From Genome Function to Biomedical Insights

Ewan Birney

Associate Director, EMBL-EBI

EMBL-EBI

Overview of the talk

- Impact of ENCODE data on understanding human biology
  - Basic Biology
  - Common Disease
  - Cancer
- Themes on how data is used
- What the human community needs over the next decade

Publications that use ENCODE data

Publication status

published	821

Category

basic biology	322
human disease	201
model organism biology	158
software tool	146
data standard	5

+ See more...

Publication year

2013	264
2014	220
2012	162
2011	111
2010	34

+ See more...

Journal

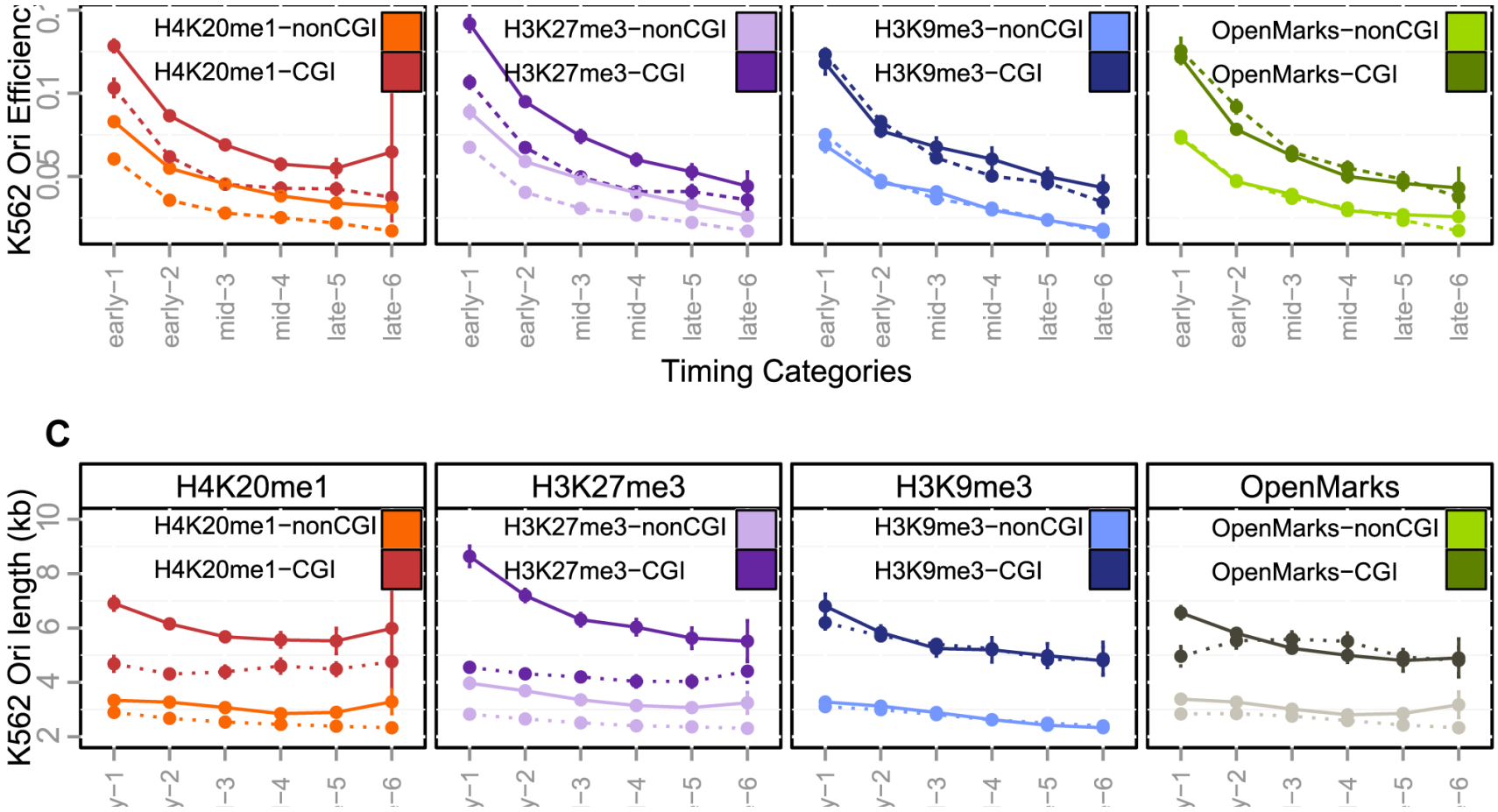
Showing 25 of 821

Discovery of new risk loci for IgA nephropathy implicates gene
against intestinal pathogens.
Kiryluk K, Li Y, Scolari F, et al.
Nature genetics. 2014 Nov;46(11):1187-96.
PMID:25305756 PMCID:PMC4213311

Mapping of transcription factor motifs in active chromatin ident
classical Hodgkin lymphoma.
Kreher S, Bouhlel MA, Cauchy P, et al.
Proceedings of the National Academy of Sciences of the United States of Americ
PMID:25288773 PMCID:PMC4210307

Defining the role of common variation in the genomic and biolo
human height.
Wood AR, Esko T, Yang J, et al.
Nature genetics. 2014 Nov;46(11):1173-86.

