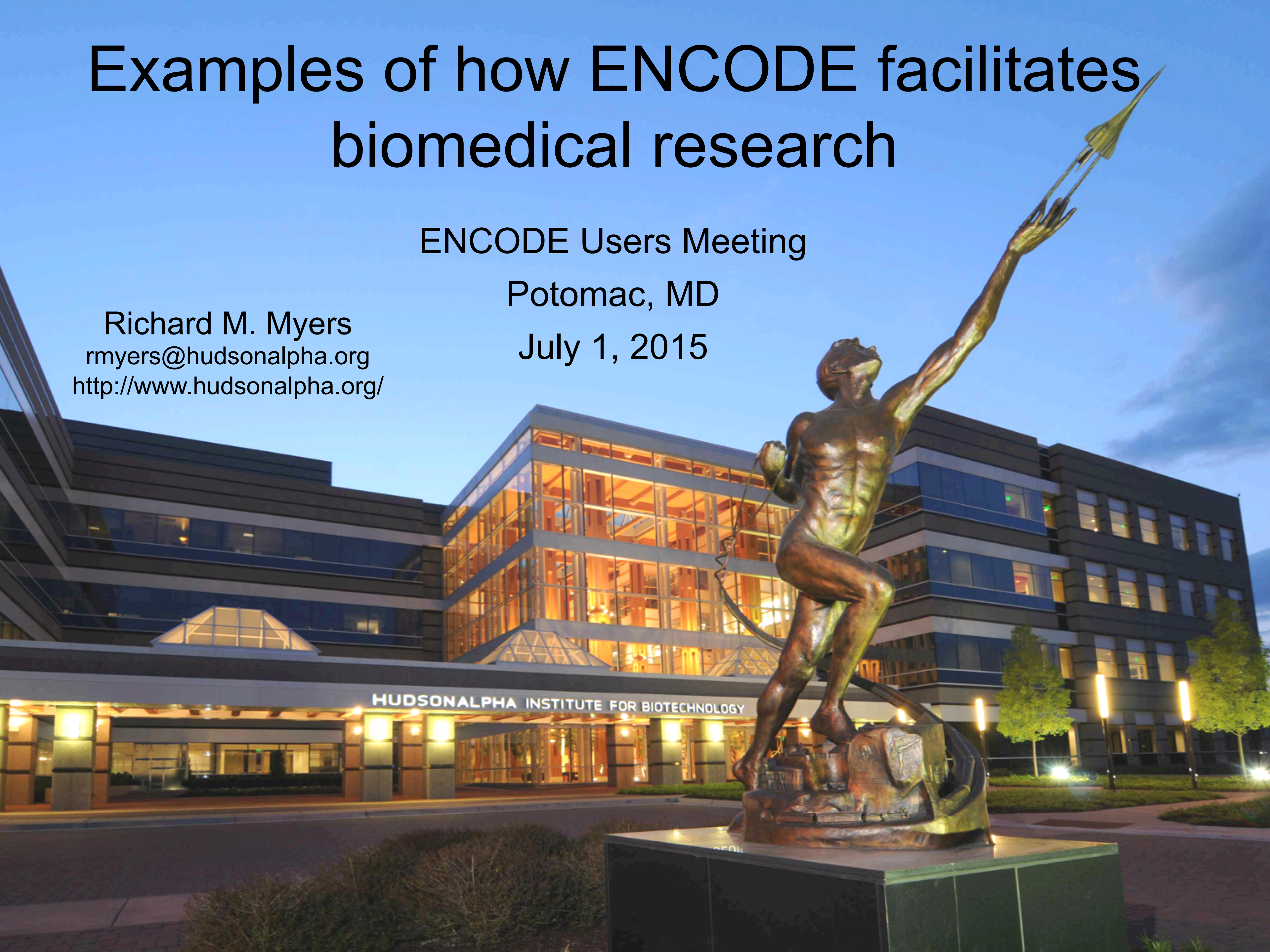
Examples of how ENCODE facilitates biomedical research

ENCODE Users Meeting

Potomac, MD

July 1, 2015

Richard M. Myers
rmyers@hudsonalpha.org
http://www.hudsonalpha.org/



Disclosure

Our group has been part of the ENCODE Consortium since it began in 2003

Rick Myers

Barbara Wold

Ross Hardison

Eric Mendenhall

Ali Mortazavi

Tim Reddy

Goals of ENCODE

Annotate the human genome

Disseminate data to researchers everywhere

5 examples of how we use ENCODE data to help in our research on human diseases

1. Discovering the causes of undiagnosed genetic diseases

Childhood genetic disorders

1.5-3% of kids worldwide are born with 1 or more of:

- intellectual disability
- developmental delay
- heart defects
- craniofacial and skeletal abnormalities
- severe autism
- seizures

The vast majority of these problems have genetic causes

Diagnostic challenges for childhood genetic disorders

Inaccurate or undetermined causes (i.e., diagnoses) are a major hardship:

Years of expensive, invasive, and futile testing

Impossible to predict disease progression, symptoms

Treatment decisions are complicated

Slows research into developing new therapies

Impacts family planning

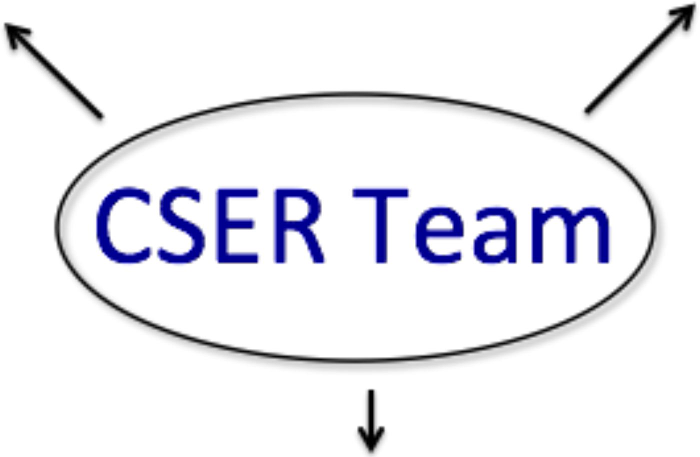
Results in feelings of parental guilt and lack of control

Thus, identifying the root genetic causes is essential

HudsonAlpha Pediatric Genetics Project

Sequence whole genomes of 500 children with developmental/ intellectual delay of unknown etiology (and both parents' genomes too)

North Alabama Children's Specialists



CSER Team

Univ. of Louisville

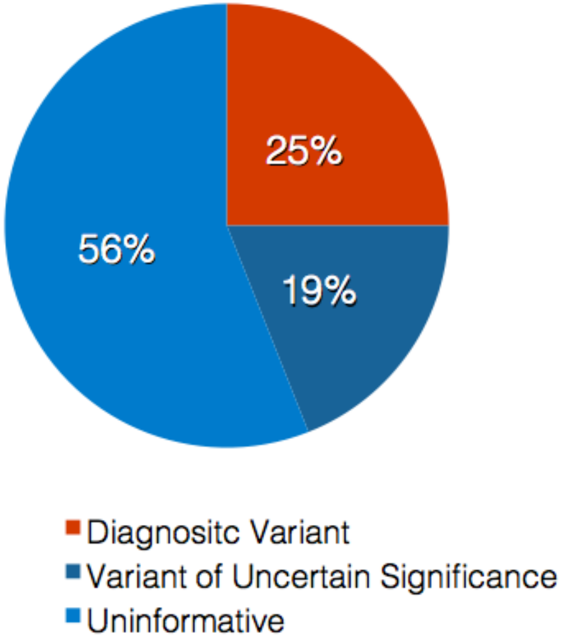
HudsonAlpha

Exome results so far

Exome sequencing completed for 171 families

Definitive genetic diagnosis in 25% of the children

- Pitt-Hopkins syndrome
- Dravet syndrome
- Rett syndrome
- Rubinstein-Taybi syndrome
- Noonan-like syndrome
- Many never-described causes



>20% of families receive uncertain genetic findings that will likely be definitively diagnostic in the future

Whole genome sequencing of trios

Illumina X Ten sequencers: $ of 30X WGS = $ of exome

We have completed WGS of 30 trios in our Childhood Genetics Project

Results:

Diagnostic rate is higher

Identified at least 3 cases where regulatory mutations were the causes

We relied heavily on ENCODE data to identify functional regulatory segments

Annotating genetic variants

Problem:

HUGE number of sequence variants in each individual

Most are not important

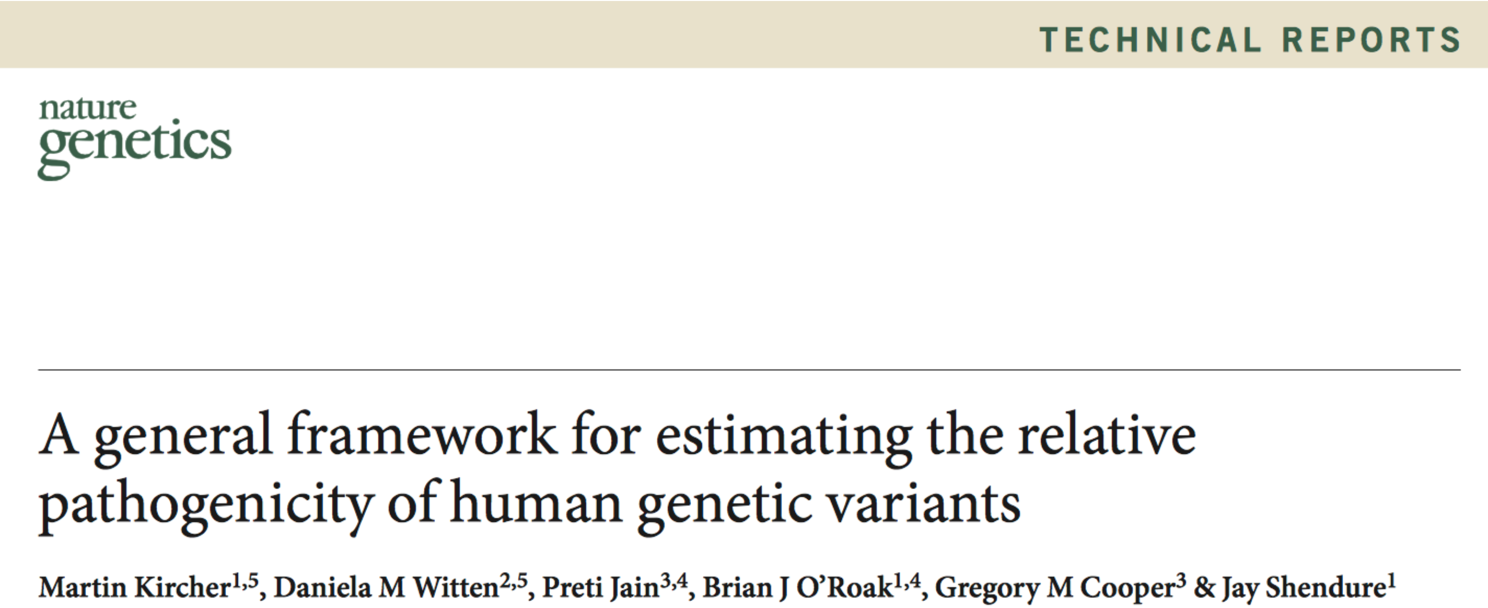
How to find which variants have an effect on:

