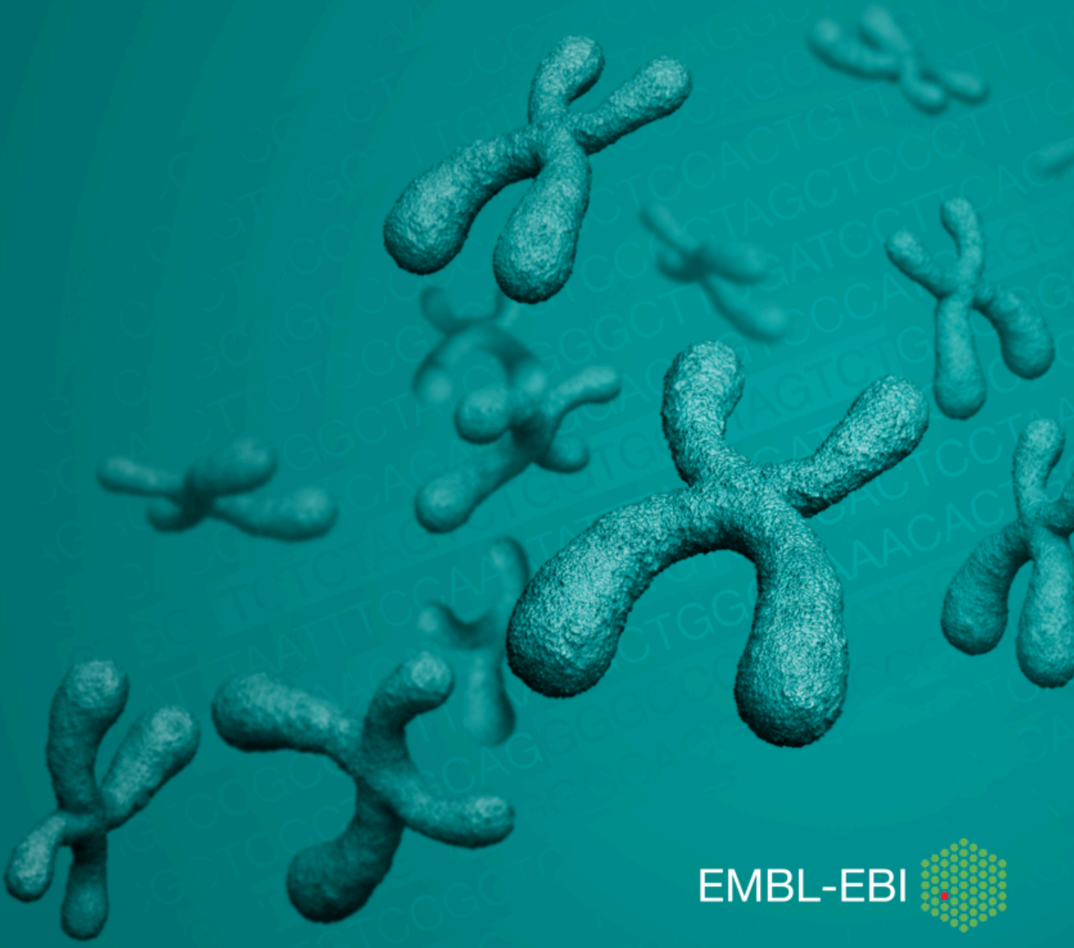
Genome Annotation

Ewan Birney (tweetable)

@ewanbirney

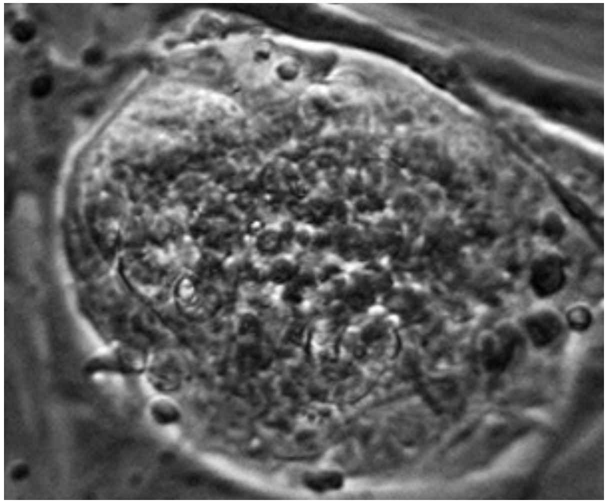


EMBL-EBI

What next?

```
GGGATTGAGAGTGATCACTCACGCTAACGTCTGCCCTGTTCCTGTATGGTGAGGCCGCAC
CACAAGCCACCACCGCCGCCGCCTTCTGCGCAACGCCAACCGCCCGCCAAAACGGATCCT
TCCCTGCGCCTGCGCAACCAATCTTGGGACCGGACCTTTTTTCTCCGCCCACTACGCATG
CGCAAAGCTAGGACAAACTCCCGCCAACACGCAGGCGCCGTAGGTTCACTGCCTACTCCT
GCCCGCCATTTCACGTGTTCTCAGAGGCAGGTGGAACTTCTTAATGCGCCTGCGCAAAAC
TCGCCATTTTACTACACGTGCGGTCAACAAGAGTTCATTGCAAAAAAATTGTTACCTCCT
AGCTGCTTGTCTAATACATAGTGTTAATCATGCTTTGCCAAGCGACTTGACTGTAATATT
TGCGCGTGGAAGATTAAAAAGATGTTAAACACCCAAGGTAGATTCAAATGTGAATGATTG
GTCGGTTGGCCAATCAGACTGGTTAACAATAACATTACTCGGGAACCAATGGACTCCAAG
GGGTGGAGACGGCGTAGAACGACCGAAGGAATGACGTTACACAGCAATGTGGCACCACAG
GCCAATAGCAGGGGGAAGCGATTTCAAGTATCCAATCAGAGCTGTTCTAGGGCGGAGTCT
ACCAATGCCGAAAGCGAGGAGGCGGGGTAAAAAAGAGAGGGCGAAGGTAGGCTGGCAGAT
ACGTTCGTCAGCTTGCTCCTTTCTGCCCGTGGACGCCGCCGAAGAAGCATCGTTAAAGTC
TCTCTTCACCCTGCCGTCATGTCTAAGTCAGAGGTGAGTTAGGCGCGCTTTCCCACTTGA
ATTTTTTCCTCTCCCTTTCCTGAATCGGTAAGATGCTGCTGGGTTTCGTTCCTTGCACCA
GCCCATTCTACAGTTCCTTCGGTCGCTGCCACGGCCTACCCCTCCCAAAGTTCAAGTCGC
CATTTTGTCCTCTTGATCGCCATGAGGCCGCTCTCCGCCAACCATGTGTTATCATGCGGG
ACTCGTTACTCGTAGCAAAATTCTTAGGCACACAGGATCTTTGTCTTTTTTTAAACCTTG
CCTTGGTGAGCGAGTTTTCTAAAGAGCGATTAGTCCCATTGTGGAGATGCACCCCTACCG
CCCAAGCCTTTGTTGCGCGTGCGTCGGAAGGCGACTAGGGACGCATGCGCTTGCGATTTC
CTAGCACTCCCAACTCCAGCATACGGCCTCCCTTGATAGGCAGAAGCACGTGTCTTGTTG
CGACCTGAACGAACAATAAGTGCTAGGTACACAGTTGGTGTCTAGTTTTTCTTTTCCTCG
ATGGAAATTGTTTCGTGTTGTAGCCCATTTAACACTTCCCCCTCCCCCCACTCTAGTCTC
```

The ultimate biological index

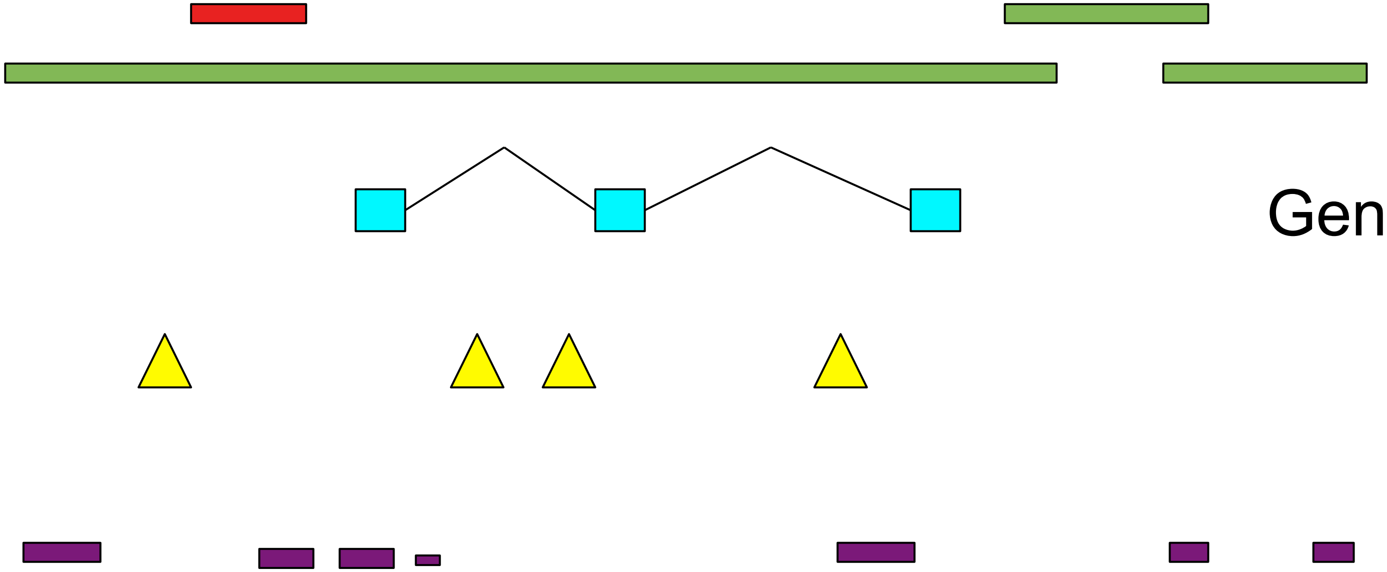


All biological molecules are encoded in some manner in the genome

…as is cellular response

…as is development

Genome Annotation



Genome

Gene Structures

Variation

Transcription Factor
Binding,
Chromatin States

Generation, Integration, Annotation

Generate

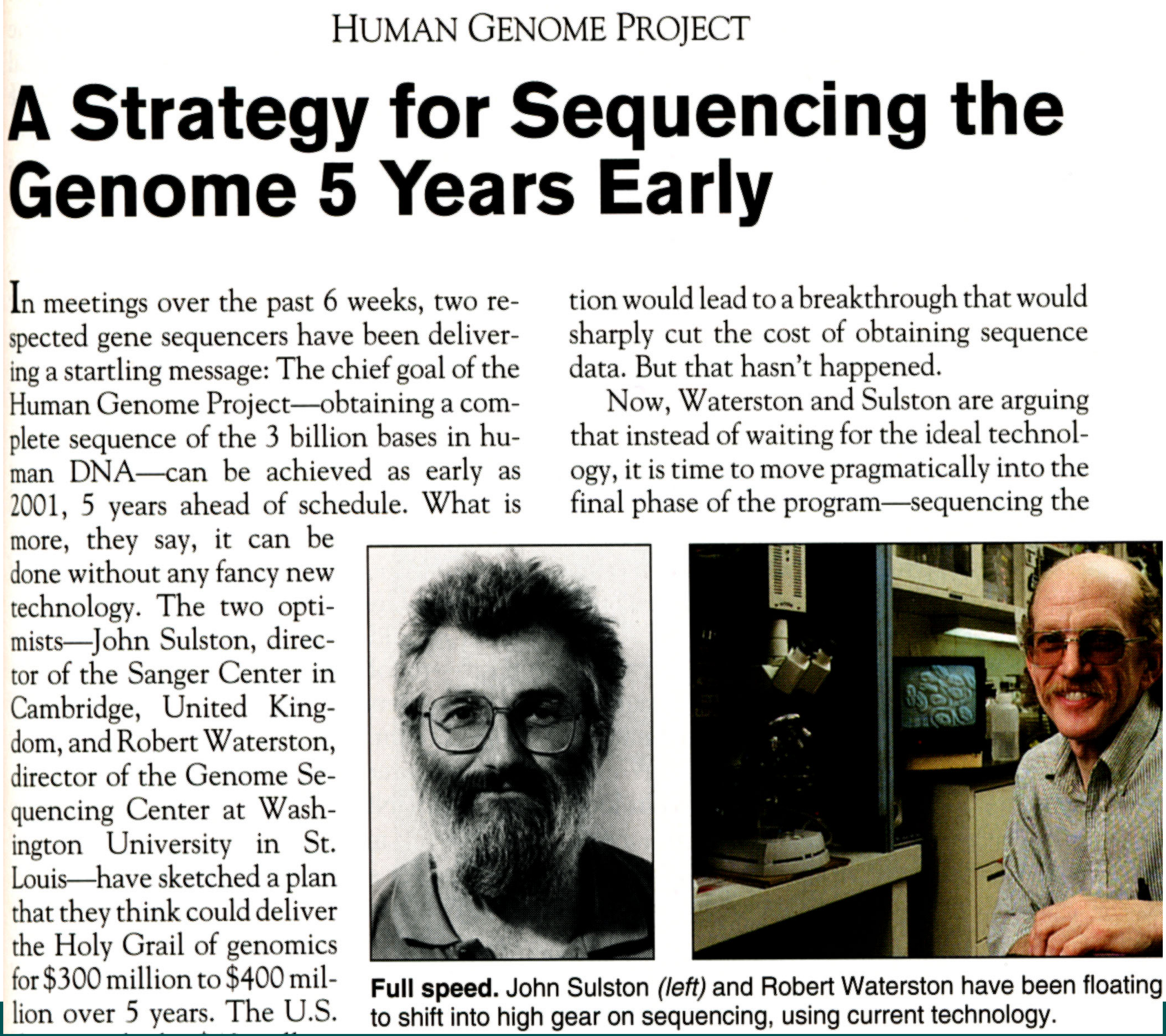
High quality data
Good data stds
Public availability(!)

Integrate

Map to genome sequence
Make non redundant
Call

Annotate

Provide initial labeling
Understand (a bit more…)



HUMAN GENOME PROJECT

A Strategy for Sequencing the Genome 5 Years Early

In meetings over the past 6 weeks, two respected gene sequencers have been delivering a startling message: The chief goal of the Human Genome Project—obtaining a complete sequence of the 3 billion bases in human DNA—can be achieved as early as 2001, 5 years ahead of schedule. What is more, they say, it can be done without any fancy new technology. The two optimists—John Sulston, director of the Sanger Center in Cambridge, United Kingdom, and Robert Waterston, director of the Genome Sequencing Center at Washington University in St. Louis—have sketched a plan that they think could deliver the Holy Grail of genomics for $300 million to $400 million over 5 years. The U.S.

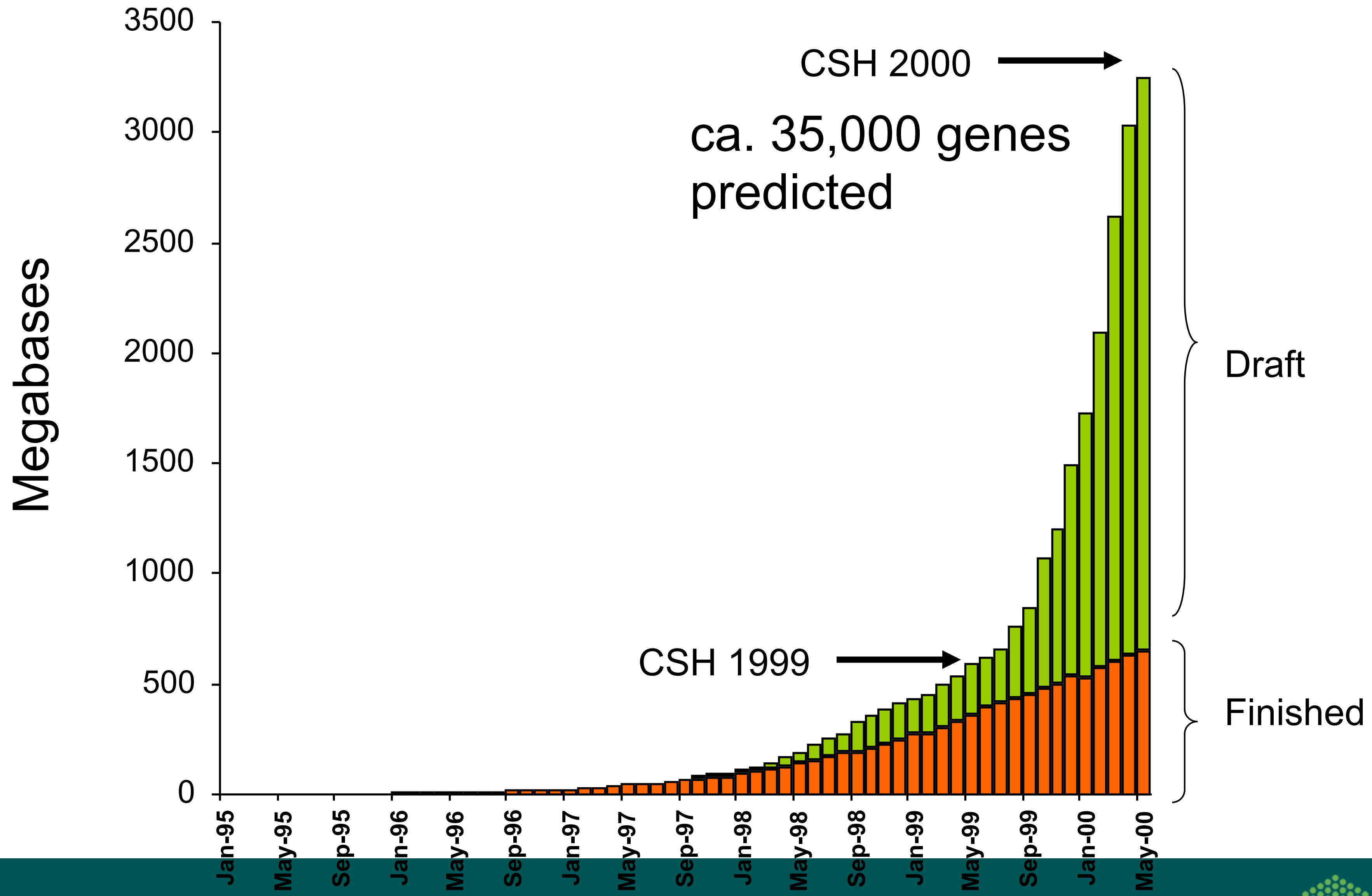
tion would lead to a breakthrough that would sharply cut the cost of obtaining sequence data. But that hasn't happened.