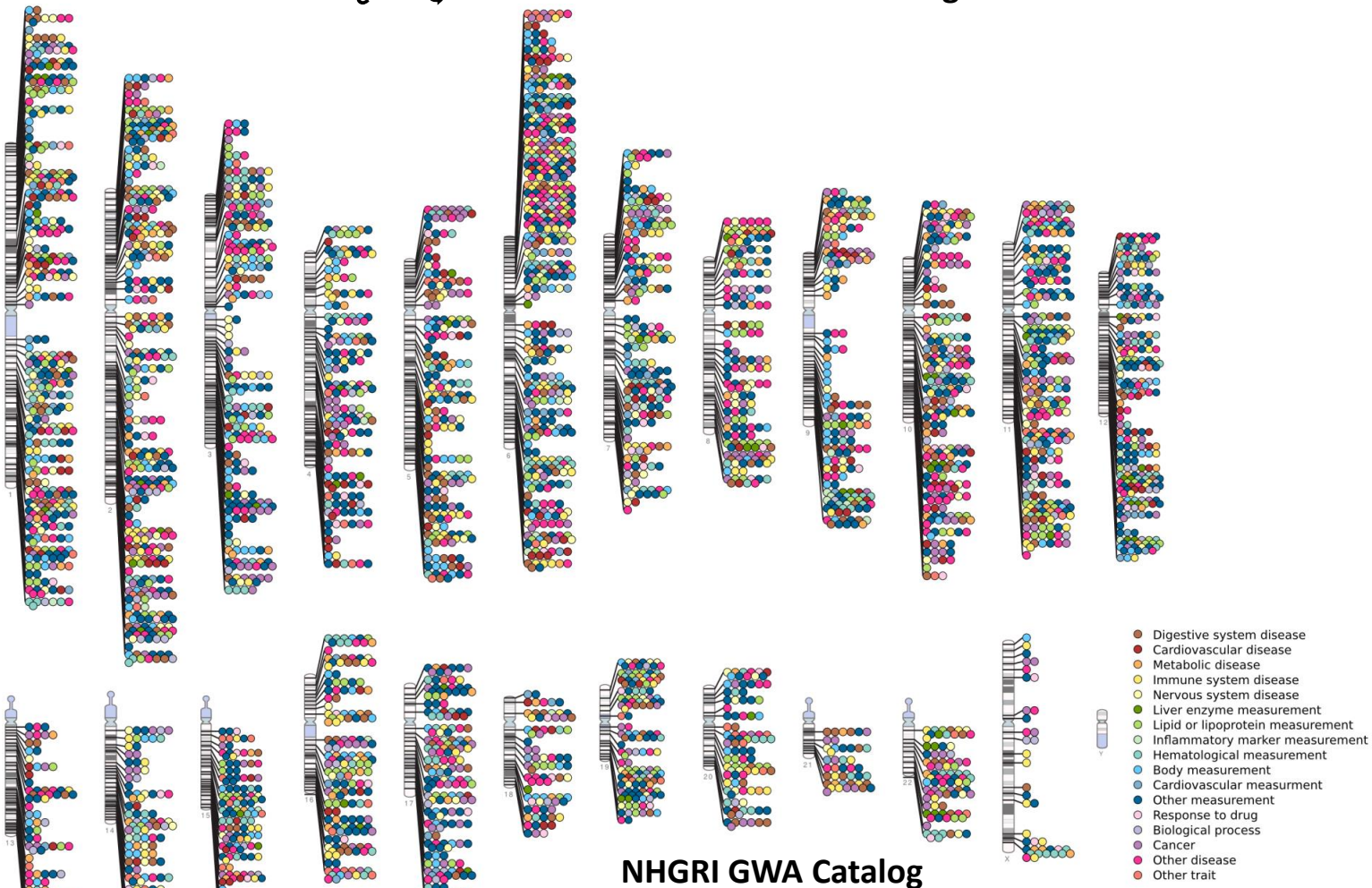
Example from Basic Biology



GWAS

Published Genome-Wide Associations through 12/2013

P for 17 trait categories