- The molecular/biochemical function of the gene
- The organism

CADD

Combined Annotation Dependent Depletion

Greg Cooper and Jay Shendure



TECHNICAL REPORTS

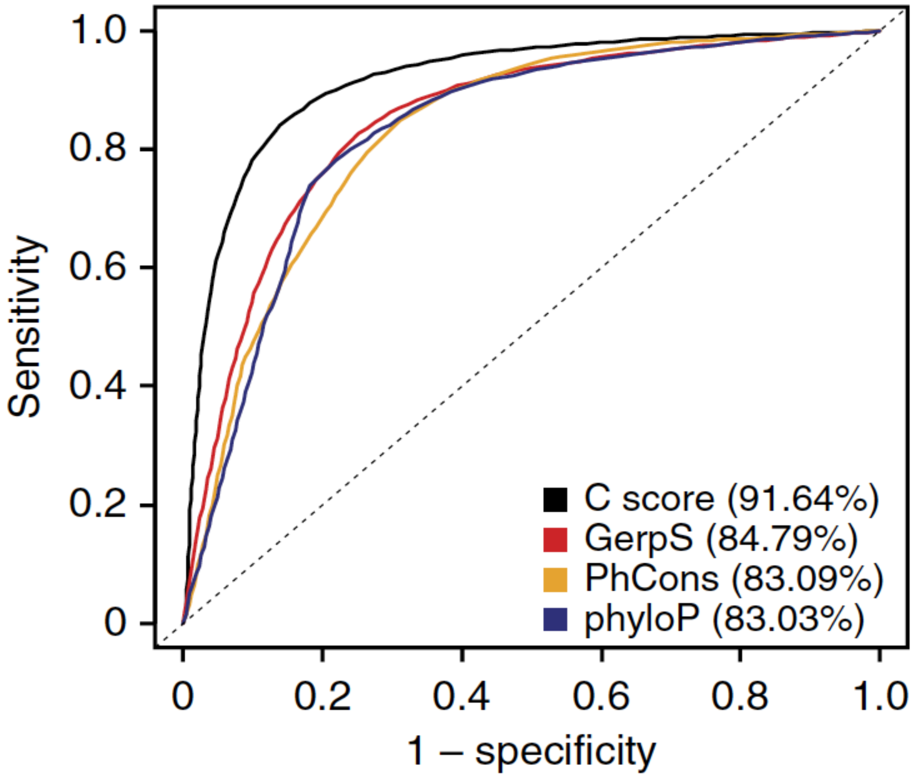
nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

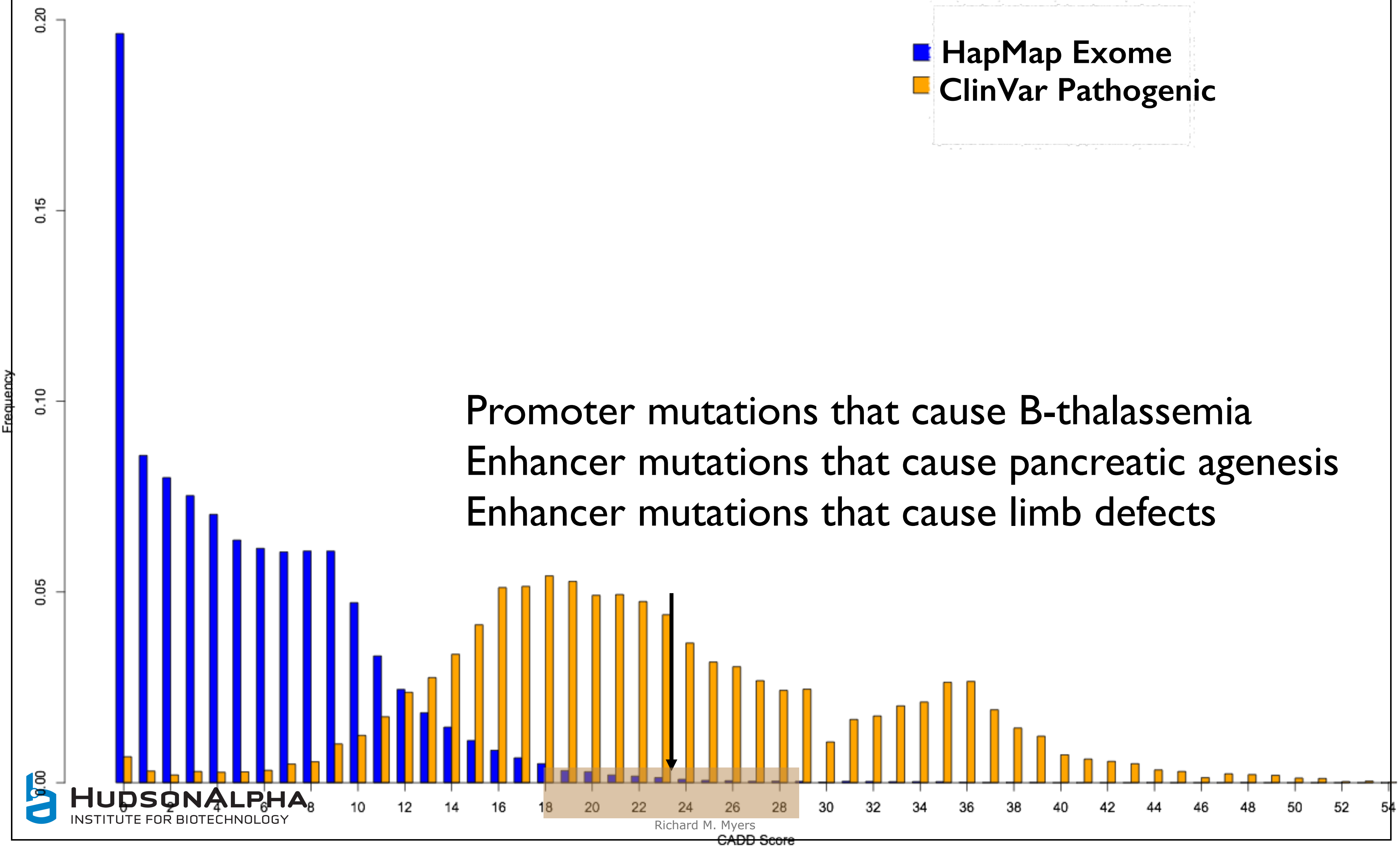
Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

Nature Genetics **46**, 310–315 (2014)

CADD integrates many features to give a single pathogenicity score



Typical vs pathogenic CADD scores



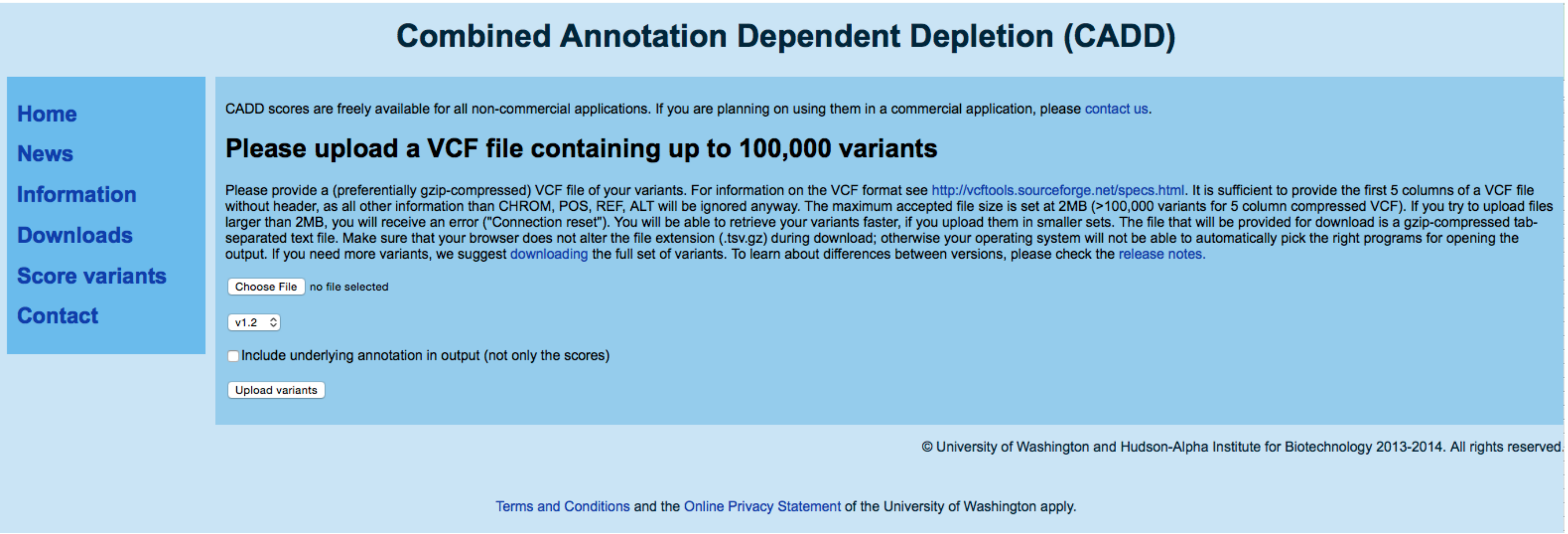
Promoter mutations that cause B-thalassemia

Enhancer mutations that cause pancreatic agenesis

Enhancer mutations that cause limb defects

Use the CADD webserver!

http://cadd.gs.washington.edu



Combined Annotation Dependent Depletion (CADD)

CADD scores are freely available for all non-commercial applications. If you are planning on using them in a commercial application, please contact us.

