Functional Genomics @ Scale

A long-term goal of functional genomics is to decipher the rules by which genomes, genes and gene networks are regulated and to understand how such regulation affects cellular function, development and disease.

Functional Genomics @ Scale

- What are the big challenges that can be solved and needs to be met relative to functional role of genomic variants in health and disease?
- What should be the role of NHGRI vs. other funders?
- What are the consequences if NHGRI decides not to pursue this area?

Functional Genomics @ Scale

- No existing _sequencing_ programs are directly pursuing functional genomics at scale*
- *The only scaled effort towards interpreting function going on in the large-scale sequencing program is computational.
- Example: associated variants found in a common disease phenotype can be linked to pathway (e.g. voltage-gated calcium channel genes and schizophrenia).This is “scaled” in the sense that only a large number of samples allows the power to attempt the clustering.
- *One can also argue that the Centers for Mendelian Genomics are doing scaled studies on function. Although the individual "solved" Mendelian disease genes are each an achievement, it is the collection of them (including allelic series/expansions) that is functionally informative about human biology.

Functional Genomics @ Scale

Existing NHGRI large functional genomics programs that have a connection to use in interpreting variants include:

ENCyclopedia Of DNA Elements (ENCODE)

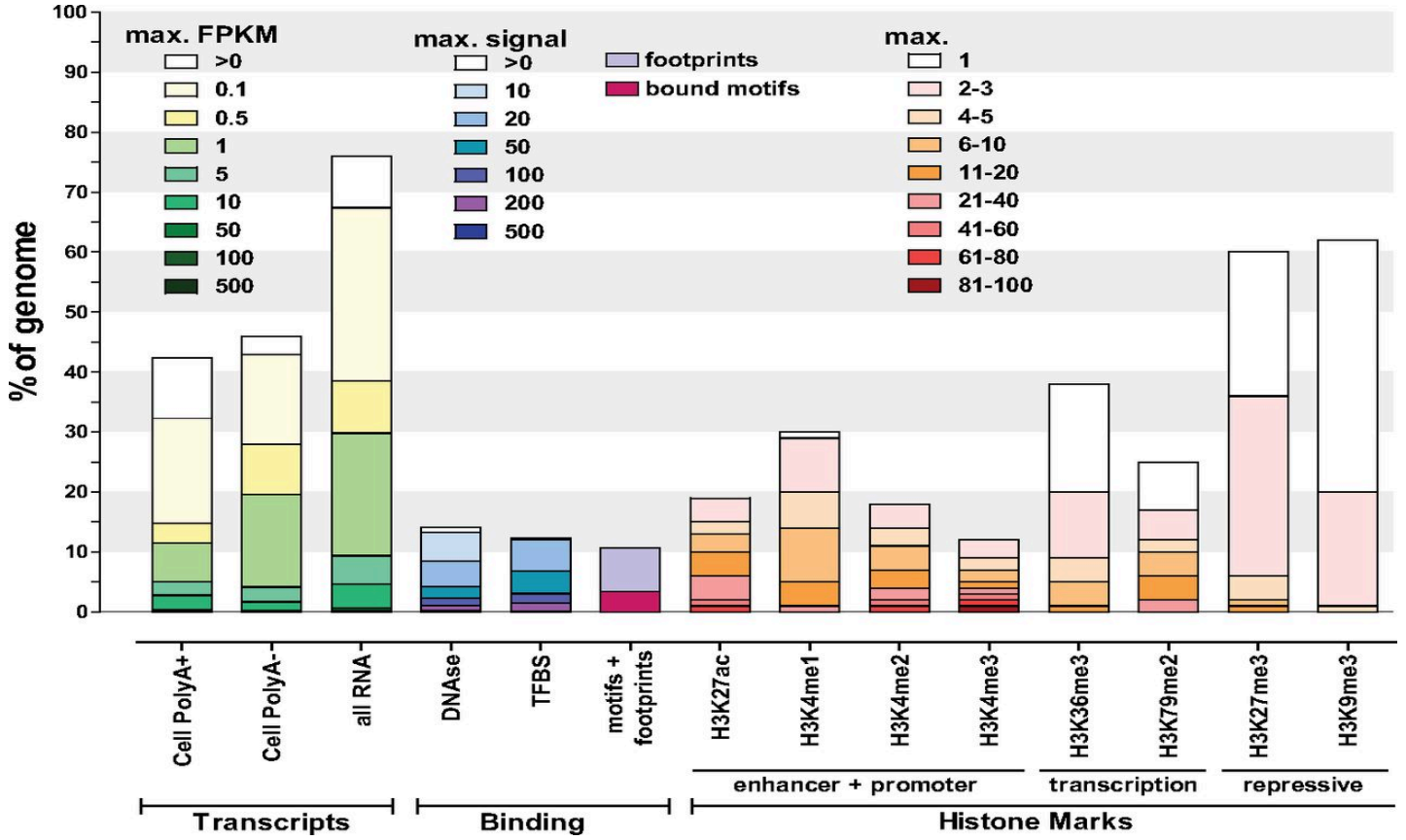
Genomics of Gene Regulation (GGR)

