

Toward linking genome dynamics to ecosystem functions

Prof. Jean-Baptiste Cazier

Chair of Bioinformatics

Director of the Centre for Computational
Biology, University of Birmingham

Prof. James Bentley Brown

Chair of Environmental Bioinformatics

Head of Molecular Ecosystems Biology,
Lawrence Berkeley National Laboratory

<http://www.birmingham.ac.uk/CCB>



@UoB_CCB

BIO / INFORMATICS ?

*Computational Biology, BioStatistics,
Statistical Genetics, Mathematical Biology, ...*

• [...]

• **Bioinformatics**

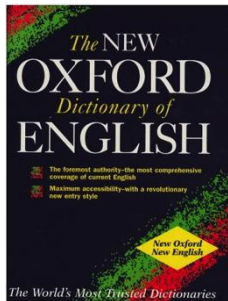
Pronunciation: /ˌbaɪəʊˌɪnfəˈmætkɪs/
plural noun [treated as singular]

the science of collecting and analysing complex biological data such as genetic codes.

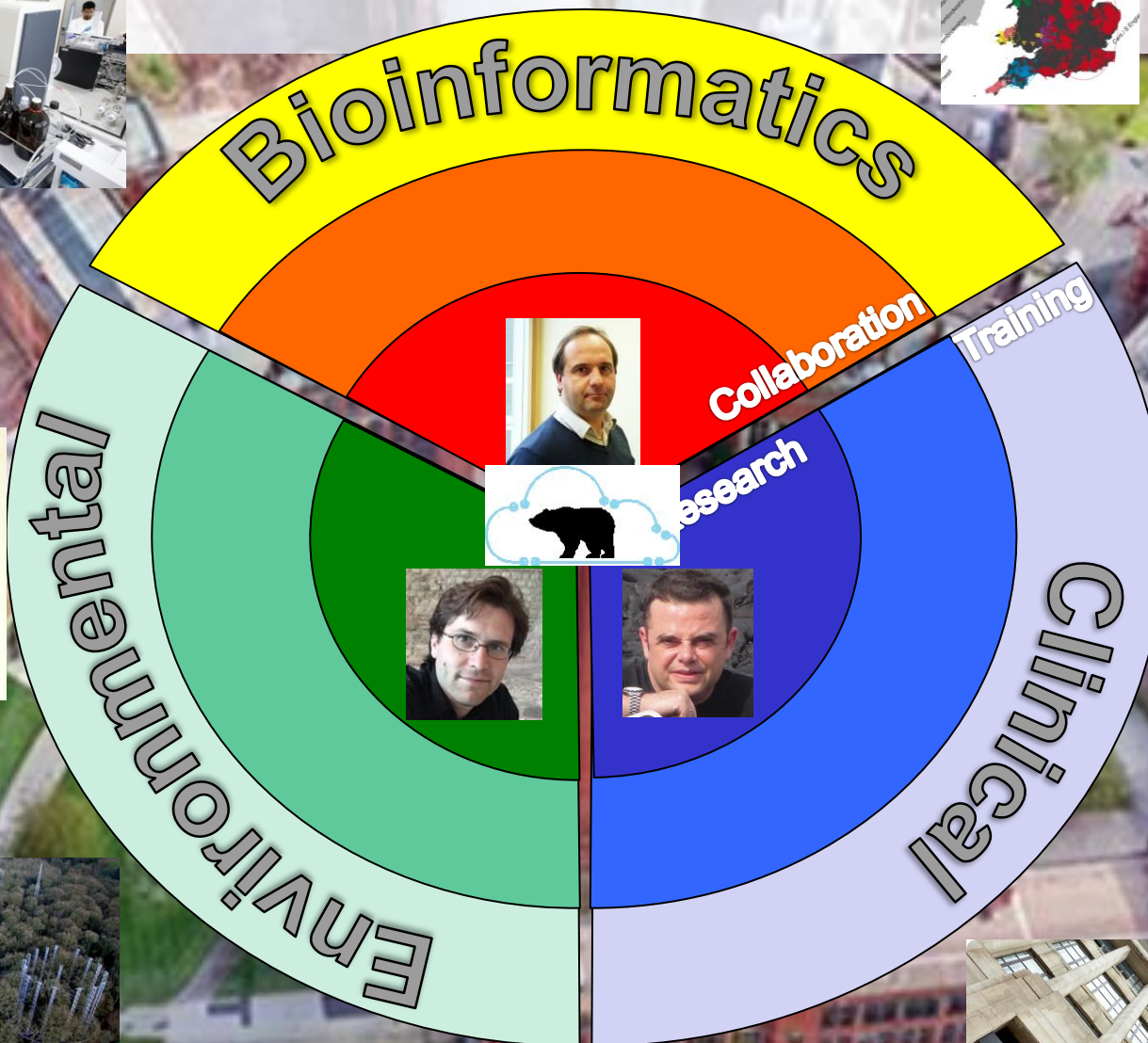
- **Bioinformatics** *i/ˌbaɪəʊˌɪnfəˈmætkɪs/* is the application of computer science and information technology to the field of biology and medicine. Bioinformatics deals with **algorithms**, **databases** and **information systems**, **web technologies**, **artificial intelligence** and soft computing, **information and computation theory**, **software engineering**, **data mining**, **image processing**, **modeling** and simulation, **signal processing**, **discrete mathematics**, control and **system theory**, **circuit theory**, and statistics, for generating new knowledge of biology and medicine, and improving & discovering new models of computation (e.g. **DNA computing**, **neural computing**, evolutionary computing, **immuno-computing**, **swarm-computing**, **cellular-computing**).



Life, the Universe and Everything



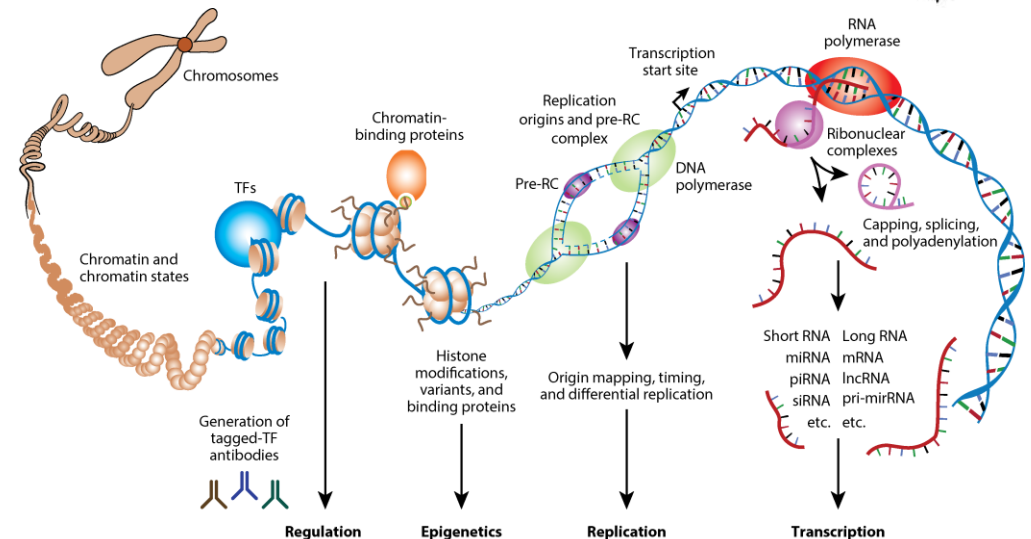
WIKIPEDIA
The Free Encyclopedia



Molecular Ecosystems Biology



Molecular Biology: understanding gene regulation in cells – extending contexts to tissues (microenvironment) and organisms (physiology) – but can we connect to populations and ecologies?



Brown JB, Celniker SE. 2015.

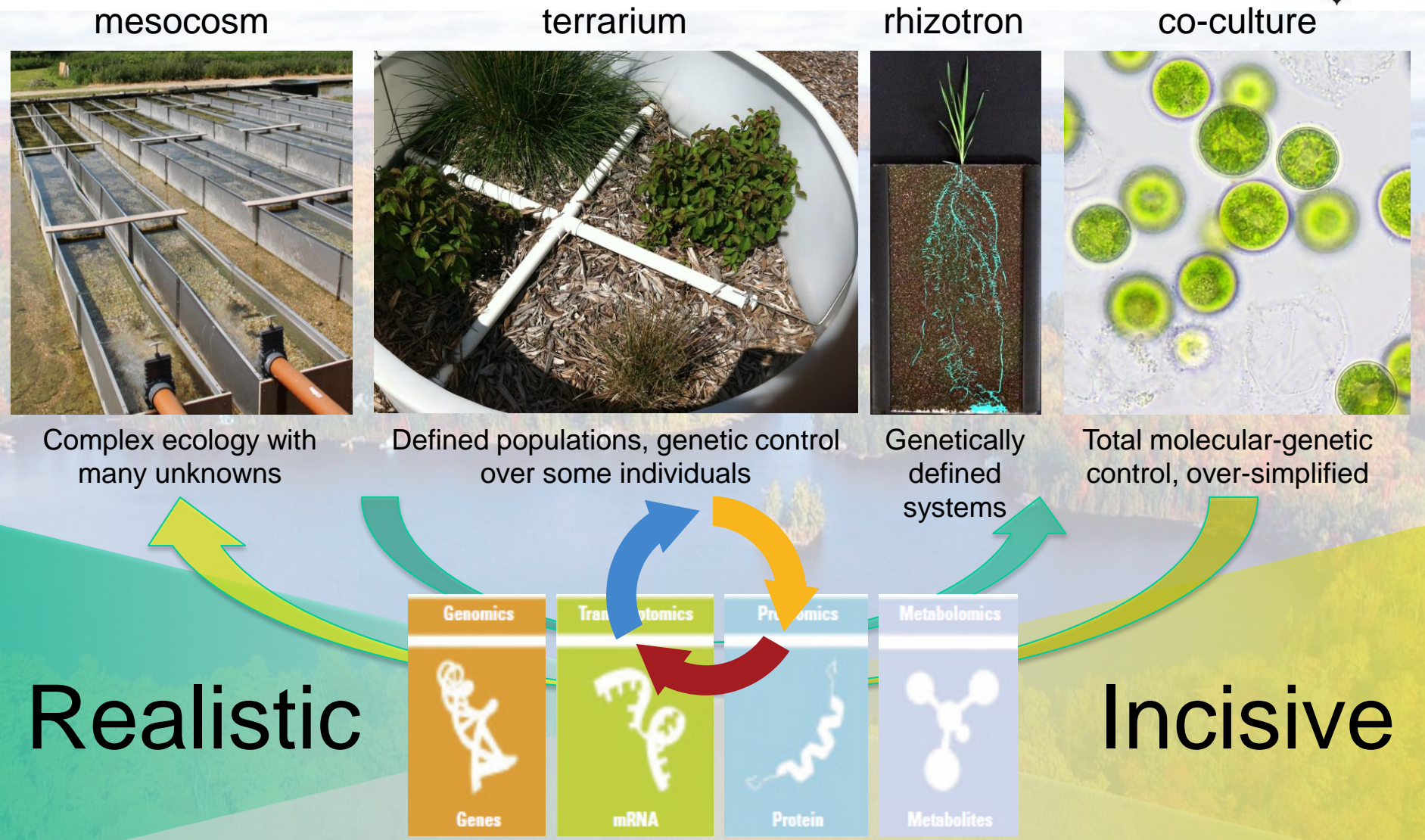
Annu. Rev. Genomics Hum. Genet. 16:31–53