Now, Waterston and Sulston are arguing that instead of waiting for the ideal technology, it is time to move pragmatically into the final phase of the program—sequencing the

Full speed. John Sulston *(left)* and Robert Waterston have been floating to shift into high gear on sequencing, using current technology.

Science, February 1995

Human Genome Sequence in INSDC



CSH 2000

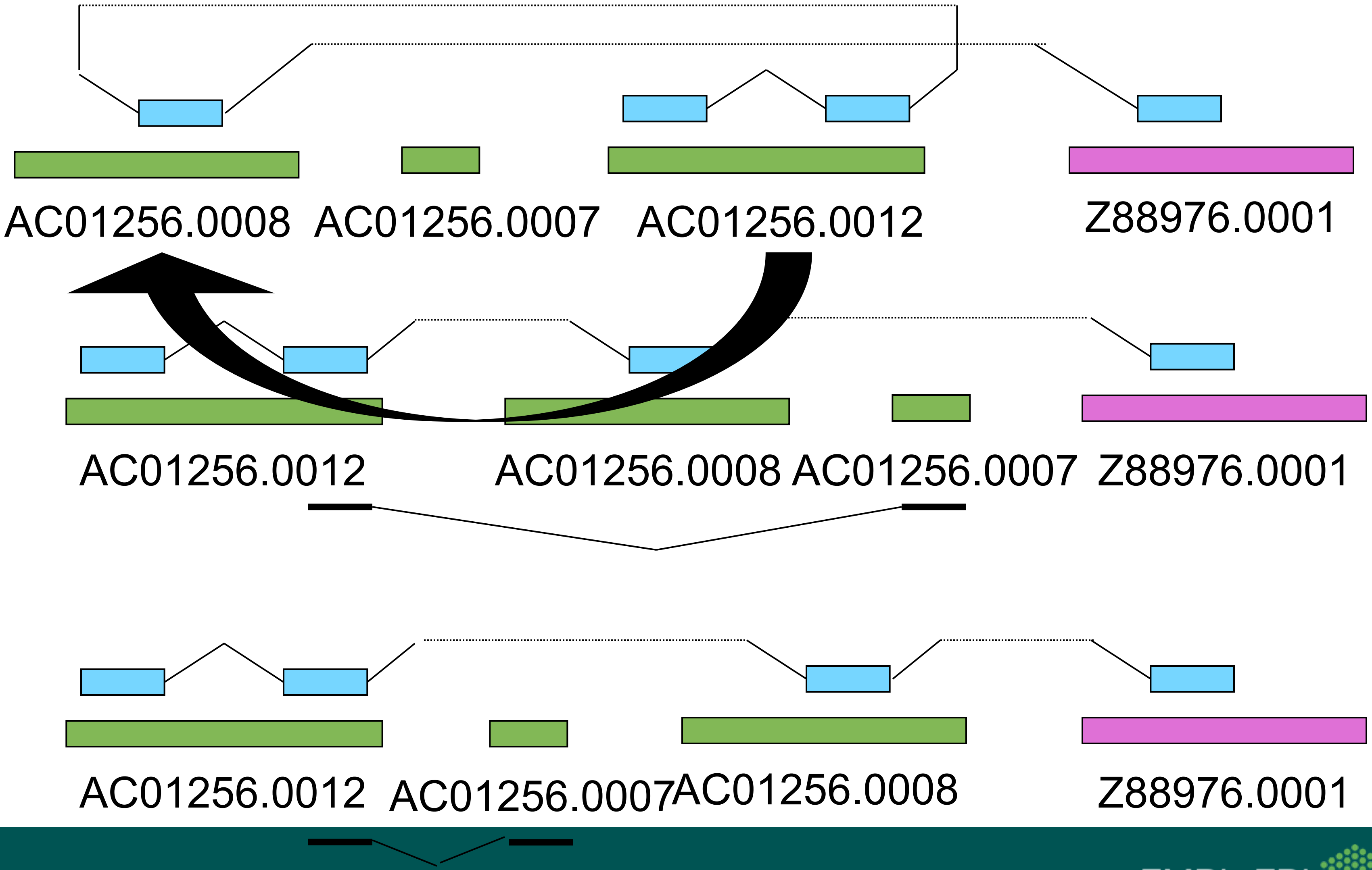
ca. 35,000 genes predicted

CSH 1999

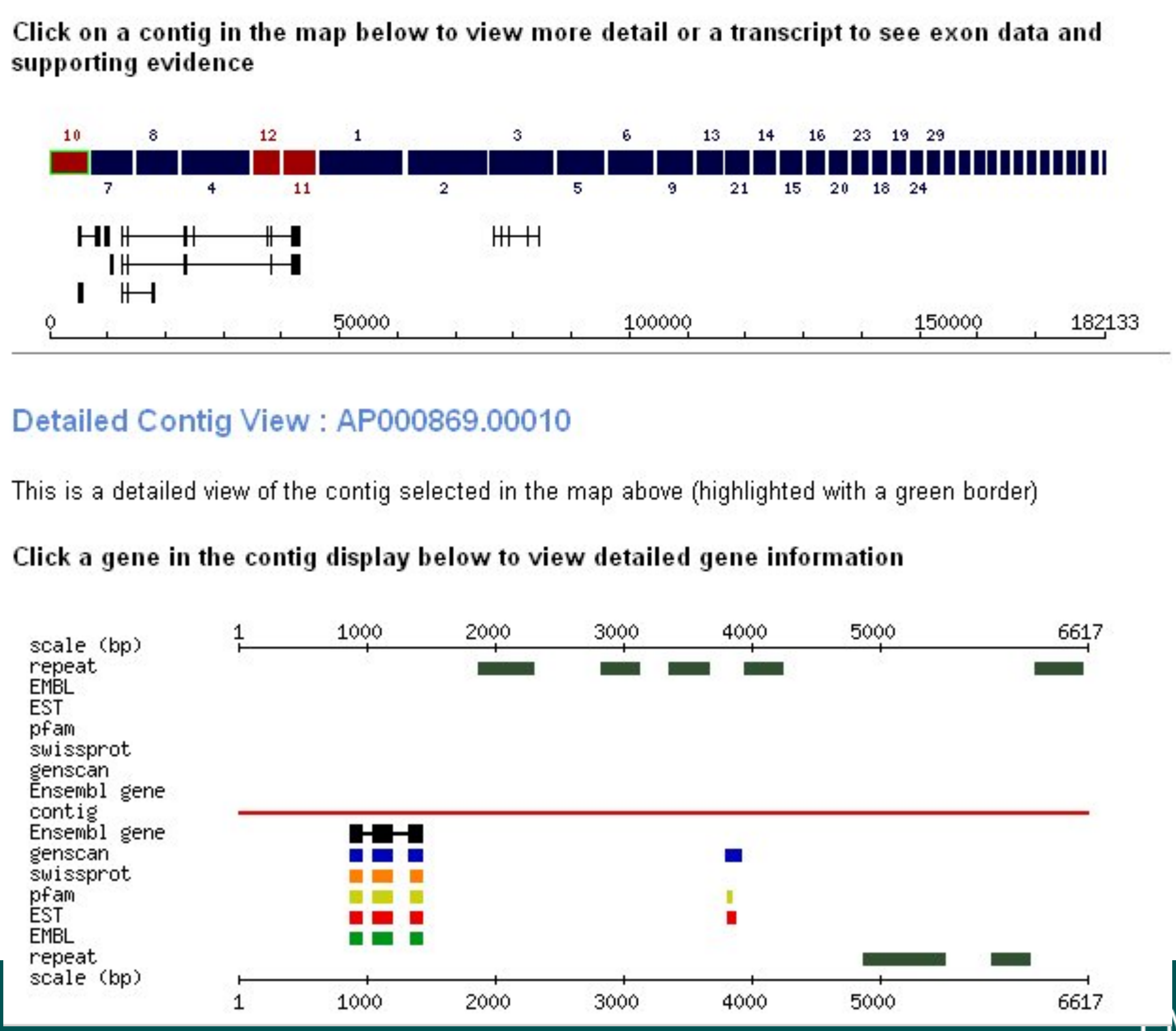
Draft

Finished

Ensembl (circa 2001!)



Web Site Contig view



Click on a contig in the map below to view more detail or a transcript to see exon data and supporting evidence

Detailed Contig View : AP000869.00010

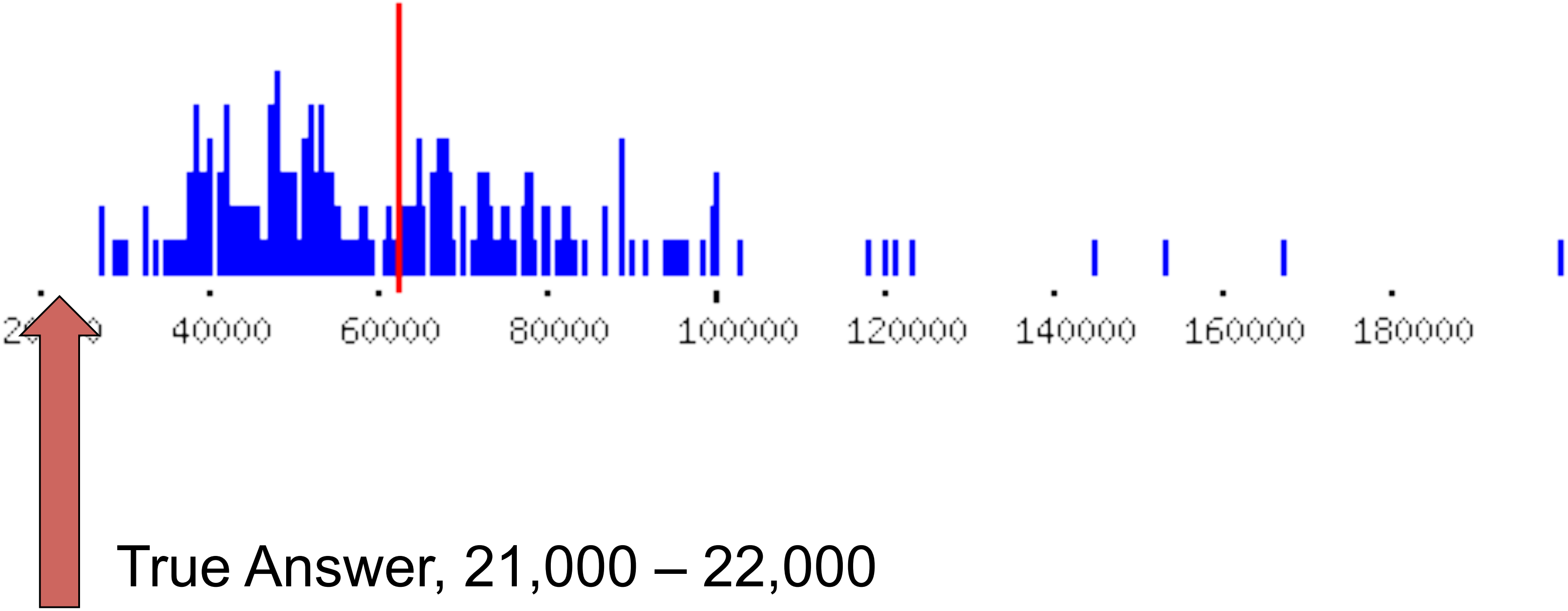
This is a detailed view of the contig selected in the map above (highlighted with a green border)

Click a gene in the contig display below to view detailed gene information

Estimated number of genes

- Chromosome 22
  - 549 + (~100) genes
  - 1.1 % of genome
  - 50,000 genes (59,000)
- Chromosome 21
  - 225 genes in Chr21
  - 1.1 % of genome
  - 20,500 genes
- Ensembl
  - 38,000 genes

Distribution of bets for the number of protein coding genes (2001)