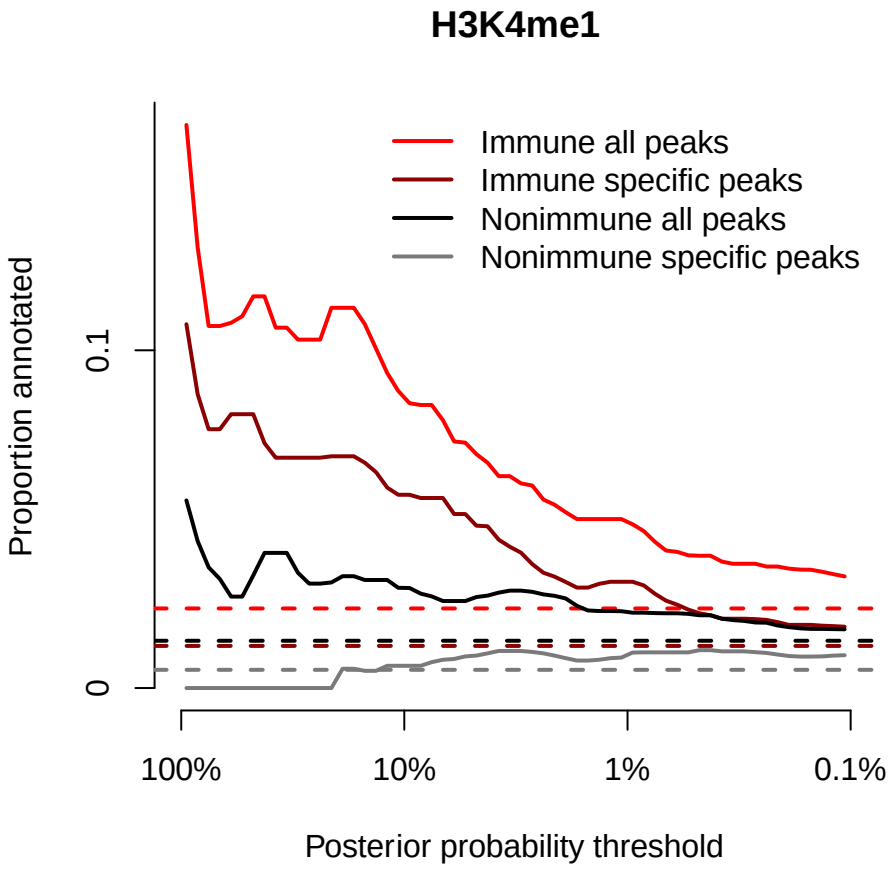


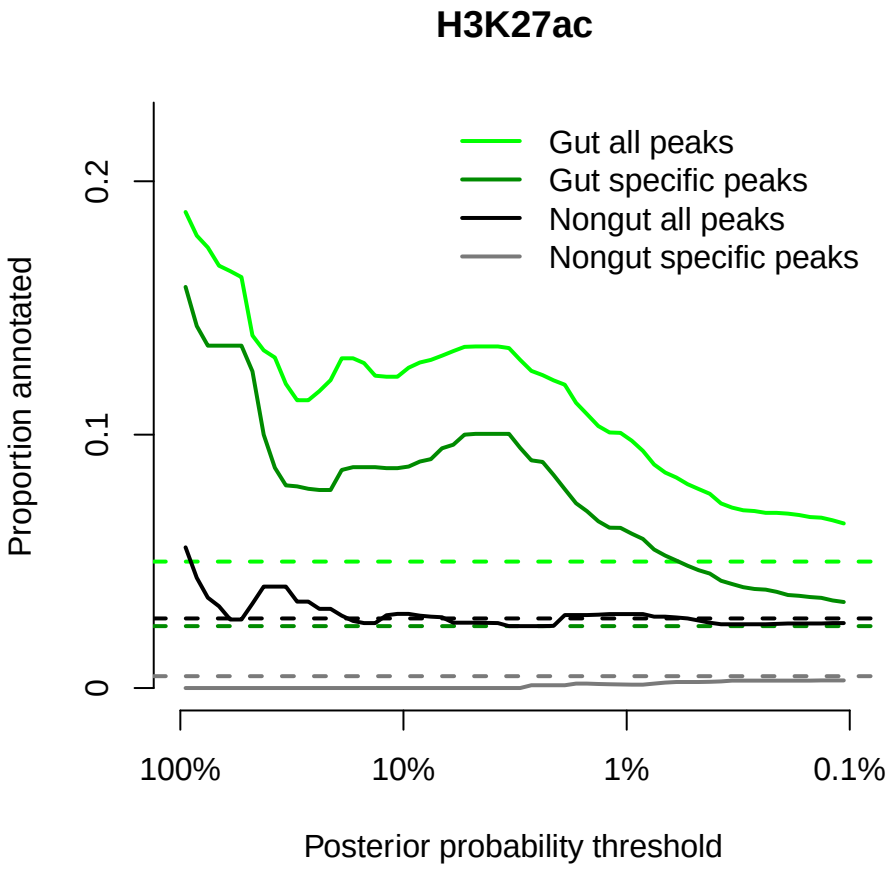
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