From Molecules to Ecologies: a grand challenge before us is to link genome dynamics to ecosystem biology – to understand how gene regulatory systems transduce, respond to, and ultimately influence populations and ecologies





Coupling multi-scale systems to link genome biology to ecosystem dynamics



Essential cross-cutting enabling technologies



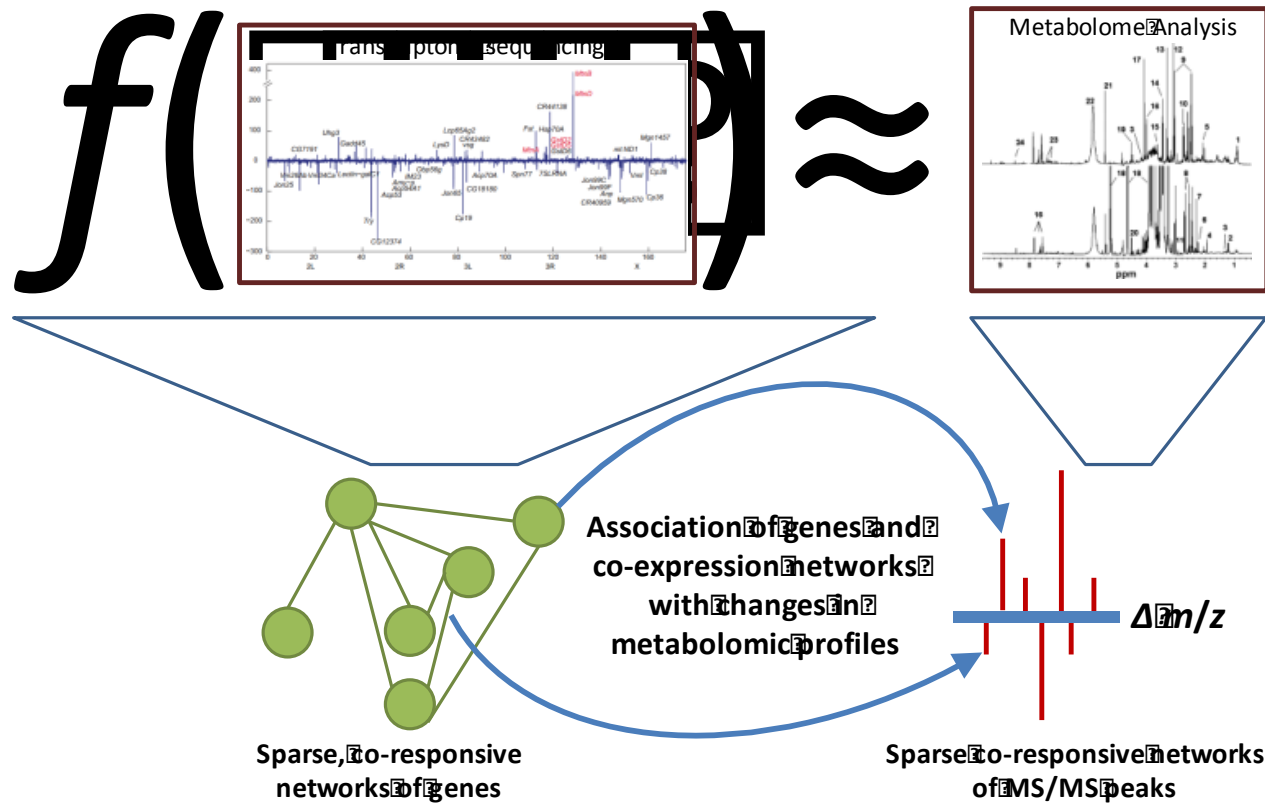
- **We have:**

- **CRISPR** enables genetic manipulation of previously intractable organisms
- **Glyco-binding-beads and Engineered Phage Libraries** enable the targeting of individual microbes in complex communities
- **Exposure biology** enables the perturbation of metabolic pathways even when those pathways are spread across consortia
- **High Performance Computing (HPC)** to fit next-generation learning machines

- **We need:**

- **Model ecologies:** self-sustaining and recapturable systems with defined trajectories
- **Phylogenomic reconstructions:** models of network evolution to enable the translation of results from model systems to natural ecologies
- **Nondestructive measurements:** molecular time-courses from individuals and consortia
- **Informative Learning Machines:** “Open Box” analytical procedures to obtain insight from multi-modal panomics datasets

Statistical Machine Learning: State of Art



Complexity & Life

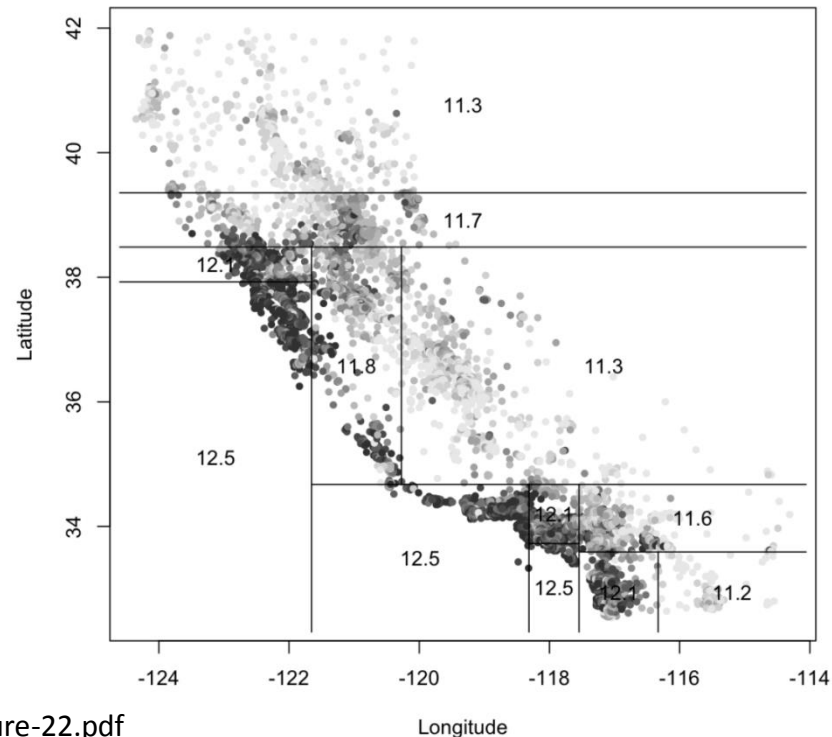
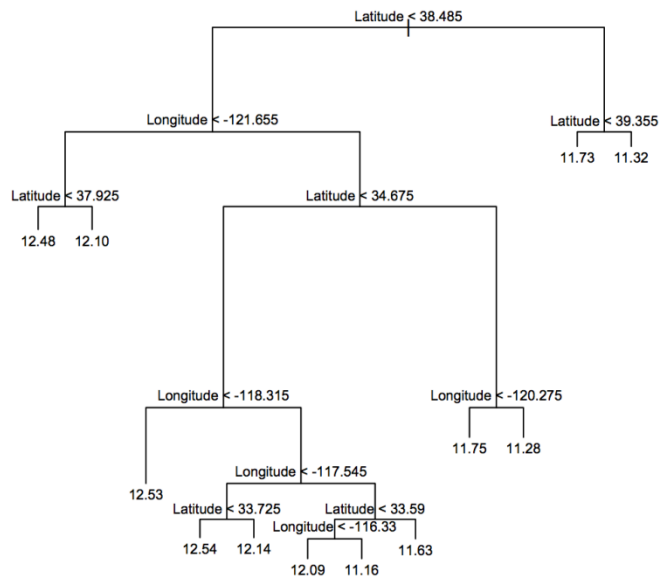


- 6,000,000 – approximate number of variants mapped in the human population
- 36,000,000,000,000 – approximate number of pairs of variants
- 21,600,000,000,000,000,000,000 – approximate number of triplets
- 1.5 Trillion CPU Hours – a comprehensive search over triplets for a complex phenotype

Decision Trees



- Hierarchically structured decision rules
- Can help us to identify simple rules that predict events
- Example: predict the price of a bushel of apples as function of latitude and longitude in California