True Answer, 21,000 – 22,000

Variation

Generate

SNP Consortium
Perlgen
HapMap
1,000 Genomes

Integrate

dbSNP RefSNP
1000 Genomes Pipeliens

Annotate

Frequency, LD
VEP, VAAST

TFs and Chromatin

Generate

ENCODE
Epigenome Roadmap
IHEC
Individual Studies

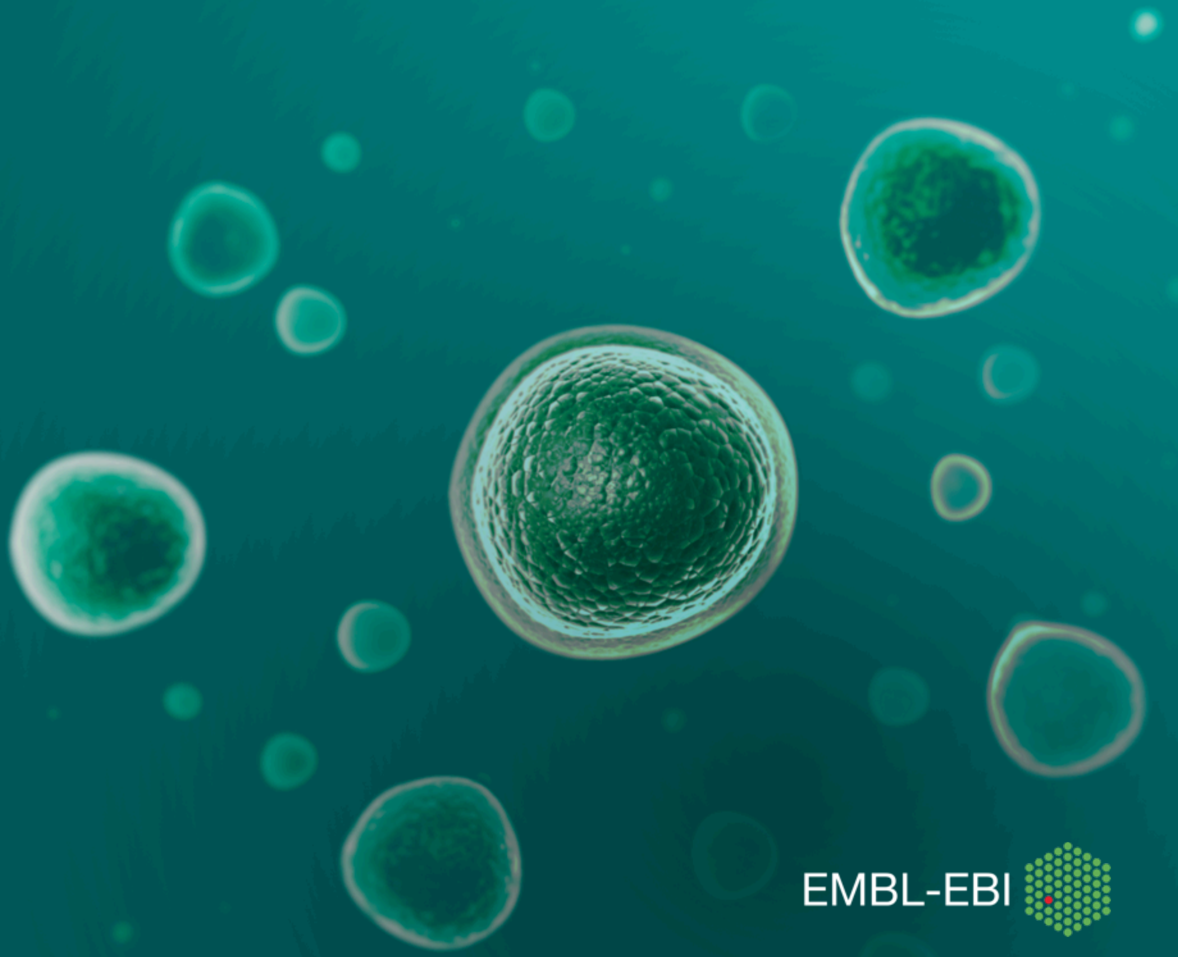
Integrate

ENCODE/Epigenome Pipelines
Ensembl Regulatory Build
RegulomeDB

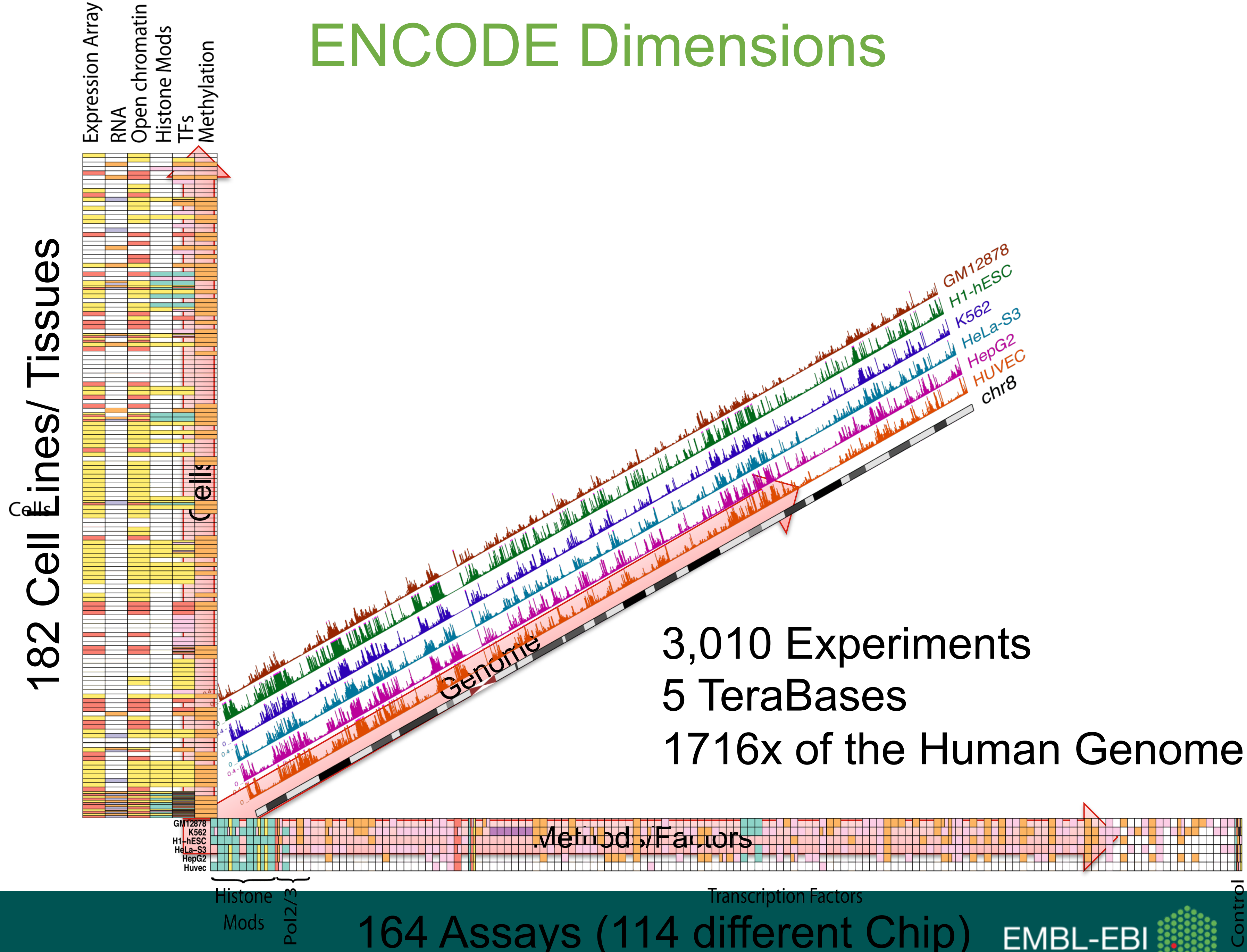
Annotate

? Chromatin State
4C/5C Experiments?

ENCODE



ENCODE Dimensions



3,010 Experiments
5 TeraBases
1716x of the Human Genome

164 Assays (114 different Chip)

EMBL-EBI

A consortium effort…

11 Main, multi Site groups
~50 Laboratories in total

10 additional groups

30 “lead” PIs

~410 Authors on the main Paper

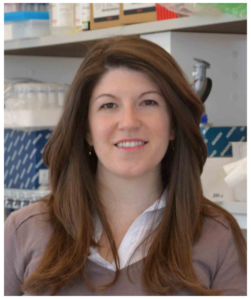
6 “high profile” papers
~25-30 companion papers

Kate Rosenbloom@UCSC

Steve Landt @Stanford

Alexias Safi @Duke

Chuck Epstein @Broad



Flo Pauli @HA

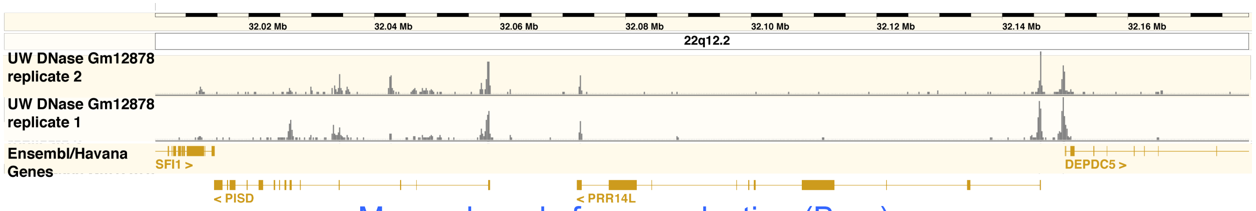
Jen Harrow @Sanger

Raj Kaul@UW

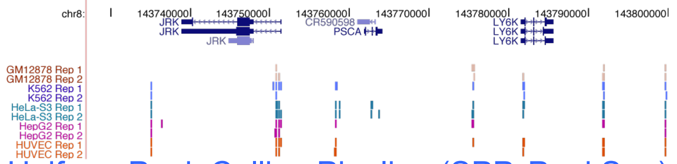
Carrie Davis@CSHL

ENCODE Uniform Analysis Pipeline

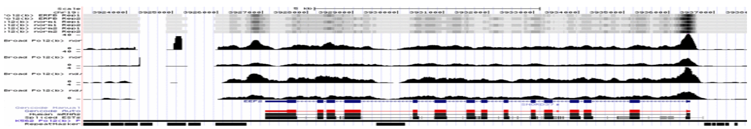
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



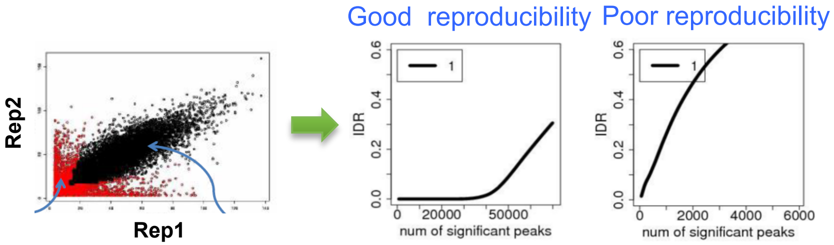
Mapped reads from production (Bam)



Uniform Peak Calling Pipeline (SPP, PeakSeq)

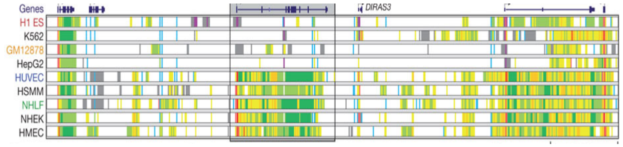


Signal Generation
(read extension and mappability correction)

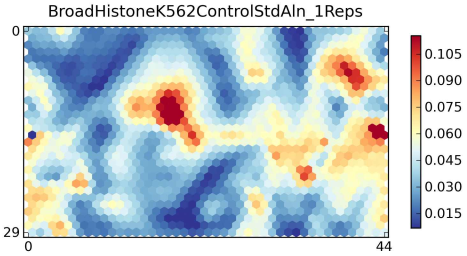


IDR Processing, QC and Blacklist Filtering

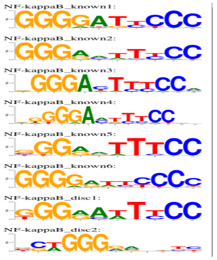
Segmentation



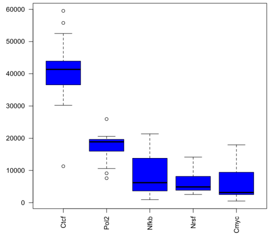
ChromHMM/Segway



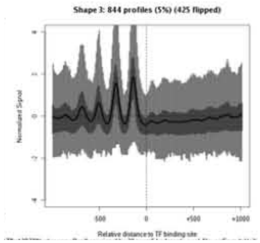
Self Organising Maps



Motif Discovery



Stats, GSC enrichments, etc.



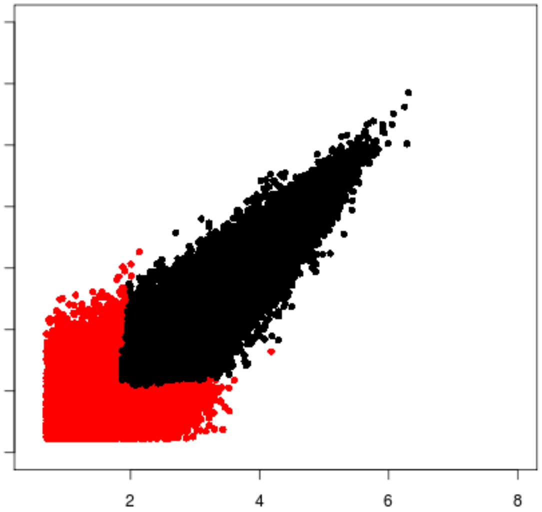
Signal Aggregation over peaks

Irreproducible Discovery Rate (IDR)

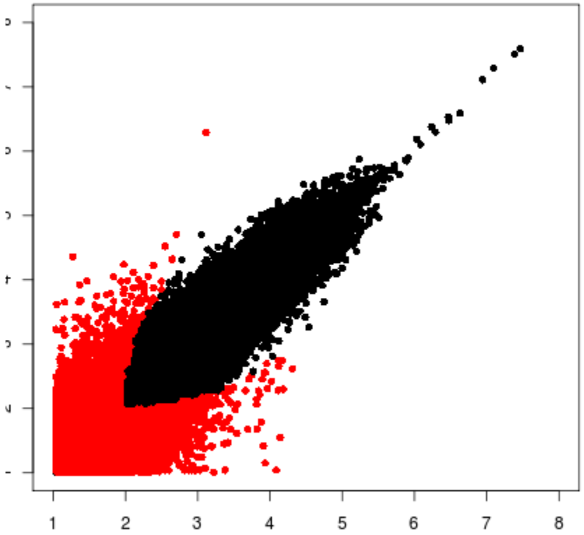
Ben Brown, Qunhau Li, Peter Bickel

If one re-ran the experiment, what is the probability one would observe the same element at this rank or better

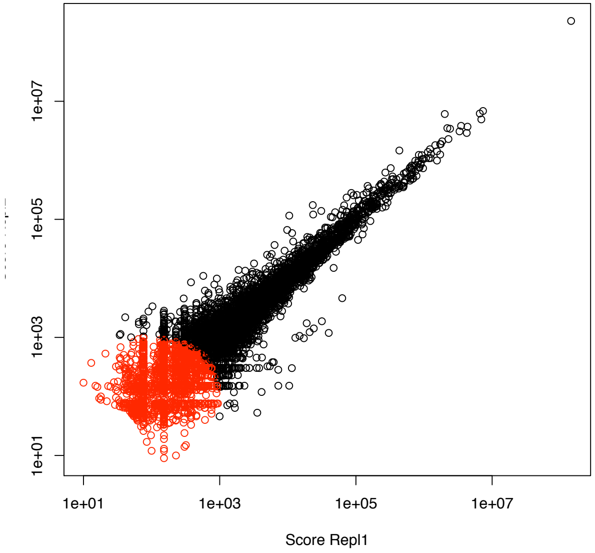
Uses ranked element lists from two replicates, and makes the assumption that there is noise at the bottom of the rank



Chip-seq



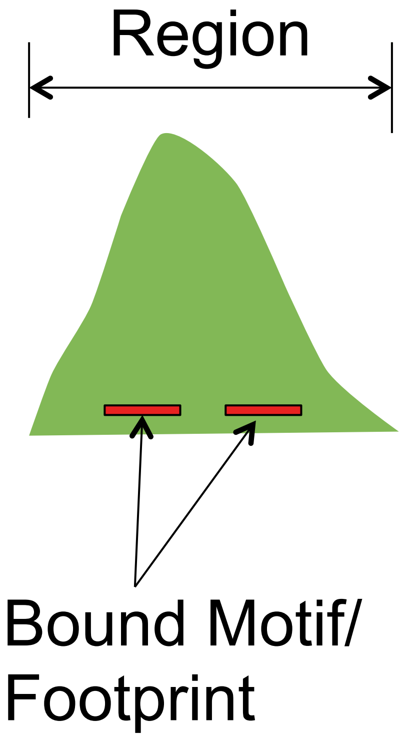
Dnase-seq



RNA-seq

Raw genome coverage of elements

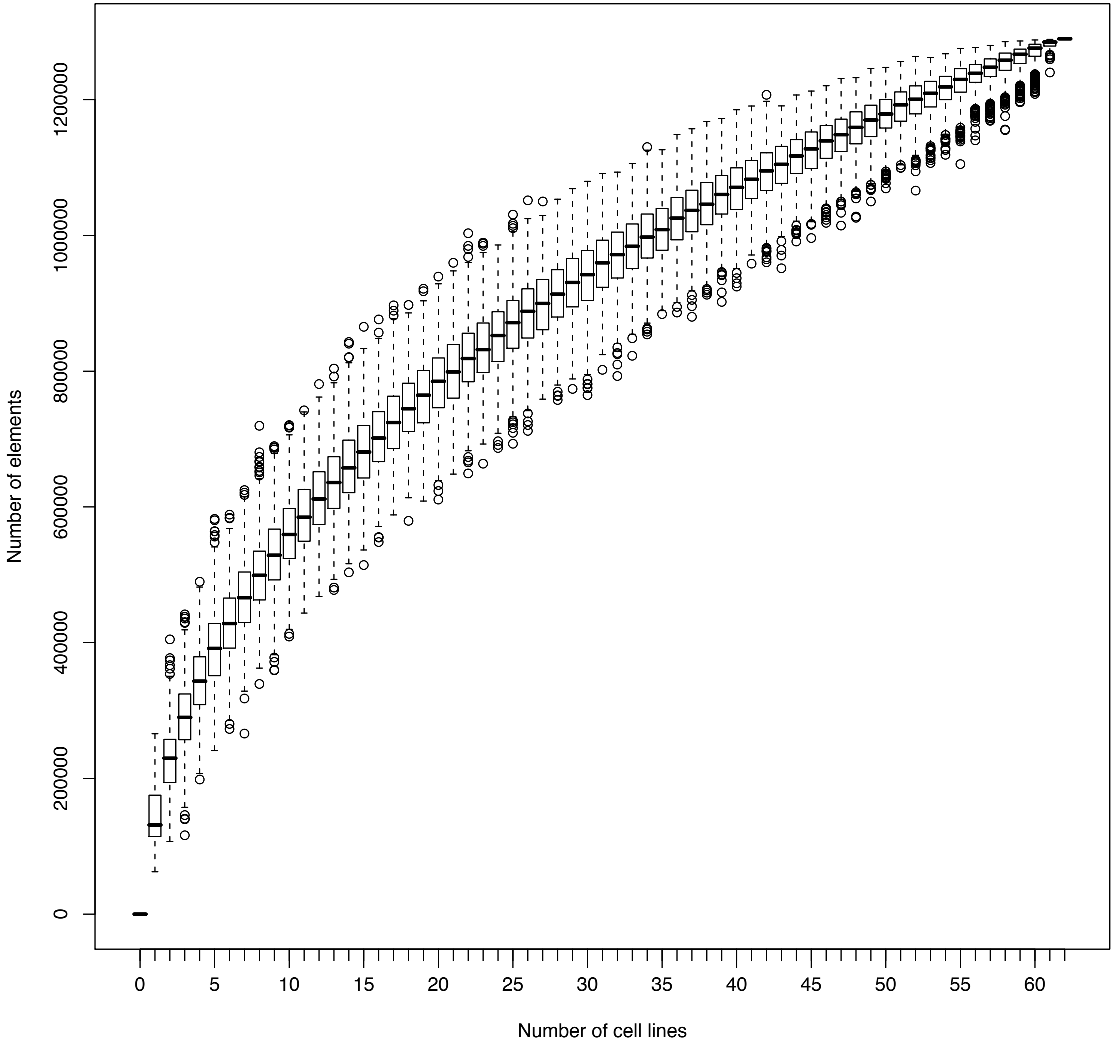
Element Type	Coverage	Cumulative Coverage
Exons	3%	3%
Chip-seq bound motifs	4.5%	5%
DNaseI Footprints	5.7%	9%
Chip-seq bound regions	8.1%	12%
DNaseI HS regions	15.2%	19.4%
Histone Modifications (*)	44%	49%
RNA	62%	80%
(* excluding broad marks)		



(Union over all experiments and cell types)