Please upload a VCF file containing up to 100,000 variants

Please provide a (preferentially gzip-compressed) VCF file of your variants. For information on the VCF format see http://vcftools.sourceforge.net/specs.html. It is sufficient to provide the first 5 columns of a VCF file without header, as all other information than CHROM, POS, REF, ALT will be ignored anyway. The maximum accepted file size is set at 2MB (>100,000 variants for 5 column compressed VCF). If you try to upload files larger than 2MB, you will receive an error ("Connection reset"). You will be able to retrieve your variants faster, if you upload them in smaller sets. The file that will be provided for download is a gzip-compressed tab-separated text file. Make sure that your browser does not alter the file extension (.tsv.gz) during download; otherwise your operating system will not be able to automatically pick the right programs for opening the output. If you need more variants, we suggest downloading the full set of variants. To learn about differences between versions, please check the release notes.

□ Include underlying annotation in output (not only the scores)

© University of Washington and Hudson-Alpha Institute for Biotechnology 2013-2014. All rights reserved.

Terms and Conditions and the Online Privacy Statement of the University of Washington apply.

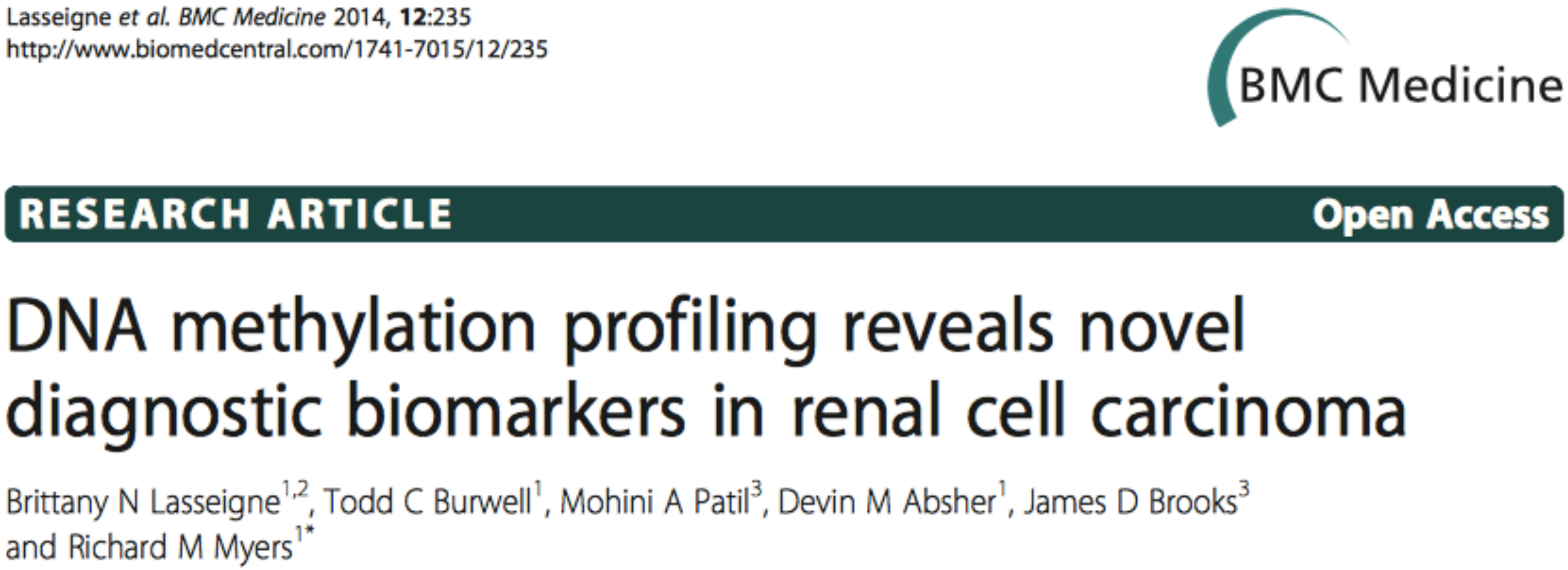
2. Understanding renal cell carcinoma

ENCODE data were instrumental in helping us identify regions of the genome that are ~100% accurate diagnostic markers for kidney cancer

(and even for prognosis of different subtypes)

Genomic signatures of renal cell carcinoma

Brittany Lasseigne, Jim Brooks, Myers Lab



Lasseigne *et al. BMC Medicine* 2014, **12**:235
http://www.biomedcentral.com/1741-7015/12/235

BMC Medicine

RESEARCH ARTICLE **Open Access**

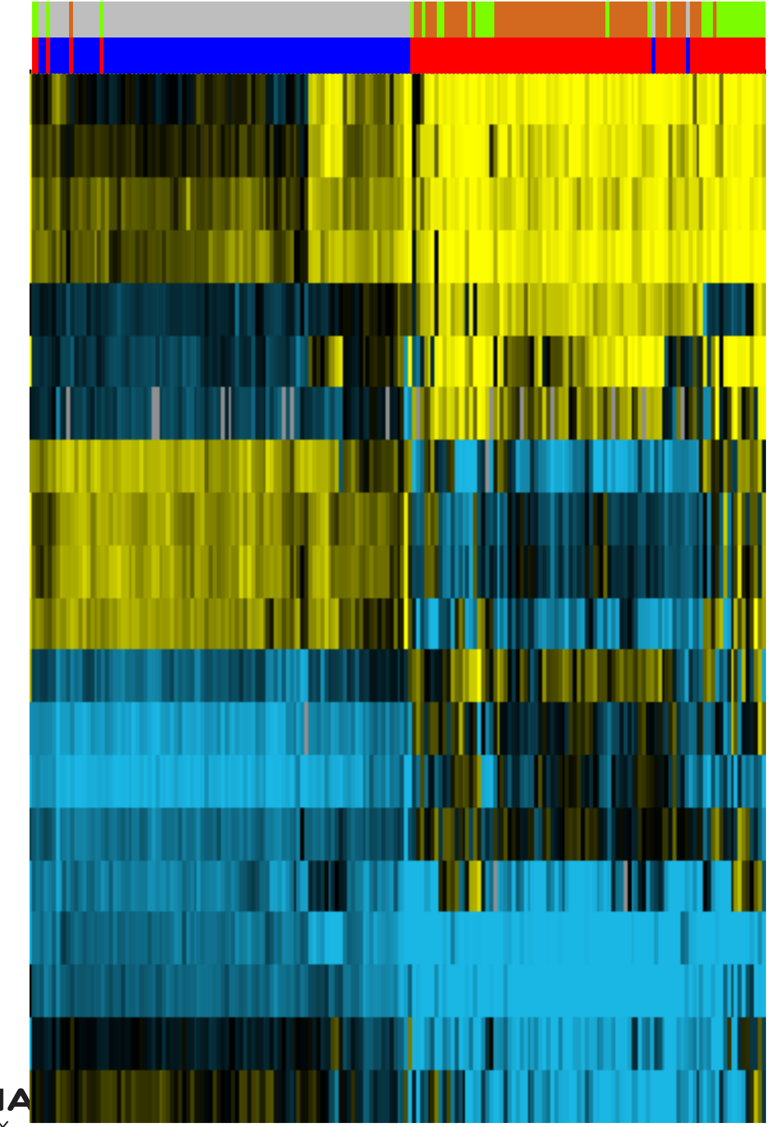
DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma

Brittany N Lasseigne[1,2], Todd C Burwell[1], Mohini A Patil[3], Devin M Absher[1], James D Brooks[3] and Richard M Myers[1*]

We measured DNA methylation and copy number variants in 135 kidney tumors and matched non-tumor kidney tissues