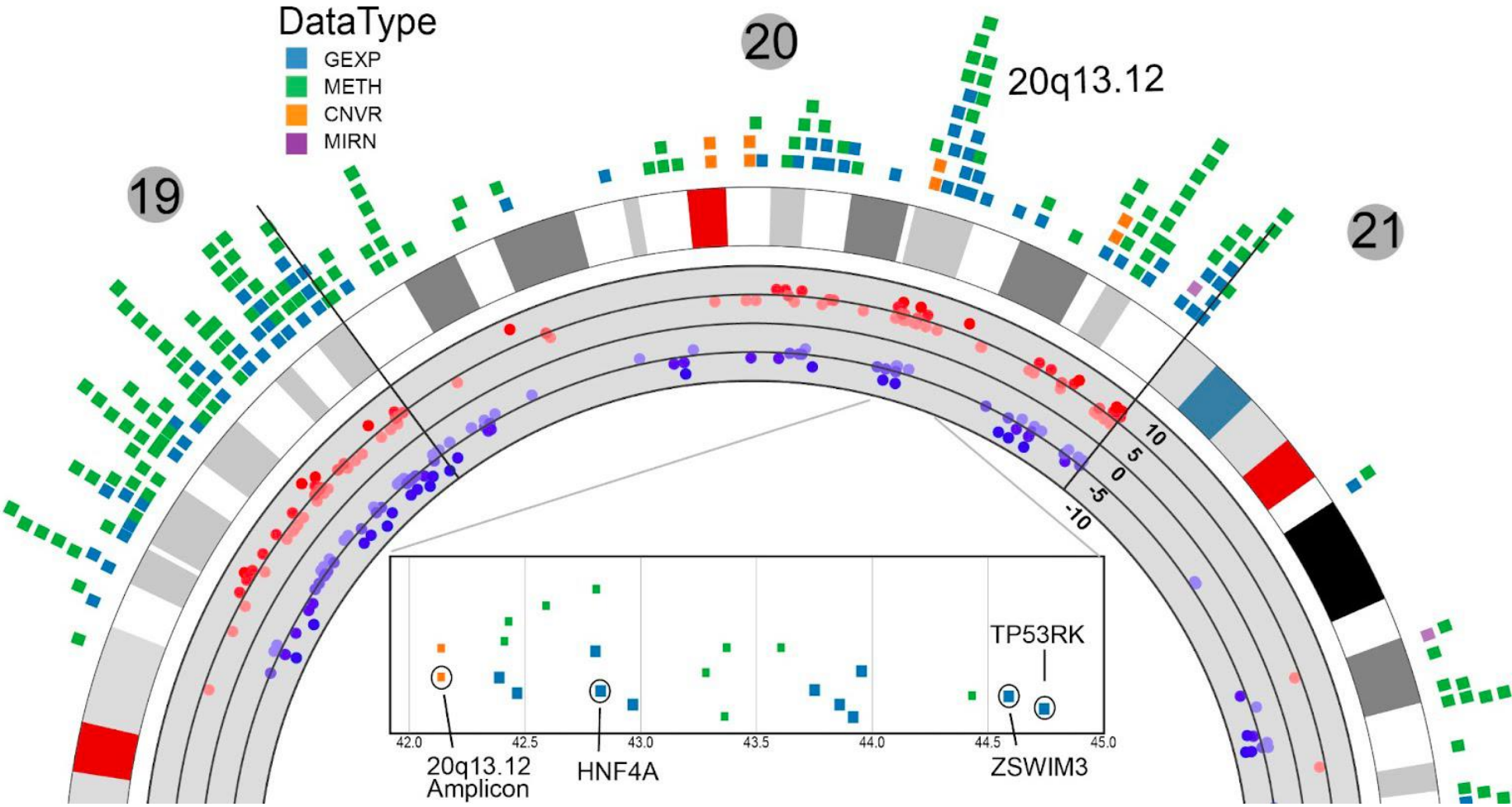
EMBL-EBI

Example from Common Disease





Somatic cancer mutations



ENCODE datasets allow exploration

- There is simply no way we would have generated the range of datasets on MCF-7 in the context of our project
  - (MCF-7 is not a perfect fit, but it's "good enough" to explore breast cancer)
- In particular technically demanding datasets
  - Annoying histone modifications
  - Repli-seq
- Lowering the barriers for exploration and discovery

