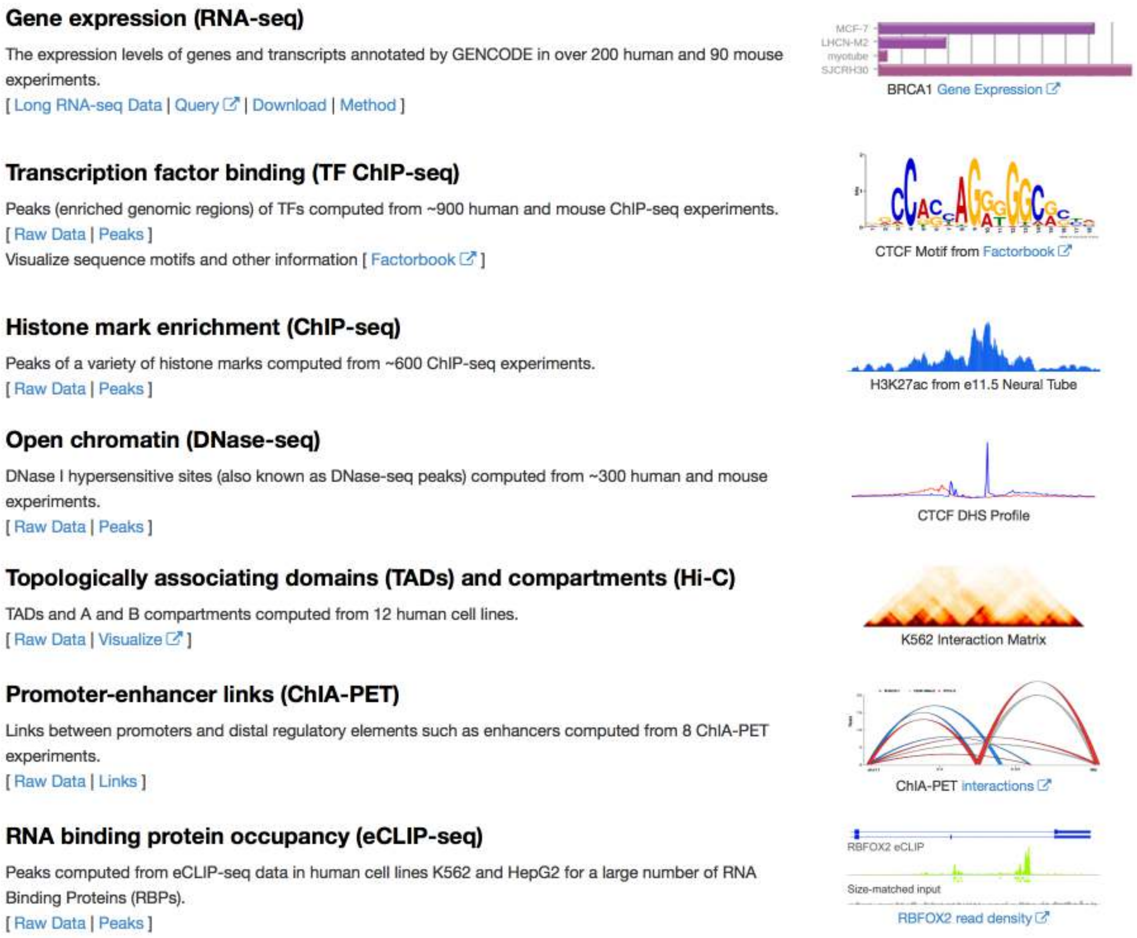


ENCODE Encyclopedia Overview

Top Level	variant annotation	chromatin states	target genes of enhancers	allele-specific events
Middle Level	promoter-like	enhancer-like	transcript expression	insulator-like silencer-like
Ground Level	DNase-seq (peaks)	Hi-C (links, TADs, compartments)	ChIA-PET (links)	RBP (peaks, motifs, target genes)
	gene expression	transcription start sites	TF ChIP-seq (peaks, motifs, motif sites)	histone mark ChIP-seq (peaks, domains)

🟦 available 🟩 under development ⬜ future plan

ENCODE: Ground Level



Gene expression (RNA-seq)

The expression levels of genes and transcripts annotated by GENCODE in over 200 human and 90 mouse experiments.

[Long RNA-seq Data | Query | Download | Method]

BRCA1 Gene Expression

Transcription factor binding (TF ChIP-seq)

Peaks (enriched genomic regions) of TFs computed from ~900 human and mouse ChIP-seq experiments.

[Raw Data | Peaks]

Visualize sequence motifs and other information [Factorbook]

CTCF Motif from Factorbook

Histone mark enrichment (ChIP-seq)

Peaks of a variety of histone marks computed from ~600 ChIP-seq experiments.

[Raw Data | Peaks]

H3K27ac from e11.5 Neural Tube

Open chromatin (DNase-seq)

DNase I hypersensitive sites (also known as DNase-seq peaks) computed from ~300 human and mouse experiments.

[Raw Data | Peaks]

CTCF DHS Profile

Topologically associating domains (TADs) and compartments (Hi-C)

TADs and A and B compartments computed from 12 human cell lines.

[Raw Data | Visualize]

K562 Interaction Matrix

Promoter-enhancer links (ChIA-PET)

Links between promoters and distal regulatory elements such as enhancers computed from 8 ChIA-PET experiments.

[Raw Data | Links]

ChIA-PET interactions

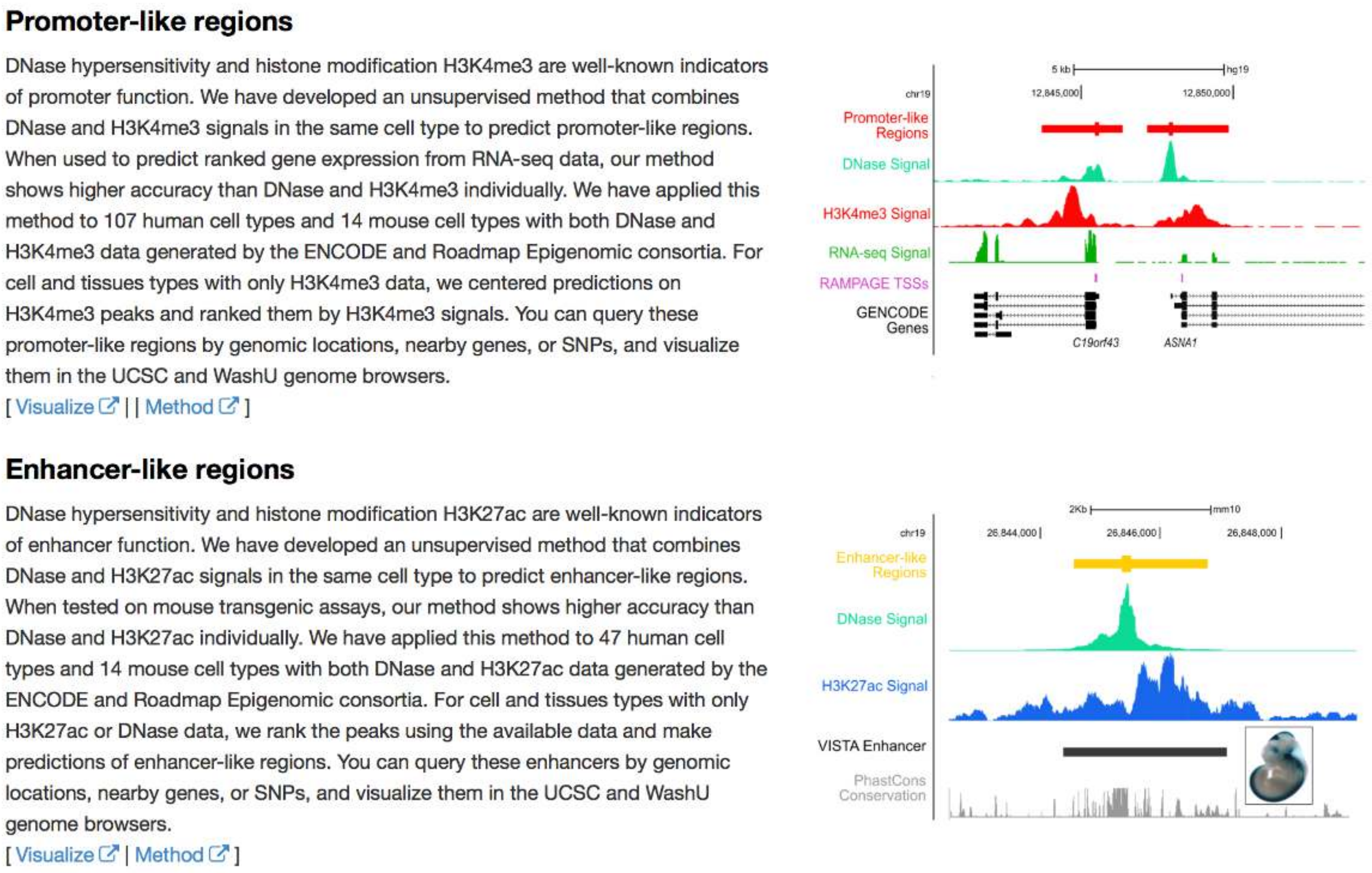
RNA binding protein occupancy (eCLIP-seq)

