A very brief introduction to bioinformatics

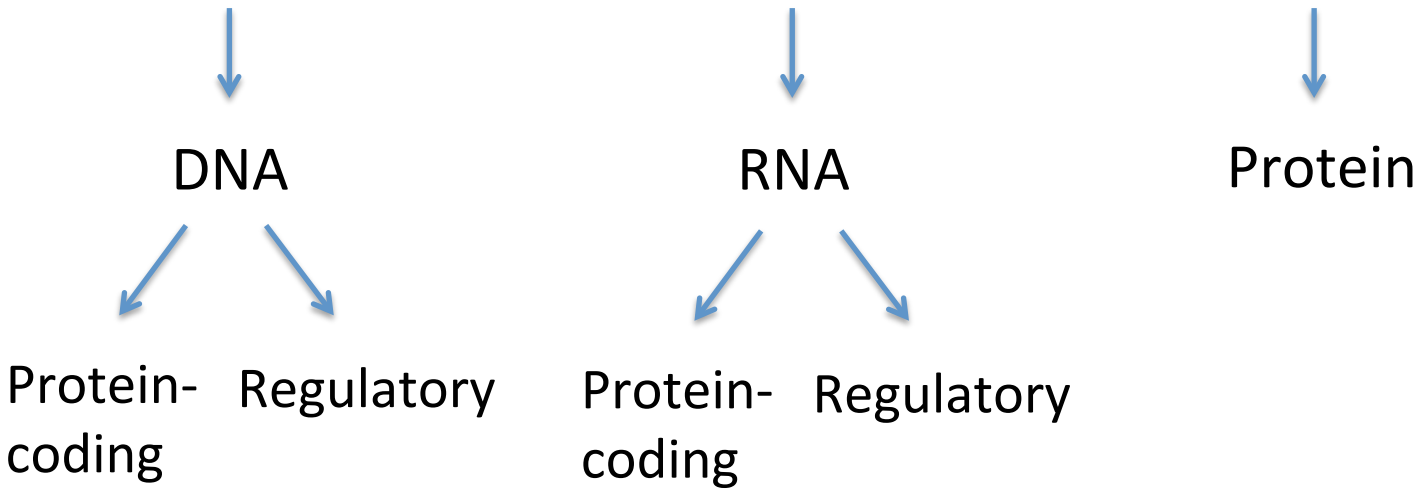
Mikhail Spivakov, PhD
European Bioinformatics Institute

What bioinformatics does?

- Cataloguing → For lab biologists to look at 'favourite' genes etc.
- Mining → OMICS (a lot of data, lower precision)
- Modelling → Currently smaller 'reductionist' subsets
  Potentially higher precision, if the data is good

What questions can bioinformatics answer?

- Relationship between sequence and function:



- Network architecture

How things are synthesized in the cells (metabolic networks)

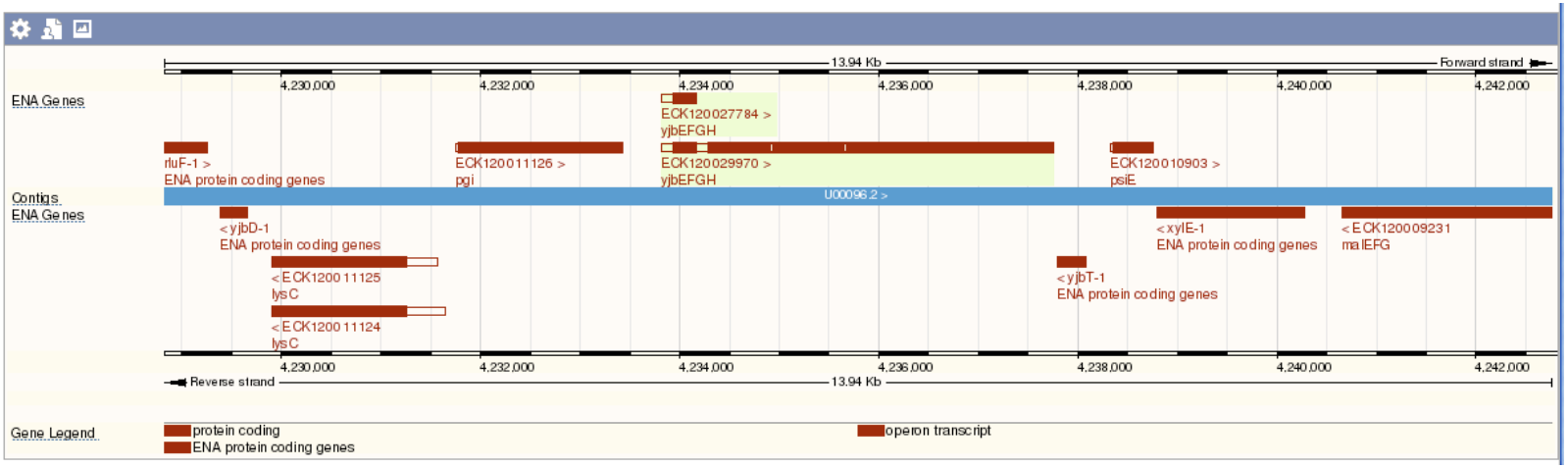
How cells make decisions (signalling/regulatory networks)