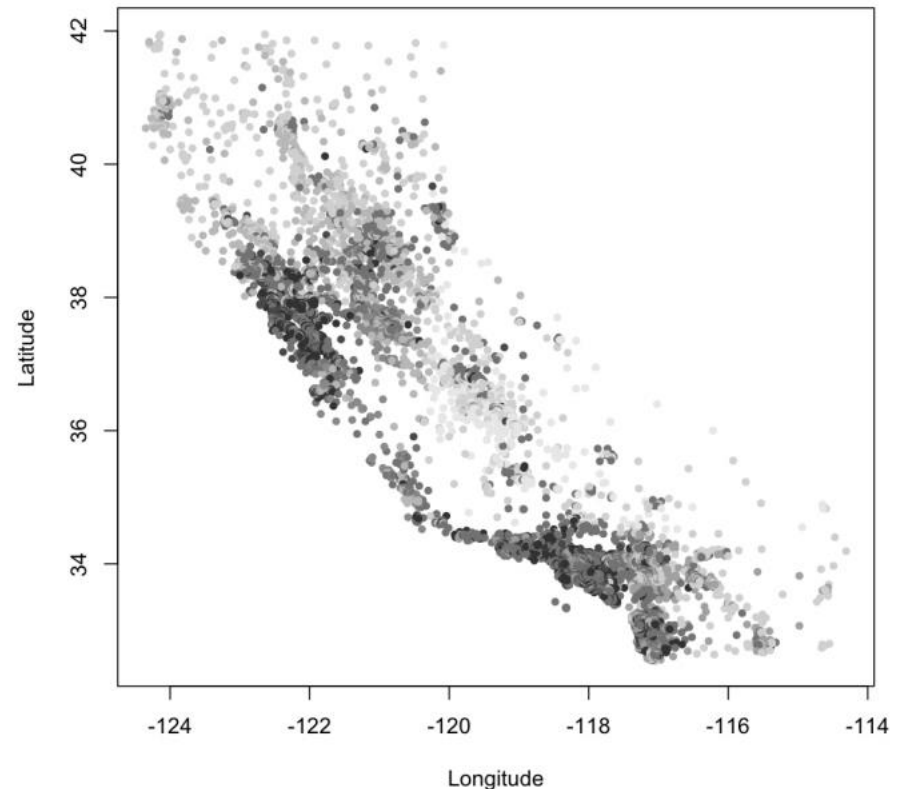
\$13

\$11

Random Forests

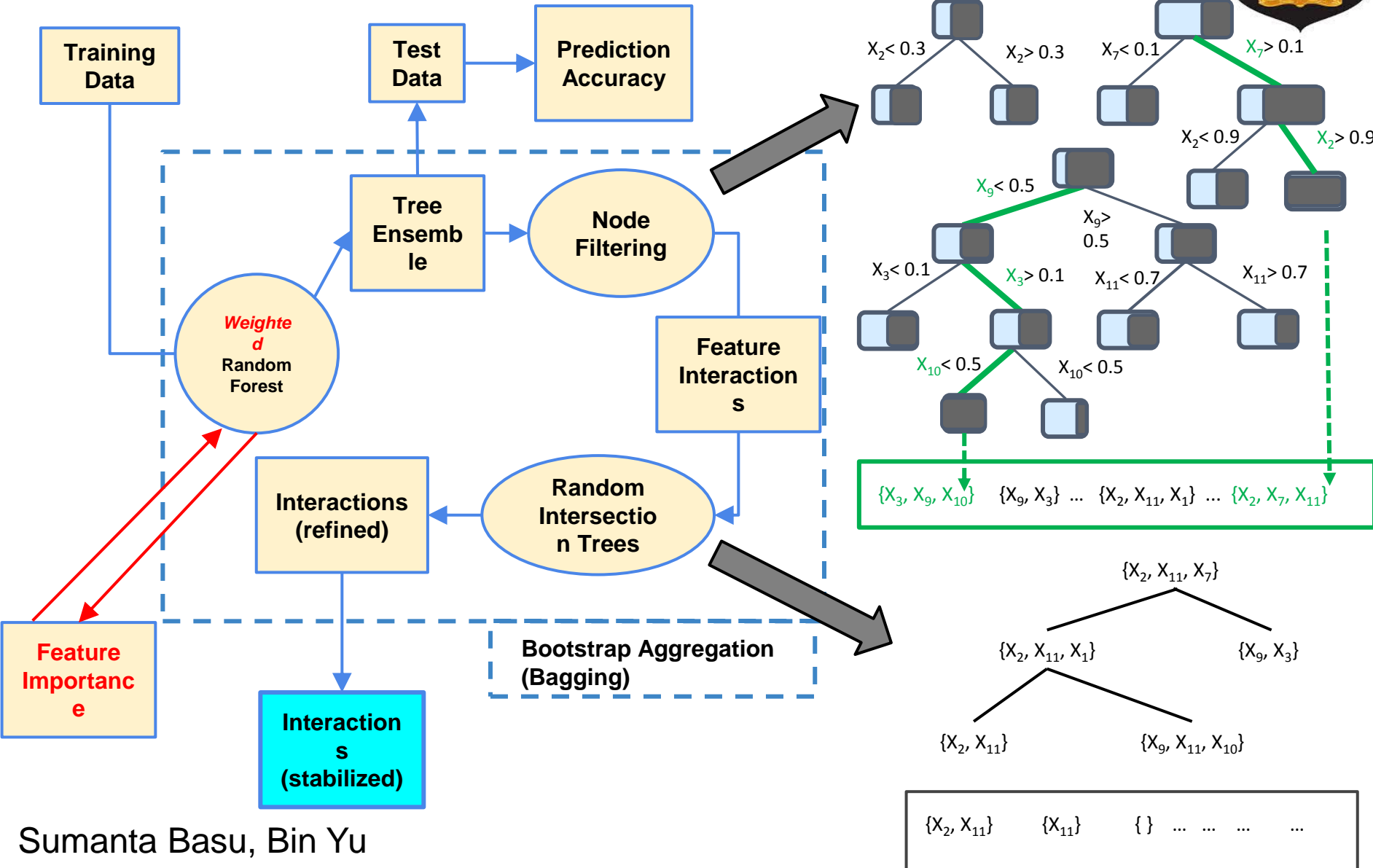


- As in the apples example, trees generate “histogram” approximations of data
- Averaging across trees “smooths” the histogram
- Any one tree is coarse, but together they can quite accurate



iterative Random Forests (iRF)

An interpretable learning machine



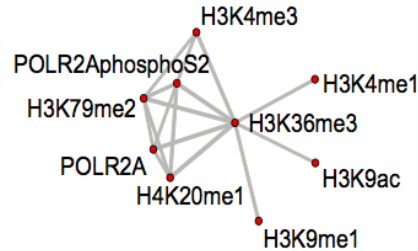
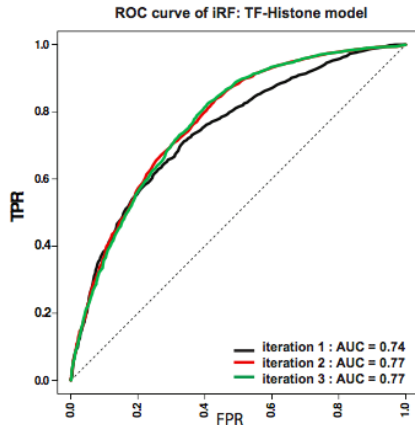
The diagram illustrates the Central Dogma of Molecular Biology. It shows three main stages in colored arrows: DNA (red), RNA (yellow), and Protein (green).

- DNA to RNA:** Labeled 'Transcription', showing a DNA double helix being unwound and an RNA strand being synthesized.
- RNA to Protein:** Labeled 'Translation', showing an RNA strand being read by a ribosome (represented by blue, green, and orange circles) to synthesize a protein.
- Protein to DNA:** Labeled 'Replication', showing a protein being used to synthesize a new DNA double helix.

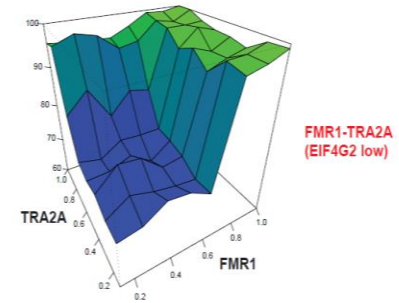
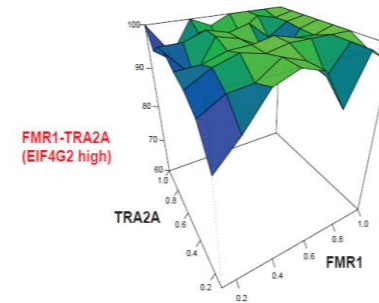
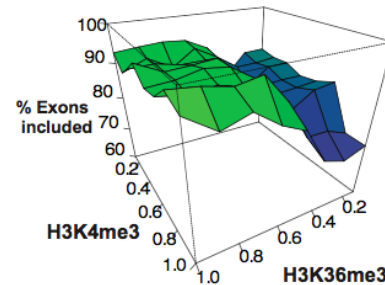
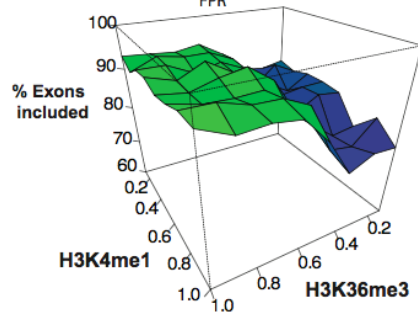
 The DNA and RNA strands are shown with their respective base pairings (A, T, C, G for DNA; U, A, C, G for RNA).

- [illegible]

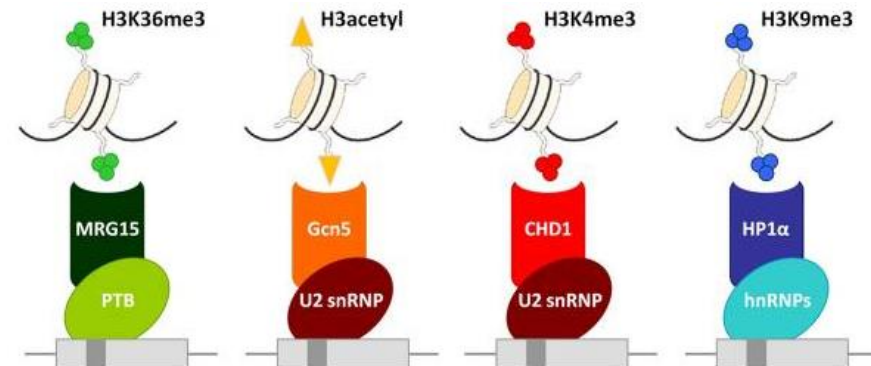
iRF reveals non-linear interactions in RNA processing



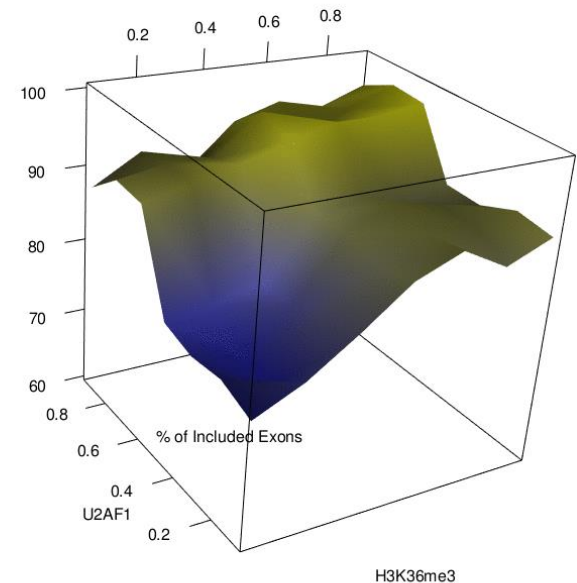
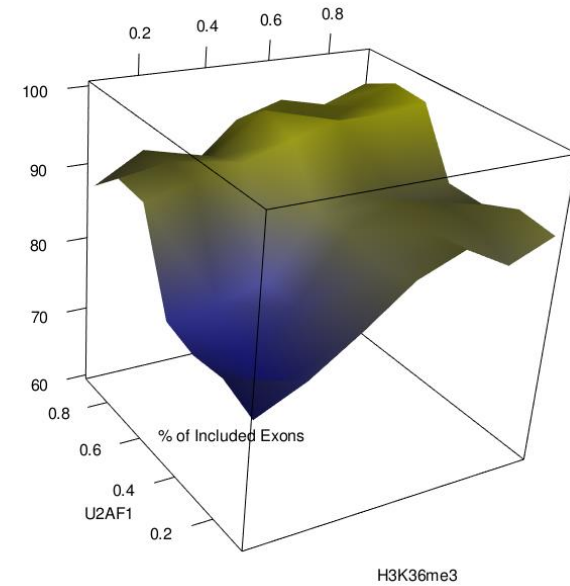
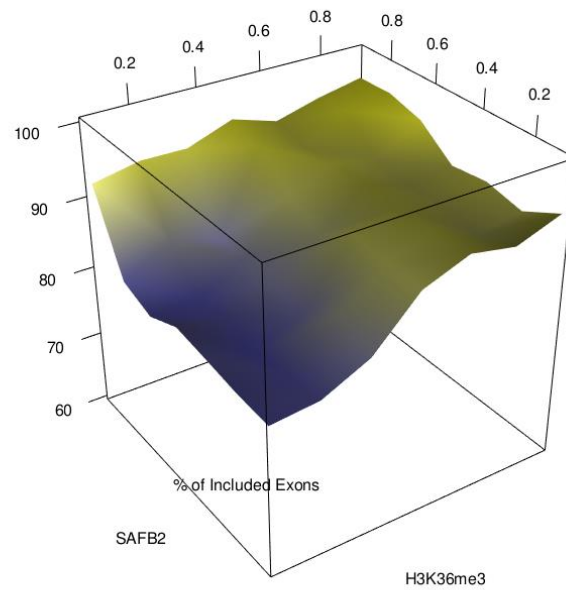
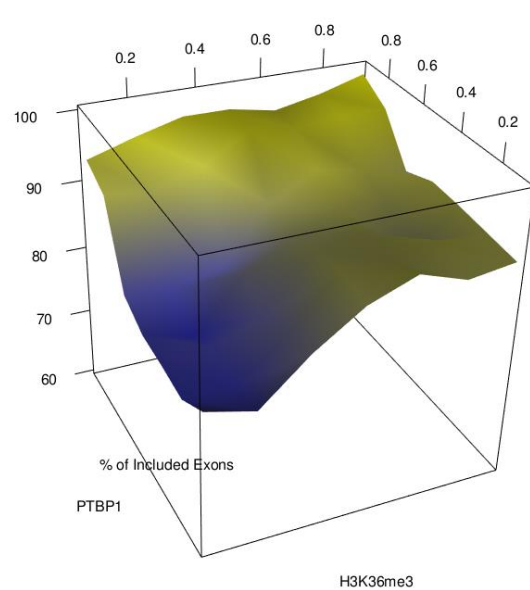
- Predict exon inclusion as a function of ChIP and eCLIP data
- Recover and parameterize complex interactions through surface mapping