Saturation

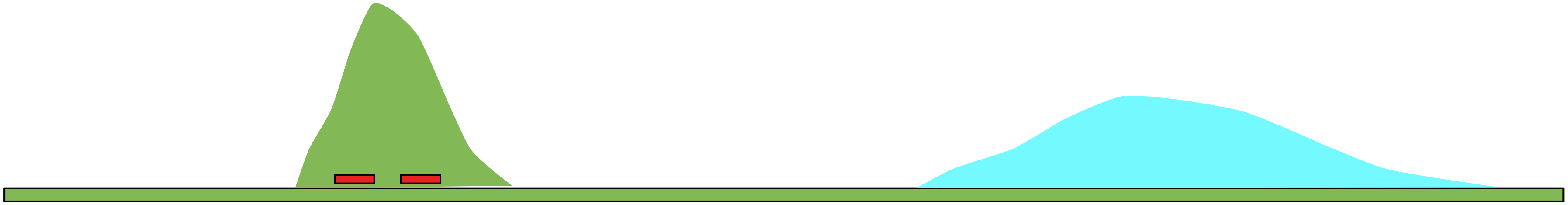
Steve Wilder



Most aggressive fit for saturation suggests a maximum of 50% of elements discovered

Likely to be lower due to inaccessible cell types etc

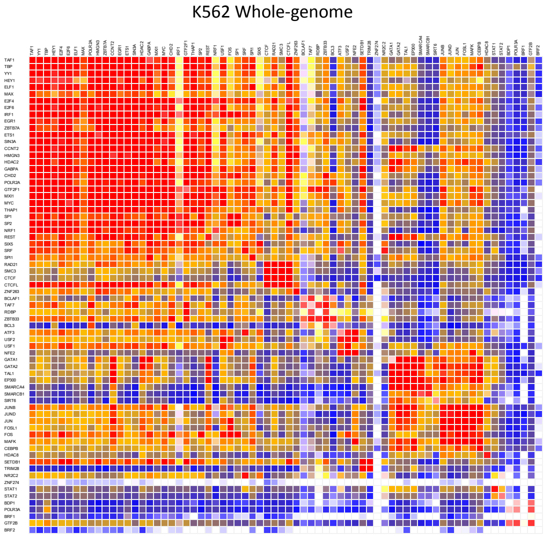
Evenly spaced over the genome



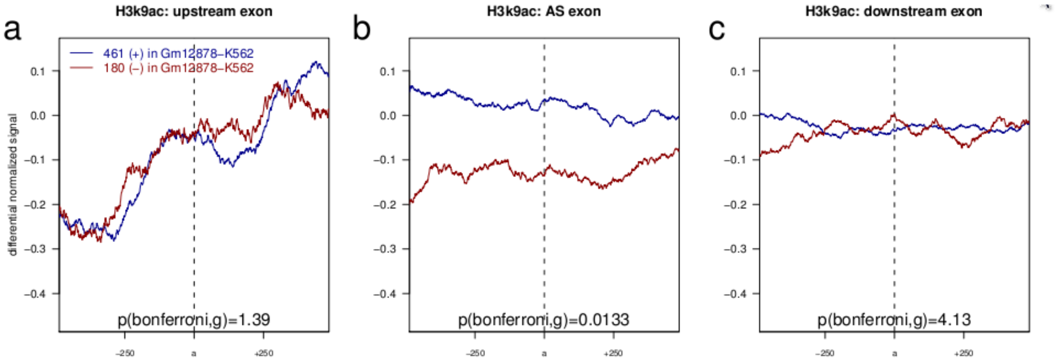
99% of the genome is within 1.7 KB of a biochemical event

95% of the genome is within 8 KB of a bound motif or footprint

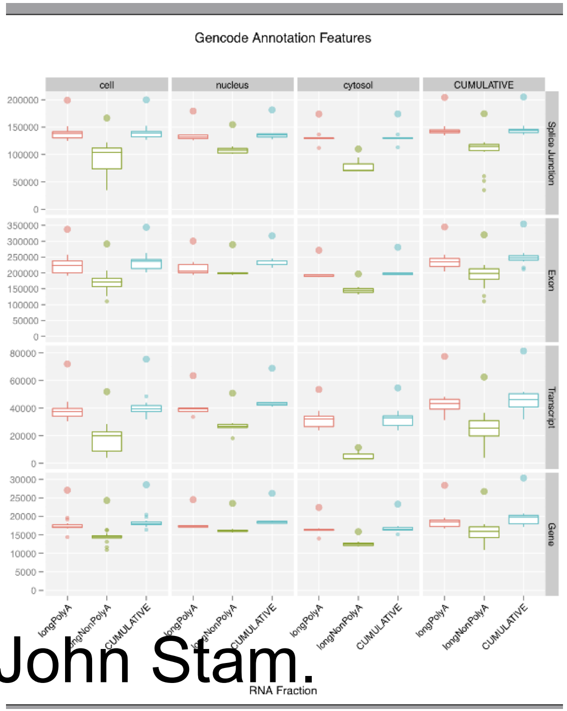
Many other stori



TF Co association,
Mike Snyder+Mark Gerstein



Splicing/Histone interaction (Roderic Guigo)

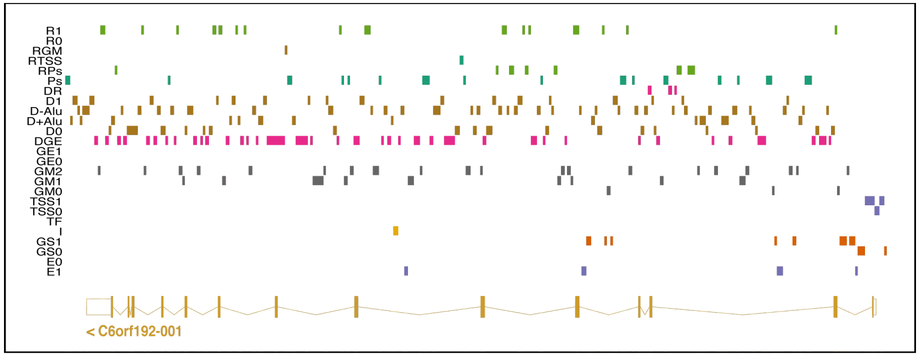


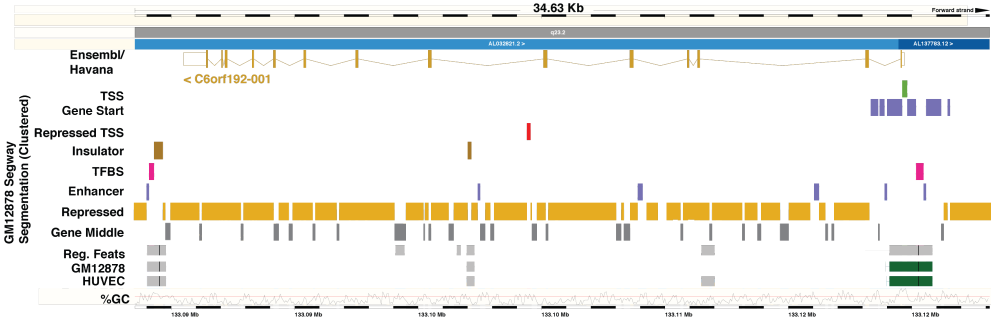
RNA landscape
Tom Gingeras

DNAseI footprints – John Stam.
DNA Methylation – Rick Myers

Discovering functional genome segments

Michael Hoffman, Jason Ernst, Bill Noble, Manolis Kellis





~7 Major flavours of genome
25 “elaborations”
1,000s of details

Well understood:
TSS, Gene Start,
Gene Bodies

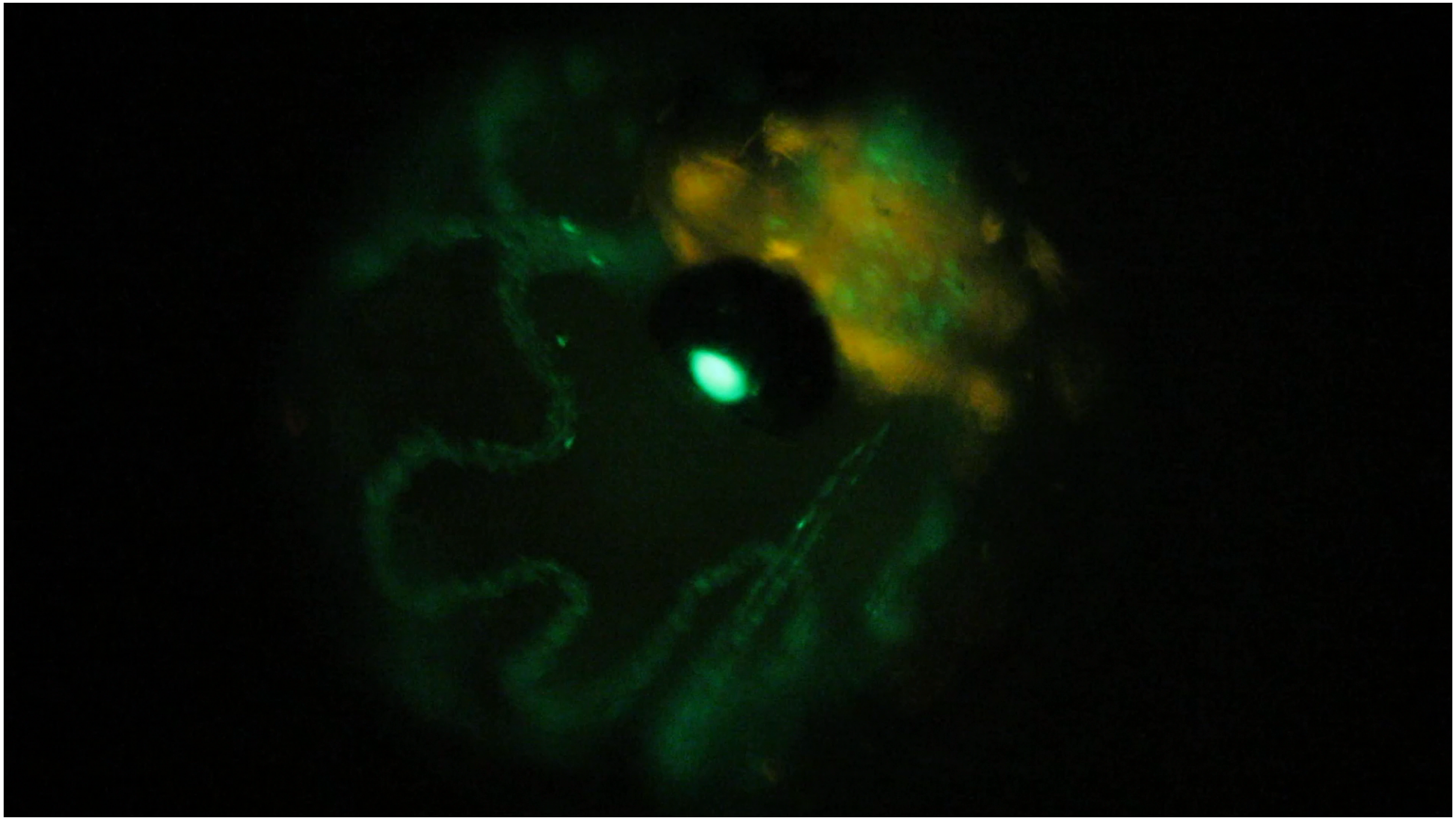
Reassuringly Interesting
“Enhancers” (2 states)
Insulators

Definitely There, Unexpected
Specific Gene End

Sub-classification of Repeats

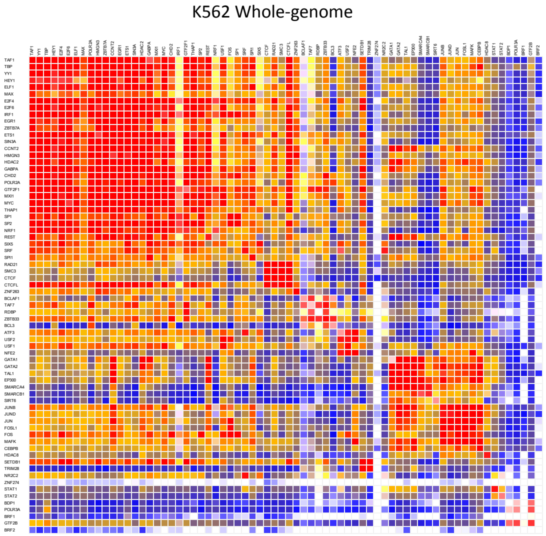
Fish Transgenics

With Wittbrodt Lab, Heidelberg

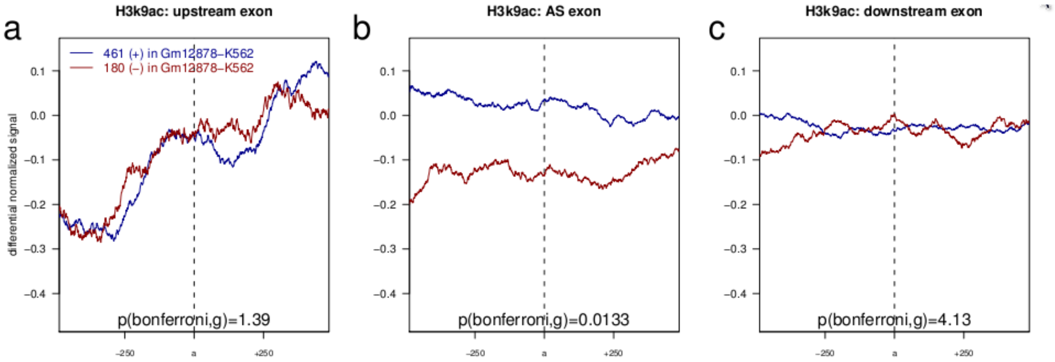


- 7 strong positives, 6 negatives from Segmentation, 3 Blood
- 2 strong positives, 4 weak, 5 negatives from Naïve picks (Fisher's Exact 0.0393)

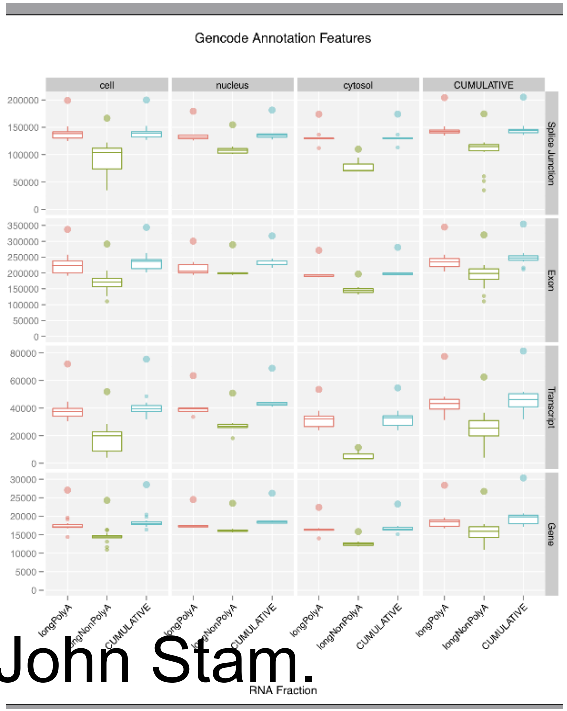
Many other stori



TF Co association,
Mike Snyder+Mark Gerstein



Splicing/Histone interaction (Roderic Guigo)



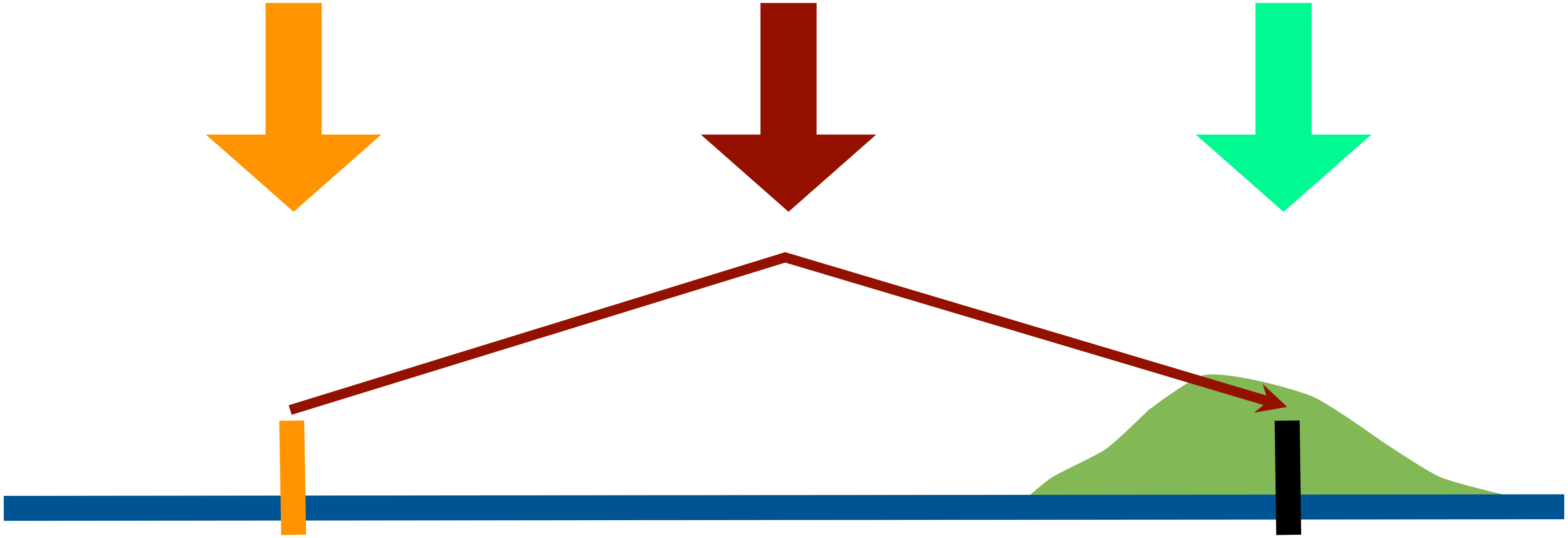
RNA landscape
Tom Gingeras

DNAseI footprints – John Stam.
DNA Methylation – Rick Myers

EMBL-EBI

Functional SNPs

Belinda Giardine, Marc Shaub, Ross Hardison, Mike Snyder, John Stam.