Types of data

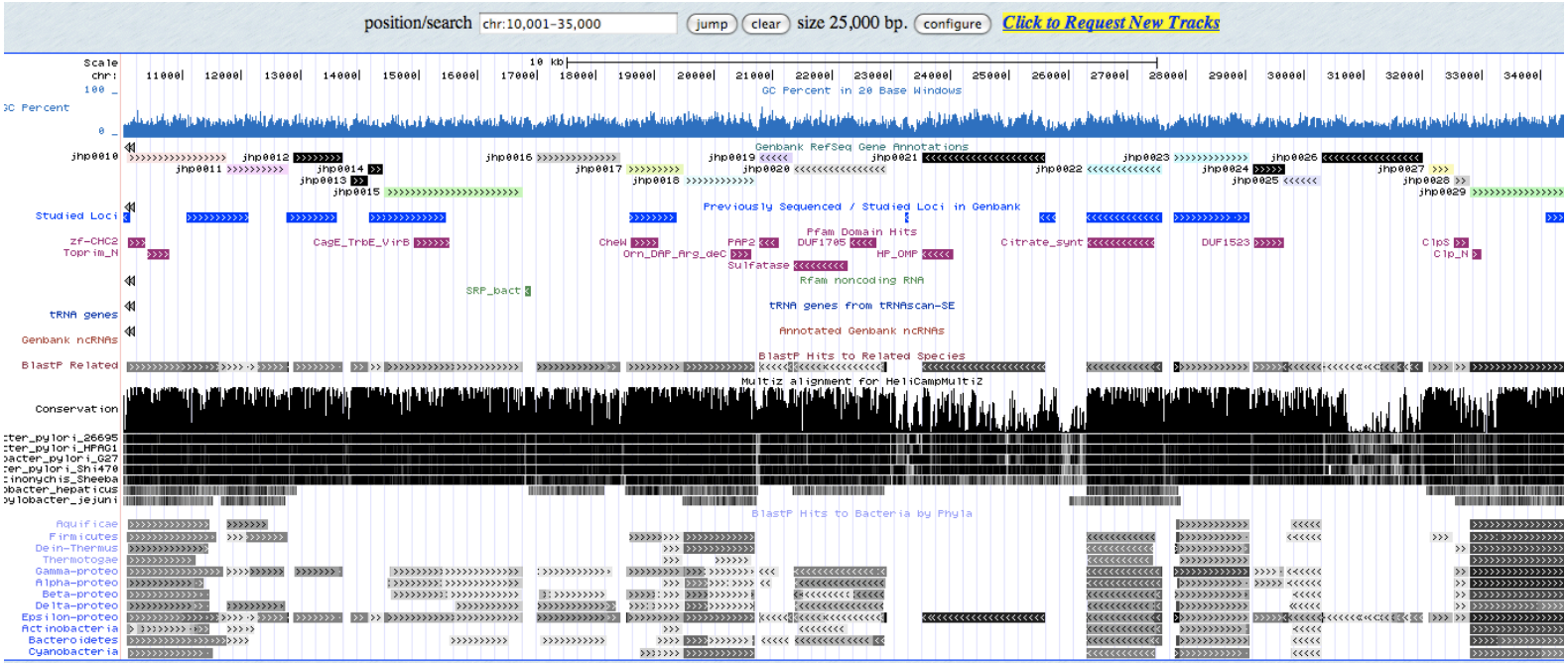
- Sequence (DNA, RNA, protein)
- Structure (RNA, protein)
- Gene expression (expression arrays, RNA-seq)
- Interactions:
  - DNA-protein (ChIP-chip, ChIP-seq)
  - RNA-protein (CLIP-seq)
  - Protein-protein (mass-spec, co-IP)
- Pathways
  - Signalling
  - Metabolic