Functional Variants (FunVar)

ENCODE

The long-term goal of the ENCyclopedia of DNA Elements (ENCODE) and ModENCODE Projects is to generate comprehensive catalogs of all functional elements in the human genome and genomes of selected model organisms

Summary of the coverage of the human genome by ENCODE data.



Kellis M et al. PNAS 2014;111:6131-6138

Genomics of Gene Regulation
(not yet funded)

- Aims to explore genomic approaches to understanding the role of genomic sequence in the regulation of gene networks.

- Aims to address the genome-proximal component of the regulation of gene networks by developing and validating models that describe how a comprehensive set of sequence-based functional elements work in concert to regulate the finite set of genes that determine a biological phenomenon, using RNA amounts, and perhaps transcript structure, as the readout.

- Aims to substantially improve the methods for developing gene regulatory network models, rather than an incremental improvement on existing methods.

- Long-term goal- to read DNA sequence and accurately predict when and at what levels a gene is expressed, in the context of a particular cell state.

Functional Variants
(not yet funded)

- FunVar aims to develop highly innovative computational approaches for interpreting sequence variants in the non-protein-coding regions of the human genome.

- Will analyze whole-genome sequence data by integrating data sets, such as ones on genome function, phenotypes, patterns of variation, and other features, to identify or substantially narrow the set of variants that are candidates for affecting organismal function leading to disease risk or other traits.

- The accuracy of the computational approaches developed will be assessed using experimental data.

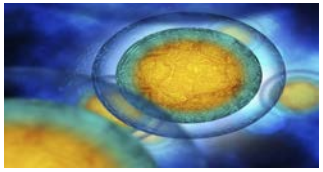
Common Fund Resources for Interpretation of Variants

Epigenomics Project

Genotype-Tissue Expression (GTEx) Project

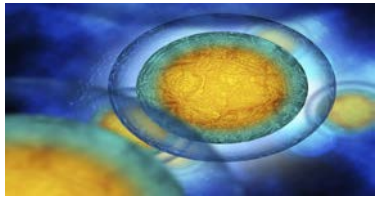
Library of Integrated Network-Based Cellular Signatures

4D Nucleome



*These are Common Fund efforts with significant NHGRI involvement

4D Nucleome



- The 4D Nucleome program will develop technologies to enable the study of how DNA is arranged within cells in space and time (the fourth dimension) and how this affects cellular function in health and disease.