Themes for the utility

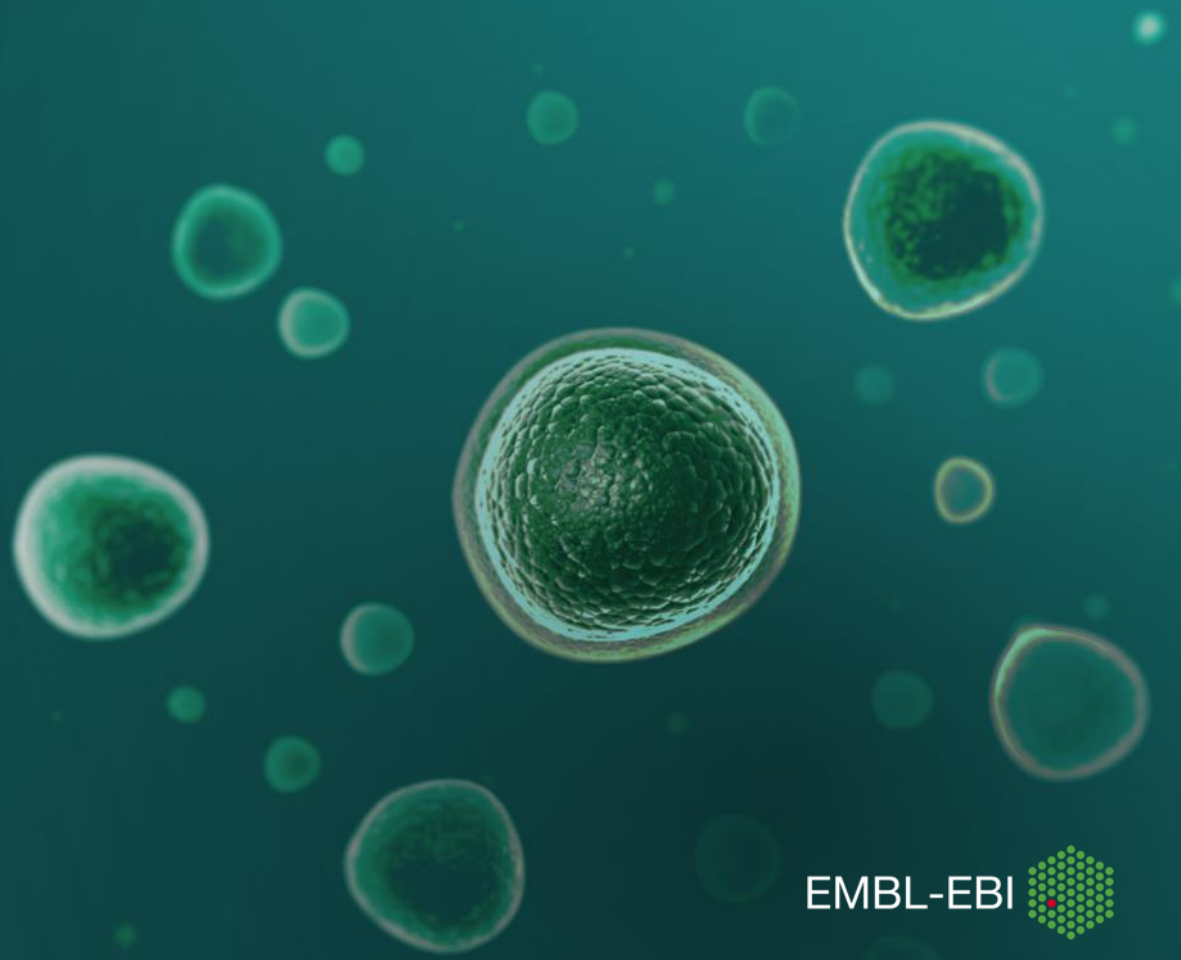
- Foundational datasets used to:
  - A design resource (eg, HiC, Promoter Capture)
  - A background resource (eg, “we tested all other histone marks…”)
  - An interpretation resource (GWAS fine mapping, Cancer)
  - For hypothesis generation

- Cell context of ENCODE data as important as chromatin state etc

(reuse details)

- Rarely work from absolutely raw data (ie, reads)
  - Though it’s nice to know this is possible!
- However quite a bit work from the first processing step (ie, signal)
- Both “called elements” (eg, Chip-seq peaks) and “overall summary” (eg, Segmentation) are used

What does biomedical science need over the next 10 years?



EMBL-EBI

Catalog-Classify-Curate

	Catalog	Classify	Curate
Protein coding genes	1980-2012	1980-now	Ongoing
ncRNA genes	2005 onwards	Active Research	Ongoing
Regulatory Elements	Minor – 1990s 2007 onwards	Active Research	
Chromosomal Structure	Minor – 1990s 2010 onwards		

Comprehensive Catalogs

- This is what distinguishes a catalog from just aggregation
- Defined quality levels, meta data
- Genomewide (it's a given these days!)
- "Countable" things
  - Every Transcription Factor (in at least 2 cell contexts?)
  - Every Histone modification (in at least 2 cell contexts?)
- "Uncountable" but bounded things
  - Expression levels, binding levels
  - Cell contexts