- Robust recovery of high order interactions in simulations
- **iRF decouples the order of interactions from the computational cost of detection**



Interactions between H3K36me3 and RBPs

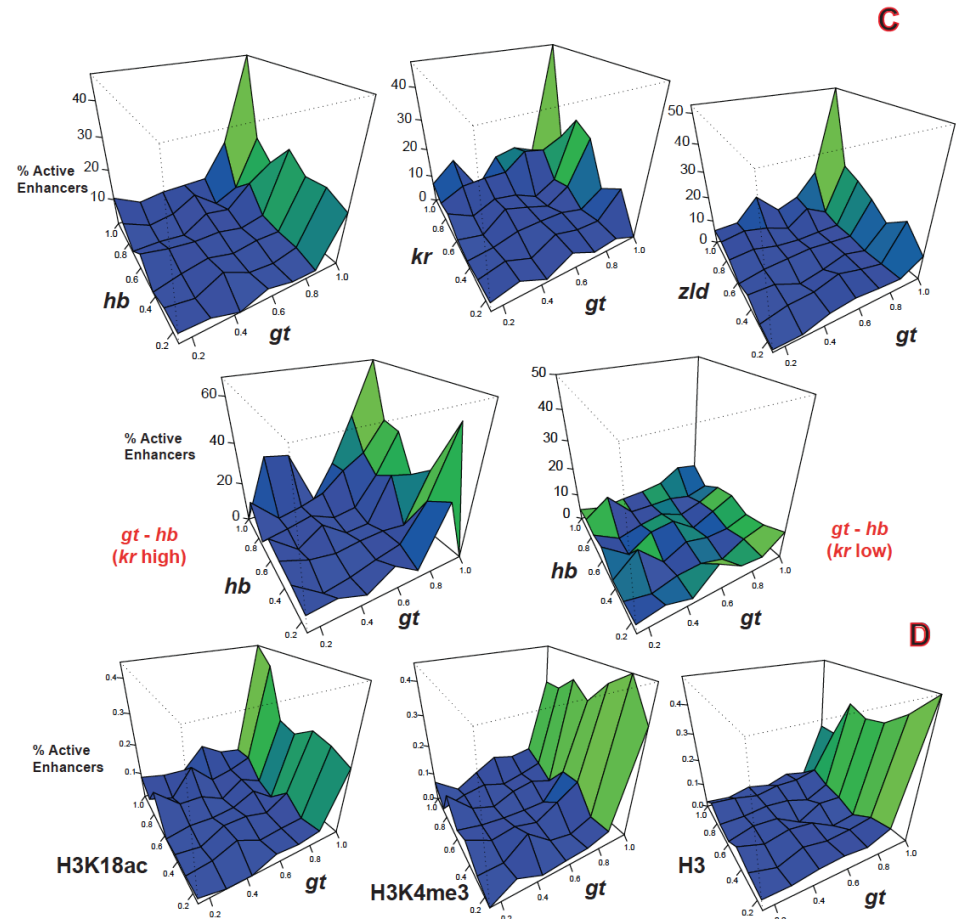
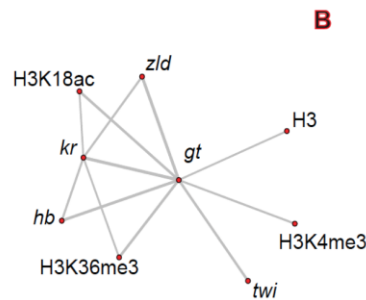
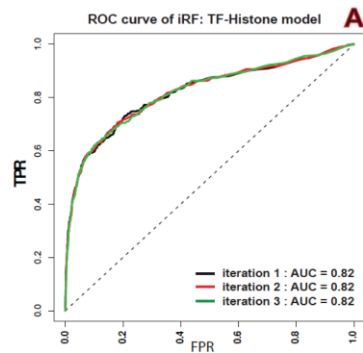


Only four eCLIP'd factors appear to interact with H3K36me3 so far:

- PTBP1 (known)
- SAFB2
- U2AF1
- HNRNPU

Potential participants in the transduction of this mark

Rules for enhancer activity in the early *Drosophila* embryo

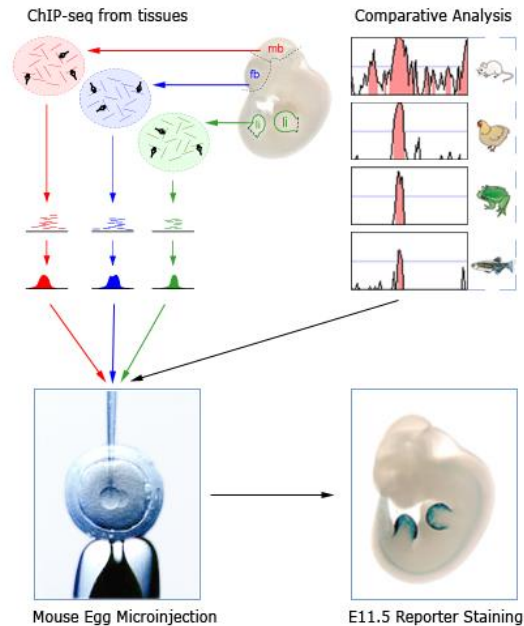


- Rules are and-like for activation, and sharper than for splicing
- Rules combine activators and repressors

Heterogeneity in a single “class” of functional elements



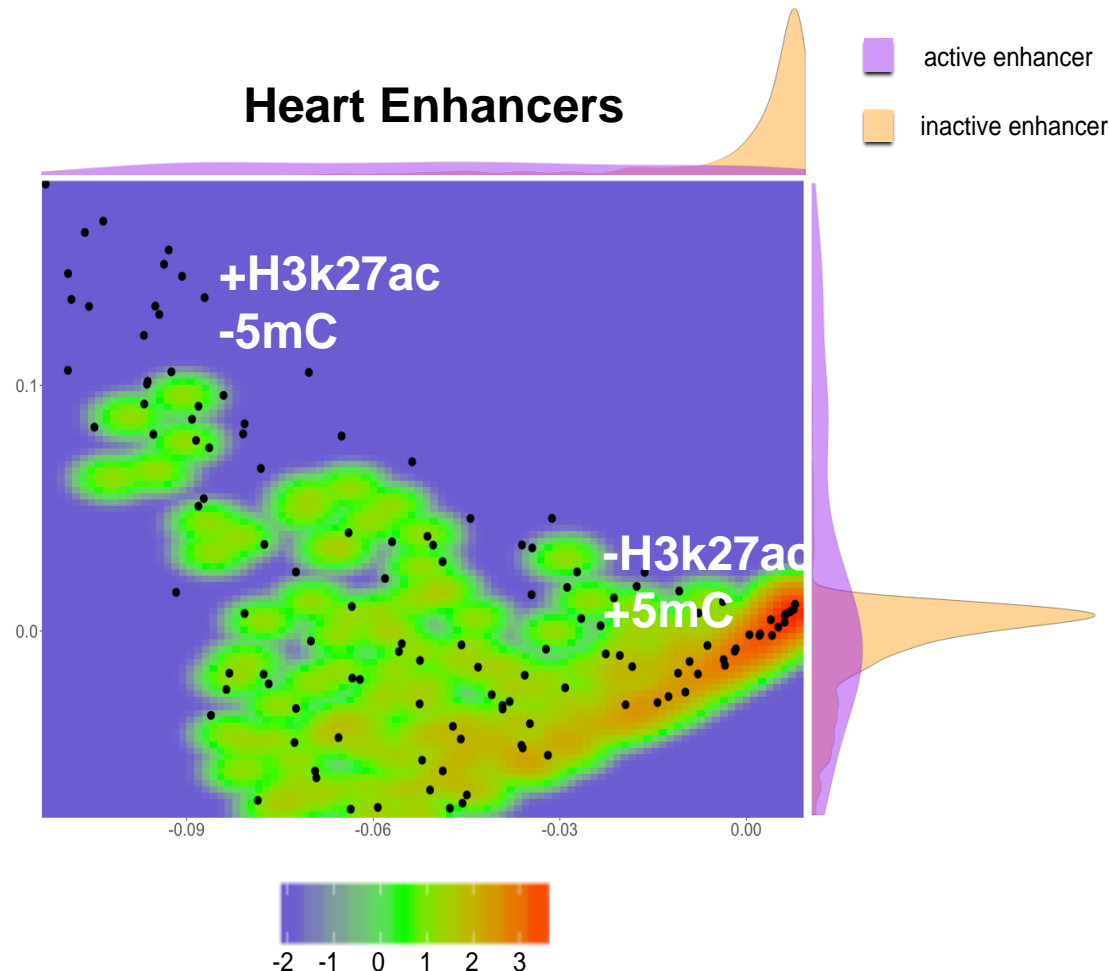
Vista Enhancer Database



- In the enhancer challenge, AUROC's were around 0.4 - 0.89 on test sets (where 20-50% of enhancers are positive)
- Good, but genome-wide FPRs likely still between 50-90%
- Learning machines rely on spread in the active enhancers – **subtypes exist**

Predict enhancer activity as a function of chromatin accessibility, chromatin marks, TF binding, and eRNAs