Top 20 DNA methylation markers

All subtypes



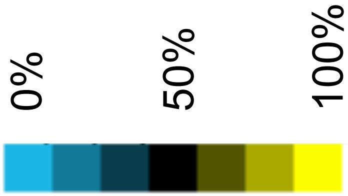
🟥 Kidney Tumor

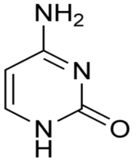
🟦 Normal Tissue

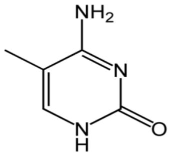
🟧 Clear Cell

🟩 Other Subtypes

⬜ Normal Tissue

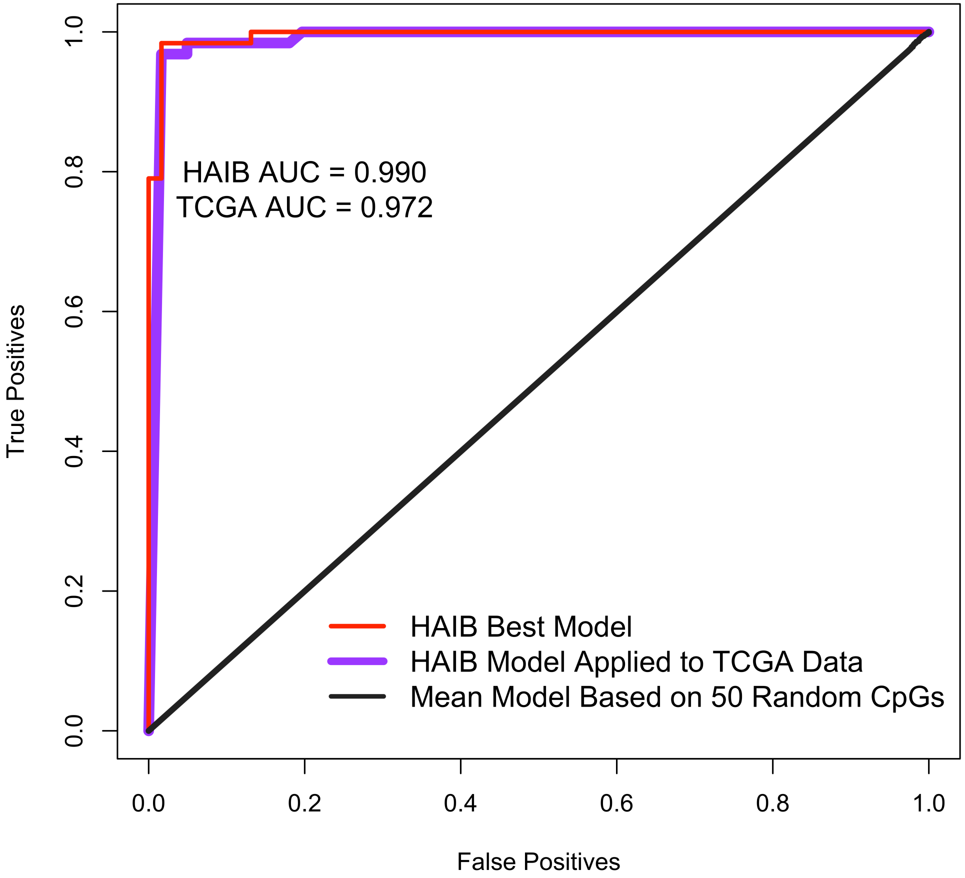



Cytosine

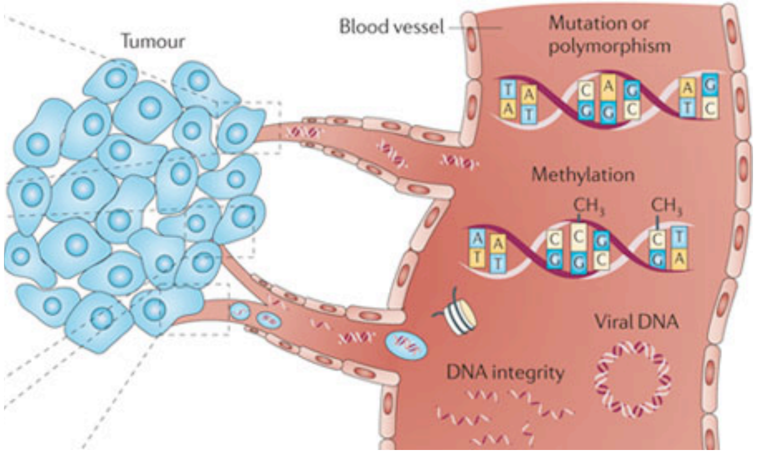

5-Methyl Cytosine

DNA methylation patterns are highly accurate at predicting patients with renal cell carcinoma

ROC curves of DNA methylation results from 135 tumor and matched non-tumor samples from RCC patients

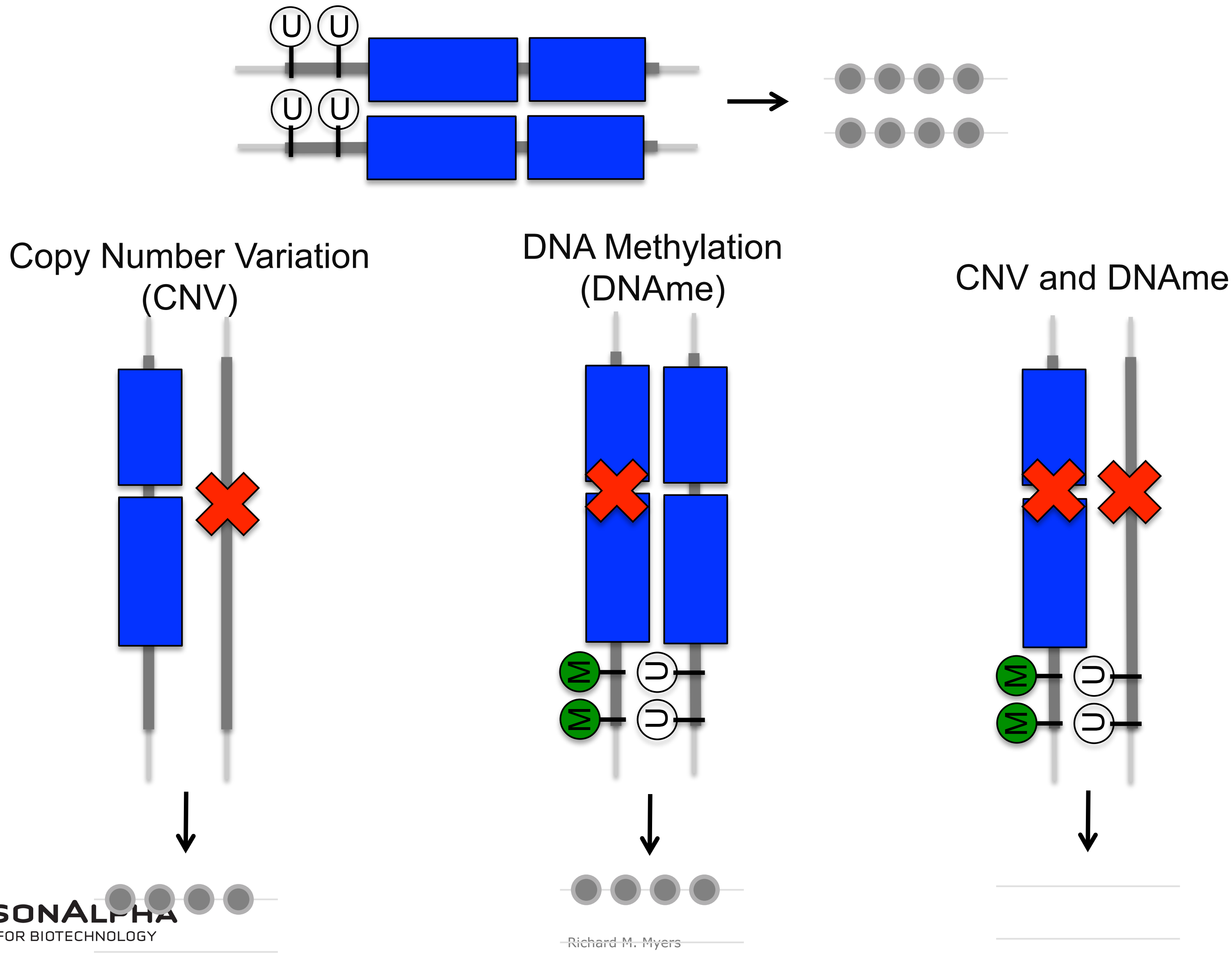


Apply these assays to urine or blood as a routine screening for early detection of kidney cancer



Schwarzenbach et al. Nature Reviews Cancer 11, 426-437

Integrating genomic signatures

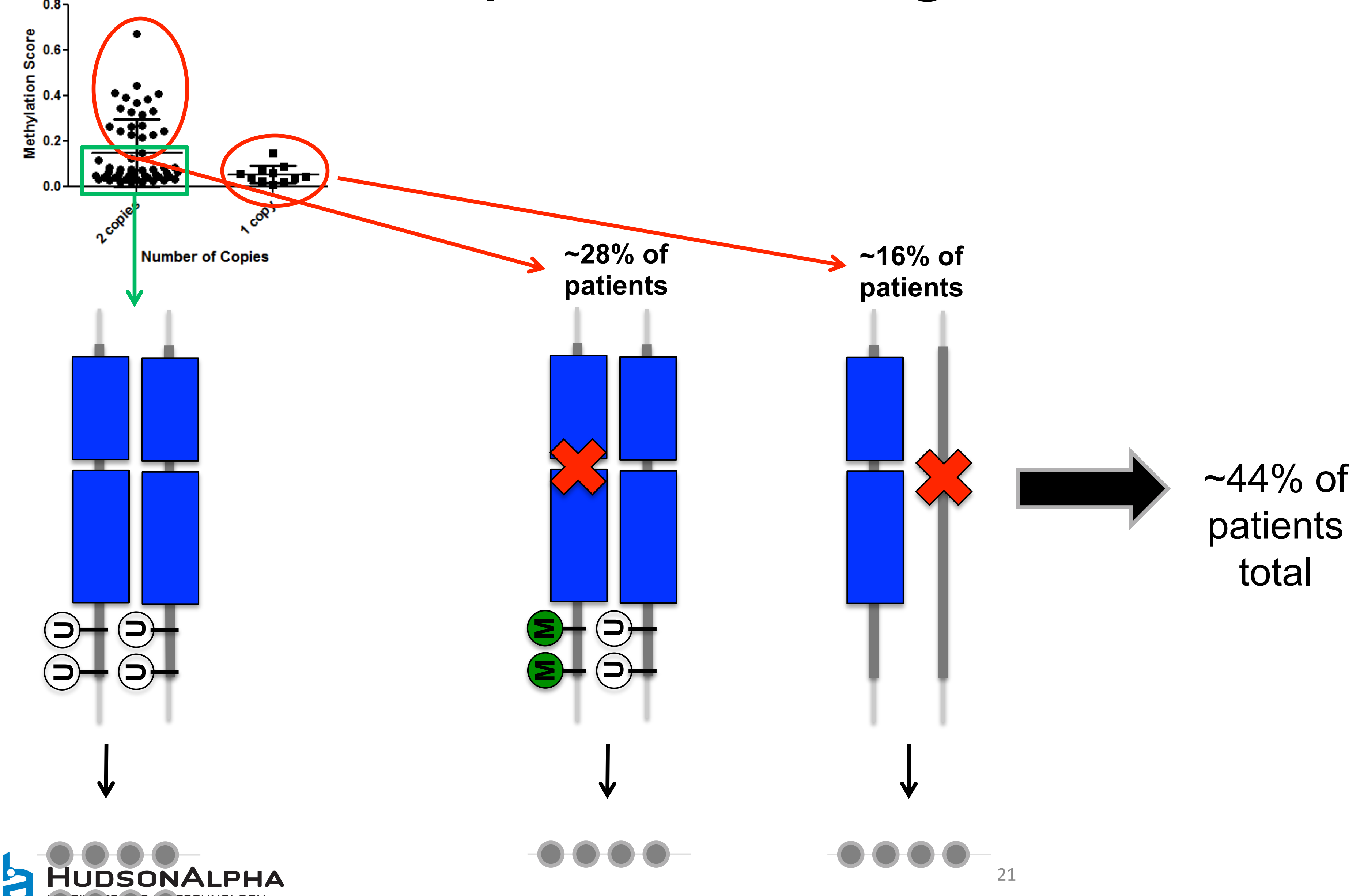


Copy Number Variation (CNV)