Genome Wide Association Studies (GWAS) Results

Linkage Disequilibrium

ENCODE Functional Region

Reported SNP

Statistically associated with the phenotype

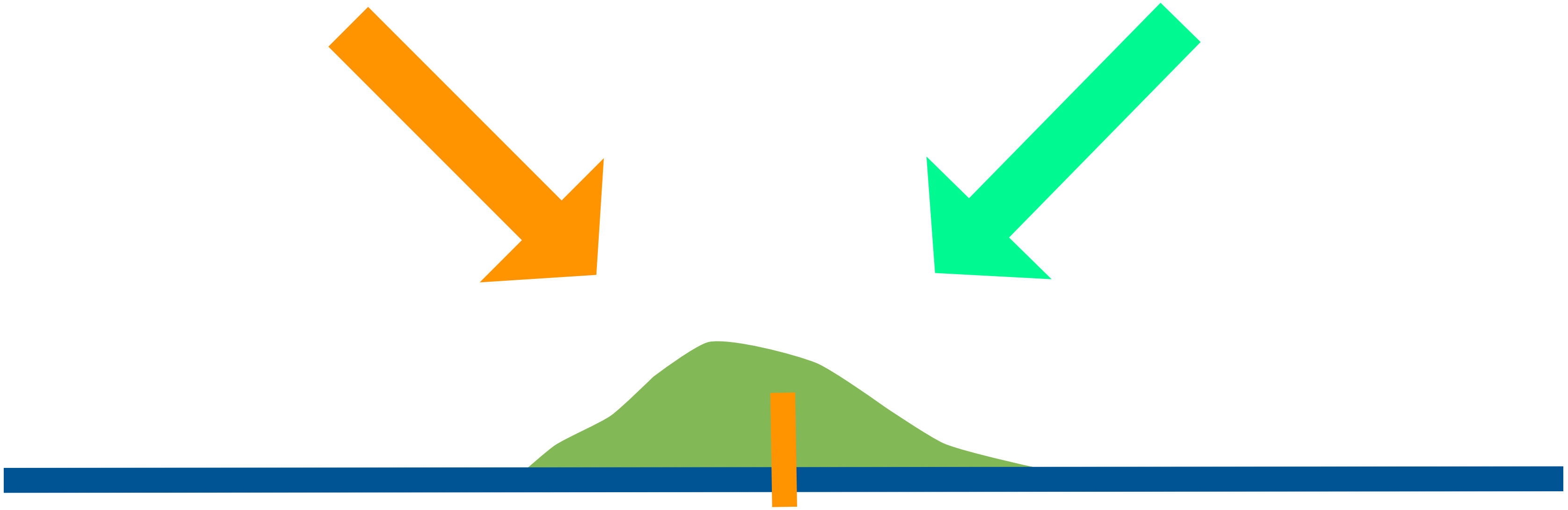
fSNP

✔ Associated with the phenotype
✔ In a functional region

Functional SNP - Direct Hit

Genome Wide Association Studies (GWAS) Results

ENCODE Functional Region

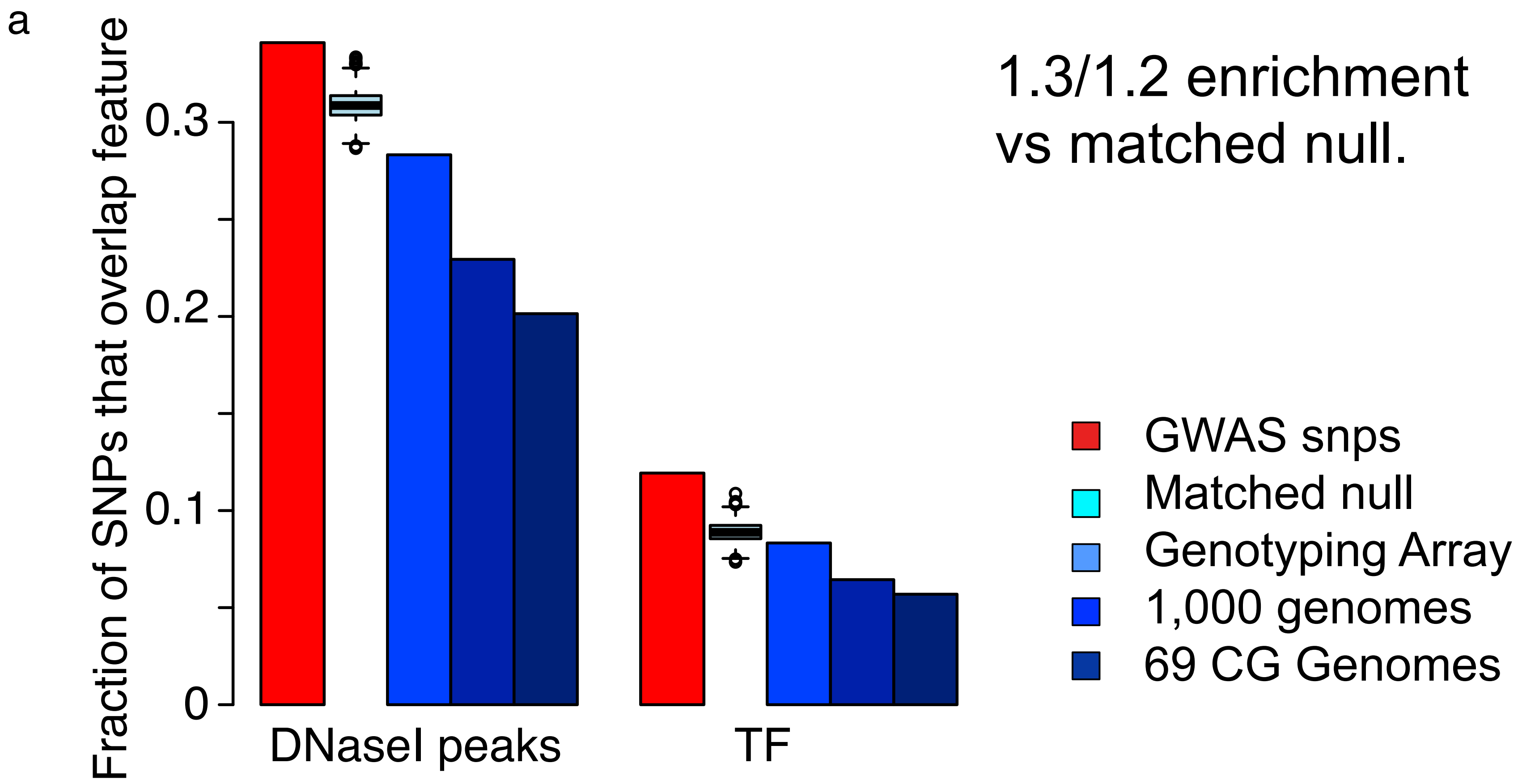


fSNP Direct Hit

✔ Association reported in a GWAS
✔ In a functional region

Direct hits

Ross Hardison, Belinda Giardine, Marc Schaub



1.3/1.2 enrichment vs matched null.

When you extend to SNPs in high LD, GWAS SNPs overlap by 80%

Association of TF or Cell Type with Disease

Ross Hardison, Belinda Giardine

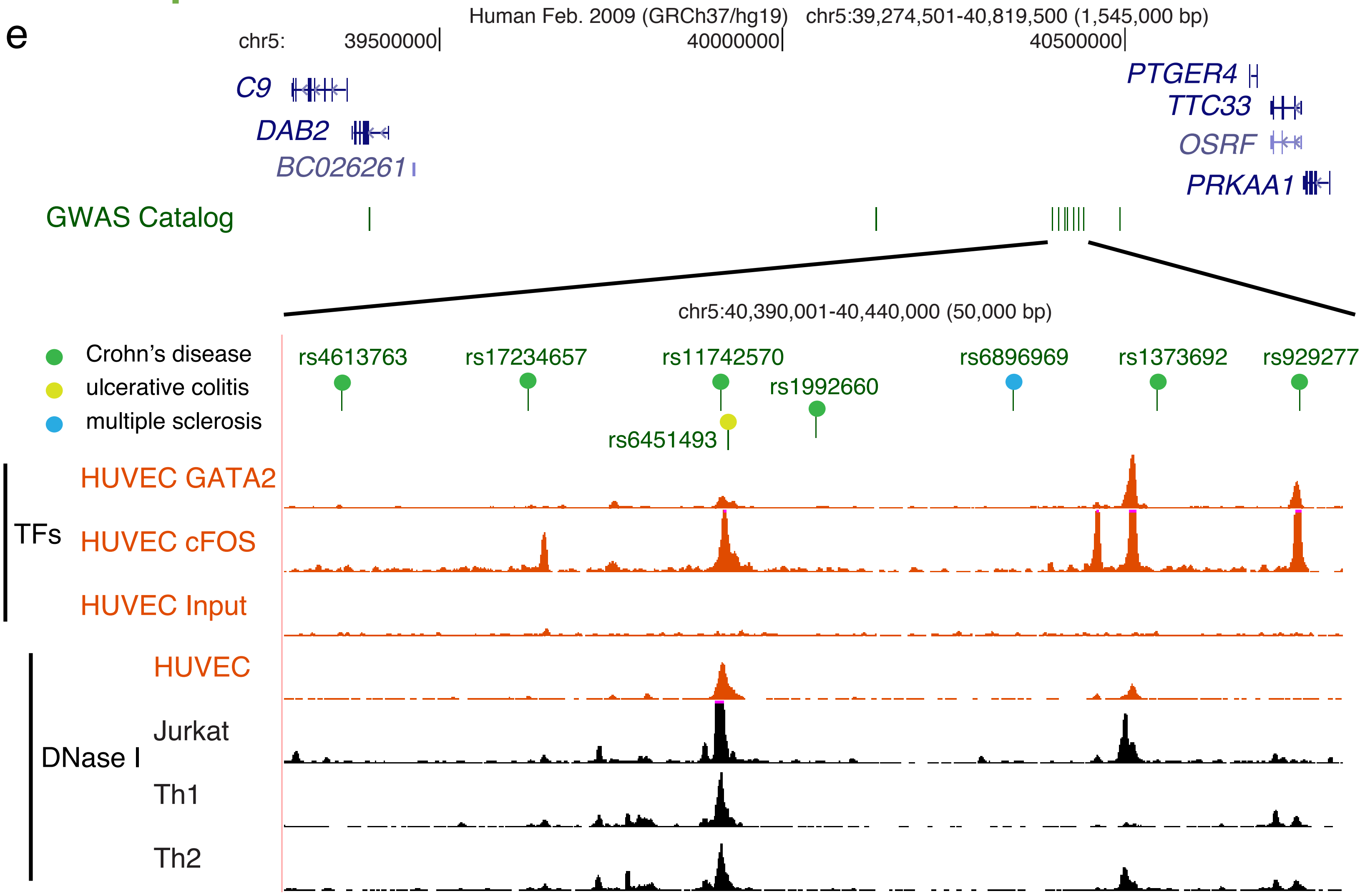
Phenotype	*SNP-Pheno associations*	*overlap any TF occupancy*	Gm12878Ebf	Gm12878Pol24h8	Gm12878Pol2	Helas3CebpbIggrab	K562Ctcf	Gm12878Pu1	Gm12878Mef2a	K562Pol2V0416101	Gm12878Pax5c20	Gm12878NfkbIggrab	Hepg2Ctcf	HuvecGata2Ucd	Gm12878Elf1sc631V0416101	Gm12878Egr1V0416101	Hepg2Mafkab50322Iggrab	HuvecCfosUcd	Gm12878Bcl11a	Gm12878Irf4	Gm12878Batf	Gm12878Tcf12	Hepg2Fosl2	K562Tal1sc12984Iggmus	K562MaxV0416102	Hepg2Tcf4Ucd	Hepg2Foxa2	Hepg2P300V0416101	Hepg2Foxa1c20	HuvecCtcf	Helas3Ctcf	Hepg2Jund	CACO2.DS8235	HUVEC.all	HepG2.all	Jurkat.DS12659	hTH1.all	hTH2.DS7842	CD34.DS16814
TOTAL	*4860*	*600*	*78*	*57*	*69*	*69*	*72*	*47*	*47*	*71*	*54*	*35*	*54*	*29*	*44*	*28*	*48*	*50*	*38*	*35*	*45*	*37*	*37*	*44*	*62*	*33*	*57*	*46*	*62*	*40*	*55*	*47*	*70*	*85*	*118*	*62*	*192*	*57*	*81*
Height	*204*	*34*	7	3	3	7	6	1	3	2	3	2	6	0	4	6	3	2	3	5	5	2	0	2	3	1	2	0	2	5	4	3	3	6	5	4	9	3	7
Systemic_lupus_erythematosus	*62*	*10*	4	6	6	2	1	1	4	0	1	4	1	1	4	2	0	1	2	3	4	2	1	0	1	0	0	0	0	1	1	1	1	2	0	0	4	2	1
Crohn's_disease	*105*	*20*	2	2	2	2	1	2	2	0	2	1	2	5	1	1	1	3	2	1	1	0	2	1	1	2	1	2	3	2	3	1	3	6	5	3	9	5	5
Ulcerative_colitis	*85*	*11*	2	3	3	0	1	2	3	1	3	3	1	2	3	2	1	1	2	1	2	2	0	2	2	1	0	2	2	0	1	1	3	2	5	3	7	2	3
Multiple_sclerosis	*71*	*15*	4	3	3	1	0	3	4	2	4	2	0	2	2	1	0	2	4	3	2	3	0	3	1	0	0	0	0	0	0	0	0	1	1	3	5	4	3
Rheumatoid_arthritis	*57*	*11*	4	2	2	1	0	4	3	0	4	4	0	0	1	1	0	0	1	0	2	2	0	1	0	0	0	0	0	0	0	0	0	2	2	1	11	3	1
LDL_cholesterol	*45*	*8*	0	0	0	2	2	1	0	4	1	0	1	0	1	0	1	0	0	0	0	0	3	0	3	2	3	2	2	1	1	3	1	0	2	1	0	1	0
Bone_mineral_density	*65*	*9*	1	1	1	1	2	2	2	1	2	1	1	0	2	2	2	0	1	2	1	1	0	0	1	0	2	2	3	1	1	1	2	2	4	3	3	2	3
Coronary_heart_disease	*107*	*17*	2	0	0	2	4	0	0	4	1	2	0	2	0	0	1	1	1	0	0	1	1	1	3	1	2	2	2	1	1	1	3	2	3	0	6	0	1
Chronic_lymphocytic_leukemia	*17*	*8*	1	4	5	0	0	3	1	0	2	1	0	0	2	0	1	0	2	1	1	2	0	1	0	1	0	0	0	0	0	0	1	0	0	0	2	0	1
Prostate_cancer	*56*	*8*	0	0	0	0	0	0	0	1	0	0	2	1	0	0	3	2	0	0	0	0	2	1	1	4	3	3	3	0	0	2	2	3	1	1	2	0	1
Triglycerides	*48*	*10*	0	0	0	1	2	0	0	2	1	0	2	1	1	0	2	2	0	0	0	0	3	1	2	1	2	2	1	2	2	3	0	2	1	0	2	1	0
Celiac_disease	*54*	*11*	4	3	3	0	2	2	1	1	1	2	0	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	2	0	1	2	0	0	0	2	2	1	2
Colorectal_cancer	*18*	*5*	0	0	0	1	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	2	3	3	3	0	0	2	2	0	1	0	2	0	1
Hematological_parameters	*85*	*12*	3	0	0	3	1	0	3	0	1	1	2	0	0	0	2	0	1	1	2	2	1	1	0	0	1	1	1	0	0	0	1	1	3	3	6	3	5
HIV-1_control	*55*	*10*	0	2	4	1	2	0	1	2	2	0	0	0	1	0	0	0	1	1	1	1	1	1	2	1	1	2	1	0	0	1	0	0	0	0	2	1	0
Protein_quantitative_trait_loci	*48*	*7*	2	2	2	0	0	2	1	1	1	0	0	1	2	2	1	0	2	1	2	1	0	0	1	0	0	0	0	0	0	1	1	1	2	1	2	1	1
Alzheimer's_disease	*42*	*5*	0	0	0	1	2	0	0	1	0	0	2	0	0	0	0	1	0	0	0	0	1	1	1	0	1	1	0	2	2	1	1	2	0	0	2	0	1
HDL_cholesterol	*55*	*8*	1	0	0	1	1	0	0	2	1	0	1	0	0	0	1	2	0	0	0	1	2	1	1	1	2	2	2	1	1	2	0	1	2	0	3	1	0
Cholesterol	*16*	*6*	1	0	0	0	2	0	0	2	2	0	1	0	1	0	0	1	0	0	0	1	1	0	1	0	0	0	0	2	1	1	0	0	2	0	1	1	0
Longevity	*30*	*5*	0	2	3	1	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	2	1	2	0	2	1	0	0	1
Attention_deficit_hyperactivity_disorder	*102*	*9*	0	0	0	1	2	0	0	1	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	2	0	1	0	0
Cognitive_performance	*111*	*8*	0	0	2	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	1	3	0	0	0	0
Type_2_diabetes	*97*	*13*	0	0	0	1	1	2	1	1	1	0	1	1	0	0	2	1	1	1	1	0	0	3	1	0	0	0	0	2	1	0	1	0	2	0	4	0	0
Conduct_disorder	*38*	*5*	0	1	1	1	3	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	2	0	0	1	2	2	1	0	1
Type_1_diabetes	*67*	*7*	2	1	1	0	0	2	1	0	1	1	0	0	1	0	0	0	1	0	1	0	0	1	1	1	1	2	2	0	1	0	1	0	2	1	5	1	1
Dialysis-related_mortality	*26*	*6*	1	0	0	1	1	1	0	0	0	0	2	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	2	1	0	0	1
Bipolar_disorder	*110*	*6*	1	0	0	2	1	0	0	1	0	0	0	1	0	0	0	3	0	0	0	0	1	0	1	0	0	0	0	0	0	1	2	3	1	0	2	0	1
Body_mass	*98*	*5*	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0	1	0	0
C-reactive_protein	*34*	*7*	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	1	0	2	1	0	0	1
Menarche_and_menopause	*62*	*6*	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	1	0	0	2	1	1	0
Breast_cancer	*43*	*6*	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	2	2	0	0
Mean_platelet_volume	*15*	*5*	1	0	0	0	0	1	0	0	1	0	0	0	0	0	2	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	2	0	1
Soluble_levels_of_adhesion_molecules	*5*	*5*	1	0	1	1	0	1	0	2	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Psoriasis	*38*	*6*	1	1	2	0	0	2	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	4	0	0
Parkinson's_disease	*46*	*5*	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0
Obesity	*36*	*5*	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fasting_glucose-related_traits	*17*	*5*	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	1	1	1	1	0	0