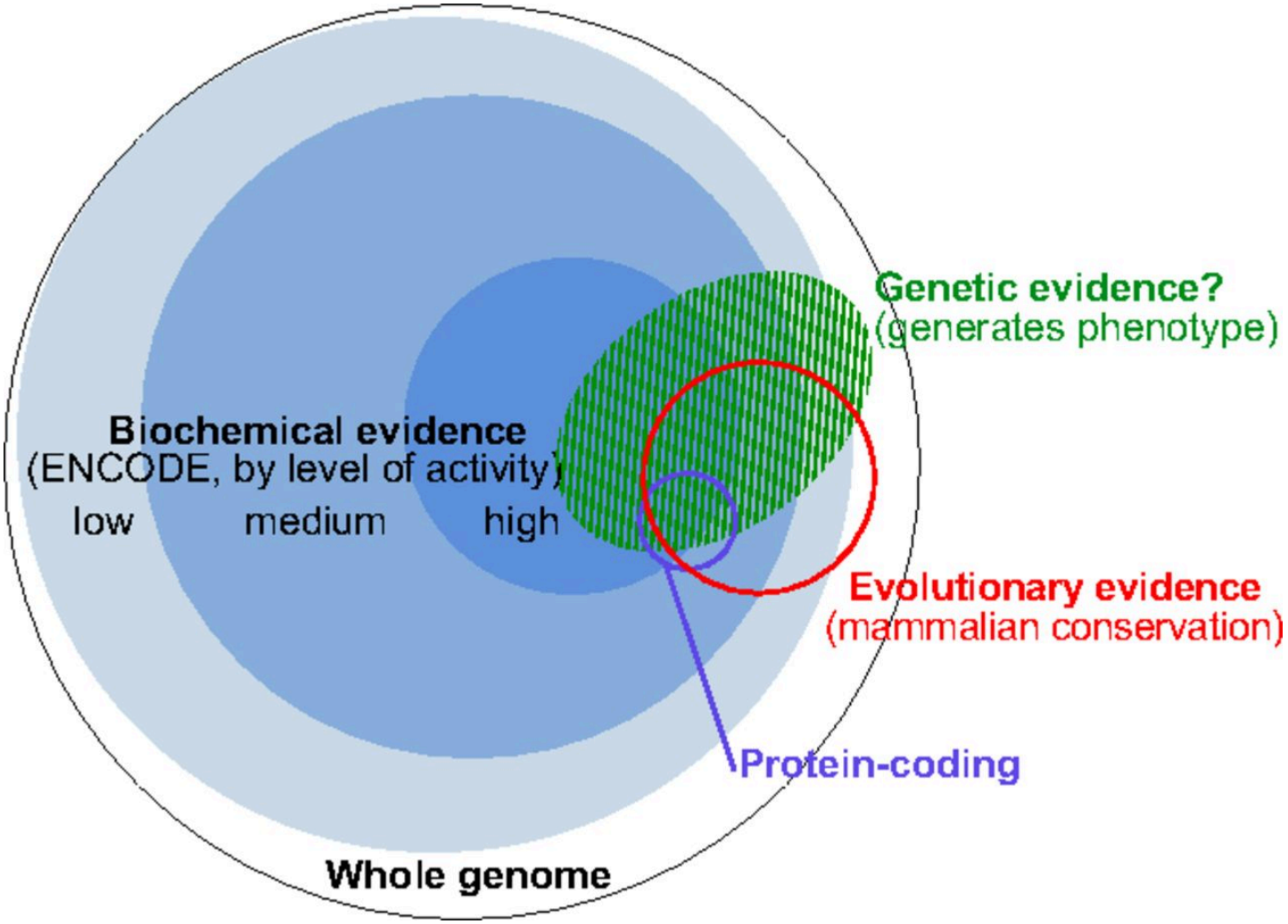
- 4D nucleome science aims to understand the principles behind the organization of the nucleus in space and time, the role that the arrangement of DNA plays in gene expression and cellular function, and how changes in nuclear organization affect health and disease.