Classification

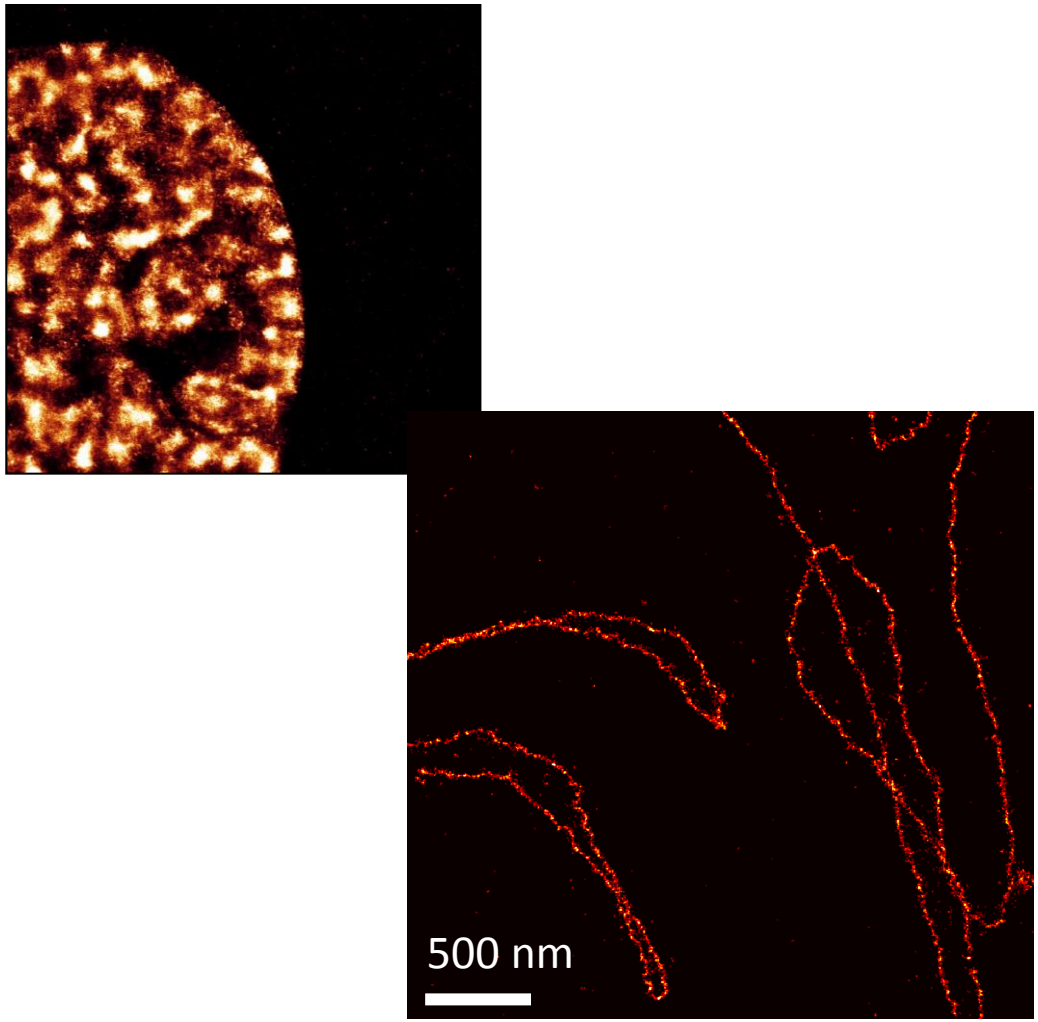
- This is a far more anarchic process
  - The genomic science
- Model organisms have a big role to play in classification
  - It's likely that many fundamental aspects of classification are the same across organisms (even if the individual elements and their positions are quite different)
- Variation as a particular need – beyond classification, towards modelling

Curation and Management

- Human Genome Reference -> Reference Consortium
- Human Protein and ncRNA set -> GenCode U41

- Regulatory Elements
  - Already a handful of known cases wrt to rare disease, TERT promoter
  - Likely to grow
  - Blend of classification at scale and curation

Opportunity



Super Resolution Microscopy is at Chromatin Scale (50nm or less)

Advances in resolution, speed labelling really blends imaging with molecular techniques

Thanks

- Mike Pazin (NHGRI), Sandro Morganella (EMBL-EBI), Serena Nik-Zainal (Sanger), Jeff Barrett (Sanger), Jan Ellenberg (EMBL-HD)