- Individual sequence variation (personal genomics)
- Metagenomics

Visualization tools: Genome browsers

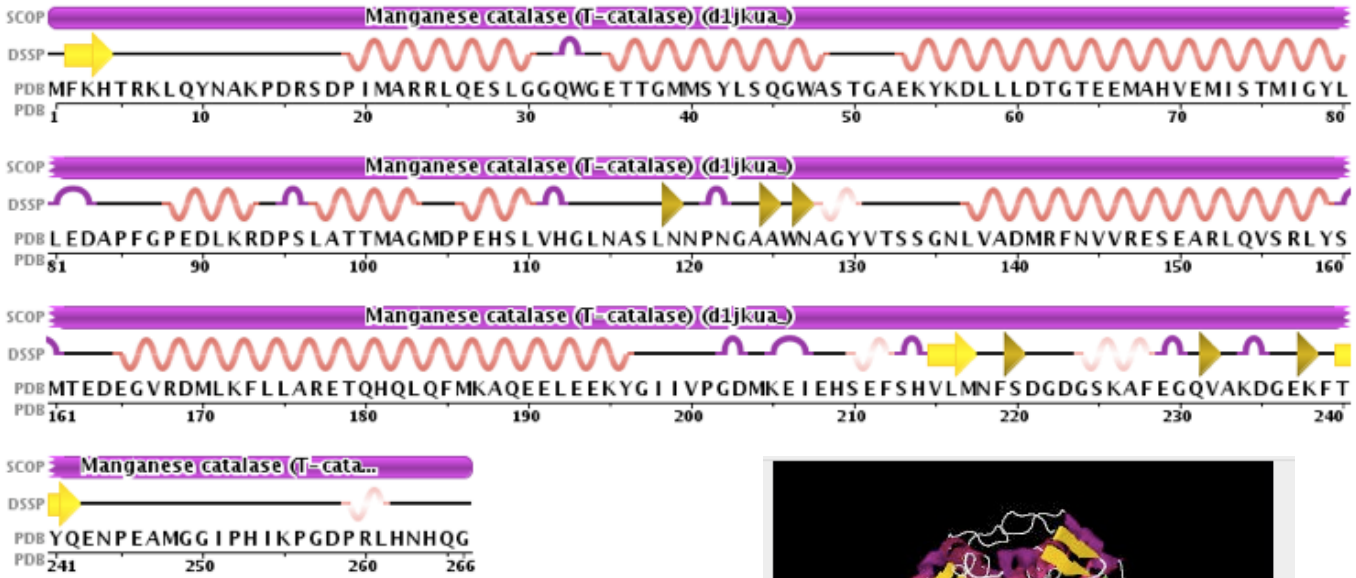


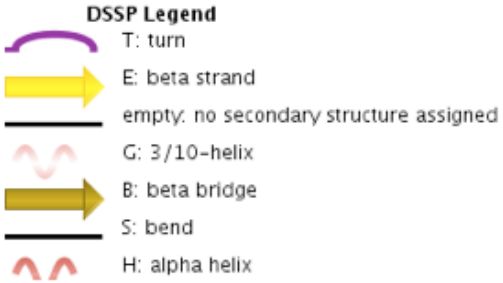
www.ensembl.org | bacteria.ensembl.org

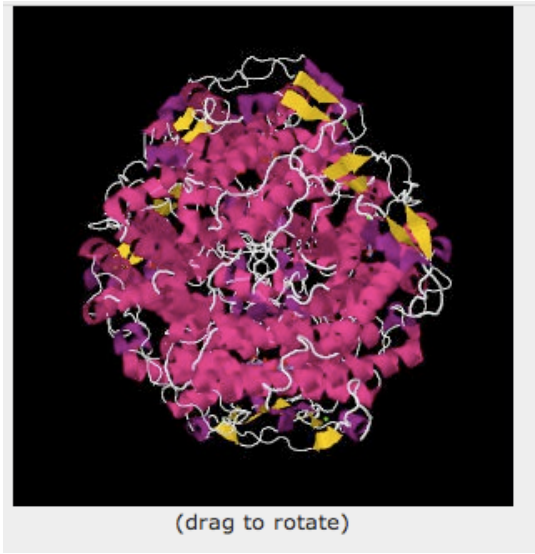


genome.ucsc.edu | microbes.ucsc.edu

Visualization tools: Protein structure



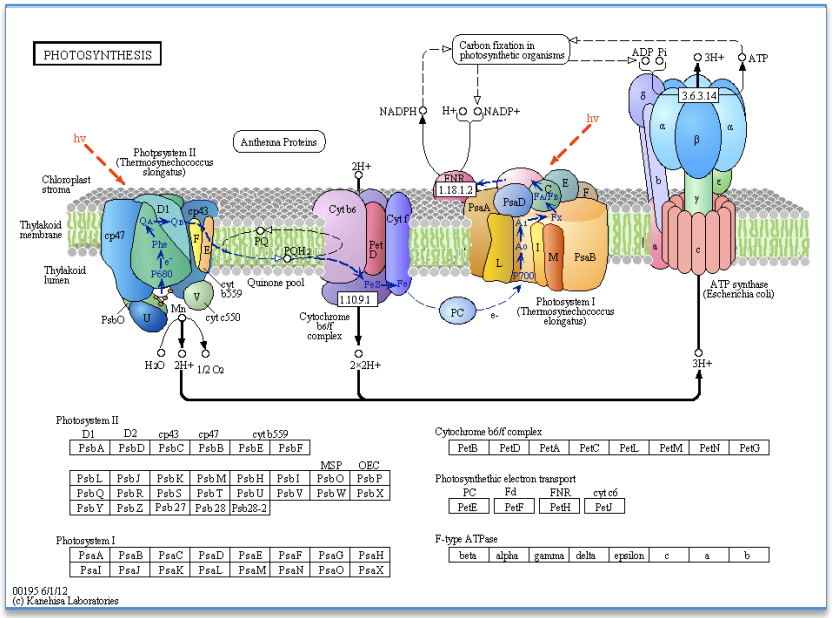




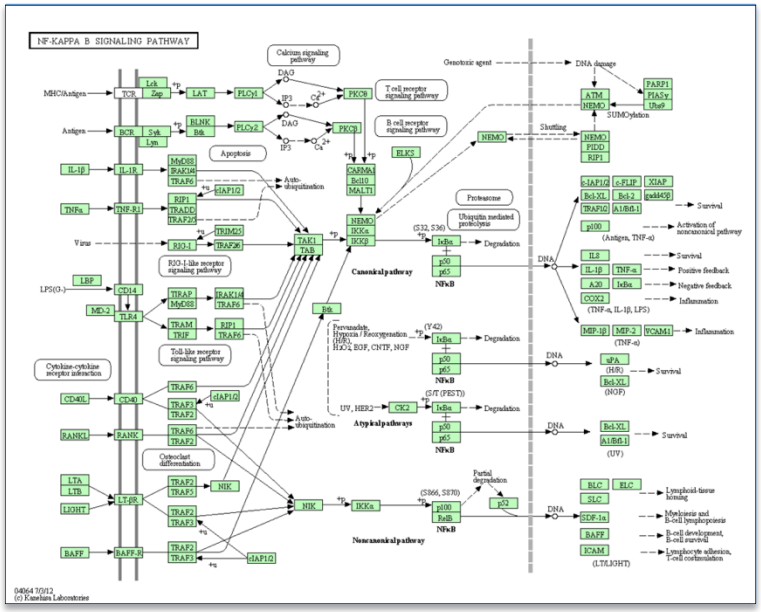
PDB: www.rcsb.org/pdb

Visualization tools: Pathways

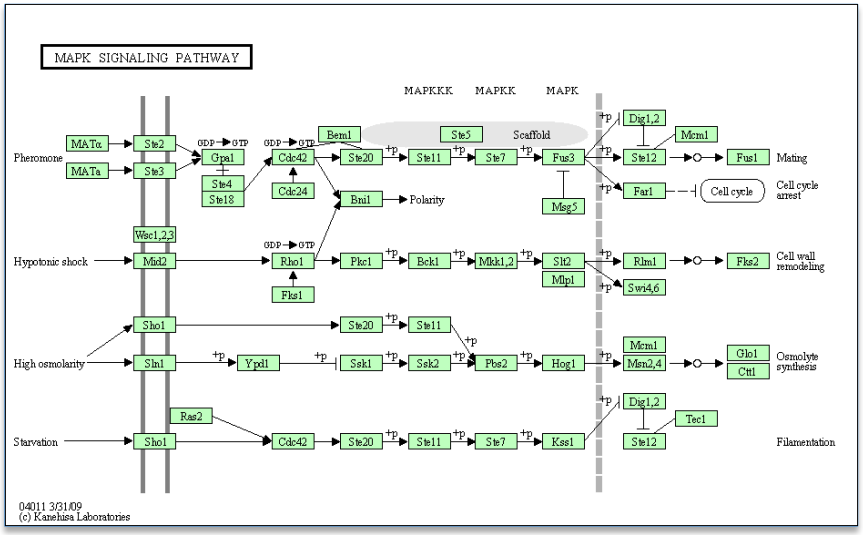
Photosynthesis



Signalling by transcription factor NFkappaB

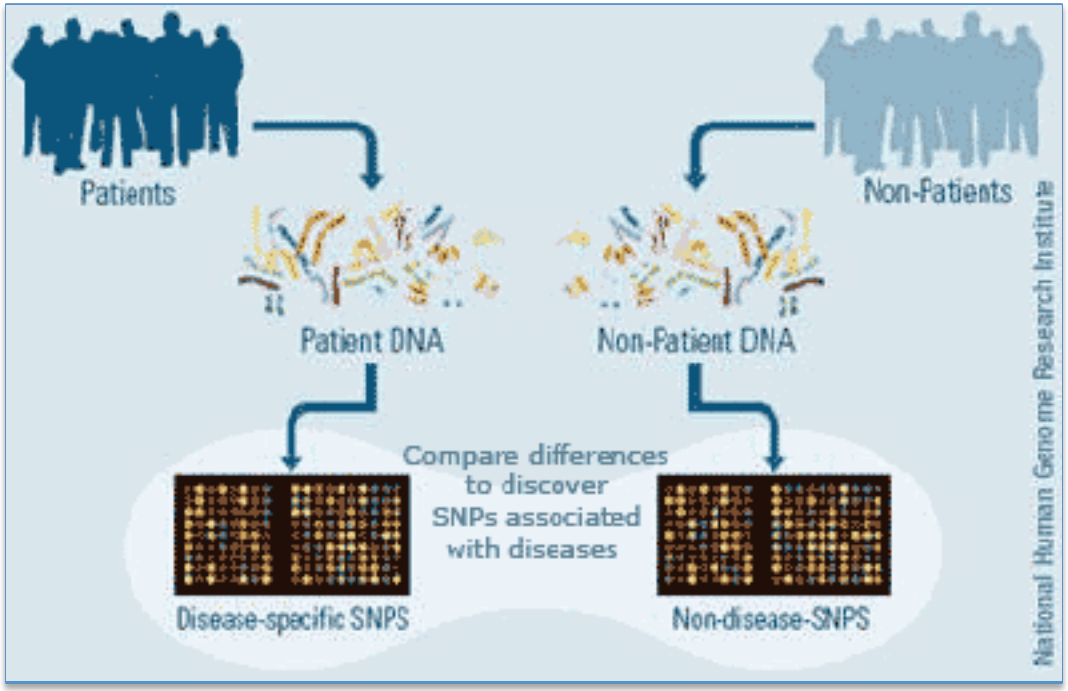


MAPK signalling pathway in yeast