Peaks computed from eCLIP-seq data in human cell lines K562 and HepG2 for a large number of RNA Binding Proteins (RBPs).

[Raw Data | Peaks]

RBFOX2 read density

ENCODE: Mid-level



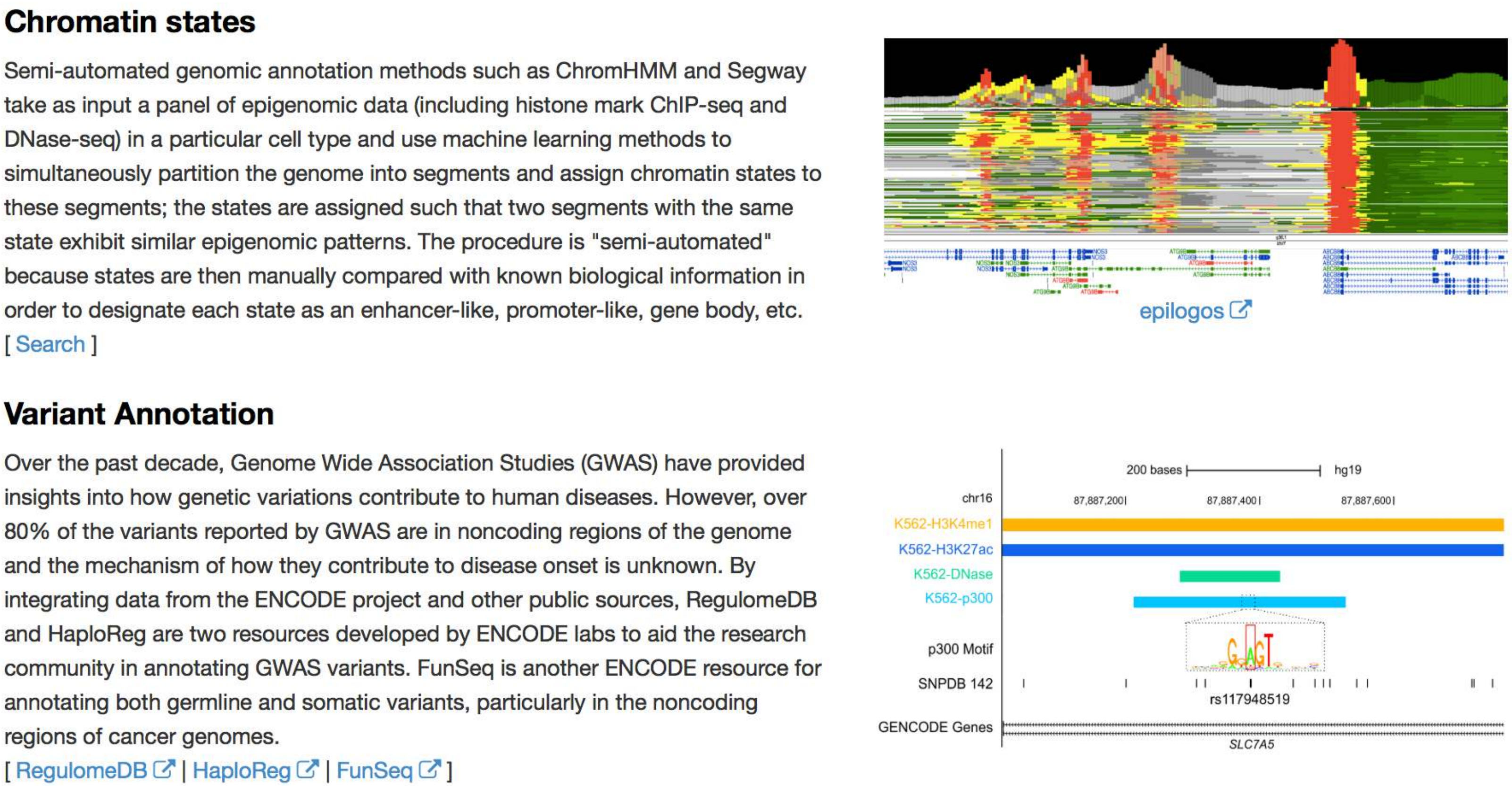
Promoter-like regions

DNase hypersensitivity and histone modification H3K4me3 are well-known indicators of promoter function. We have developed an unsupervised method that combines DNase and H3K4me3 signals in the same cell type to predict promoter-like regions. When used to predict ranked gene expression from RNA-seq data, our method shows higher accuracy than DNase and H3K4me3 individually. We have applied this method to 107 human cell types and 14 mouse cell types with both DNase and H3K4me3 data generated by the ENCODE and Roadmap Epigenomic consortia. For cell and tissues types with only H3K4me3 data, we centered predictions on H3K4me3 peaks and ranked them by H3K4me3 signals. You can query these promoter-like regions by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers.
[Visualize ↗ | | Method ↗]

Enhancer-like regions

DNase hypersensitivity and histone modification H3K27ac are well-known indicators of enhancer function. We have developed an unsupervised method that combines DNase and H3K27ac signals in the same cell type to predict enhancer-like regions. When tested on mouse transgenic assays, our method shows higher accuracy than DNase and H3K27ac individually. We have applied this method to 47 human cell types and 14 mouse cell types with both DNase and H3K27ac data generated by the ENCODE and Roadmap Epigenomic consortia. For cell and tissues types with only H3K27ac or DNase data, we rank the peaks using the available data and make predictions of enhancer-like regions. You can query these enhancers by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers.
[Visualize ↗ | Method ↗]

ENCODE: Top-Level



Chromatin states

Semi-automated genomic annotation methods such as ChromHMM and Segway take as input a panel of epigenomic data (including histone mark ChIP-seq and DNase-seq) in a particular cell type and use machine learning methods to simultaneously partition the genome into segments and assign chromatin states to these segments; the states are assigned such that two segments with the same state exhibit similar epigenomic patterns. The procedure is "semi-automated" because states are then manually compared with known biological information in order to designate each state as an enhancer-like, promoter-like, gene body, etc.
[Search]