Heart Enhancers

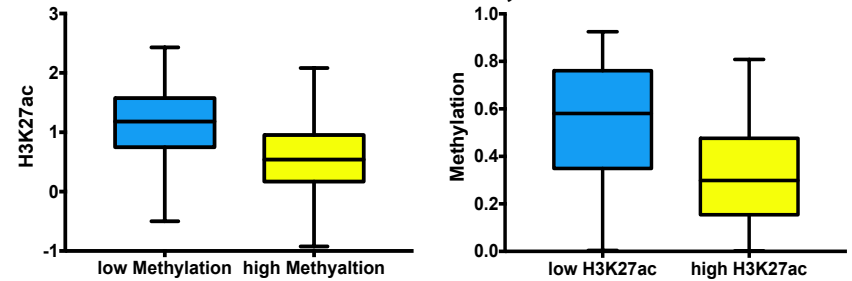




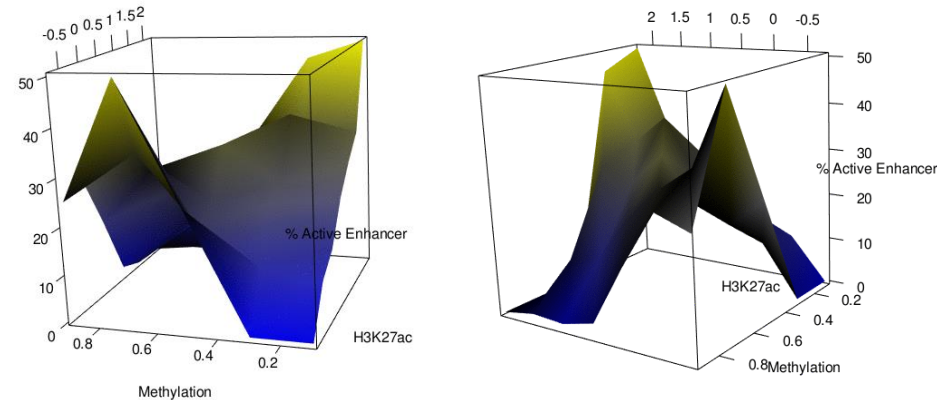
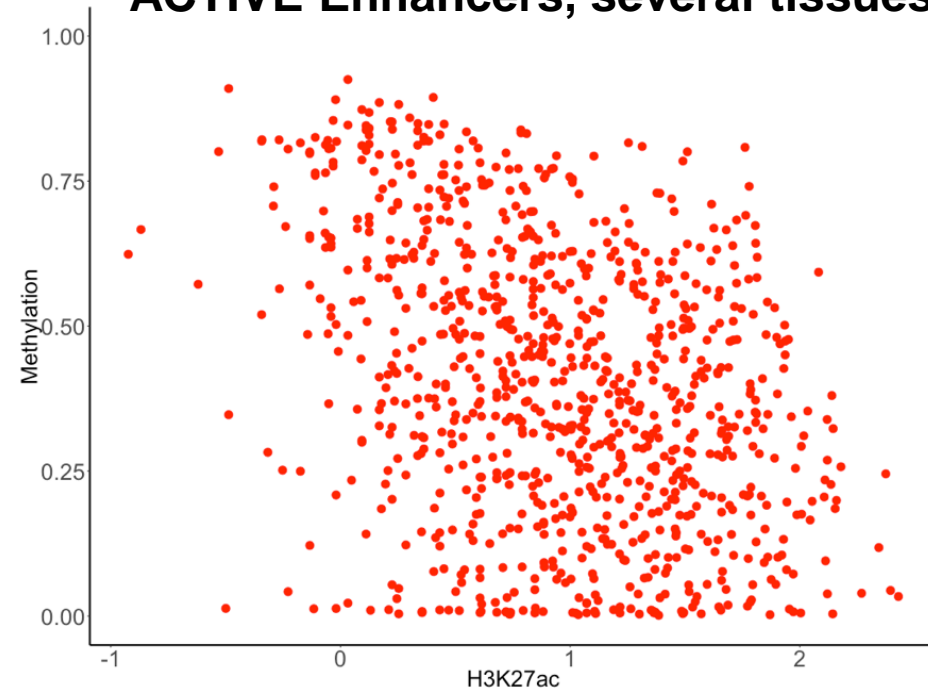
Heterogeneity: 5mC+/-H3K27ac enhancers

- **Classical:** Hypomethylated enhancers tend to have high levels of H3K27ac
- **Novel:** Hypermethylated enhancers tend to have low levels of H3K27ac – often below genomic background

ACTIVE Enhancers, several tissues



ACTIVE Enhancers, several tissues



- Different mechanisms may enable the action of 5mC+/-H3K27ac enhancers, additional assays are needed
- Test more enhancers, and other assays
- Need factors with 5mC binding domains ChIP'd & eCLIP'd + seq'd & MS/MS'd

Hypermethylated enhancers in Zebrafish



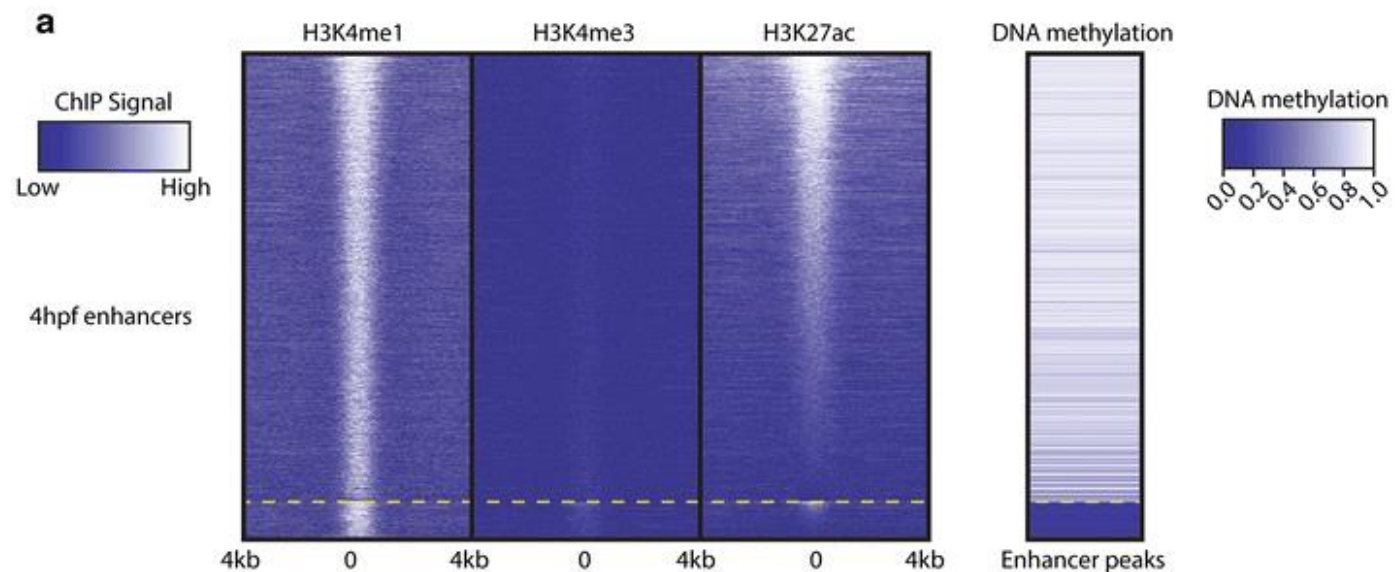
Enhancers reside in a unique epigenetic environment during early zebrafish development

Lucas J. T. Kaaij, Michal Mokry[†], Meng Zhou[†], Michael Musheev, Geert Geeven, Adrien S. J. Melquiond, António M. de Jesus Domingues, Wouter de Laat, Christof Niehrs, Andrew D. Smith and René F. Ketting ✉

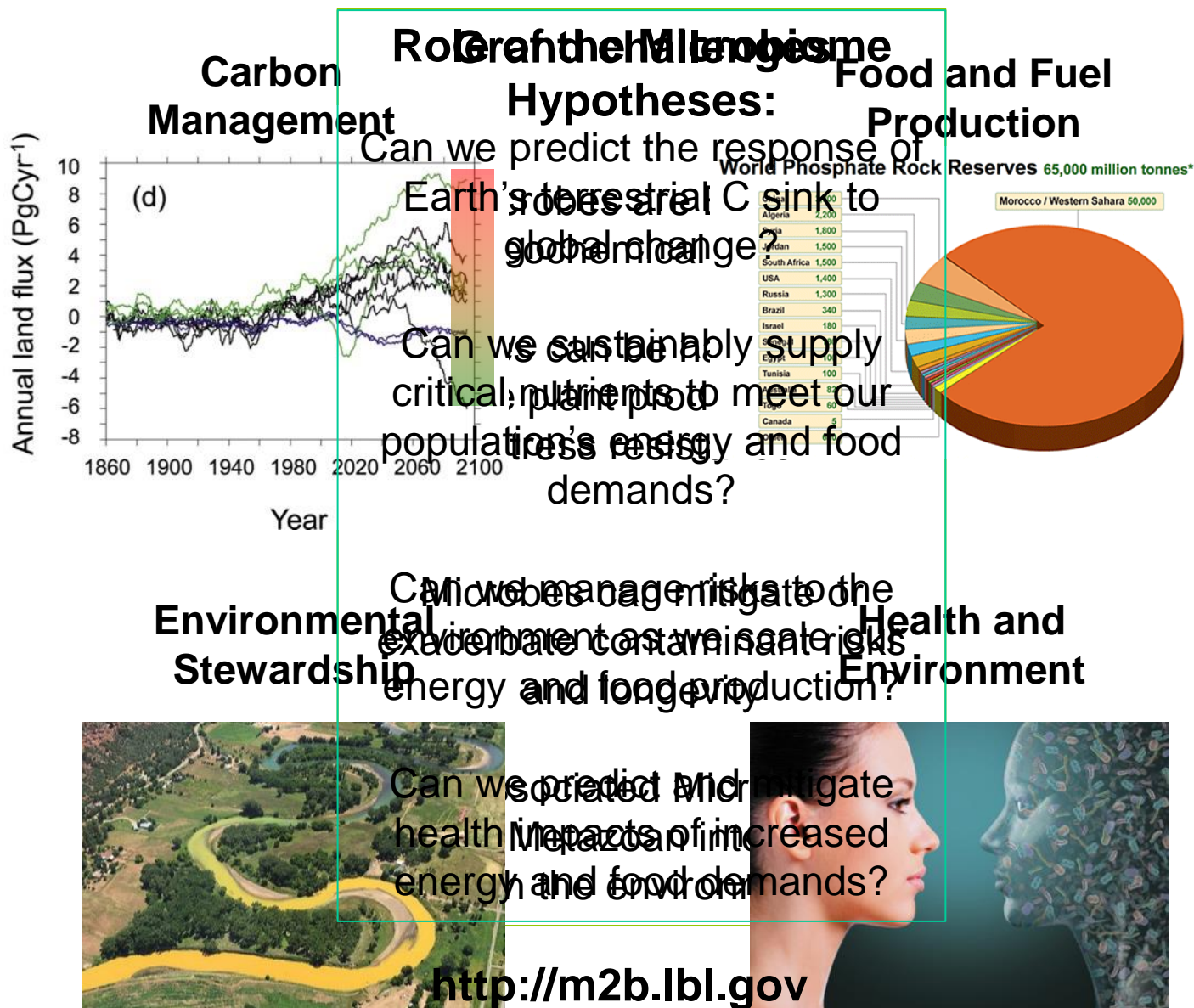
[†] Contributed equally

Genome Biology 2016 17:146 | DOI: 10.1186/s13059-016-1013-1 | © The Author(s). 2016

Received: 25 May 2016 | Accepted: 20 June 2016 | Published: 5 July 2016



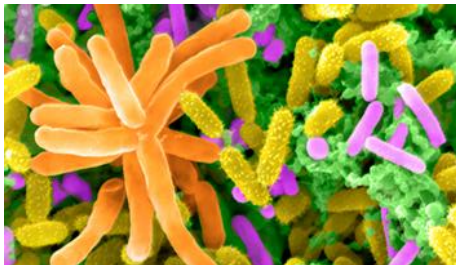
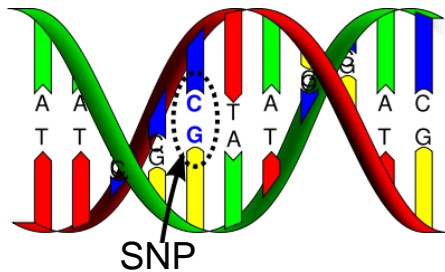
The Microbes to Biomes Initiative