Zoom in…

Phenotype	*SNP-Pheno associations*	*overlap any TF occupancy*	Gm12878Mef2a	Gm12878Pol2	Gm12878Ebf	Gm12878Pol24h8	Gm12878NfkbIggrab	Gm12878Irf4	Gm12878Pax5c20	Gm12878Pu1	Gm12878Batf	HuvecGata2Ucd	Gm12878Egr1V0416101	Helas3CebpbIggrab	Hepg2Ctcf	HUVEC.all	hTH1.all	hTH2.DS7842
TOTAL	*4860*	*600*	*47*	*69*	*78*	*57*	*35*	*35*	*54*	*47*	*45*	*29*	*28*	*69*	*54*	*85*	*192*	*57*
Multiple_sclerosis	*71*	*15*	4	3	4	3	2	3	4	3	2	2	1	1	0	1	5	4
Systemic_lupus_erythematosus	*62*	*10*	4	6	4	6	4	3	1	1	4	1	2	2	1	2	4	2
Height	*204*	*34*	3	3	7	3	2	5	3	1	5	0	6	7	6	6	9	3
Rheumatoid_arthritis	*57*	*11*	3	2	4	2	4	0	4	4	2	0	1	1	0	2	11	3
Chronic_lymphocytic_leukemia	*17*	*8*	1	5	1	4	1	1	2	3	1	0	0	0	0	0	2	0
Celiac_disease	*54*	*11*	1	3	4	3	2	1	1	2	1	0	0	0	0	0	2	1
Ulcerative_colitis	*85*	*11*	3	3	2	3	3	1	3	2	2	2	2	0	1	2	7	2
Crohn’s_disease	*105*	*20*	2	2	2	2	1	1	2	2	1	5	1	2	2	6	9	5

Example loci

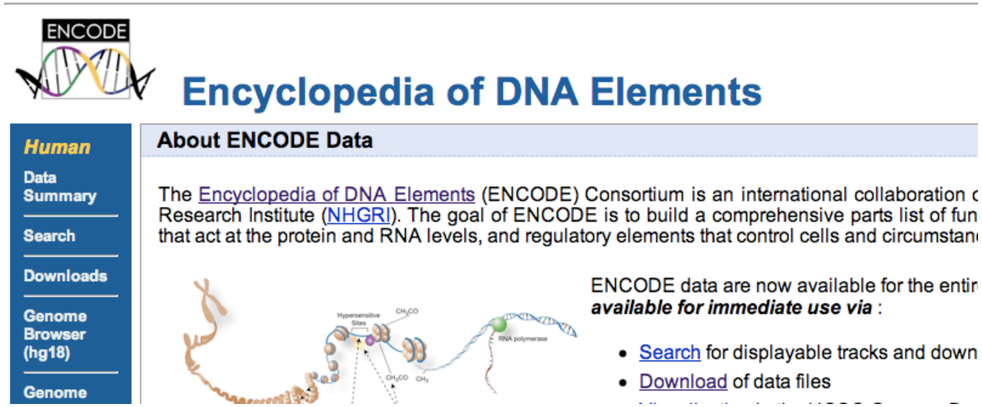
Ross Hardison, Belinda Giardine



Organisation+Access for the data

wwww.encodeproject.org (UCSC)

“Factorbook”



ENCODE

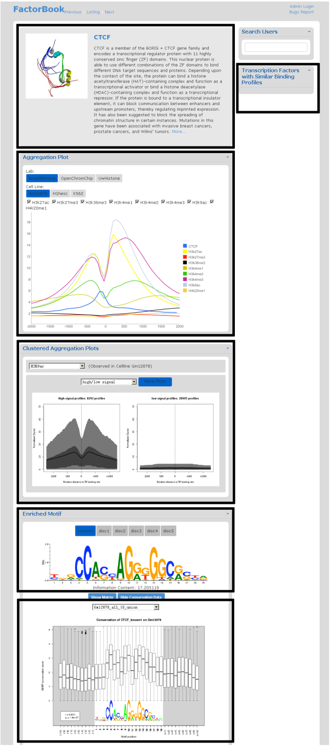
Encyclopedia of DNA Elements

About ENCODE Data

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration c
Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of fun
that act at the protein and RNA levels, and regulatory elements that control cells and circumstan

ENCODE data are now available for the entir
available for immediate use via :

- Search for displayable tracks and down
- Download of data files



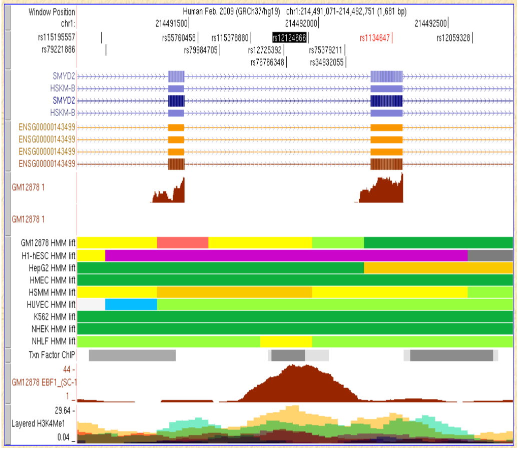
FactorBook
Previous | Listing | Next
Admin Login
Bugs Report
Search Users
Transcription Factors with Similar Binding Profiles

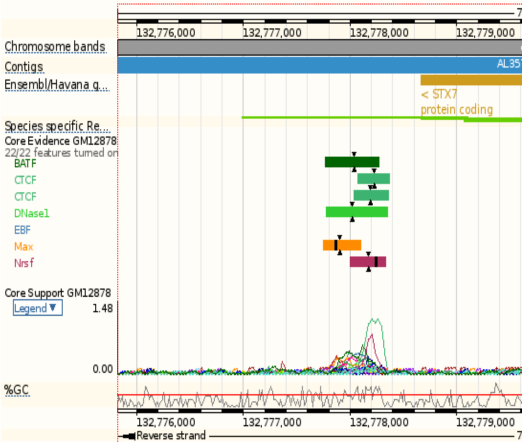
CTCF

CTCF is a member of the BORIS + CTCF gene family and encodes a transcriptional regulator protein with 11 highly conserved zinc finger (ZF) domains. This nuclear protein is able to use different combinations of the ZF domains to bind different DNA target sequences and proteins. Depending upon the context of the site, the protein can bind a histone acetyltransferase (HAT)-containing complex and function as a transcriptional activator or bind a histone deacetylase (HDAC)-containing complex and function as a transcriptional repressor. If the protein is bound to a transcriptional insulator element, it can block communication between enhancers and upstream promoters, thereby regulating imprinted expression. It has also been suggested to block the spreading of chromatin structure in certain instances. Mutations in this gene have been associated with invasive breast cancers, prostate cancers, and Wilms' tumors. More...

UCSC Genome Browser

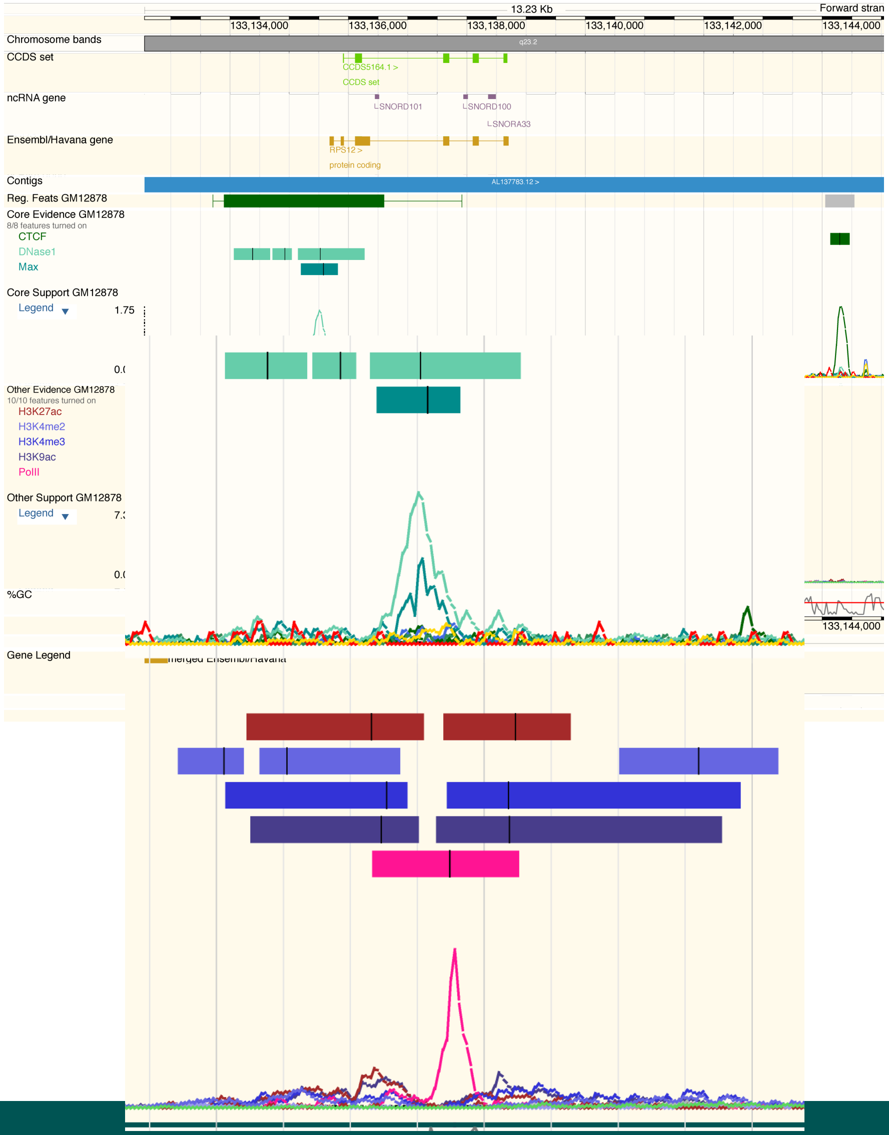
Ensembl





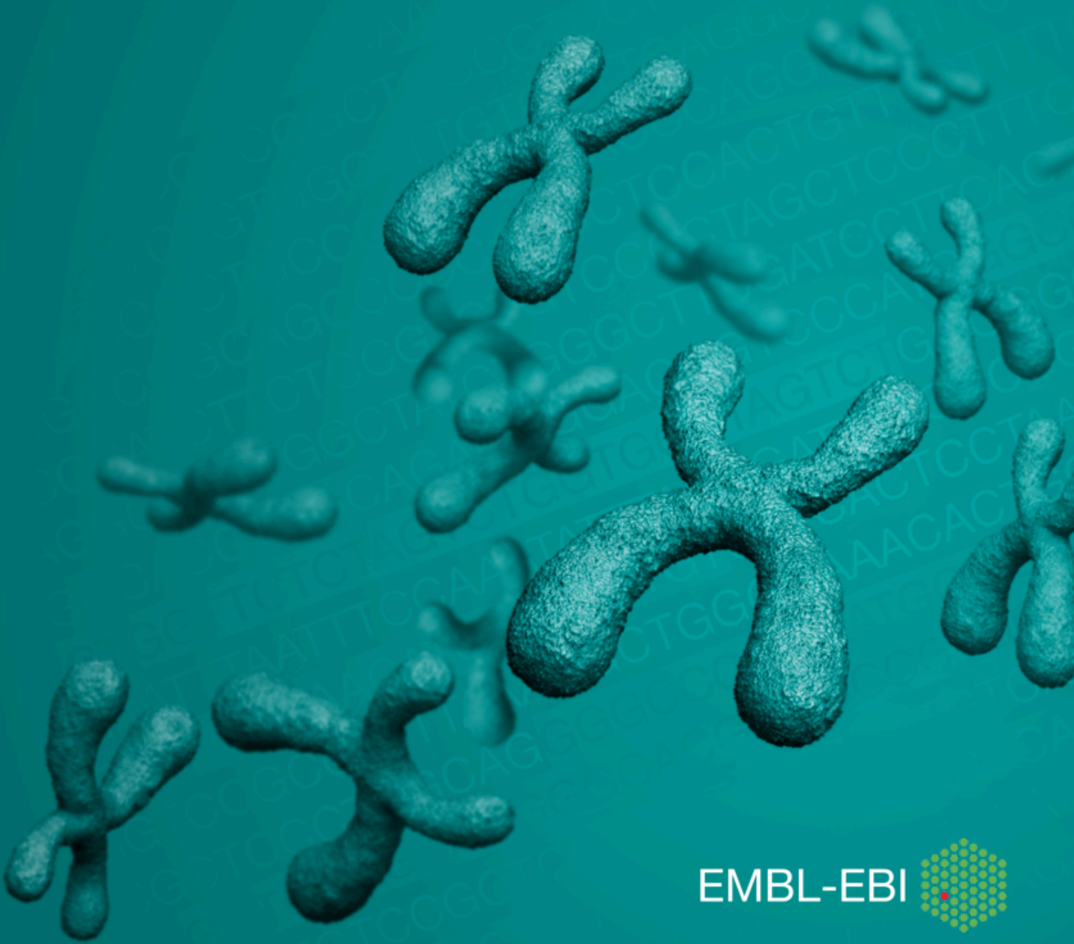
Raw Data in GEO/SRA

Ensembl Regulation



- Integration over ENCODE, NIH Epigenome Roadmap, IHEC, some individual lab
- “Experiment”, not per-replicate level view, only QC passed data
- Progressive integration into “one track” on the genome
- Integration into VEP - variant effect predictor (with SIFT and POLYPHEN as well!)
- Future:
  - Linking of genes to regulatory element, Linking of tfs to phenotypes
- Ensembl workshops (we come to you)
- Ensembl course (you come to us)
- helpdesk@ensembl.org