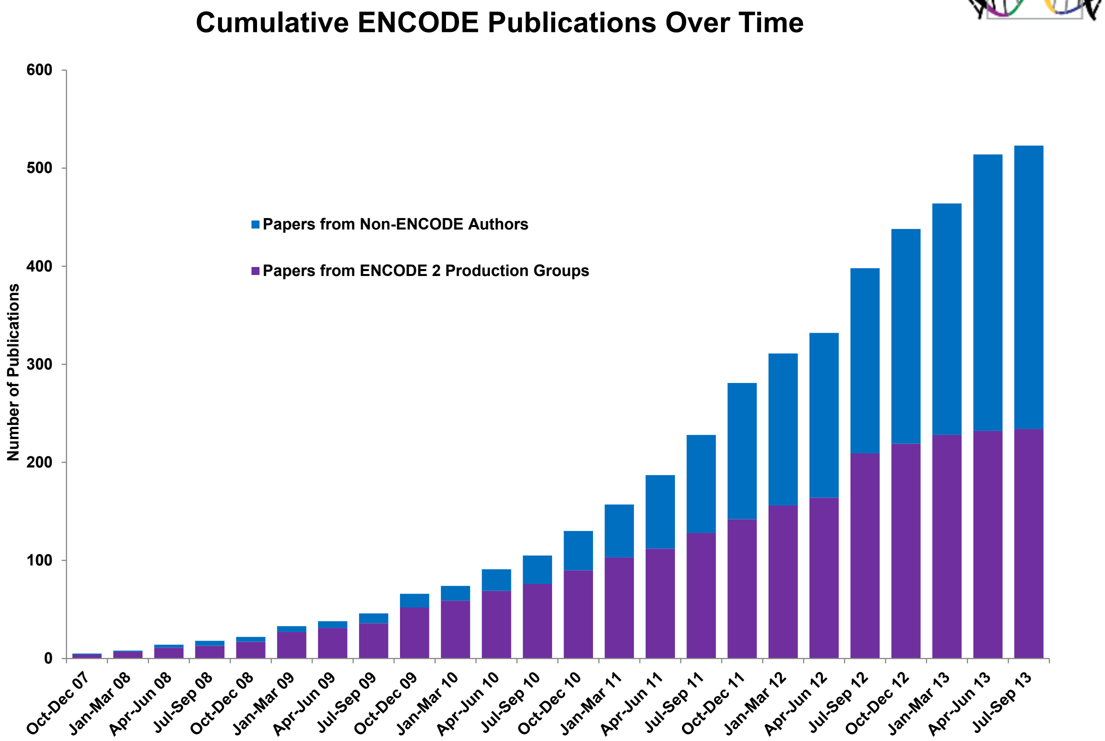
epilogos

Variant Annotation

Over the past decade, Genome Wide Association Studies (GWAS) have provided insights into how genetic variations contribute to human diseases. However, over 80% of the variants reported by GWAS are in noncoding regions of the genome and the mechanism of how they contribute to disease onset is unknown. By integrating data from the ENCODE project and other public sources, RegulomeDB and HaploReg are two resources developed by ENCODE labs to aid the research community in annotating GWAS variants. FunSeq is another ENCODE resource for annotating both germline and somatic variants, particularly in the noncoding regions of cancer genomes.
[RegulomeDB | HaploReg | FunSeq]

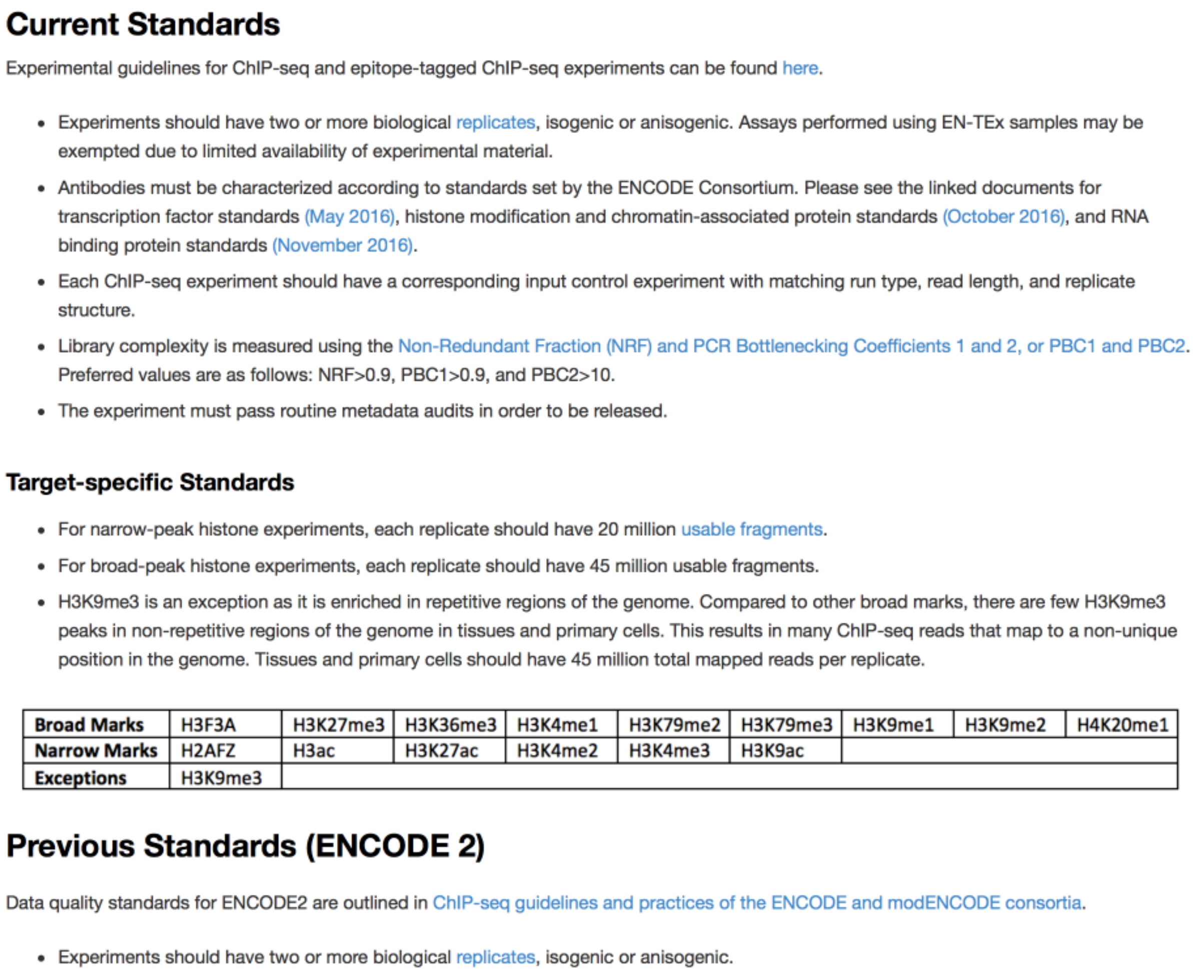
ENCODE publications

- Of course, one of the products is publicaitons!



ENCODE standards

> Data Standards



Current Standards

Experimental guidelines for ChIP-seq and epitope-tagged ChIP-seq experiments can be found here.

- Experiments should have two or more biological replicates, isogenic or anisogenic. Assays performed using EN-TEx samples may be exempted due to limited availability of experimental material.
- Antibodies must be characterized according to standards set by the ENCODE Consortium. Please see the linked documents for transcription factor standards (May 2016), histone modification and chromatin-associated protein standards (October 2016), and RNA binding protein standards (November 2016).
- Each ChIP-seq experiment should have a corresponding input control experiment with matching run type, read length, and replicate structure.
- Library complexity is measured using the Non-Redundant Fraction (NRF) and PCR Bottlenecking Coefficients 1 and 2, or PBC1 and PBC2. Preferred values are as follows: NRF>0.9, PBC1>0.9, and PBC2>10.
- The experiment must pass routine metadata audits in order to be released.

Target-specific Standards

- For narrow-peak histone experiments, each replicate should have 20 million usable fragments.
- For broad-peak histone experiments, each replicate should have 45 million usable fragments.
- H3K9me3 is an exception as it is enriched in repetitive regions of the genome. Compared to other broad marks, there are few H3K9me3 peaks in non-repetitive regions of the genome in tissues and primary cells. This results in many ChIP-seq reads that map to a non-unique position in the genome. Tissues and primary cells should have 45 million total mapped reads per replicate.