DNA Methylation (DNAme)

CNV and DNAme

Example: MSH4 gene



~28% of patients

~16% of patients

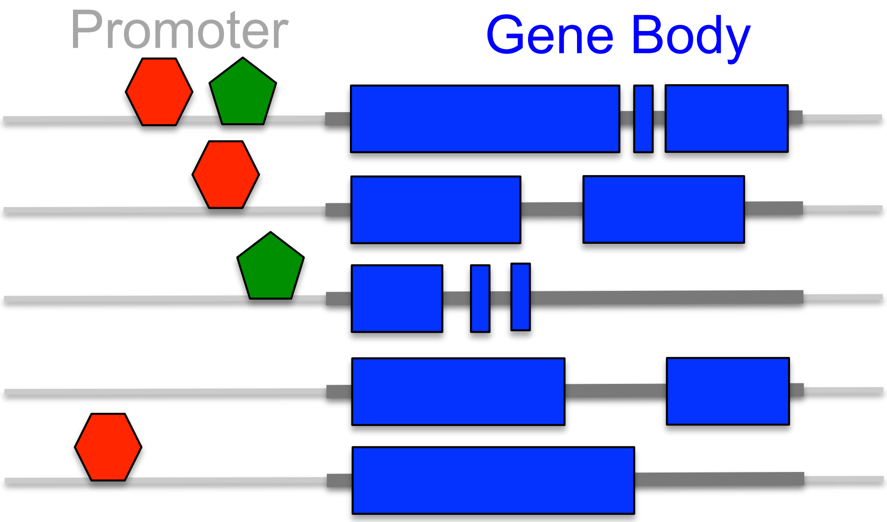
~44% of patients total

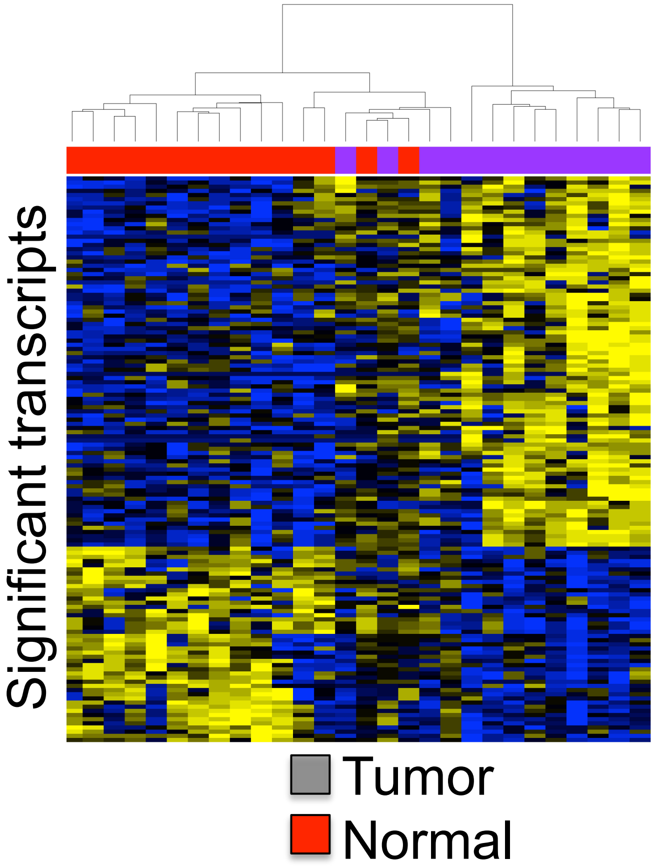
3. Using ENCODE TF data to prioritize cancer genetics and functional genomics data

Using ENCODE data to find cancer regulators

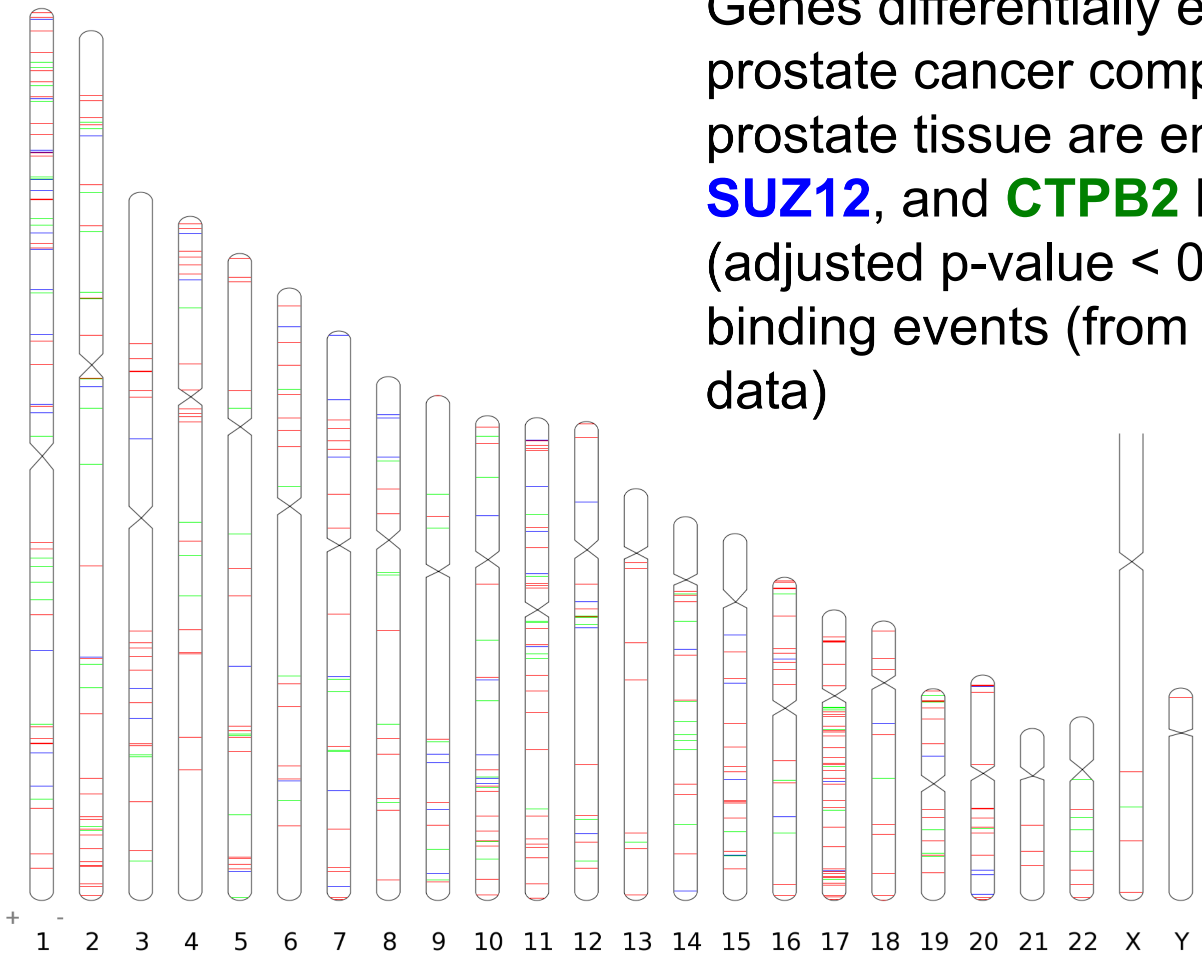
Genomic assays often reveal thousands of dysregulation events in cancer

These widespread genomic changes may be regulated by a few key transcription factors