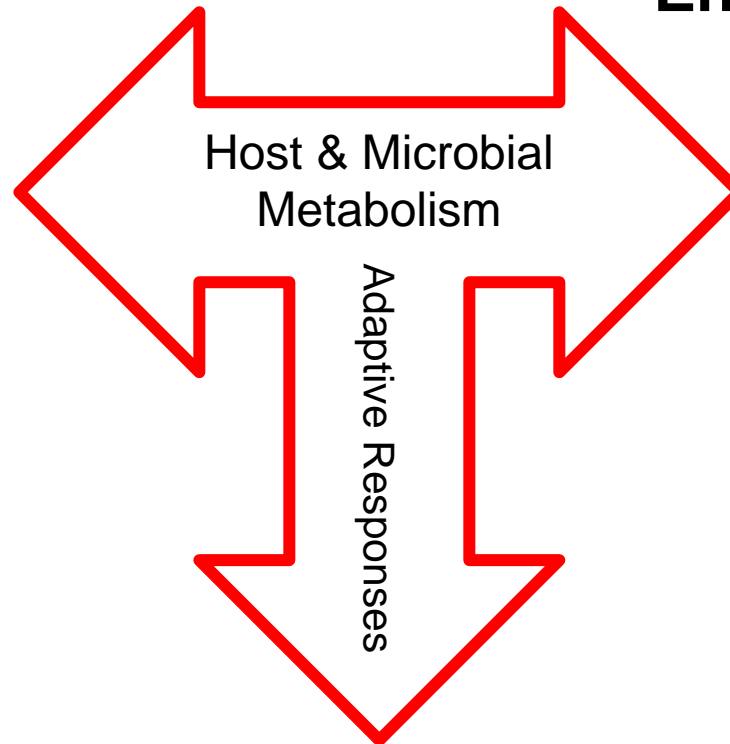
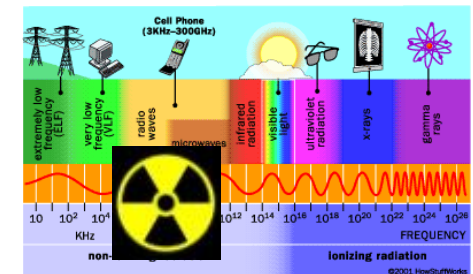
The Microbes to Biomes: Health



Host & Microbial Interactions



Environmental factors



Factors that influence robustness and susceptibility to challenges

How can we identify factors that control individual, population, and ecosystem susceptibility to environmental challenges?

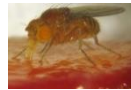
Experimental Design



Measure immediate and long term host and microbial responses to toxicants using **'omics, imaging, and phenotyping**



wild type

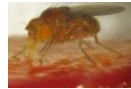


Screen 20 environmental toxicants in flies

Age (days)

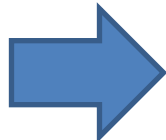
3 5 30

gnotobiotic



Gut: Histology,
Metagenomics,
Metatranscriptomics,
16S classification
Behavior

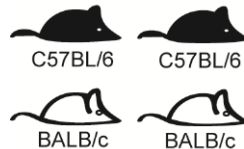
Study toxicants with
the most significant host-
microbiome interactions
in pre-pubescent mice



specified
microbiome

wild type

Screen 3 high-priority toxicants in mice

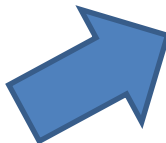


Age (wks)

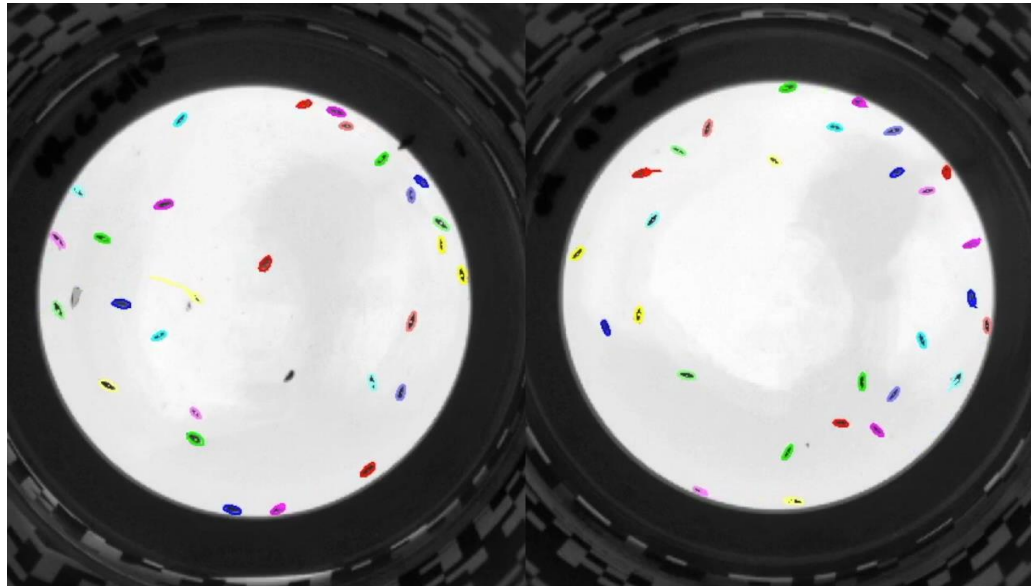
4 5 7 11 15

Fecal : Metagenomics,
Metatranscriptomics,
16S classification
Behavior

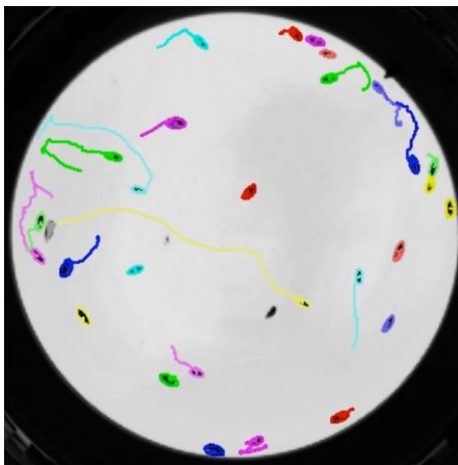
Gut: Histology,
transcriptome,
Metagenomics,
Metatranscriptomics,
16S classification.
Blood: metabolites



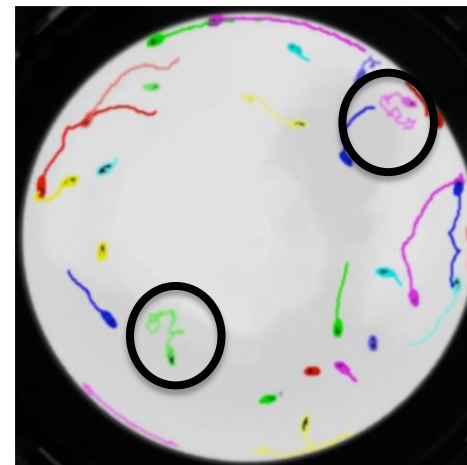
Atrazine induces locomotor phenotypes in OreR flies



Wild-type



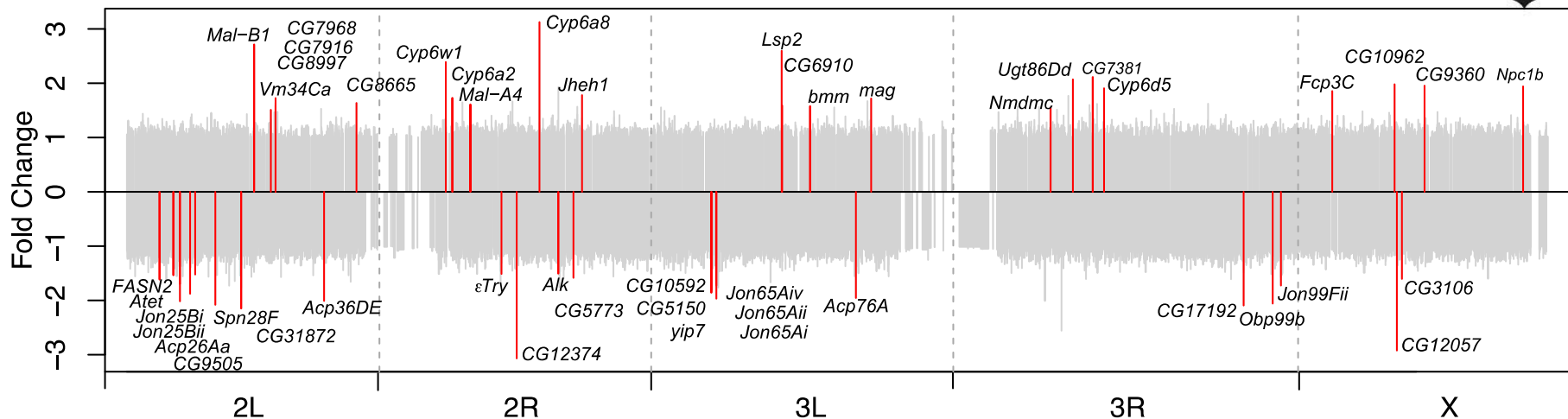
Atrazine treated



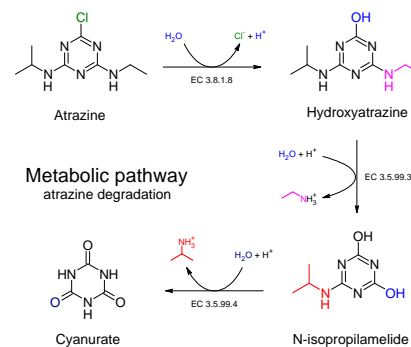
Host response to pesticide exposure



Genome-wide Transcriptional Response to Atrazine



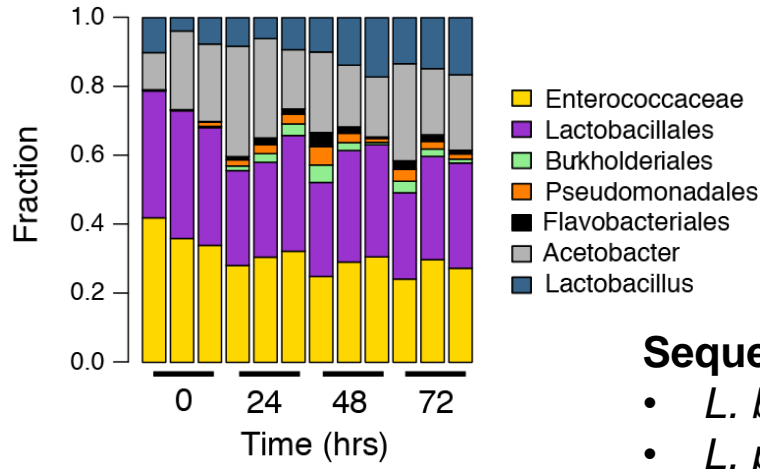
Fly Gene	Human	Function	p - value
<i>Ncp1b</i>	<i>NCP1</i>	Cholesterol trafficking	1e-6
<i>FASN2</i>	<i>FASN</i>	Fatty Acid Synthase	1e-5
<i>bmm</i>	<i>PNPLA2</i>	Triglyceride hydrolysis	1e-5



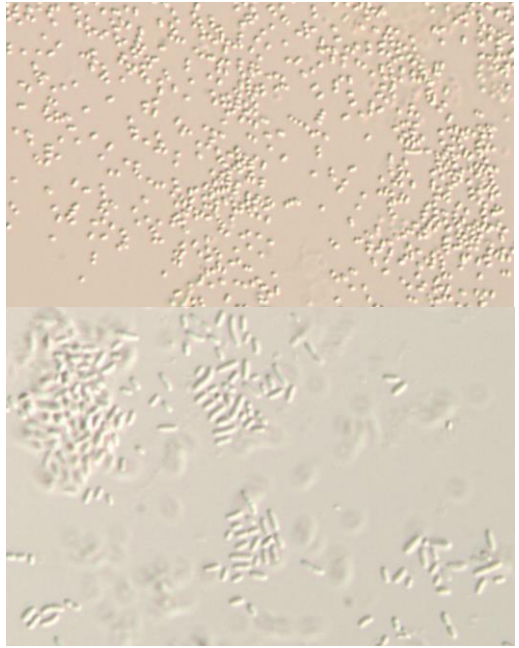
- Atrazine induces an adaptive response associated largely with the transport of lipids, but is not deadly at extremely high doses
- The fruit fly lacks the genes to metabolize Atrazine – how is it surviving the exposure?