Broad Marks	H3F3A	H3K27me3	H3K36me3	H3K4me1	H3K79me2	H3K79me3	H3K9me1	H3K9me2	H4K20me1
Narrow Marks	H2AFZ	H3ac	H3K27ac	H3K4me2	H3K4me3	H3K9ac			
Exceptions	H3K9me3								

Previous Standards (ENCODE 2)

Data quality standards for ENCODE2 are outlined in ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.

- Experiments should have two or more biological replicates, isogenic or anisogenic.

BluePrint

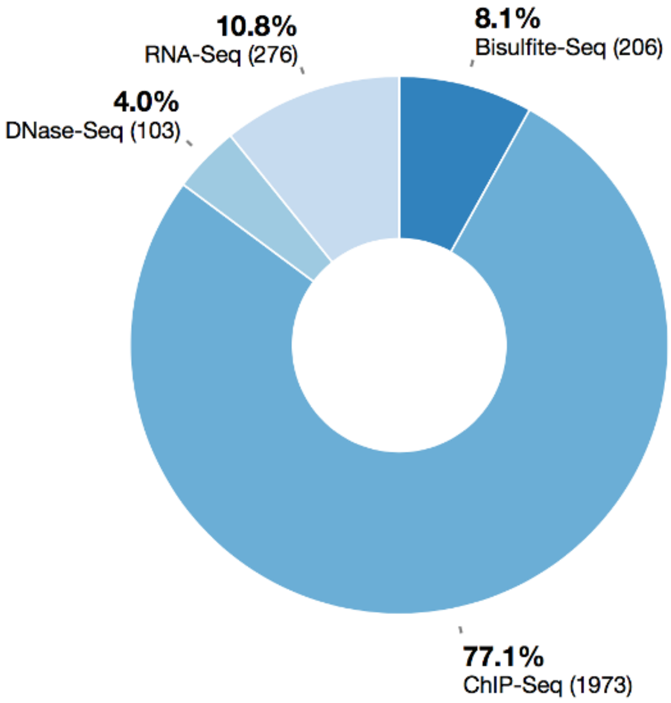
- > “BLUEPRINT is a large-scale research project receiving close to 30 million euro funding from the EU.”
- > 42 leading European scientific centers
- > The aim to further the understanding of how genes are activated or repressed in both healthy and diseased human cells
- > Focus on distinct types of haematopoietic cells from healthy individuals and on their malignant leukaemic counterparts

BluePrint

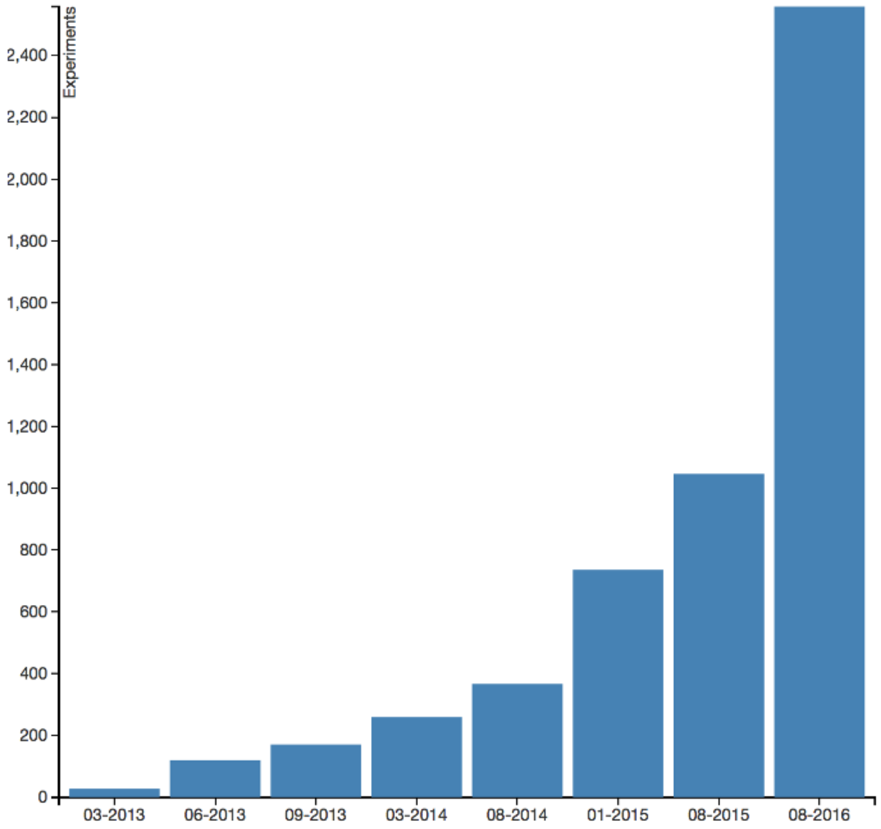
- > http://www.blueprint-epigenome.eu
- > Publications (Cell Papers) & Data Portal

BluePrint

> http://dcc.blueprint-epigenome.eu/#/home

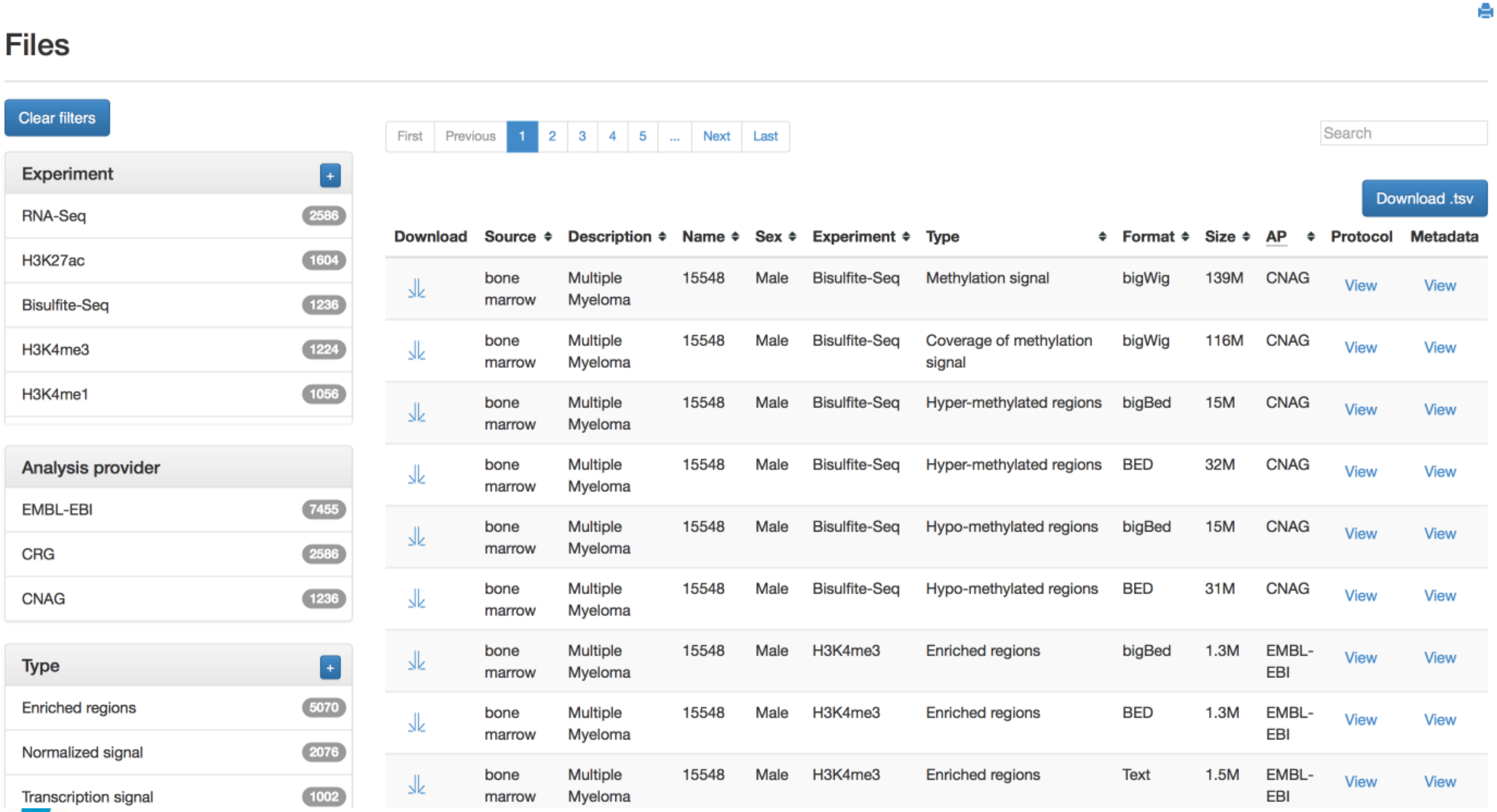


Release composition by assay type



Number of experiments per release

BluePrint



Files

Clear filters

Experiment	
RNA-Seq	2586
H3K27ac	1604
Bisulfite-Seq	1236
H3K4me3	1224
H3K4me1	1056