Differentially expressed genes in cancer are enriched for particular TFs

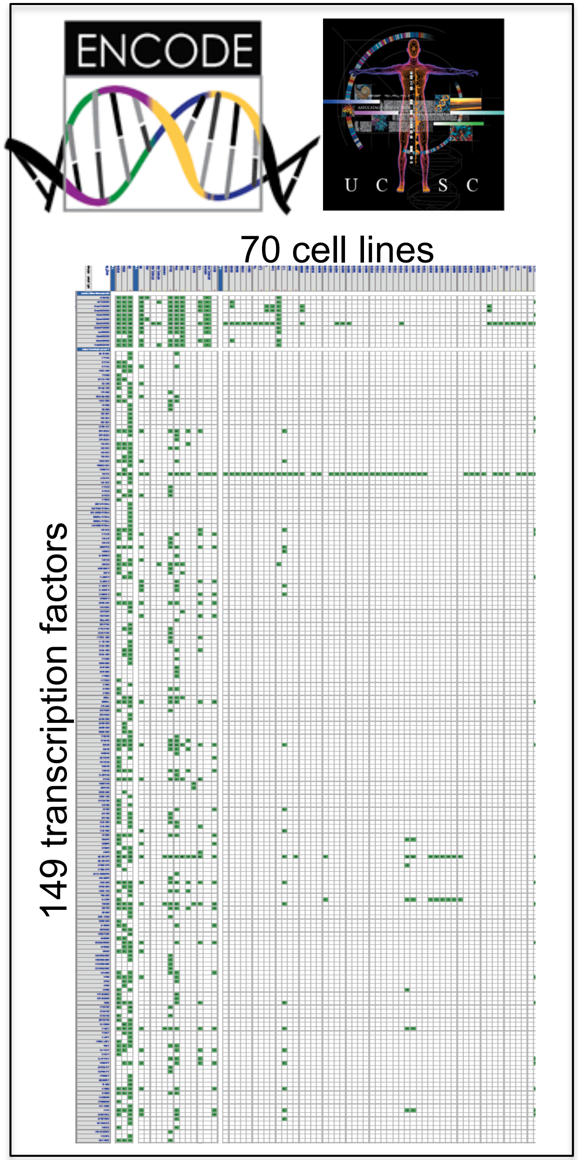


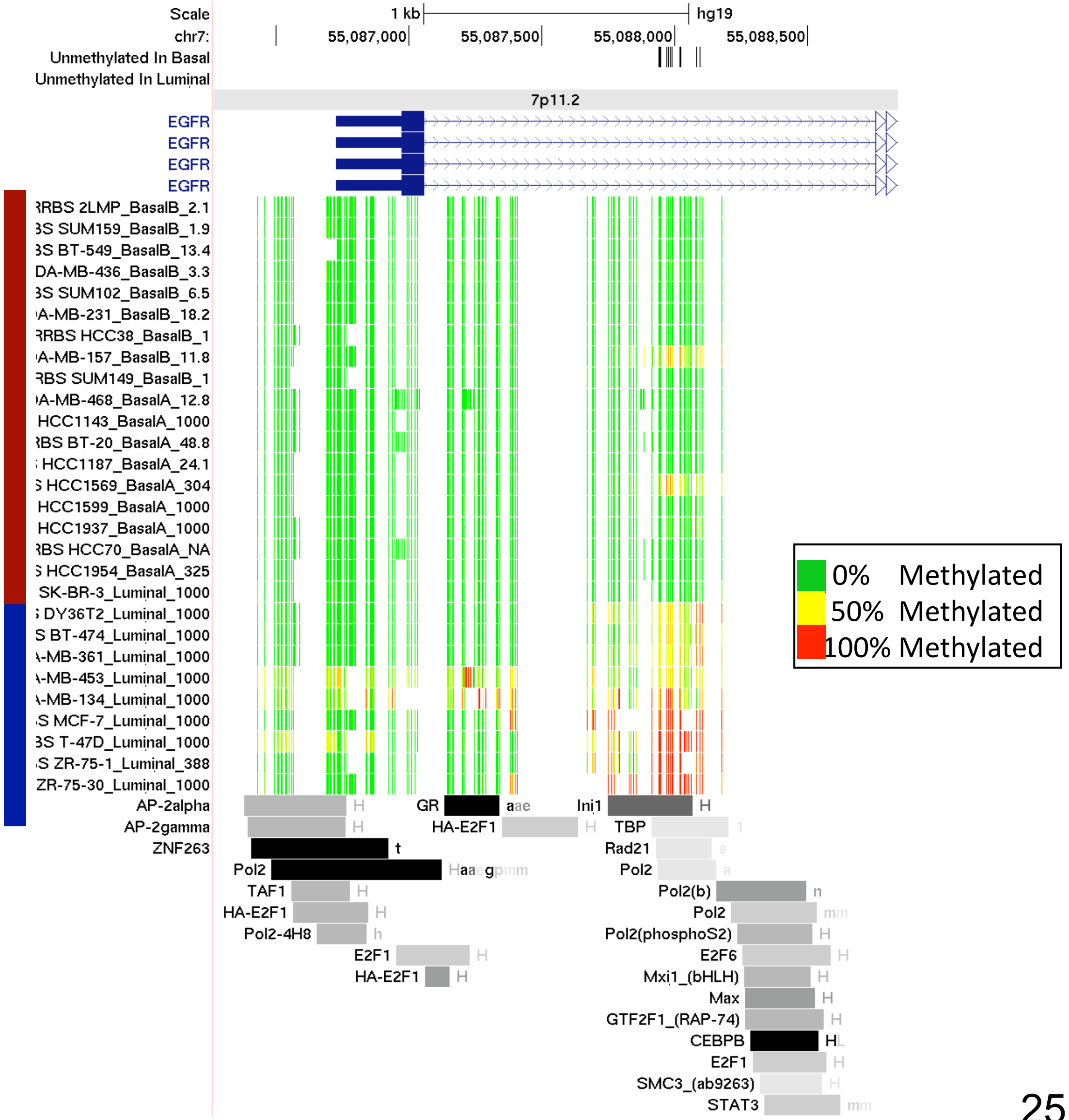
Genes differentially expressed in prostate cancer compared to normal prostate tissue are enriched for **EZH2**, **SUZ12**, and **CTPB2** binding sites (adjusted p-value < 0.05) and actual binding events (from ENCODE ChIP data)

HudsonAlpha INSTITUTE FOR BIOTECHNOLOGY

Richard M. Myers

Intersect transcription factor binding sites from the ENCODE Project with genomic regions specifically unmethylated in basal breast cancer



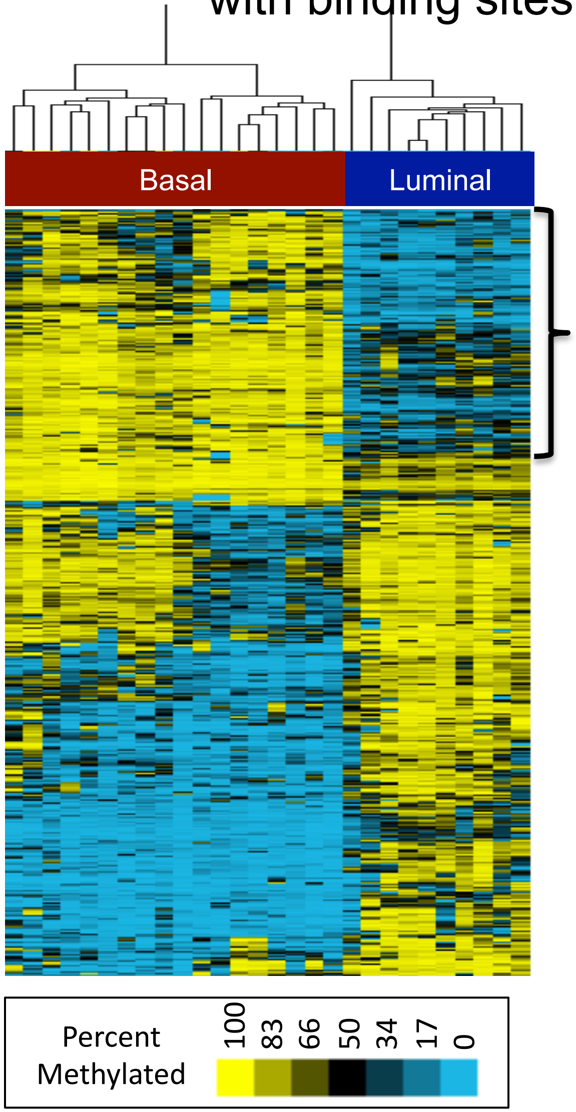


🟩 0% Methylated
🟨 50% Methylated
🟥 100% Methylated

Richard M. Myers

Master regulators (?) of different breast cancer subtypes

Intersect gene regulatory regions containing subtype-associated methylation with binding sites of 149 transcription factors in ENCODE datasets



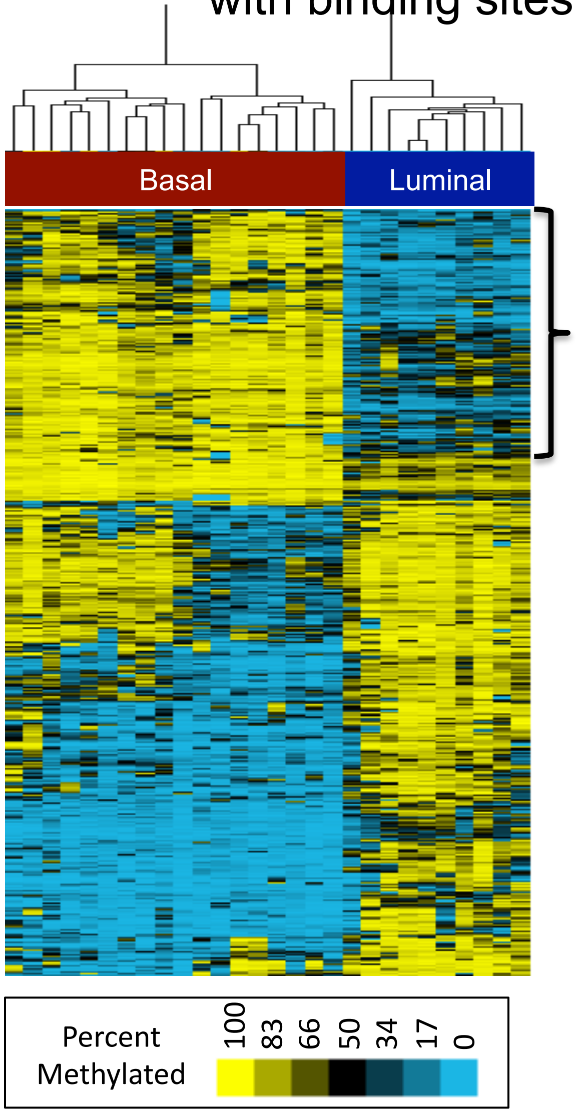
Significantly enriched binding sites:

Transcription Factor	Fold Enrichment
Estrogen Receptor	6.9
FOXA1	8.1
GATA3	10.3

Richard M. Myers

Master regulators (?) of different breast cancer subtypes

Intersect gene regulatory regions containing subtype-associated methylation with binding sites of 149 transcription factors in ENCODE datasets