Mapping the basal microbiome of laboratory OreR



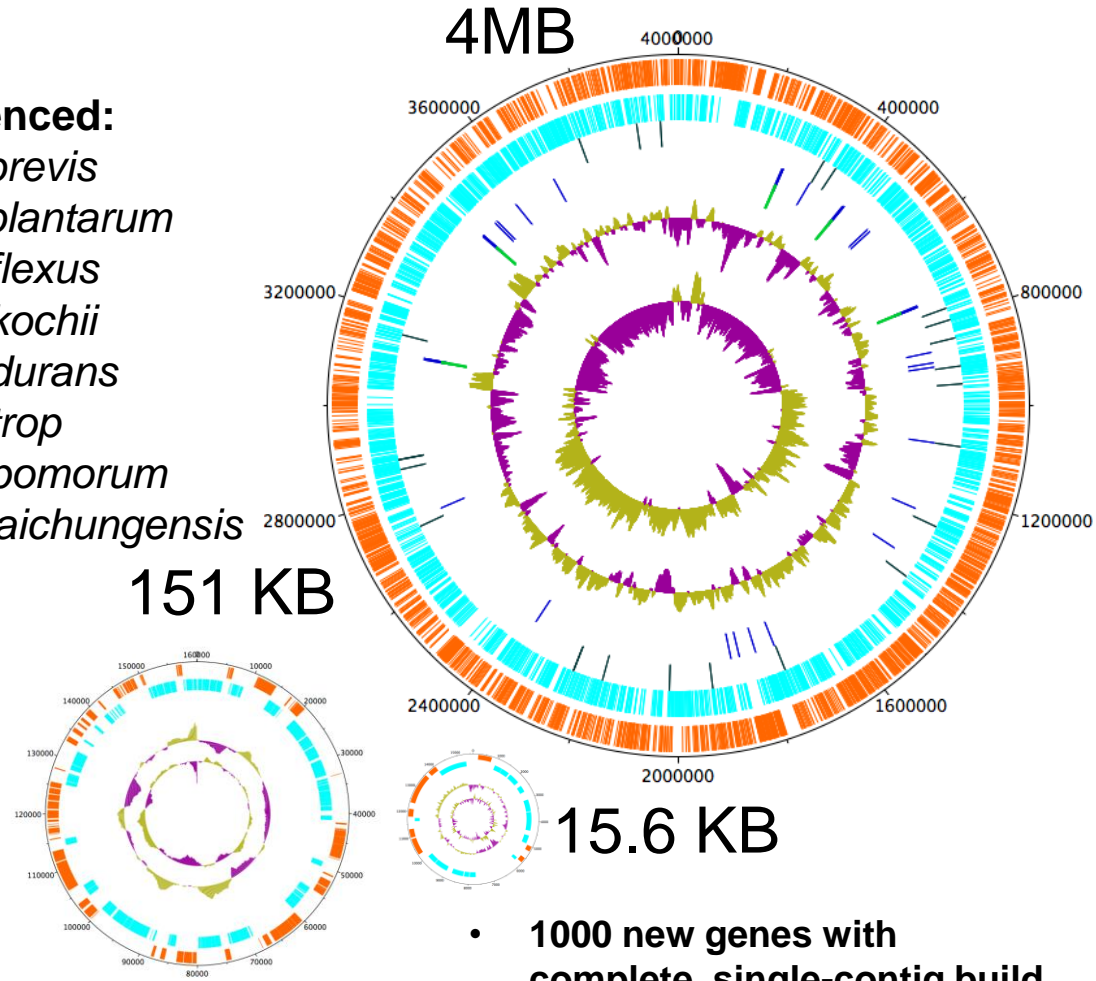
All major species isolated (>95% total microbiome)



Sequenced:

- *L. brevis*
- *L. plantarum*
- *B. flexus*
- *B. kochii*
- *E. durans*
- *A. trop*
- *A. pomorum*
- *P. taichungensis*

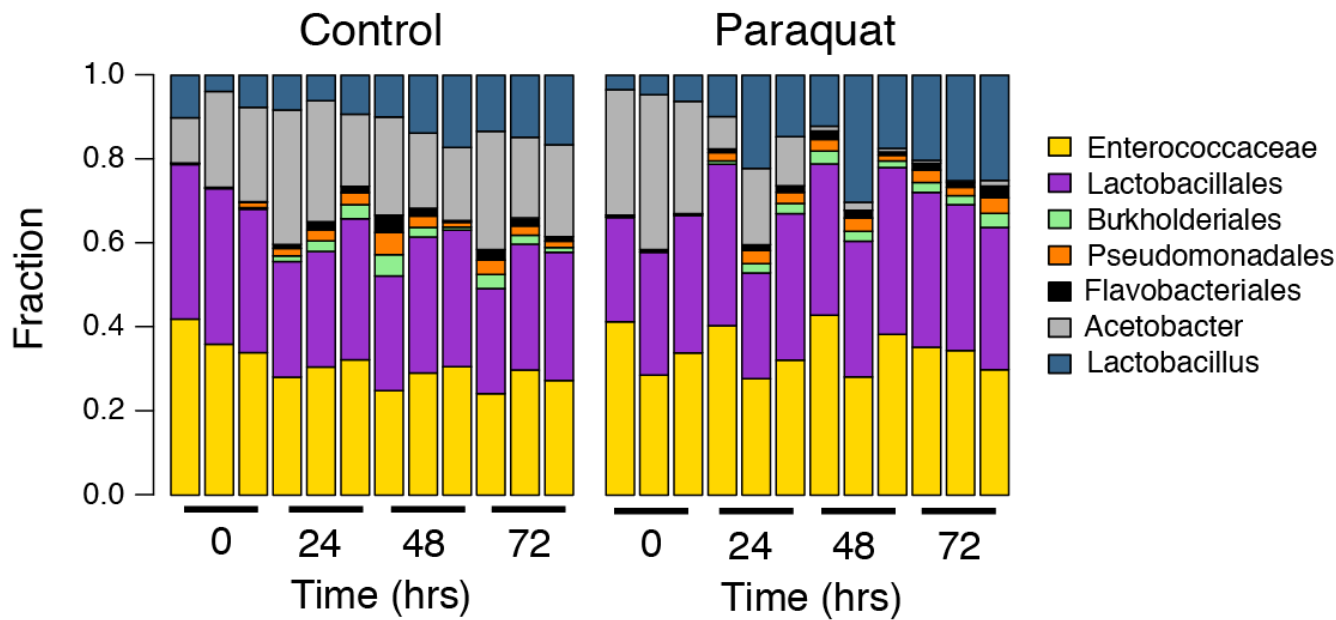
Acetobacter tropicalis



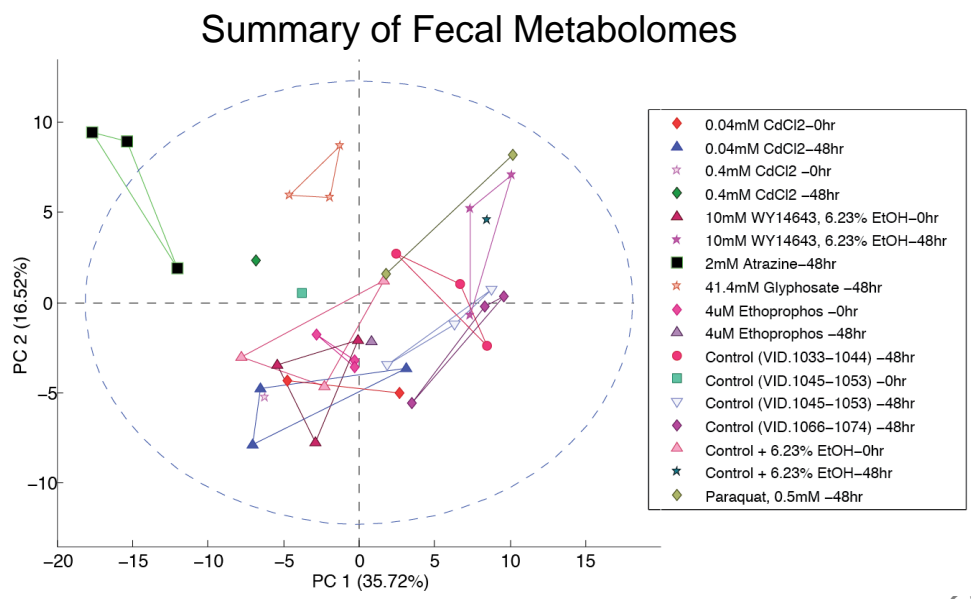
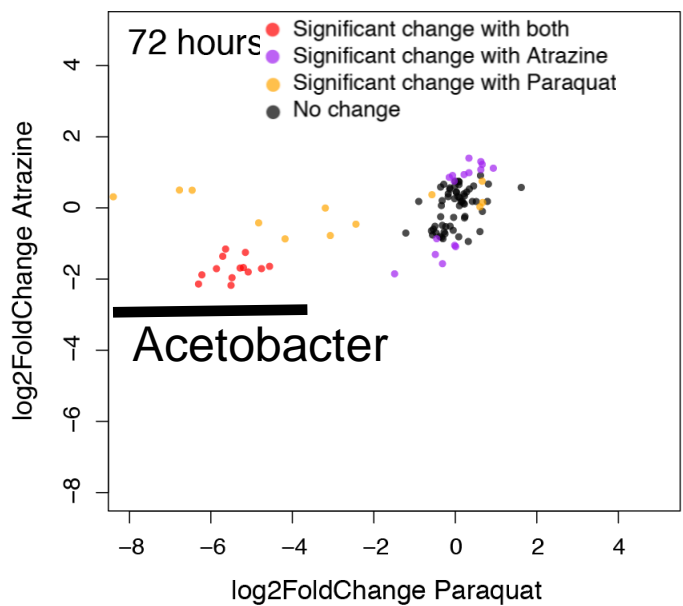
- 1000 new genes with complete, single-contig build