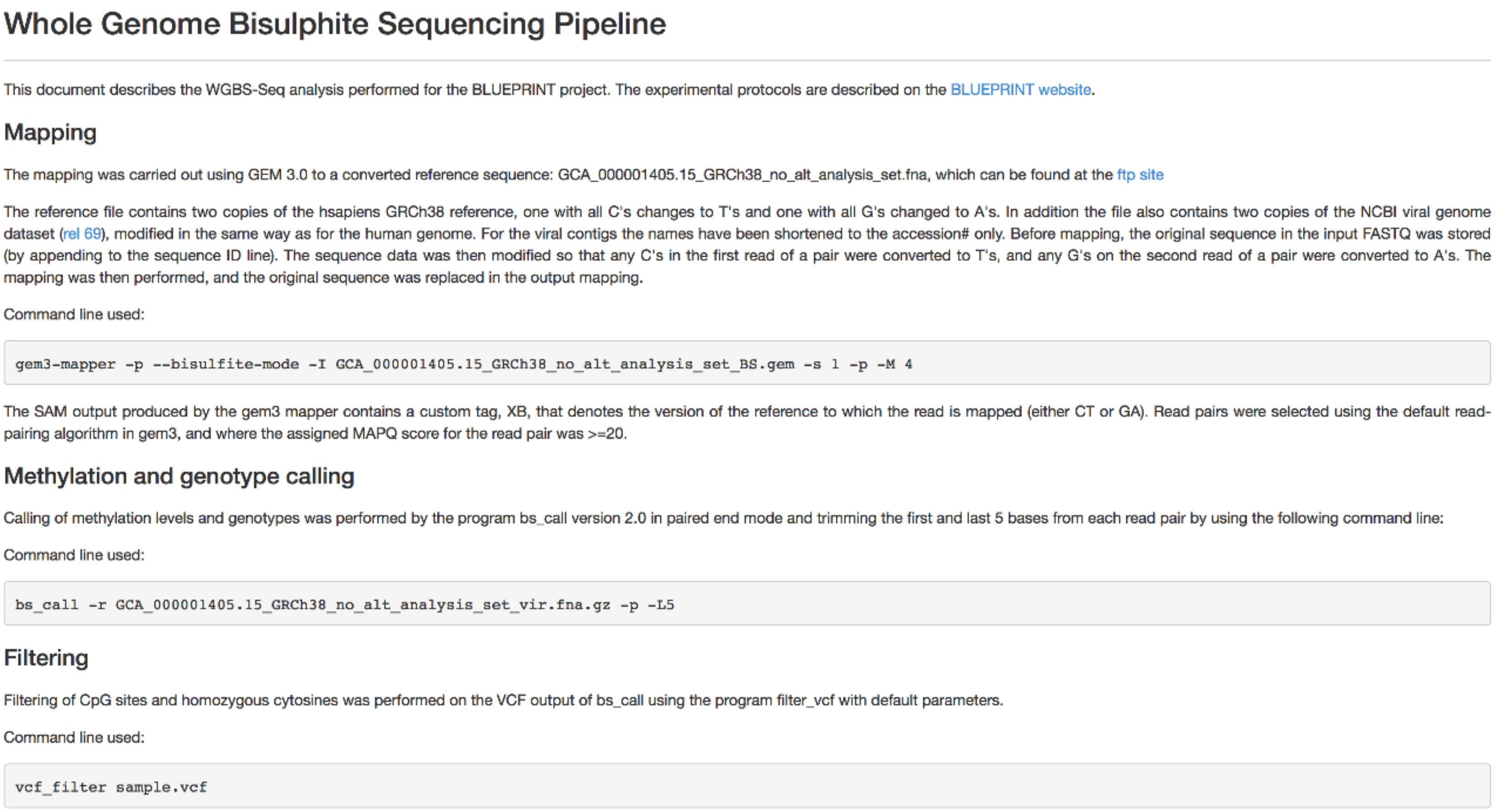
Analysis provider	
EMBL-EBI	7455
CRG	2586
CNAG	1236

Type	
Enriched regions	5070
Normalized signal	2076
Transcription signal	1002

Download	Source	Description	Name	Sex	Experiment	Type	Format	Size	AP	Protocol	Metadata
↓	bone marrow	Multiple Myeloma	15548	Male	Bisulfite-Seq	Methylation signal	bigWig	139M	CNAG	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	Bisulfite-Seq	Coverage of methylation signal	bigWig	116M	CNAG	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	Bisulfite-Seq	Hyper-methylated regions	bigBed	15M	CNAG	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	Bisulfite-Seq	Hyper-methylated regions	BED	32M	CNAG	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	Bisulfite-Seq	Hypo-methylated regions	bigBed	15M	CNAG	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	Bisulfite-Seq	Hypo-methylated regions	BED	31M	CNAG	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	H3K4me3	Enriched regions	bigBed	1.3M	EMBL-EBI	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	H3K4me3	Enriched regions	BED	1.3M	EMBL-EBI	View	View
↓	bone marrow	Multiple Myeloma	15548	Male	H3K4me3	Enriched regions	Text	1.5M	EMBL-EBI	View	View

Centre de Regulació Genòmica

BluePrint



Whole Genome Bisulphite Sequencing Pipeline

This document describes the WGBS-Seq analysis performed for the BLUEPRINT project. The experimental protocols are described on the BLUEPRINT website.

Mapping

The mapping was carried out using GEM 3.0 to a converted reference sequence: GCA_000001405.15_GRCh38_no_alt_analysis_set.fna, which can be found at the ftp site

The reference file contains two copies of the hsapiens GRCh38 reference, one with all C's changes to T's and one with all G's changed to A's. In addition the file also contains two copies of the NCBI viral genome dataset (rel 69), modified in the same way as for the human genome. For the viral contigs the names have been shortened to the accession# only. Before mapping, the original sequence in the input FASTQ was stored (by appending to the sequence ID line). The sequence data was then modified so that any C's in the first read of a pair were converted to T's, and any G's on the second read of a pair were converted to A's. The mapping was then performed, and the original sequence was replaced in the output mapping.

Command line used:

```
gem3-mapper -p --bisulfite-mode -I GCA_000001405.15_GRCh38_no_alt_analysis_set_BS.gem -s 1 -p -M 4
```

The SAM output produced by the gem3 mapper contains a custom tag, XB, that denotes the version of the reference to which the read is mapped (either CT or GA). Read pairs were selected using the default read-pairing algorithm in gem3, and where the assigned MAPQ score for the read pair was >=20.

Methylation and genotype calling

Calling of methylation levels and genotypes was performed by the program bs_call version 2.0 in paired end mode and trimming the first and last 5 bases from each read pair by using the following command line:

Command line used:

```
bs_call -r GCA_000001405.15_GRCh38_no_alt_analysis_set_vir.fna.gz -p -L5
```

Filtering

Filtering of CpG sites and homozygous cytosines was performed on the VCF output of bs_call using the program filter_vcf with default parameters.