KEGG:
www.genome.jp/kegg

Genomic variation data:
Genome-wide association studies

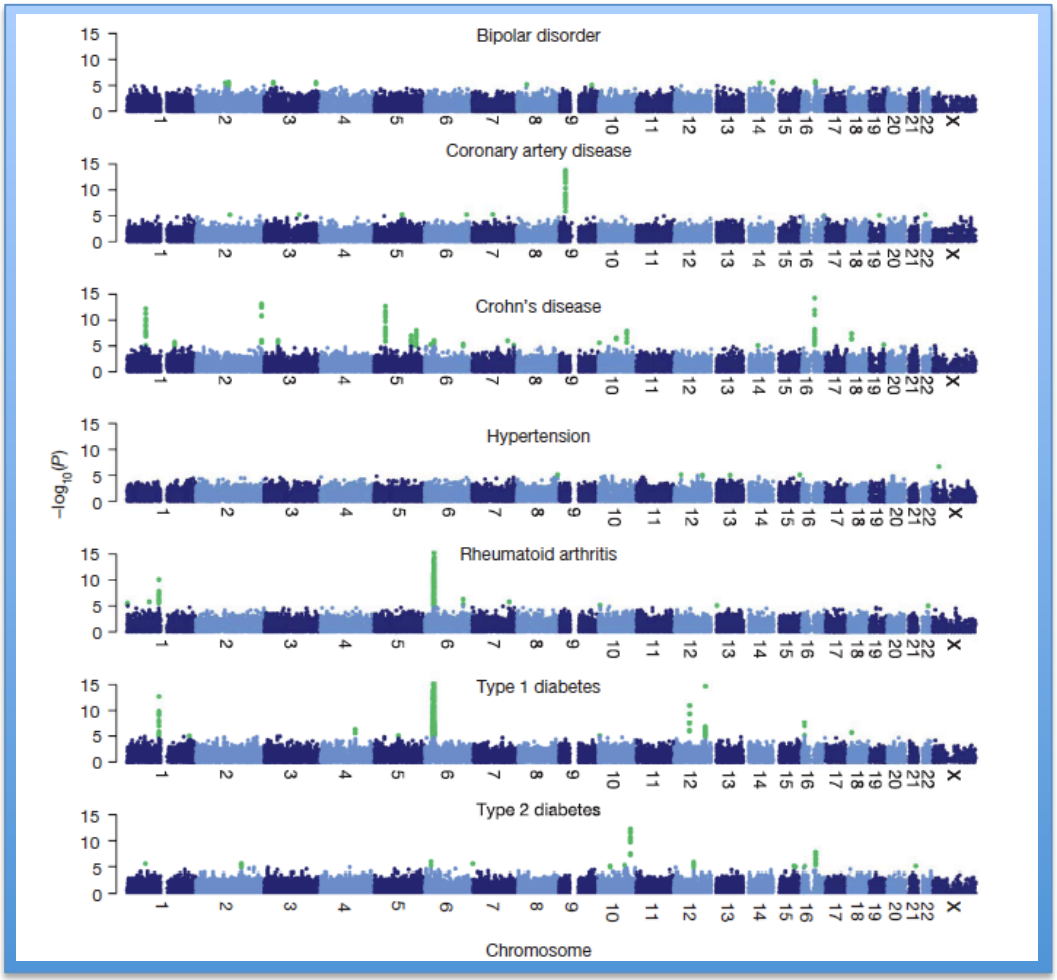


Goal: find genotypes associated with phenotypes:

- disease
- metabolism (including drug sensitivity)
- height, weight, eye colour, etc…

…

Example: Genome-wide associated study for 7 common diseases



2,000 individuals for each disease and a shared set of 3,000 'healthy' controls.

Challenges:

- The majority of associations are outside of protein-coding genes
- Most are only found in a fraction of patients, and are also present in people with disease

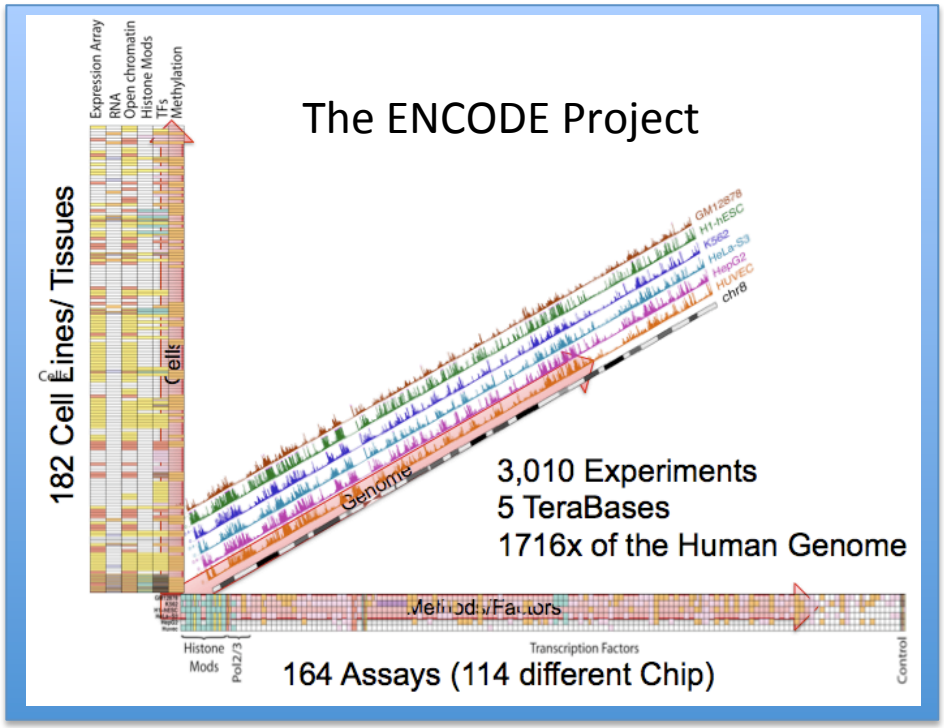
Age-related Macular Degeneration	★★★★	8.4%	6.5%	1.29x	
Colorectal Cancer	★★★★	7.7%	5.6%	1.38x	
Chronic Kidney Disease	★★★★	4.2%	3.4%	1.22x	
Restless Legs Syndrome	★★★★	2.5%	2.0%	1.25x	
Ulcerative Colitis	★★★★	1.3%	0.8%	1.73x	
Esophageal Squamous Cell Carcinoma (ESCC)	★★★★	0.4%	0.4%	1.21x	
Stomach Cancer (Gastric Cardia Adenocarcinoma)	★★★★	0.3%	0.2%	1.22x	
Psoriasis	★★★★	7.1%	11.4%	0.62x	
Melanoma	★★★★	2.2%	2.9%	0.75x	
Rheumatoid Arthritis	★★★★	0.8%	2.4%	0.34x	
Crohn's Disease	★★★★	0.4%	0.5%	0.66x	
Multiple Sclerosis	★★★★	0.2%	0.3%	0.59x	
Exfoliation Glaucoma	★★★★	0.2%	0.7%	0.22x	
Type 1 Diabetes	★★★★	0.2%	1.0%	0.15x	
Celiac Disease	★★★★	0.07%	0.12%	0.58x	
Primary Biliary Cirrhosis	★★★★	0.05%	0.08%	0.66x	