Microbiome remodeling during exposure



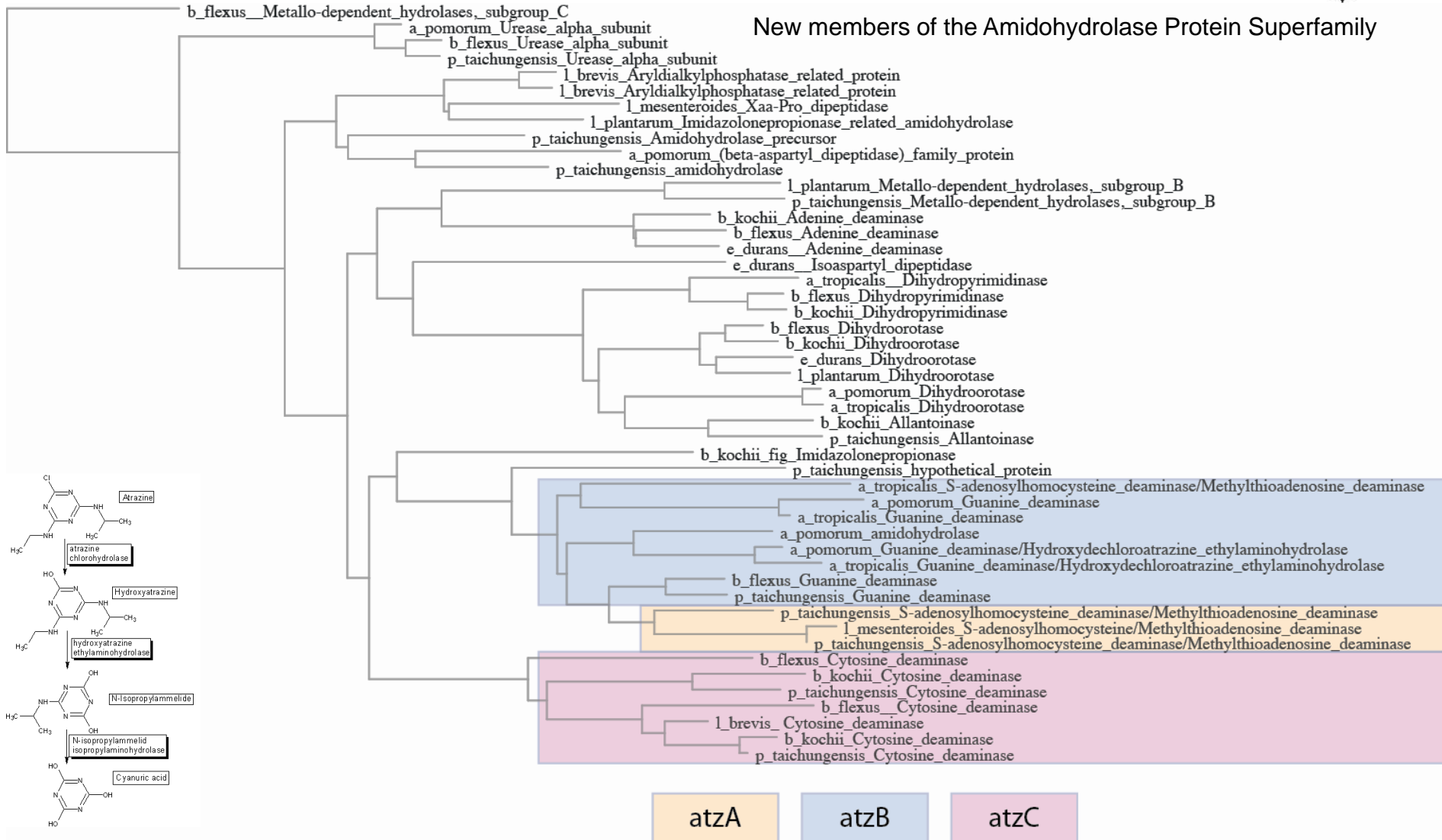
- Paraquat and Atrazine induce the collapse of *Acetobacter* in the insect gut
- Atrazine is rapidly and completely metabolized in the host gut, yet still induces a hyperactive phenotype



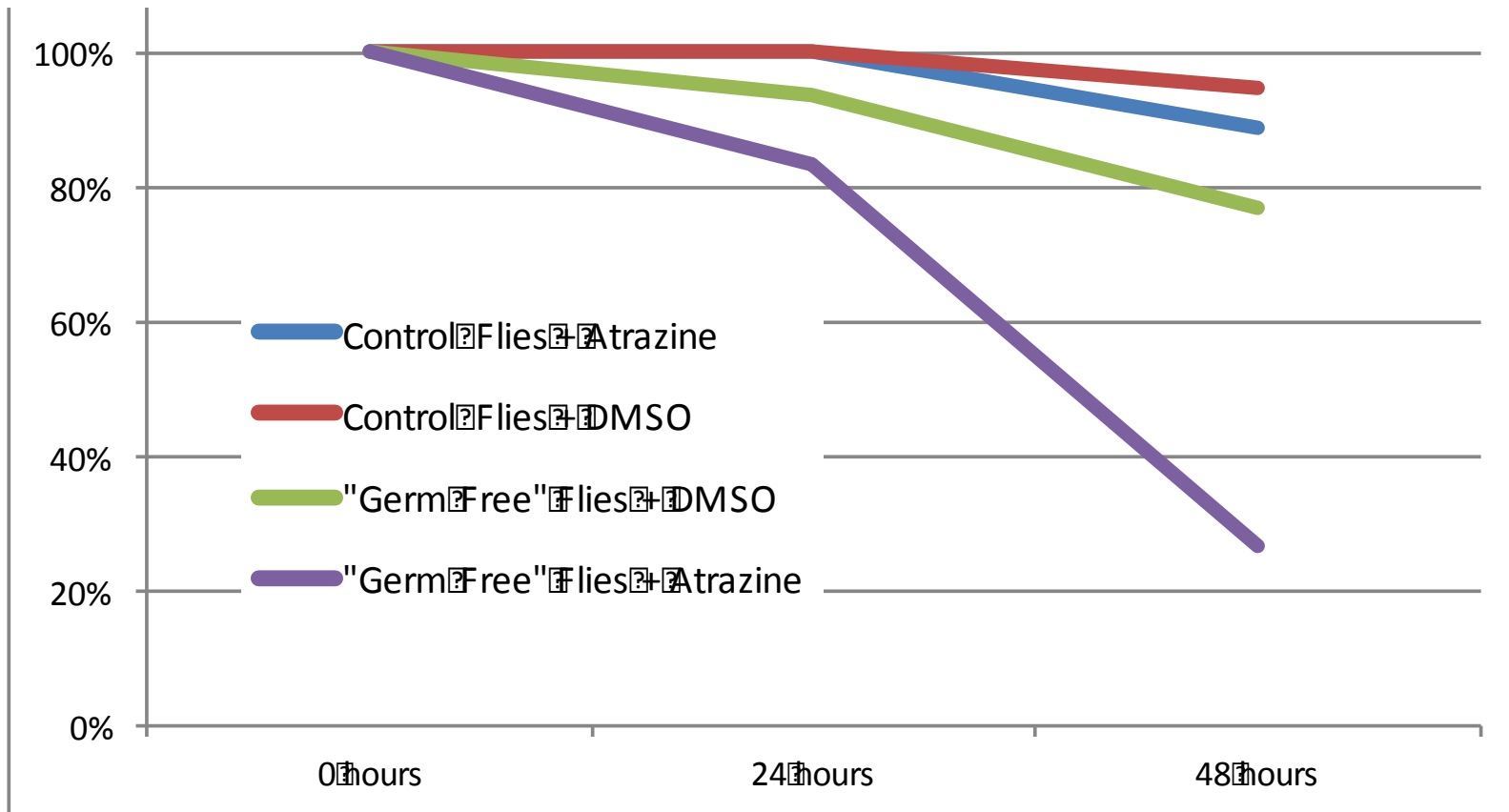
Candidate atrazine metabolizing genes from sequenced fly gut microbes



New members of the Amidohydrolase Protein Superfamily



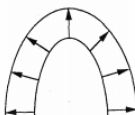
The gut microbiome is protective



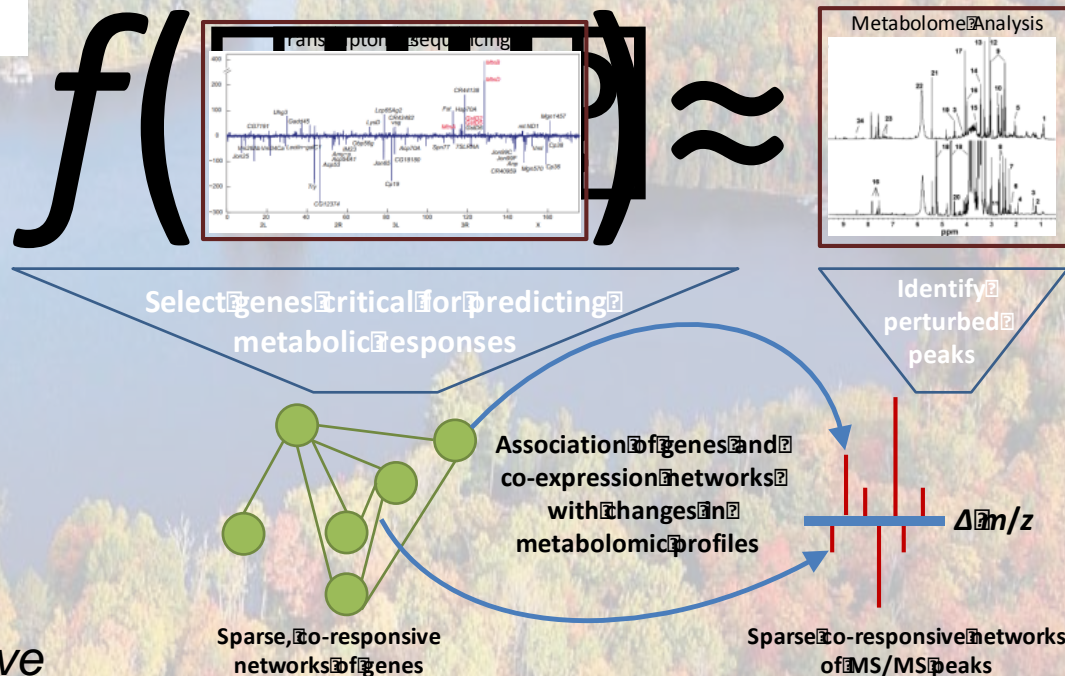
Machine learning to achieve predictive molecular ecology by 2022



$$\partial_t h = - \underbrace{\frac{1}{2}(\partial_x h)^2}_{\text{lateral growth}} + \underbrace{\frac{1}{2}\partial_x^2 h}_{\text{relaxation}} + \underbrace{\xi}_{\text{space-time white noise}} \underbrace{h(t, x)}_{\text{height at time } t \text{ position } x}$$

$$F(\partial_x h) \sim F(0) + F'(0)\partial_x h + \frac{1}{2}F''(0)(\partial_x h)^2 + \dots$$


- **We have iterative Random Forests (iRF):** Interactions of any order at the same computational cost as pairwise



- **We need AutoFit Dynamics (AFD):** By 2020, it may be possible to infer dynamics in high-dimensional systems

The Health M2B Team



Lawrence Berkeley National Lab

- Susan Celniker
- Ken Wan
- Sarah Morris
- Jian-Hua Mao
- Antoine Snijders
- Gary Karpen
- **Sasha Langley**

University of Birmingham

- Mark Viant
- Jennifer Kirwan

University of Oklahoma

- Jan Sunner
- Vincent Bonifay
- Iwona Beech

Pacific Northwest National Lab

- Jannet Jansson
- Young-Mo Kim
- Thomas Metz
- Colin J. Brislawn

Cornell University

- Sumanta Basu