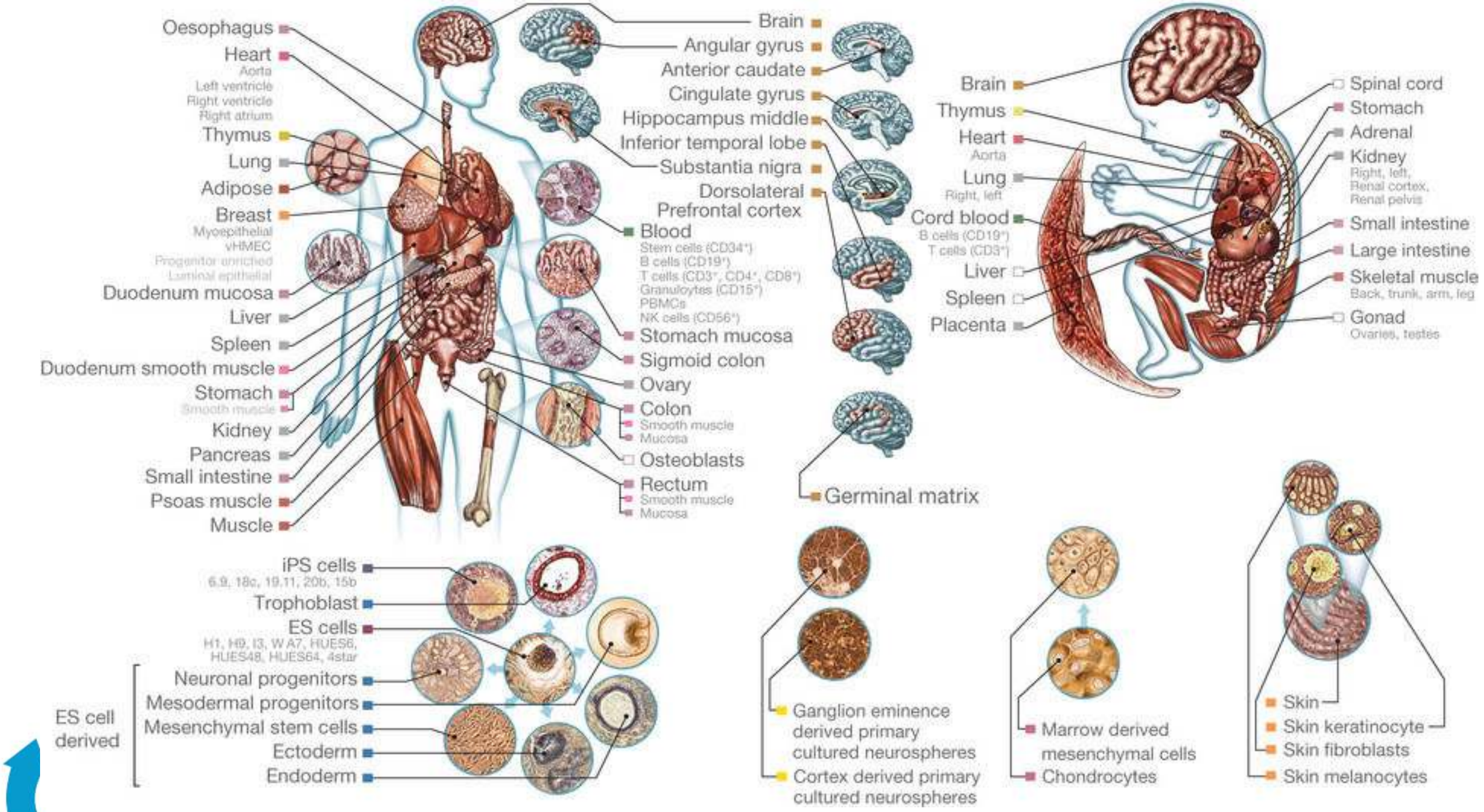
Command line used:

```
vcf_filter sample.vcf
```

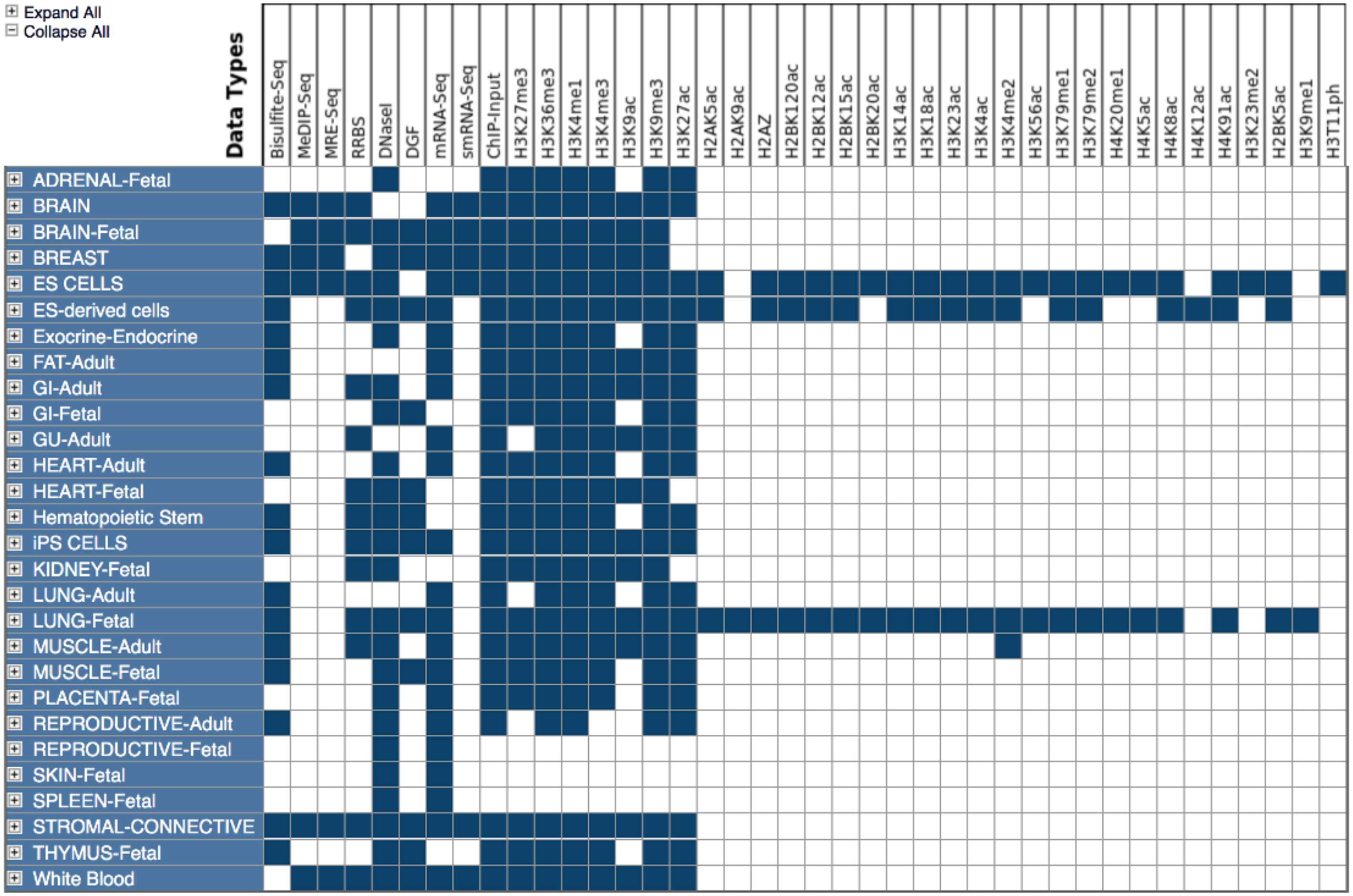
RoadMap Epigenomics

- The NIH Roadmap Epigenomics Research to transform our understanding of how epigenetics contributes to disease
- The Consortium leverages experimental pipelines built around next-generation sequencing technologies to map DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in **stem cells** and **primary ex vivo tissues** selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease

RoadMap Epigenomics



RoadMap Epigenomics

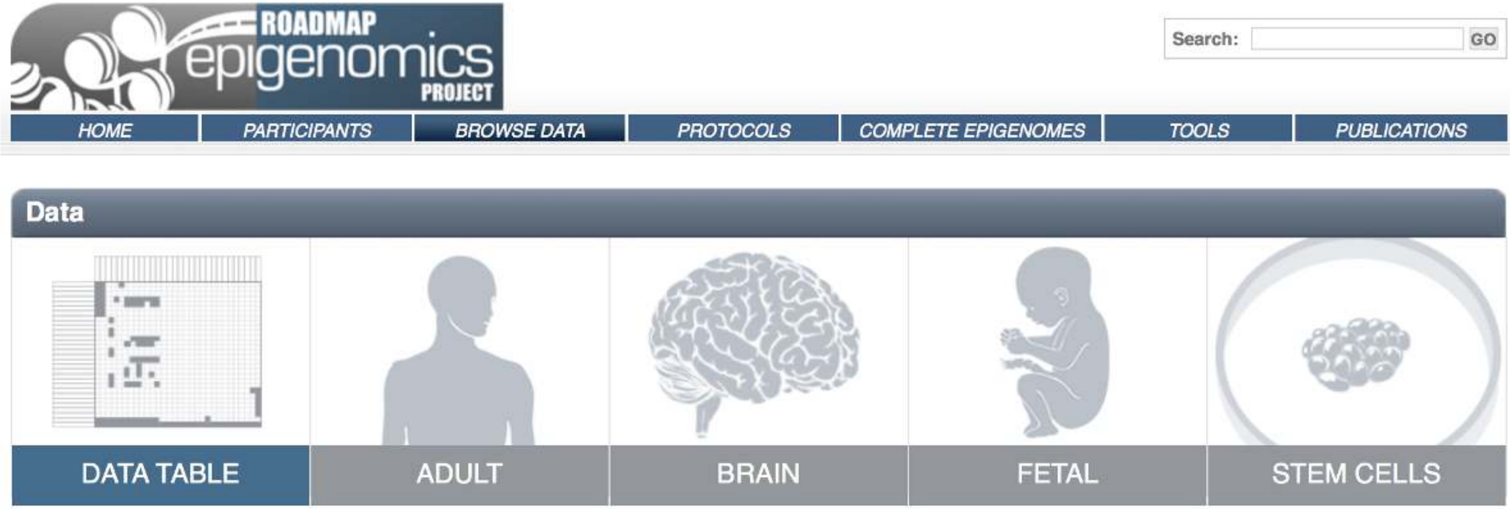


⊞ Expand All
⊟ Collapse All

Data Types	Bisulfite-Seq	MeDIP-Seq	MRE-Seq	RRBS	DNaseI	DGF	mRNA-Seq	smRNA-Seq	ChIP-Input	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9ac	H3K9me3	H3K27ac	H2AK5ac	H2AK9ac	H2AZ	H2BK120ac	H2BK12ac	H2BK15ac	H2BK20ac	H3K14ac	H3K18ac	H3K23ac	H3K4ac	H3K4me2	H3K56ac	H3K79me1	H3K79me2	H4K20me1	H4K5ac	H4K8ac	H4K12ac	H4K91ac	H3K23me2	H2BK5ac	H3K9me1	H3T11ph
⊞ ADRENAL-Fetal					■				■	■	■	■	■		■	■																								
⊞ BRAIN	■	■	■	■			■	■	■	■	■	■	■	■	■	■																								
⊞ BRAIN-Fetal		■	■	■	■	■	■	■	■	■	■	■	■	■	■																									
⊞ BREAST	■	■	■		■	■	■	■	■	■	■	■	■	■	■																									
⊞ ES CELLS	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■
⊞ ES-derived cells	■			■	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	■		■	■	■	■	■		■	■			■	■	■		■		
⊞ Exocrine-Endocrine	■				■		■		■	■	■	■	■		■	■																								
⊞ FAT-Adult	■						■		■	■	■	■	■	■	■	■																								
⊞ GI-Adult	■			■	■		■		■	■	■	■	■	■	■	■																								
⊞ GI-Fetal					■	■			■	■	■	■	■		■	■																								
⊞ GU-Adult				■			■		■		■	■	■	■	■	■																								
⊞ HEART-Adult	■				■		■		■	■	■	■	■		■	■																								
⊞ HEART-Fetal				■	■	■			■	■	■	■	■	■	■																									
⊞ Hematopoietic Stem	■			■	■	■			■	■	■	■	■		■	■																								
⊞ iPS CELLS	■			■	■	■	■		■	■	■	■	■	■	■	■																								
⊞ KIDNEY-Fetal				■	■				■	■	■	■	■	■	■																									
⊞ LUNG-Adult	■						■		■		■	■	■		■	■																								
⊞ LUNG-Fetal	■			■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	
⊞ MUSCLE-Adult	■			■	■		■		■	■	■	■	■	■	■	■												■												
⊞ MUSCLE-Fetal	■				■	■	■		■	■	■	■	■		■	■																								
⊞ PLACENTA-Fetal					■		■		■	■	■	■	■		■	■																								
⊞ REPRODUCTIVE-Adult	■				■		■		■		■	■			■	■																								
⊞ REPRODUCTIVE-Fetal					■		■																																	
⊞ SKIN-Fetal					■		■																																	
⊞ SPLEEN-Fetal					■		■																																	
⊞ STROMAL-CONNECTIVE	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																								
⊞ THYMUS-Fetal	■				■	■			■	■	■	■	■		■	■																								
⊞ White Blood		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																								

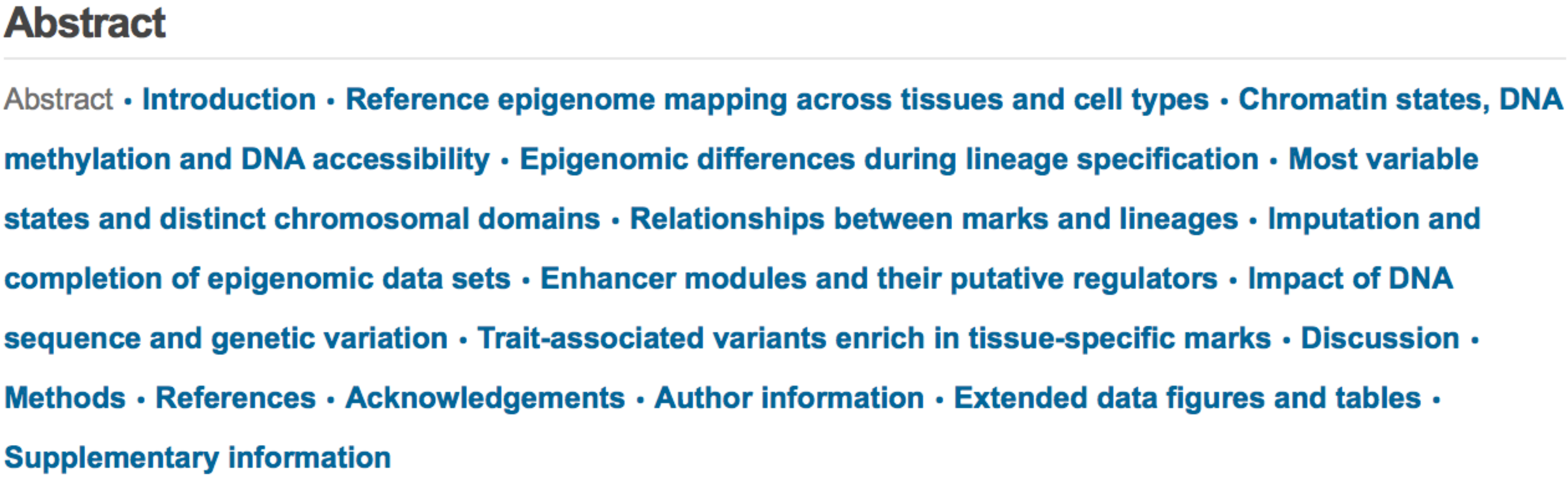
RoadMap Epigenomics

It looks like we can get Protocols clicking on the link, however, there are not a lot of them there. The protocols are super outdated! (eg REMC STANDARDS AND GUIDELINES FOR CHIP-SEQDEC. 2, 2011 — V1.0)



RoadMap Epigenomics

- If you wanna to work with these data – read the paper “Integrative analysis of 111 reference human epigenomes” (+16 ENCODE2012, do not print the paper!)