www.23andme.com

An open access personal genomes project: http://www.personalgenomes.org/

Catalog of Published Genome-Wide Association Studies:
https://www.genome.gov/26525384

Gene expression + DNA-protein interactions data: Functional genomics

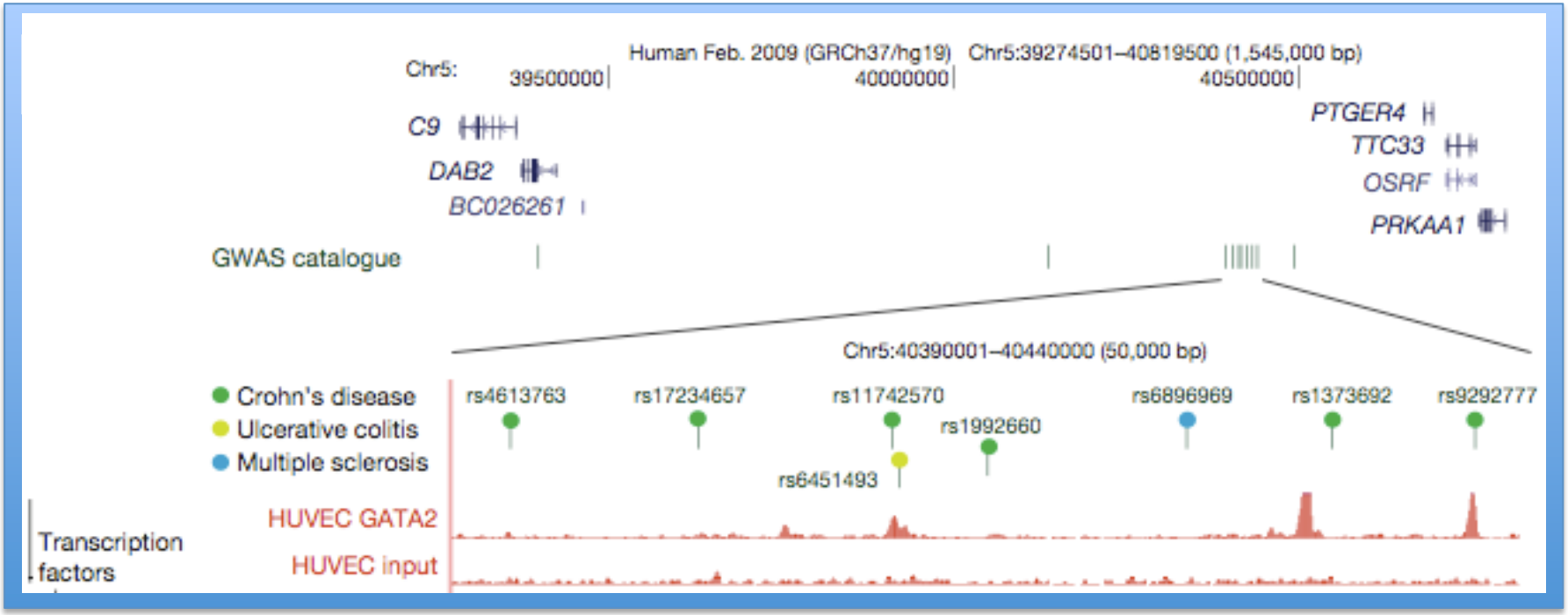


- The majority of the genome participates in a biochemical event
- Transcription factors co-bind the DNA in complex, combinatorial ways
- We can classify the genome into 7 very broad classes of states, and 1,000s of microstates
- We can inform >50% of the non coding genotype-disease associations

Open access: Genome.ucsc.edu/ENCODE

Ewan Birney

Example: Region associated with Crohn's disease contains multiple binding sites for transcription factor *GATA2*