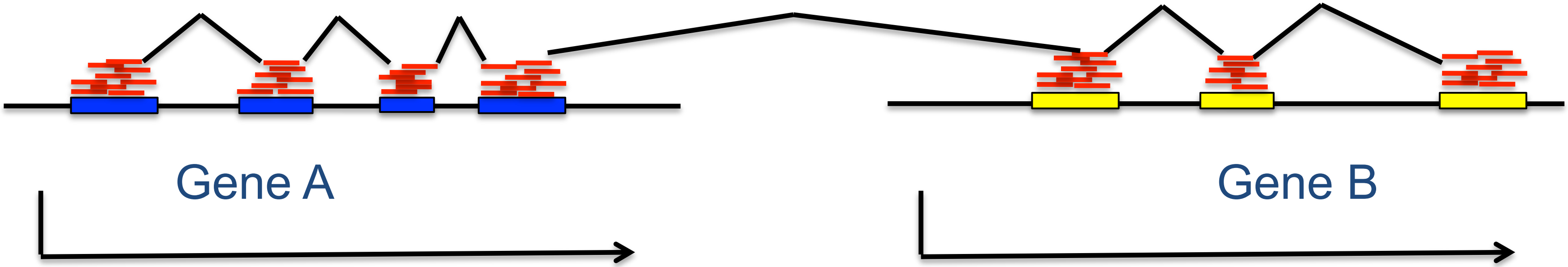
Significantly enriched binding sites:

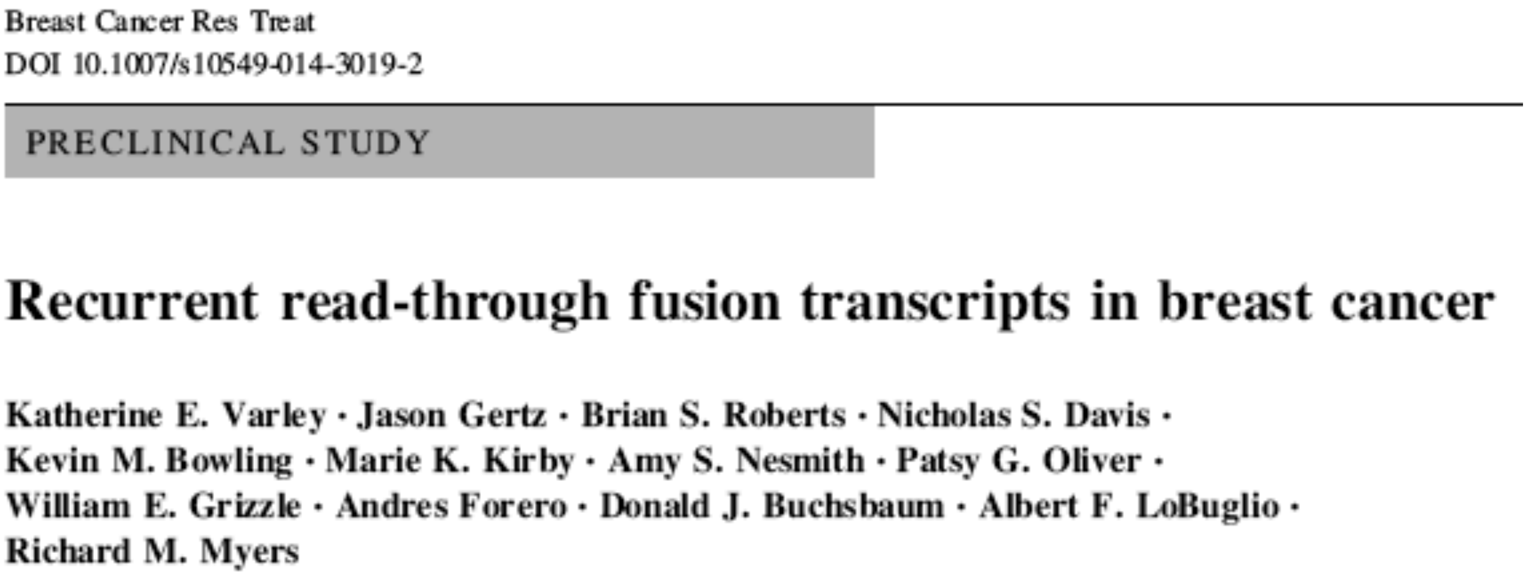
Transcription Factor	Fold Enrichment
Estrogen Receptor	6.9
FOXA1	8.1
GATA3	10.3

Transcription Factor	Fold Enrichment
STAT3	4.8
GR (glucocorticoid receptor)	4.2

4. Using RNA-seq to identify drug targets

Transcript fusions in cancer





Breast Cancer Res Treat
DOI 10.1007/s10549-014-3019-2

PRECLINICAL STUDY

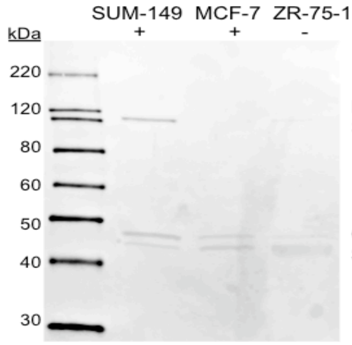
Recurrent read-through fusion transcripts in breast cancer

Katherine E. Varley · Jason Gertz · Brian S. Roberts · Nicholas S. Davis · Kevin M. Bowling · Marie K. Kirby · Amy S. Nesmith · Patsy G. Oliver · William E. Grizzle · Andres Forero · Donald J. Buchsbaum · Albert F. LoBuglio · Richard M. Myers

K-T Varley with collaborators at UAB Comprehensive Cancer Center

3 fusion transcripts produce fusion proteins located in the cell membrane

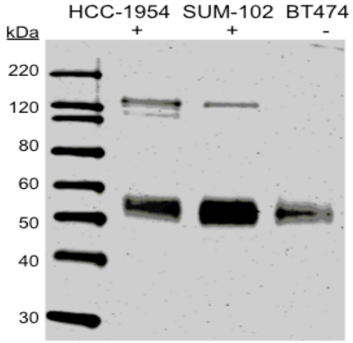
IL17RC-CRELD1



Predicted Fusion
Size = 102 kDa

CRELD1 Canonical
Size = 46 kDa

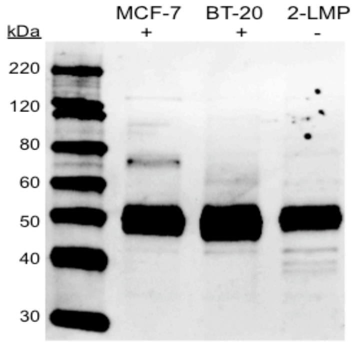
SCNN1A-TNFRSF1A



Predicted Fusion
Size= 112 kDa

TNFRSF1A Canonical
Size = 51 kDa

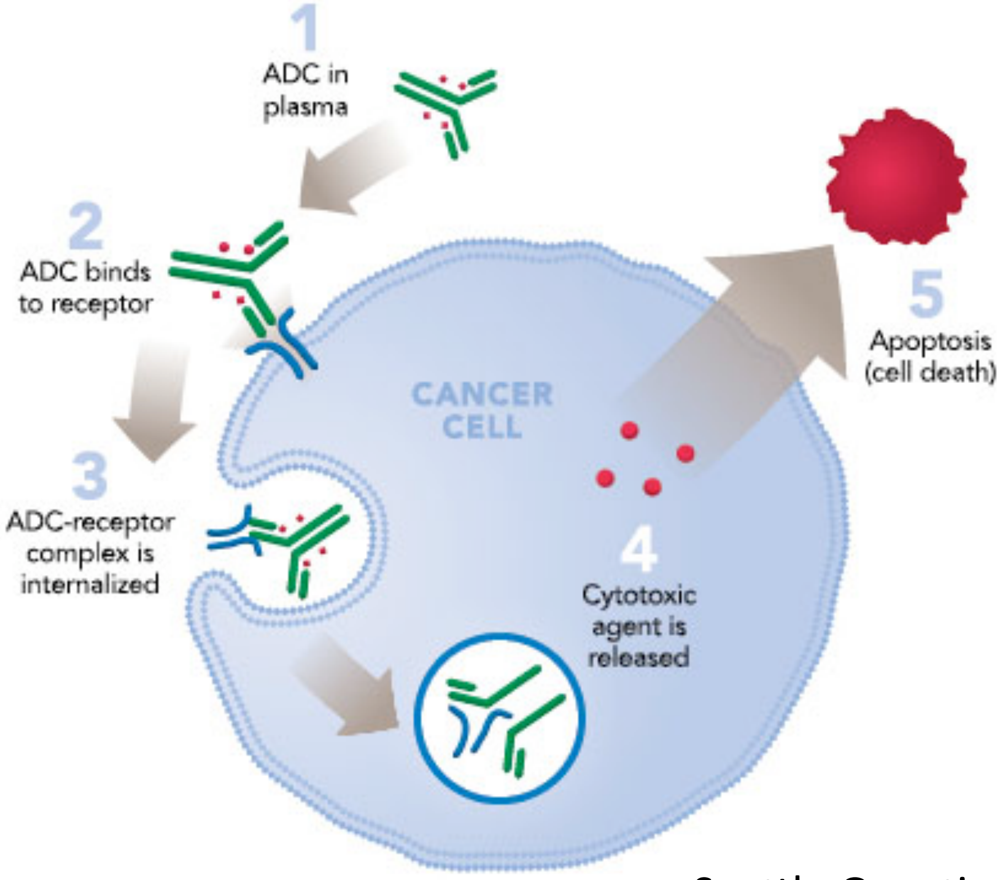
CTSD-IFITM10



Predicted Fusion
Size = 60 kDa

CTSD Canonical
Size = 45 kDa

Potential therapeutic: Use drug-antibody complexes to direct a cellular toxin exclusively to cancer cells