Abstract

Abstract • Introduction • Reference epigenome mapping across tissues and cell types • Chromatin states, DNA methylation and DNA accessibility • Epigenomic differences during lineage specification • Most variable states and distinct chromosomal domains • Relationships between marks and lineages • Imputation and completion of epigenomic data sets • Enhancer modules and their putative regulators • Impact of DNA sequence and genetic variation • Trait-associated variants enrich in tissue-specific marks • Discussion • Methods • References • Acknowledgements • Author information • Extended data figures and tables • Supplementary information

- Go through the “Publications” list

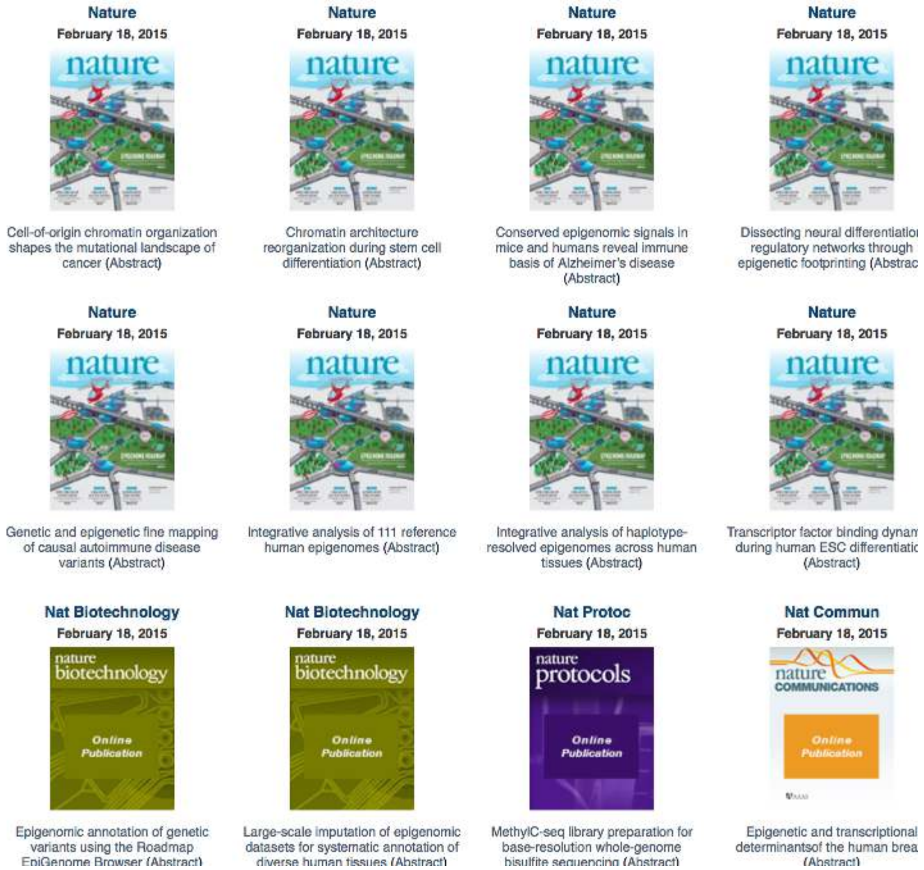
RoadMap Epigenomics

The most useful section is **Methods**:

- RNA-seq uniform processing and quantification for consolidated epigenomes
- ChIP-seq and DNase-seq uniform reprocessing for consolidated epigenomes
- Methylation data cross-assay standardization and uniform processing for consolidated epigenomes
- Chromatin state learning
- Etc.

RoadMap Epigenomics

> Publications



Nature
February 18, 2015
Cell-of-origin chromatin organization shapes the mutational landscape of cancer (Abstract)

Nature
February 18, 2015
Chromatin architecture reorganization during stem cell differentiation (Abstract)

Nature
February 18, 2015
Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease (Abstract)

Nature
February 18, 2015
Dissecting neural differentiatio regulatory networks through epigenetic footprinting (Abstrac

Nature
February 18, 2015
Genetic and epigenetic fine mapping of causal autoimmune disease variants (Abstract)

Nature
February 18, 2015
Integrative analysis of 111 reference human epigenomes (Abstract)

Nature
February 18, 2015
Integrative analysis of haplotype-resolved epigenomes across human tissues (Abstract)

Nature
February 18, 2015
Transcriptor factor binding dynam during human ESC differentiatic (Abstract)

Nat Biotechnology
February 18, 2015
Epigenomic annotation of genetic variants using the Roadmap EpiGenome Browser (Abstract)

Nat Biotechnology
February 18, 2015
Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues (Abstract)

Nat Protoc
February 18, 2015
MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing (Abstract)

Nat Commun
February 18, 2015
Epigenetic and transcriptional determinantsof the human brea (Abstract)

RoadMap Epigenomics

- Histone mark combinations show distinct levels of DNA methylation and accessibility, and **predict differences in RNA expression levels** that are not reflected in either accessibility or methylation.
- Megabase-scale regions with distinct epigenomic signatures show strong differences in activity, gene density and nuclear lamina associations, suggesting **distinct chromosomal domains**.
- Approximately 5% of each reference epigenome shows **enhancer and promoter signatures**, which are twofold enriched for evolutionarily conserved non-exonic elements on average.
- Epigenomic data sets can be **imputed** at high resolution from existing data, completing missing marks in additional cell types, and providing a more robust signal even for observed data sets.
- Dynamics of epigenomic marks in their relevant chromatin states allow a **data-driven approach to learn biologically meaningful relationships** between cell types, tissues and lineages.

Working in Consortia

Working with Data

- *Getting Raw Data*
- *Working with the data from different consortia simultaneously: different QCs, different data analysis pipeline*
- *Versions of tools missed or outdated/ unsupported tools – failure of replication!*

Working in Consortia I

- *When your Server gets down or all your data were accidentally removed*
- *Deadlines – add 3-6 months to expected date!*
- *Communication: teleconferences*
- *Passwords renewal, permissions to access*
- *Efficient data sharing – speed, reliability, confidentiality*

Working in Consortia II

- *Different naming of the same samples in different working groups / labs*
- *Wrong/Missing Identifiers (eg wrong cancer type or population) – case: normal and somatic were actually swapped*
- *The same, but from clinicians*
- *Different labs - different library preparation (eg coverage depths after PCR-free and PCR-based WGS)*
- *Several tools can be used for the analysis – establishment of the best tool or generation of joint callset*
- *Multiple blacklist or outlier lists (every lab/group has its own and they do not completely overlap)*

Working in Consortia III

- *Unbalanced Population Structure*
- *Mix of different effects (eg Cancer vs. Population)*
- *Is your Germline really Germline?*

Slide from AgENCODE, Ewan Birney

- Many scientists (PIs) are “cats” (who like to walk alone) not “dogs” (who hunt in packs)
- These resource generation plus integration and display is a “hunt in a pack” moment
  - If you have a strong cat mentality, just orientate your research to take advantage of these resources when it emerges
  - If you are part dog, go for it

Спасибо за внимание!

CRG

Centre
de Regulació
Genòmica