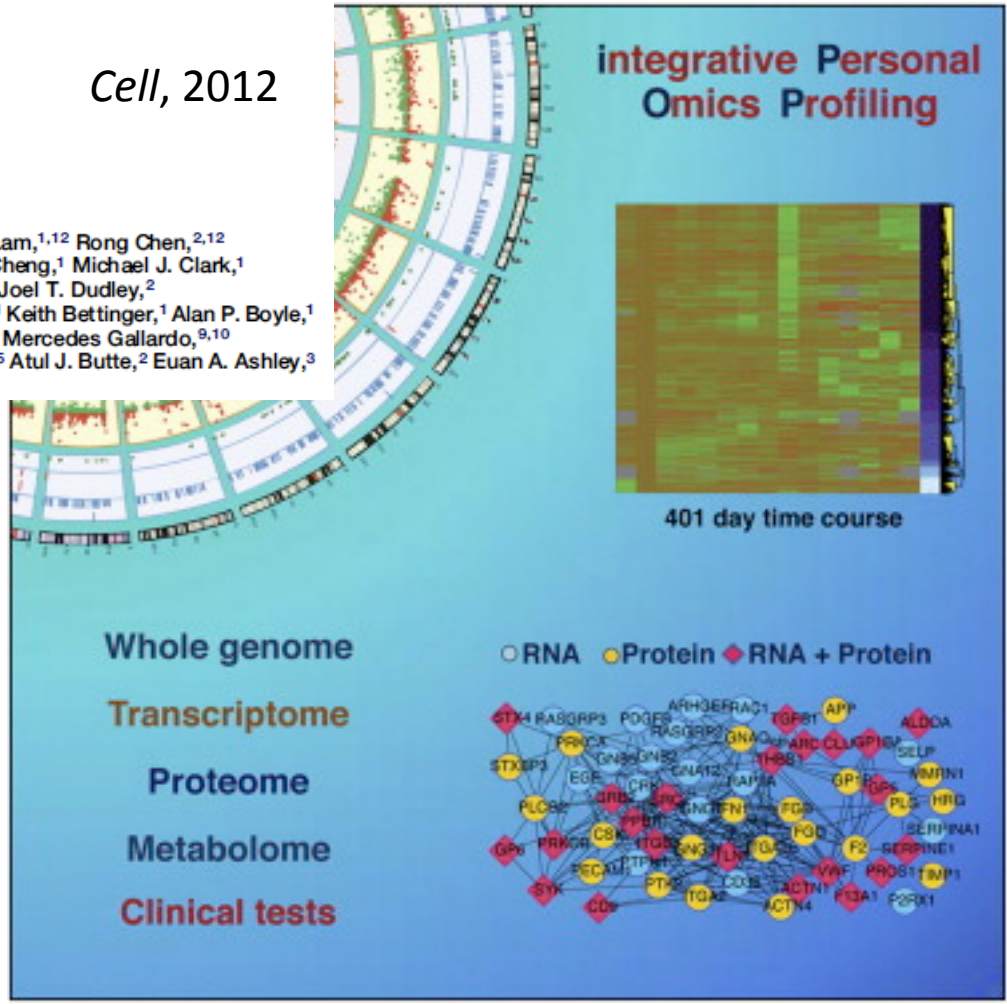
ENCODE, *Nature* 2012
To appear in Septembr

Testing the power of personal genomics: Prof Snyder's blood

Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Cell, 2012

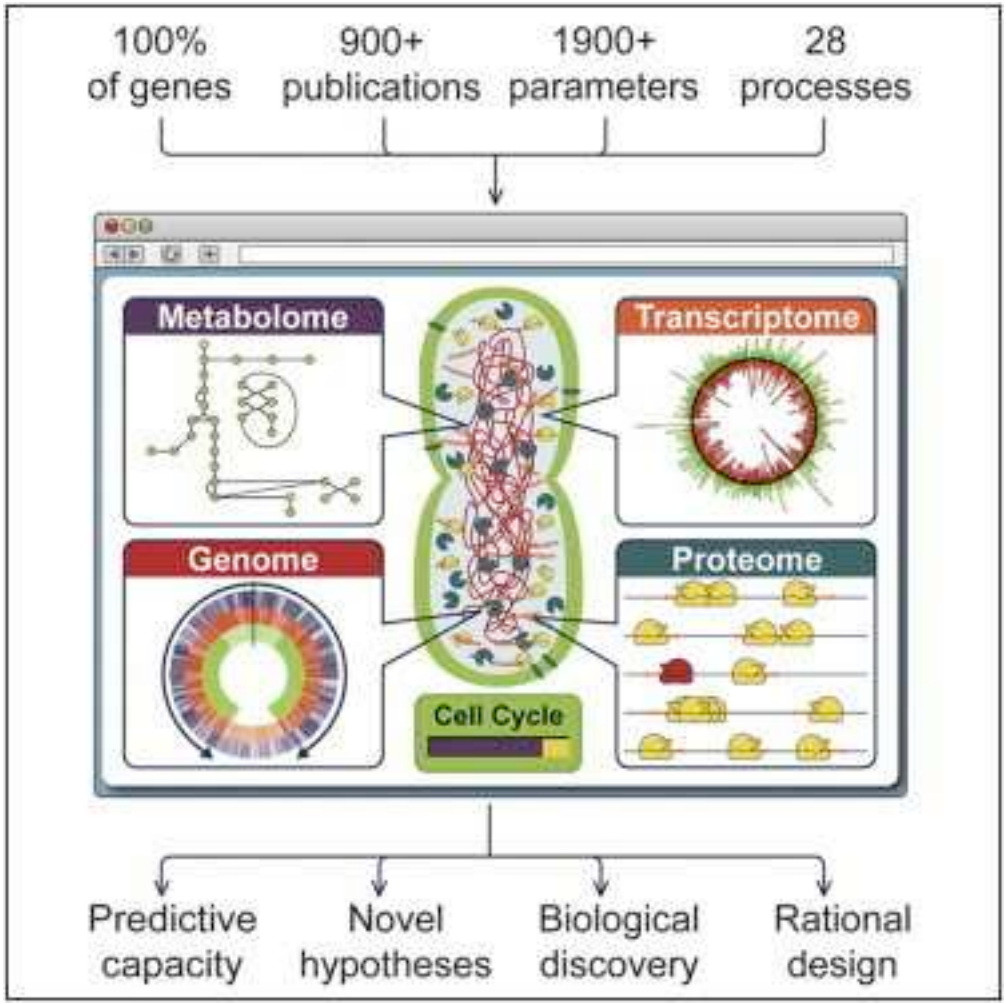
Rui Chen,[1,11] George I. Mias,[1,11] Jennifer Li-Pook-Than,[1,11] Lihua Jiang,[1,11] Hugo Y.K. Lam,[1,12] Rong Chen,[2,12] Elana Miriami,[1] Konrad J. Karczewski,[1] Manoj Hariharan,[1] Frederick E. Dewey,[3] Yong Cheng,[1] Michael J. Clark,[1] Hogune Im,[1] Lukas Habegger,[6,7] Suganthi Balasubramanian,[6,7] Maeve O'Huallachain,[1] Joel T. Dudley,[2] Sara Hillenmeyer,[1] Rajini Haraksingh,[1] Donald Sharon,[1] Ghia Euskirchen,[1] Phil Lacroute,[1] Keith Bettinger,[1] Alan P. Boyle,[1] Maya Kasowski,[1] Fabian Grubert,[1] Scott Seki,[2] Marco Garcia,[2] Michelle Whirl-Carrillo,[1] Mercedes Gallardo,[9,10] Maria A. Blasco,[9] Peter L. Greenberg,[4] Phyllis Snyder,[1] Teri E. Klein,[1] Russ B. Altman,[1,5] Atul J. Butte,[2] Euan A. Ashley,[3] Mark Gerstein,[6,7,8] Kari C. Nadeau,[2] Hua Tang,[1] and Michael Snyder[1,*]