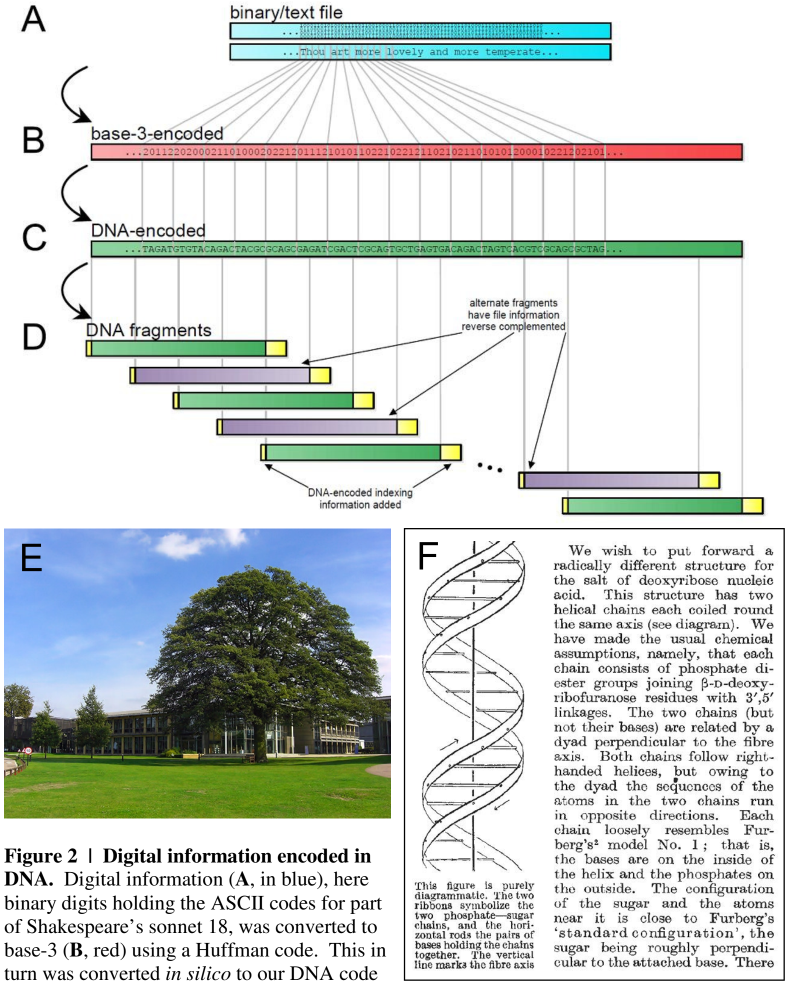
And… just for fun…



Over a beer…

Ha! At some point all the data we
Store is going to be DNA…

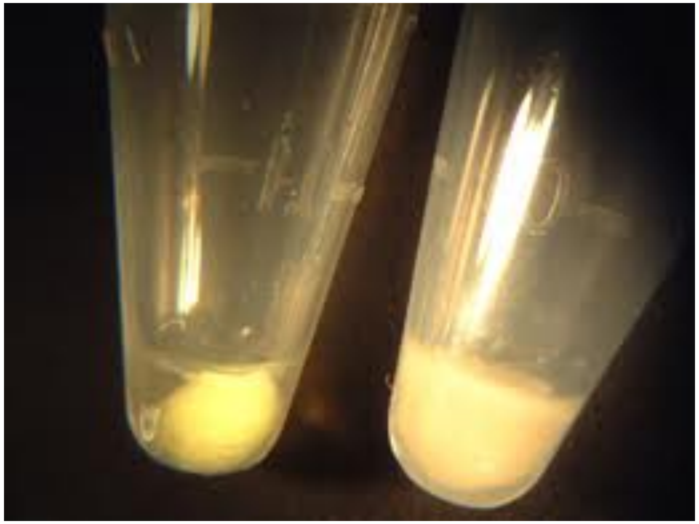
Of course, the cost effective way
To store this would be as DNA…


A
binary/text file
...Thou art more lovely and more temperate...
B
base-3-encoded
...20112202000211010002022120111210101102210221211021021101010120001022120210 1...
C
DNA-encoded
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
D
DNA fragments
alternate fragments have file information reverse complemented
DNA-encoded indexing information added
E
F

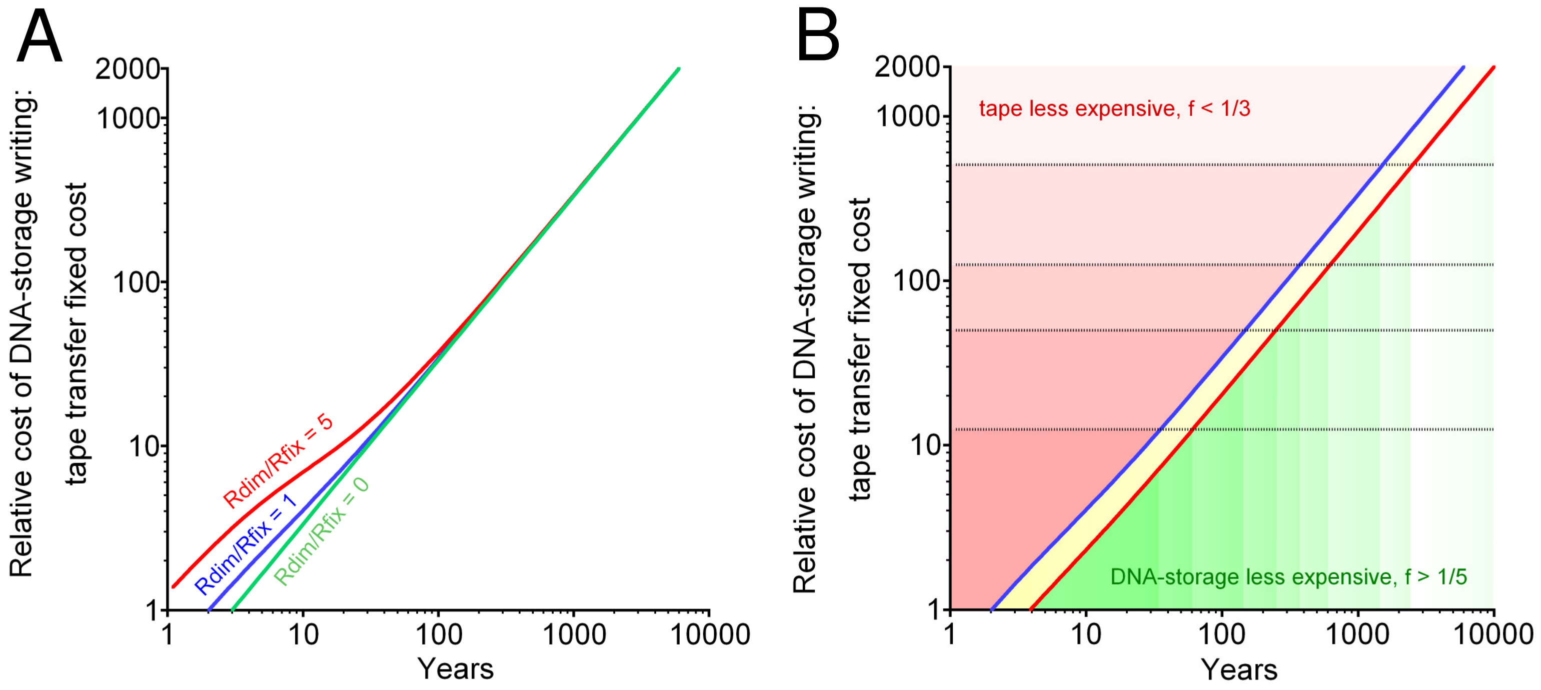
We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Figure 2 | Digital information encoded in DNA. Digital information (**A**, in blue), here binary digits holding the ASCII codes for part of Shakespeare's sonnet 18, was converted to base-3 (**B**, red) using a Huffman code. This in turn was converted *in silico* to our DNA code (**C**, green), with no homopolymers, which formed the basis for a large number of overlapping DNA segments each containing 100 bases of encoded information (**D**, green or, with alternate segments reverse complemented for added data security, violet) and with orientation and indexing DNA codes added (yellow, as described in the text). These strings were synthesised, sequenced and decoded. **E**, A digital photograph of the EMBL-European Bioinformatics Institute (JPEG 2000 format) and **F**, an extract of the Watson and Crick (1953) paper[10] (PDF format) that were among the files encoded in DNA and successfully recovered in this study.



Cost effective?



Dave Simonds

ENCODE Authors

Ian Dunham,
Anshul Kundaje

Shelley F. Aldred Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Vishwanath R. Iyer, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein

Funded by NIH-NHGRI

*** Overall Coordination ***
Ian Dunham (1), Anshul Kundaje (2).

*** Data Production Leads ***
Shelley F. Aldred (3), Patrick J. Collins (3), Carrie A. Davis (4), Francis Doyle (5), Charles B. Epstein (6), Seth Frietze (7), Jennifer Harrow (8), Vishwanath R. Iyer (9), Rajinder Kaul (10), Jainab Khatun (11), Bryan R. Lajoie (12), Stephen G. Landt (13), Bum-Kyu Lee (9), Florencia Pauli (14), Kate R. Rosenbloom (15), Peter Sabo (16), Alexias Safi (17), Amartya Sanyal (12), Noam Shoresh (6), Jeremy M. Simon (18), Lingyun Song (17), Nathan D. Trinklein (3).

*** Lead Analysts ***
Robert C. Altshuler (19), E
Ian Dunham (1), Jason Ern
Ross C. Hardison (26), Rol
Kellis (19), Jainab Khatun
Xinying Lin (23), Angelika
Rozowsky (21), Felix Schle
Steven P. Wilder (1), Weish

*** Writing Group ***
Bradley E. Bernstein (33),

*** NHGRI Project Manager
Leslie B. Adams (36), Laur
Kelly (36), Rebecca F. Low

*** Principal Investigators
Bradley E. Bernstein (33),
Farnham (7), Mark Gerstein
Guig<F3> (40), Ross C. Ha
(18), Elliott H. Margulies (3
Tenenbaum (5), Zhiping We

*** Broad Institute Group *
Bradley E. Bernstein (33),
Mikkelsen (6), Shawn Gillespie (43), Alon Goren (33), Oren Ram (33), Xiaolan Zhang (6), Li Wang (6), Robbyn Issner (6),
Michael J. Coyne (6), Timothy Durham (6), Manching Ku (33), Thanh Truong (6), Lucas D. Ward (19), Robert C. Altshuler
(19), Matthew Eaton (19), Manolis Kellis (19).

*** Boise State University Proteomics Group ***
Jainab Khatun (11), Yanbao Yu (44), John Wrobel (11), Brian A. Risk (11), Harsha Gunawardena (44), Heather C. Kuiper (44), Christopher W. Maier (44), Ling Xie (44), Xian Chen (44), Morgan C. Giddings (11).

*** Data Coordinatio
Katrina Learned (15) enkat S. Malladi (15), Kate R. Rosenblo n (15), Cricket A. Slo (15), M hew C. Wong (15), Galt
P. Barber (15), Melis
Kent (15), Vanessa M Kirkup (1 Lau nce Me r ( ), Je y Mor n Maddr (15)
(15).

*** Sanger Institute, Washington University, Yale University, Center for Genomic Regulation, Barcelona, UCSC, MIT, University of Lausanne, CNIO Group ***
Bronwen Aken (8), Roger P. Alexander (21), Suganthi Balasubramanian (21), Daniel Barrell (8), Gemma Barson (8), Andrew Berry (8), Nitin Bhardwaj (21), Alexandra Bignell (8), Veronika Boychenko (8), Michael Brent (45), Giovanni Bussotti (22), Chao Cheng (21), Jacqueline Chrast (46), Claire Davidson (8), Thomas Derrien (47), Gloria Despacio-Reyes (8), Mark Diekhans (48), Jason Ernst (19), Iakes Ezkurdia (49), Julio Fernandez Banet (8), Adam Frankish (8), Mark Gerstein (21), James Gilbert (8), Jose Manuel Gonzalez (8), Ed Griffiths (8), Roderic Guig<F3> (40), Lukas Habegger (21), Jennifer Harrow (8), Rachel Harte (48), David Haussler (50), C<E9>dric Howald (51), Timothy J. Hubbard (8), Toby Hunt (8), Mike Kay (8), Manolis Kellis (19), Pouya Kheradpour (19), j Khurana (21), Felix Kokocinski (8), Jing (Jane) Leng (21), Michael F. Lin (19), Lucas Lochovsky (21), Jane Loveland (8), Zhi Lu (52), Deepa Manthravadi (8), Marco Mariotti (22), Renqiang Min (21), Xinmeng (Jasmine) Mu (21), Jonathan Mudge (8), Gaurab Mukherjee (8), Cedric Notredame (22), Baikang Pei (21), Alexandre Reymond (51), Jose Manuel Rodriguez (49), Joel Rozowsky (21), Gary Saunders (8), Andrea Sboner (53), Stephen Searle (8), Cristina Sisu (21), Catherine Snow (8), Charlie Steward (8), Andrea Tanzer (54), Electra Tapanari (8), Michael L. Tress (49), Alfonso Valencia (49), Marijke J. van Baren (55), Nathalie Walters (46), Laurens Wilming (8), Koon-Kiu Yan (21), Kevin (Yuk-Lap) Yip (21), Amonida Zadissa (8), Zhengdong Zhang (56), Arif Harmanci (21), Alexej Abyzov (21).

*** Genome Institute of Singapore Group ***

*** HudsonAlpha
Devin M. Absher
Bowling (14), Ma
(14), DeSalvo Gi
Levy (14), Max W
(30), Michael A.
Newberry (14), S
Barbara Pusey (
Vielmetter (42), E

*** University of
Job Dekker (12), Gaurav Jain (12), Bryan R. Lajoie (12), Amartya Sanyal (12).

*** University of Massachusetts Medical School Weng Group ***
Zhiping Weng (23), Troy W. Whitfield (23), Jie Wang (23), Patrick J. Collins (3), Shelley F. Aldred (3), Nathan D. Trinklein (3), E. Christopher Partridge (14), Richard M. Myers (14).

David Gonzalez (22), Assaf Gordon (4), Harsha Gunawardena (44), C<E9>dric Howald (51), Sonali Jha (4), Rory Johnson (22), Philipp Kapranov (68), Brandon King (59), Colin Kingswood (70), Guoliang Li (71), Oscar J. Luo (57), Eddie Park (30), Jonathan B. Preall (4), Kimberly Presaud (4), Paolo Ribeca (67), Brian A. Risk (11), Daniel Robyr (72), Xiaoan Ruan (57), Michael Sammeth (67), Kuljeet Singh Sandhu (57), Lorain Schaeffer (59), Lei-Hoon See (4), Atif Shahab (57), Jorgen Skancke (22), Ana Maria Suzuki (66), Hazuki Takahashi (66), Hagen Tilgner (73), Diane Trout (59), Nathalie Walters (46), Huaien Wang (4), John Wrobel (11), Yanbao Yu (44), Yoshihide Hayashizaki (66), Jennifer Harrow (8), Mark Gerstein (21), Timothy J. Hubbard (8), Alexandre Reymond (51), Stylianos E. Antonarakis (72), Gregory J. Hannon (4), Morgan C. Giddings (11), Yijun Ruan (57), Barbara Wold (42), Piero Carninci (66), Roderic Guig<F3> (40), Thomas R. Gingeras (39).

*** University of W ngt U ers sachusett ica Cen Gro ***
Job Dekker (12), M rg Dieg (1 Erica st R. co ser 4), Er Ha en (ch
(16 urav Jain), A K nns (16) ajinder Ka (10 tya V. Ku avi 6), Br R Lajoie
Kris Lee (16), P rick as), ane J eph (16), denc V. N (16), ex Re (16), c Ryn
Sabo (16), Richard S. Sandstrom (16), Daniel L. Bates (16), Theresa K. Canfield (16), Amartya Sanyal (12),
Anthony O. Shafer (16), George Stamatoyannopoulos (75), John A. Stamatoyannopoulos (16), Sean Thomas (16),
Bob Thurman (16), Shinny Vong (16), Hao Wang (16), Molly A. Weaver (16).

*** Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UC Davis Group ***
Alexej Abyzov (21), Nick Addleman (13), Roger P. Alexander (21), Raymond K. Auerbach (76), Suganthi Balasubramanian (21), Keith Bettinger (13), Nitin Bhardwaj (21), Alan P. Boyle (13), Alina R. Cao (77), Philip Cayting (13), Alexandra Charos (78), Chao Cheng (21), Yong Cheng (13), Catharine Eastman (13), Ghia Euskirchen (13), Peggy J. Farnham (7), Joseph D. Fleming (79), Seth Frietze (7), Mark Gerstein (21), Fabian Grubert (13), Lukas Habegger (21), Manoj Hariharan (13), Arif Harmanci (21), Sushma Iyengar (80), Victor X. Jin (81), Konrad J. Karczewski (13), Maya Kasowski (13), j Khurana (21), Phil Lacroute (13), Hugo Lam (13), Nathan Lamarre-Vincent (79), Stephen G. Landt (13), Jing (Jane) Leng (21), Jin Lian (82), Marianne Lindahl-Allen (79), Lucas Lochovsky (21), Zhi Lu (52), Renqiang Min (21), Benoit Miotto (79), Hannah Monahan (78), Zarmik Moqtaderi (79), Xinmeng (Jasmine) Mu (21), Henriette O'Geen (77), Zhengqing Ouyang (13), Dorrelyn Patacsil (13), Baikang Pei (21), Debasish Raha (78), Lucia Ramirez (13), Brian Reed (78), Joel Rozowsky (21), Andrea Sboner (53), Minyi Shi (13), Cristina Sisu (21), Teri Slifer (13), Michael Snyder (35), Kevin Struhl (79), Sherman M. Weissman (82), Heather Witt (83), Linfeng Wu (13), Xiaoqin Xu (77), Koon-Kiu Yan (21), Xinqiong Yang (13), Kevin (Yuk-Lap) Yip (21), Zhengdong

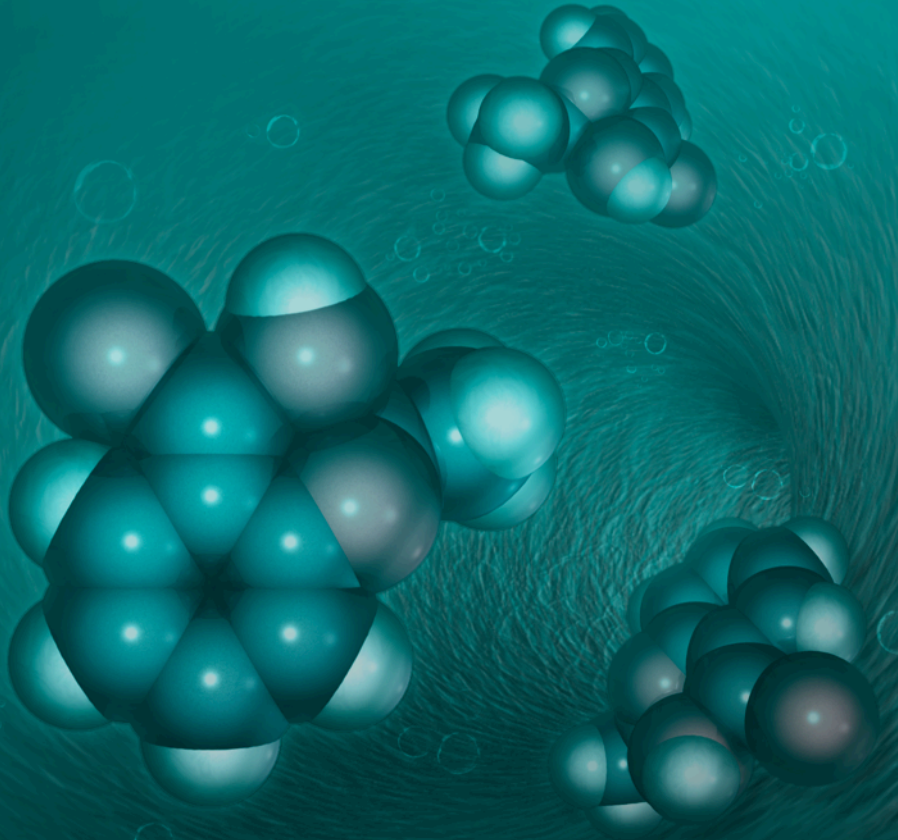
***
lva
roup
na
Lijia
Michael Snyder (35), Kevin P. White (41).