Seattle Genetics

5. Which TF binding events are functionally important?

Using expression assays to identify functional transcription elements

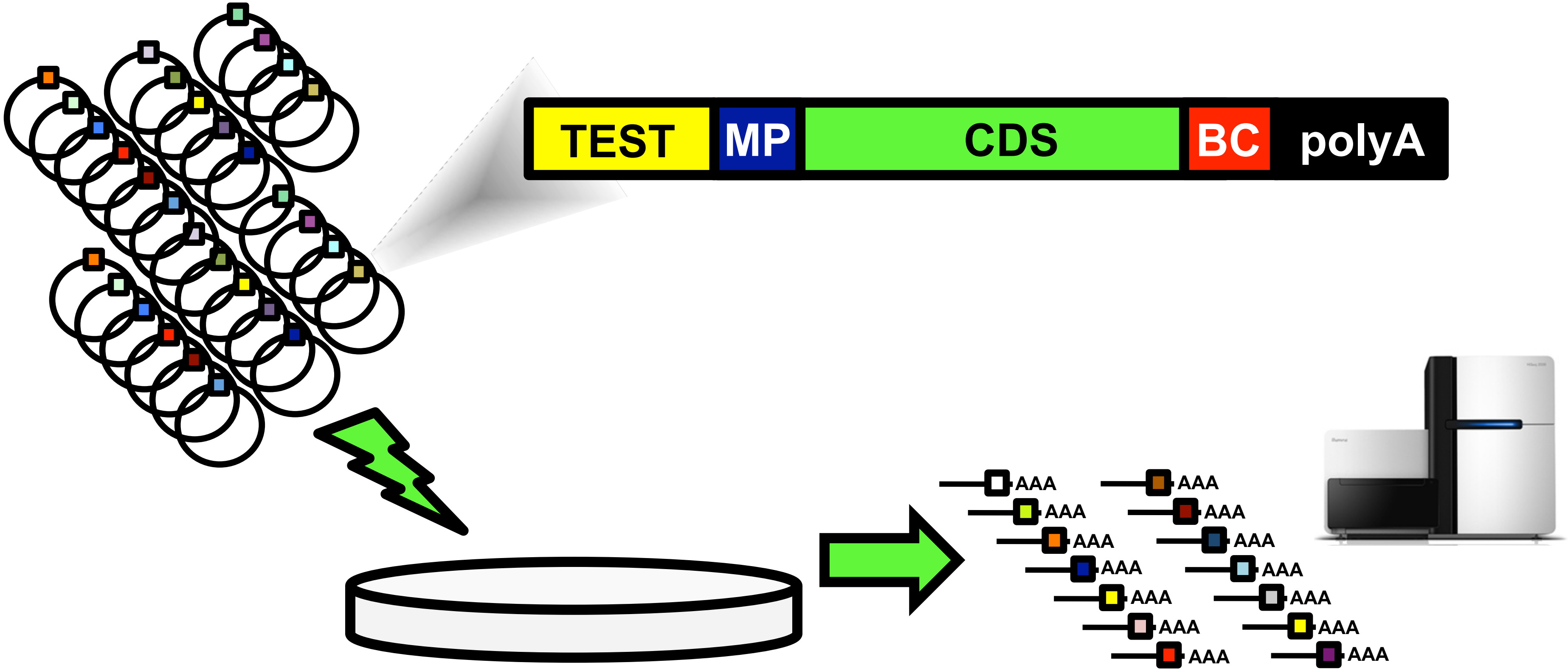
(especially long-distance ones)

Dan Savic, Brian Roberts, Chris Partridge, Barak Cohen, Greg CooperJay Gertz, Rick Myers

Test thousands of ENCODE-identified putative elements (based on TF binding, chromatin marks, etc.) in an ultra-high throughput reporter assay

Massively parallel reporter assay

Cis-Regulatory Element sequencing (CRE-seq)

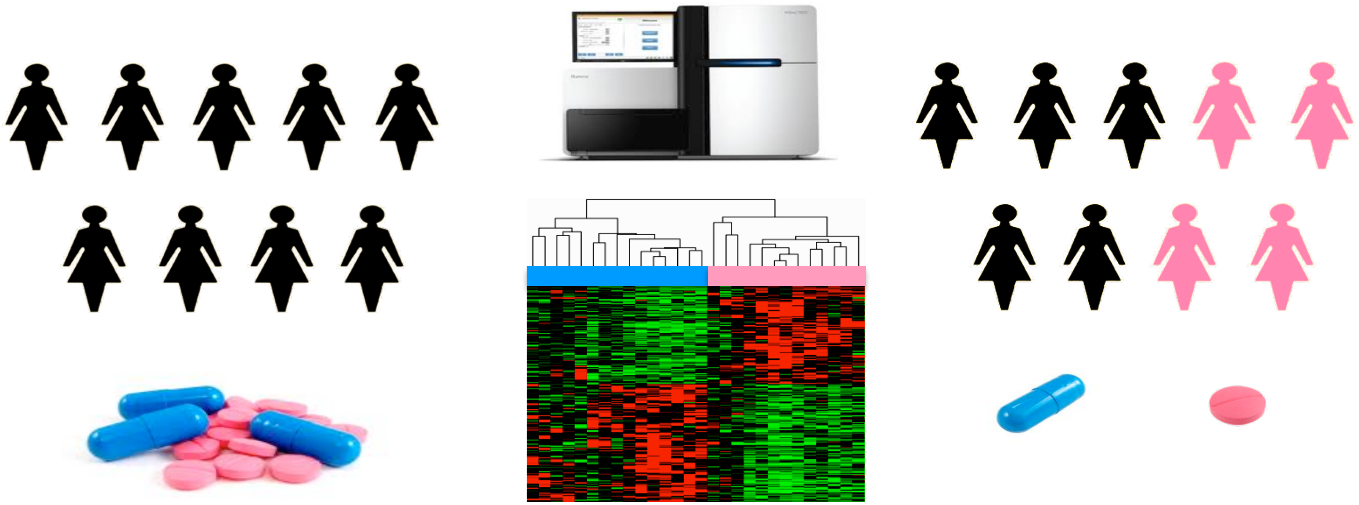


Barcode abundance (sequence count)
is a proxy for test sequence activity

Findings

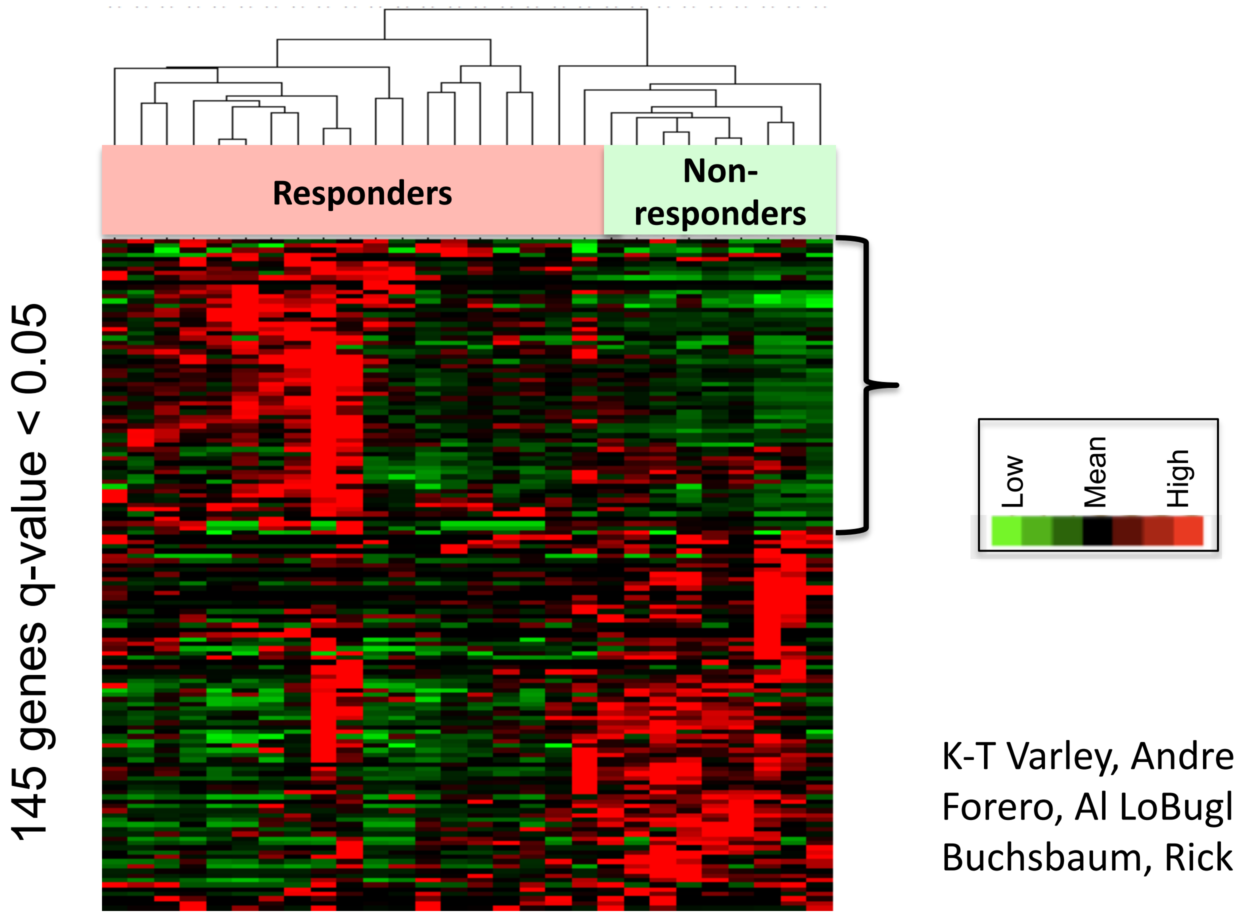
RNAP2 at promoter-distal TF sites is a very strong mark of active regulatory elements

3. Using genomics to predict which patients will respond to various treatments



Clinical trial of a novel combination of drugs in ER+ breast cancer

Gene expression patterns in **responders** and **non-responders** during clinical trial of Letrozole (anti-estrogen) and Avastin (anti-angiogenesis)



K-T Varley, Andres Forero, Al LoBuglio, Don Buchsbaum, Rick Myers