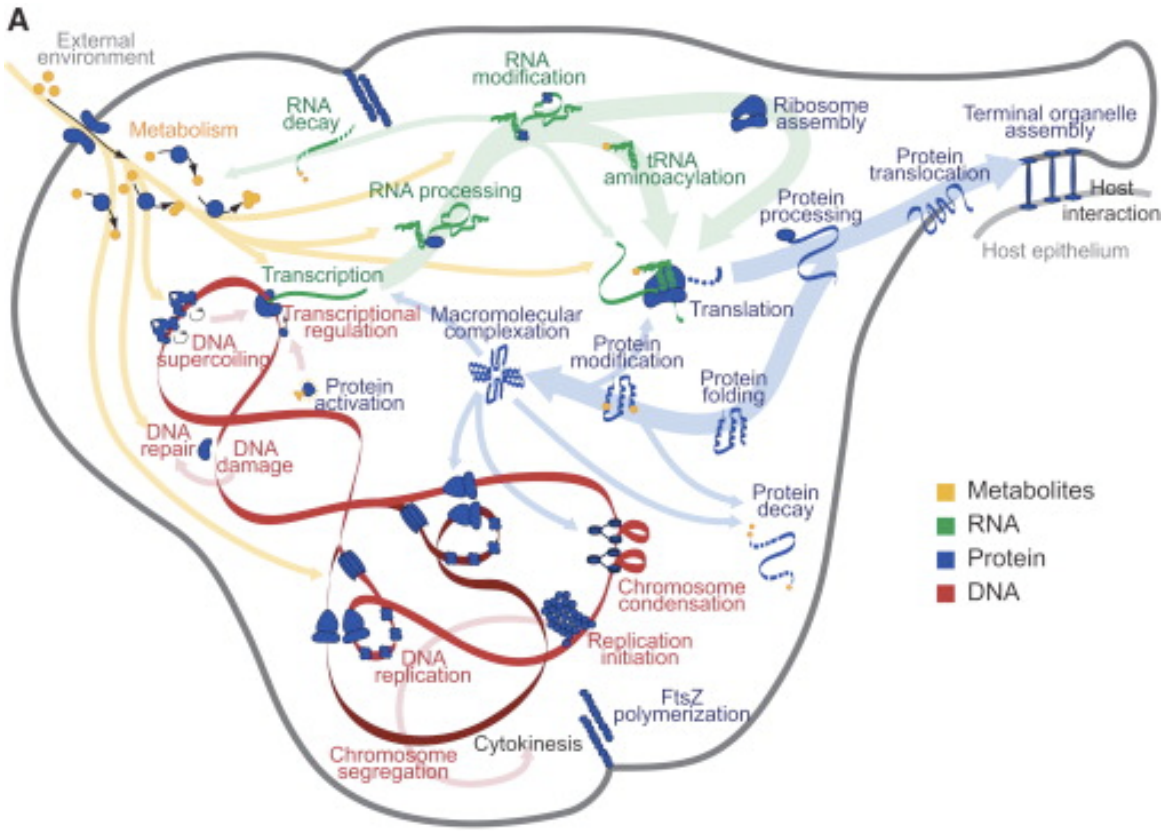
- Blood samples taken every day for 14 months, including 2 viral infections
- Found mutations associated with disease risk
- Type 2 diabetes suspected and glucose levels controlled by diet
- Saw systemic response to viral infection

Modelling: a full dynamic model for *M. genitalia*

- Smallest free-living bacteria (525 genes)



Jonathan Karr
Jayodita Sanghvi
Markus Covert et al.,
Cell 20 July 2012



- Depending on the available knowledge, used models with different levels of abstraction for each module and developed ways to 'pierce' them together
- Prediction of effect of permutations on growth rate: ~70% precision.

Big, noisy data

(Smaller) precise data

OMICS (data mining, statistical models)

Functional genomics / proteomics
Genome-wide association studies
Personal genomics
Metagenomics
Interactomics (interaction networks)
…

Dynamic systems modelling

Intracellular signalling
Metabolism
Transcriptional regulation
Neurotransmission
…