*** University of Heidelberg Group ***
Thomas Auer (85), Lazaro Centanin (85), Michael Eichenlaub (85), Franziska Gruhl (85), Stephan Heermann (85), Daigo Inoue
5),
(85),
Sinn

lvis
Herrero
efe (1),
combe
uerizas
dine
echt (2),
e (2),
(25),
rris
kel
Boley
Huang
Li (88),
William Stafford Noble (89), Jeffrey A. Bilmes (90), Orion J. Buske (16), Michael M. Hoffman (16), Avinash D. Sahu (16), Peter V. Kharchenko (91), Peter J. Park (91), Zhiping Weng (23), Sowmya Iyer (27), Xianjun Dong (23), Melissa Greven (23), Xinying Lin (23), Jie Wang (23), Hualin S. Xi (32), Jiali Zhuang (23), Alexej Abyzov (21), Roger P. Alexander (21), Suganthi Balasubramanian (21), Nitin Bhardwaj (21)
, Lukas Habegger (21),
1), j Khurana (21), Jing
Lucas Lochovsky (21), Zhi Lu (52), Renqiang Min (21), Xinmeng (Jasmine) Mu (21), Baikang Pei (21), Andrea Sboner (53), Cristina Sisu (21), Koon-Kiu Yan (21), Kevin (Yuk-Lap) Yip (21), Mark Gerstein (21), Joel Rozowsky (21), Zhengdong Zhang (56), Ewan Birney (1).

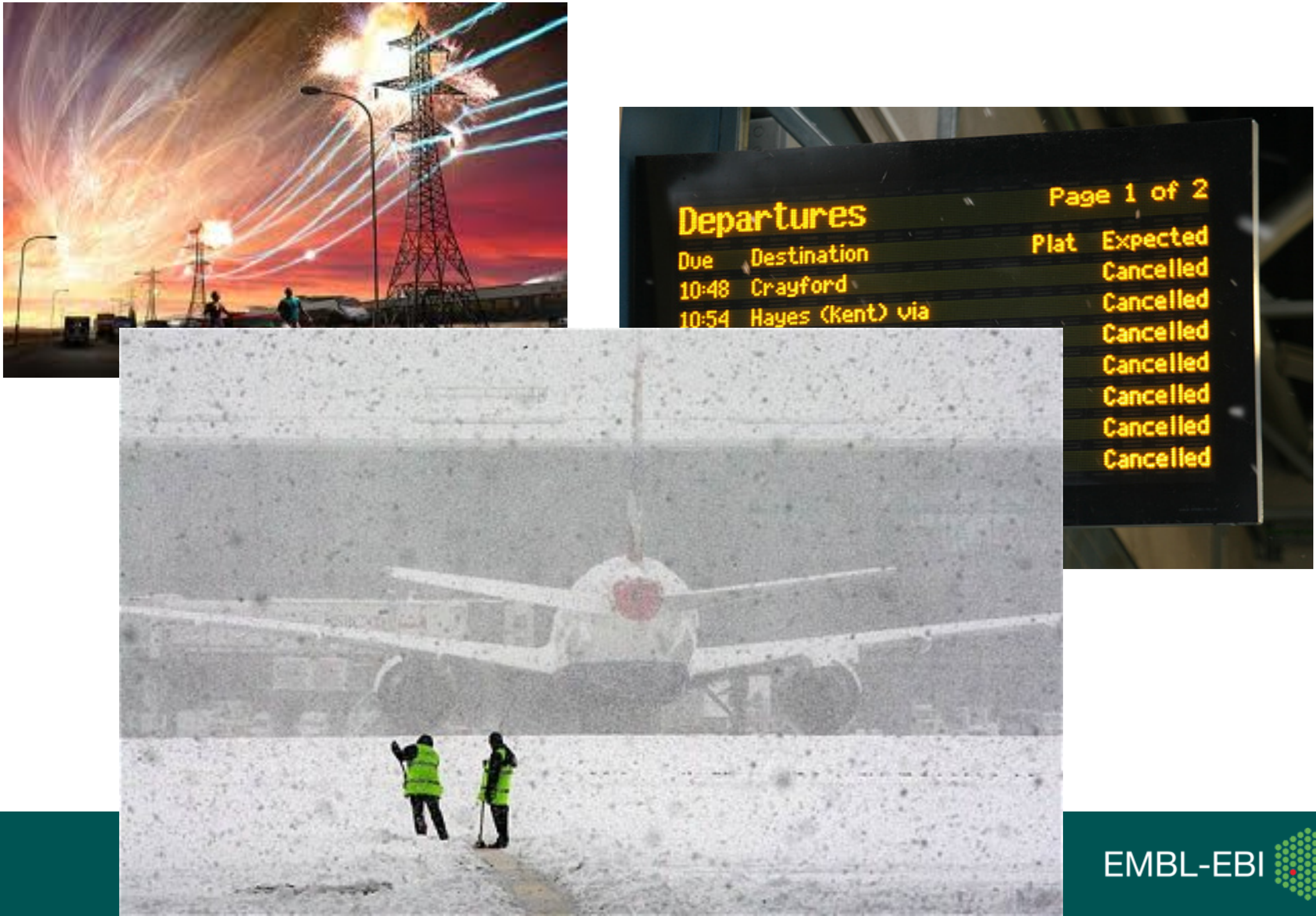
EMBL-EBI

Why we need a infrastructure



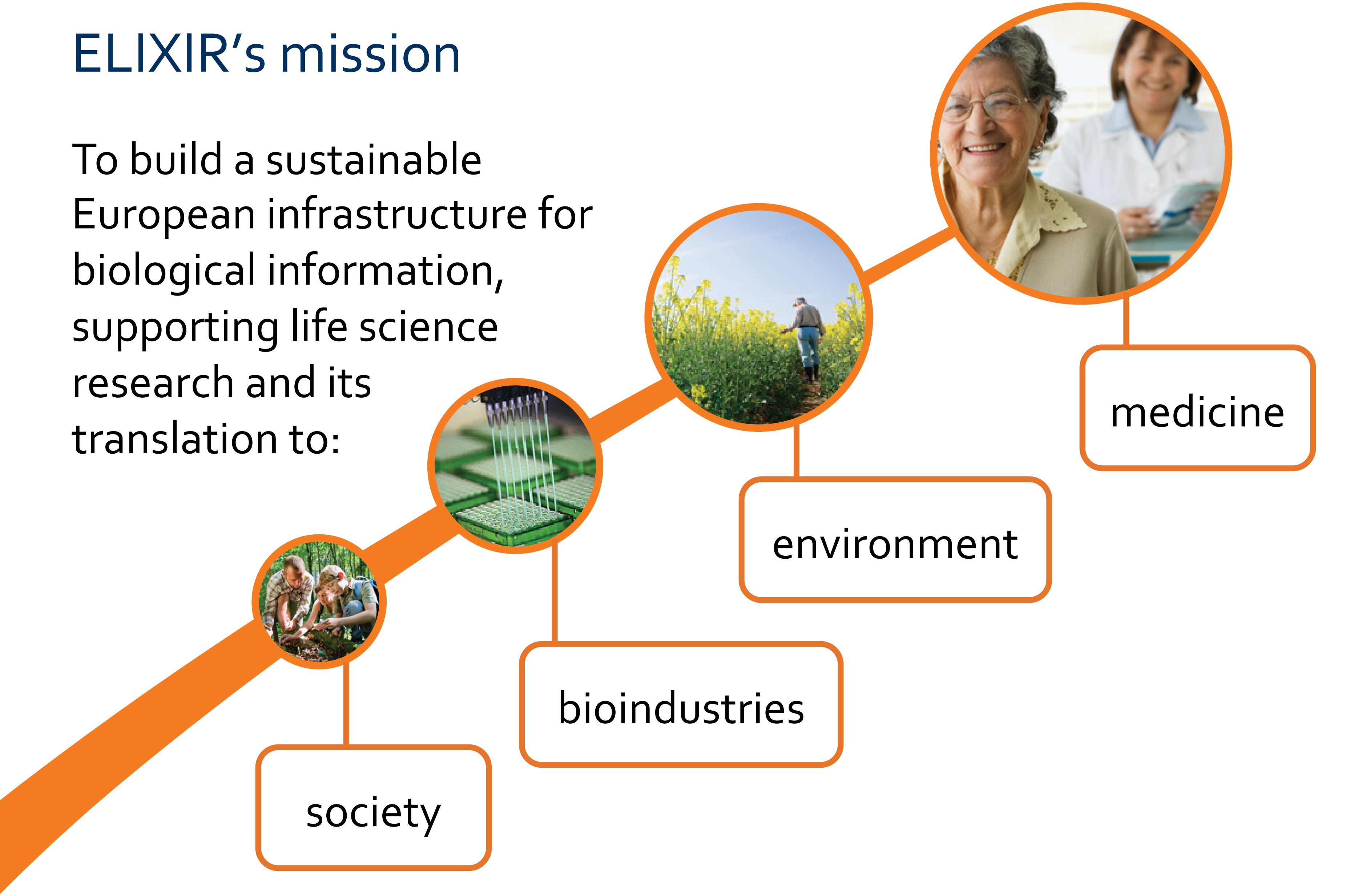
Infrastructures are critical…

But we only notice them when they go wrong



ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:



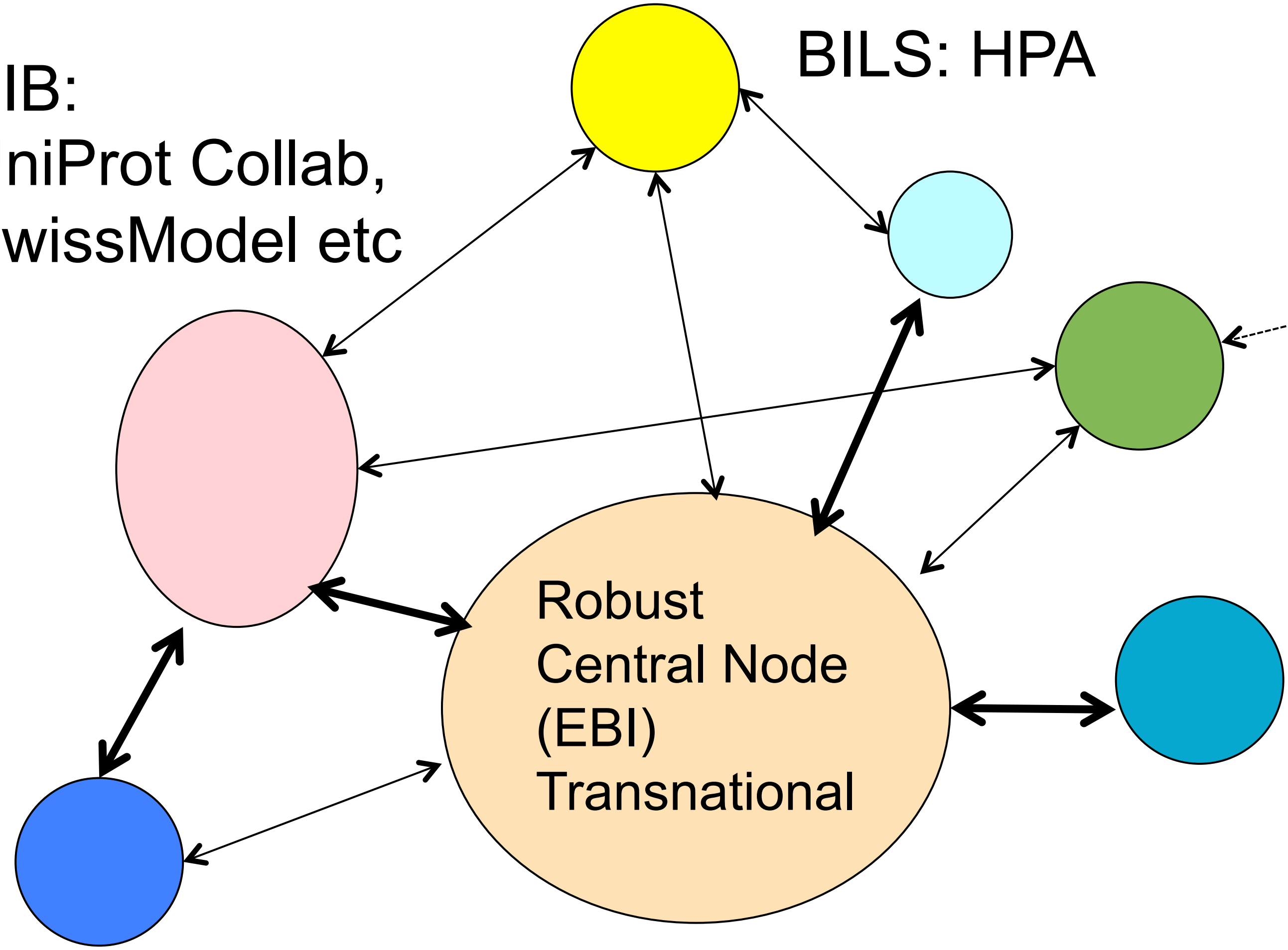
society

bioindustries

environment

medicine

Robust network with a strong hub



SIB:
UniProt Collab,
SwissModel etc

BILS: HPA

Node

“Domain
Specific hub”

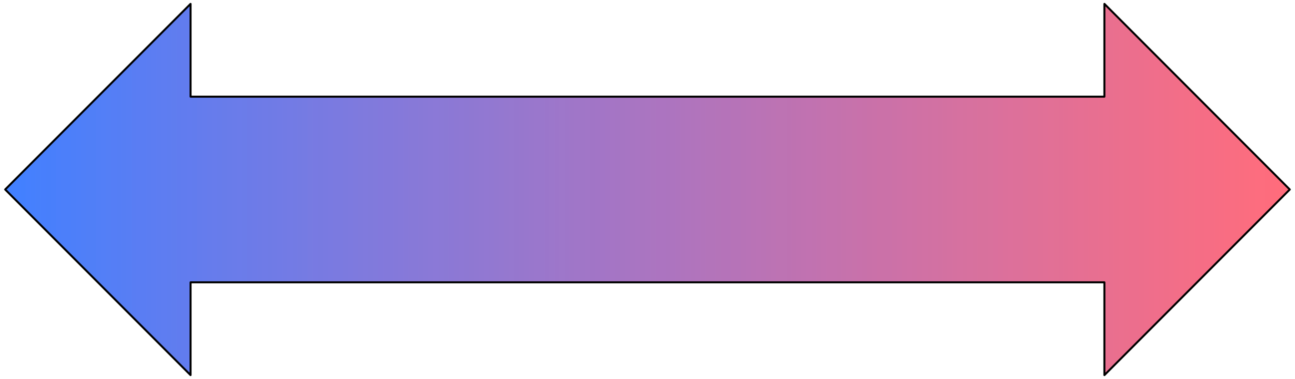
“National”

Robust
Central Node
(EBI)
Transnational

CRG Provides
EGA collaboration

Research

Healthcare



International
EBI / Elixir
English
Low legalities

National
Healthcare
National Language
Complex legalities

Other infrastructures needed for biology

- EuroBioImaging
  - Cellular and whole organism Imaging
- BioBanks (BBMRI)
  - We need numbers – European populations – in particular for rare diseases, but also for specific sub types of common disease
- Mouse models and phenotypes (Infrafrontier)
  - A baseline set of knockouts and phenotypes in our most tractable mammalian model
  - (it’s hard to *prove* something in human)
- Robust molecular assays in a clinical setting (EATRIS)
  - The ability to reliably use state of the art molecular techniques in a clinical research setting

Questions?

(you can follow me on twitter @ewanbirney)
I blog and update this on Google Plus publically

EMBL-EBI is funded by the 20 member states of EMBL, Wellcome Trust, European Union FP7, NIH, BBSRC, MRC and over 20 other funding agencies