Functional Genomics @ Scale

- Resources for Interpretation of variants
- Functional validation of variants

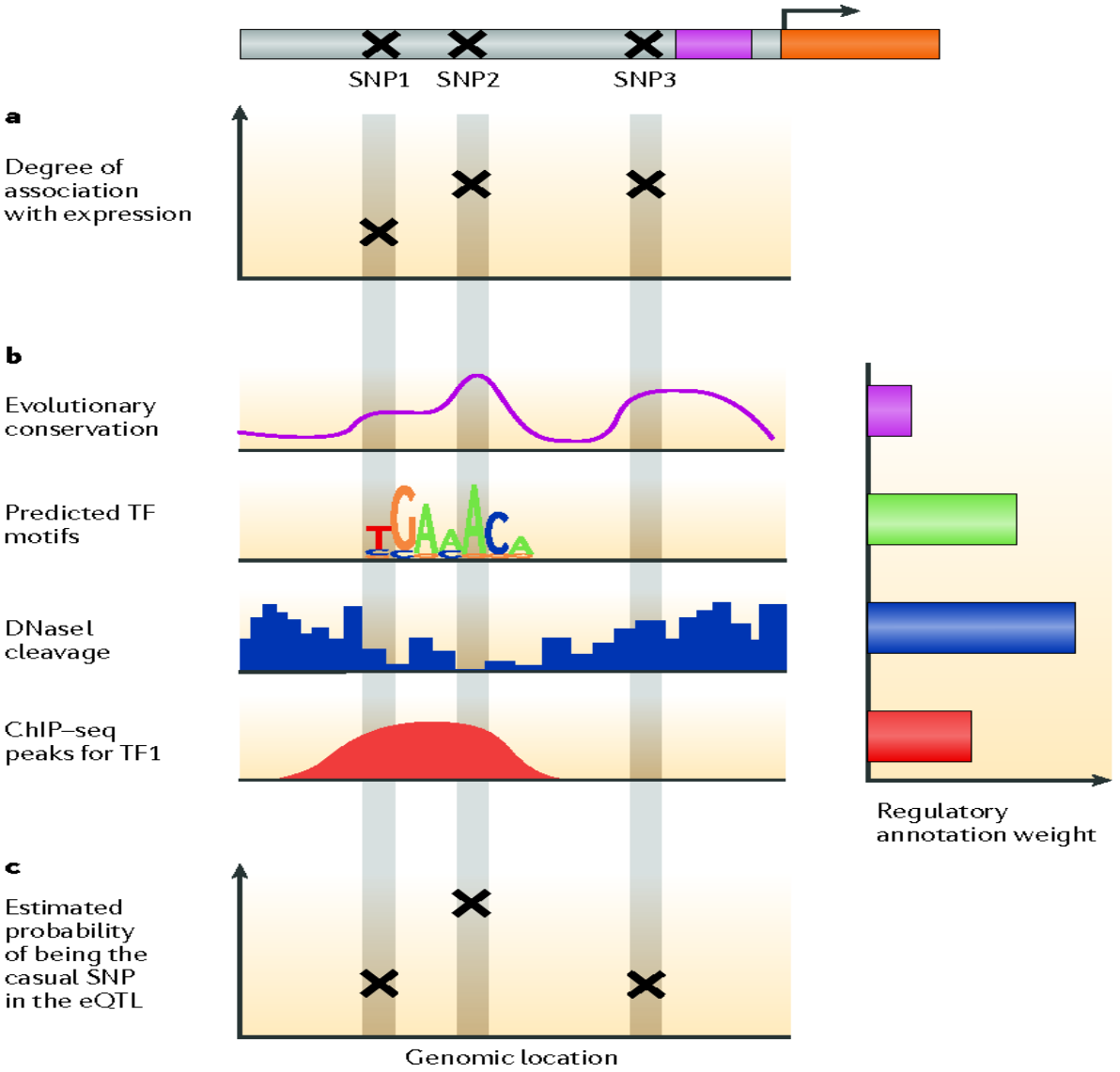
Incorporating conservation and regulatory annotations to prioritize SNVs

The complementary nature of evolutionary, biochemical, and genetic evidence.



Kellis M et al. PNAS 2014;111:6131-6138

Incorporating conservation and regulatory annotations to prioritize SNVs



Levo and Segal (2014) NATURE REVIEWS | **GENETICS**

Enhancers can act over a long range, making it challenging to define their targets

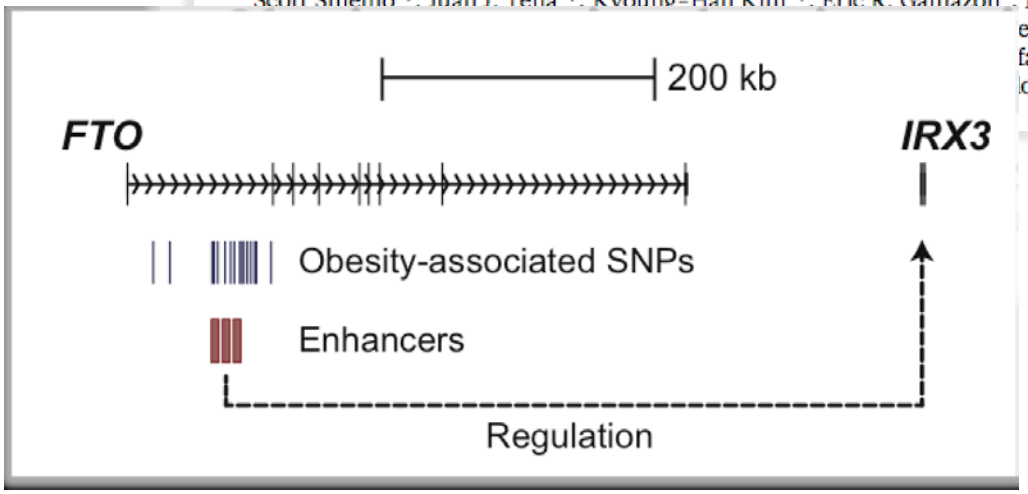
Nature, 2014

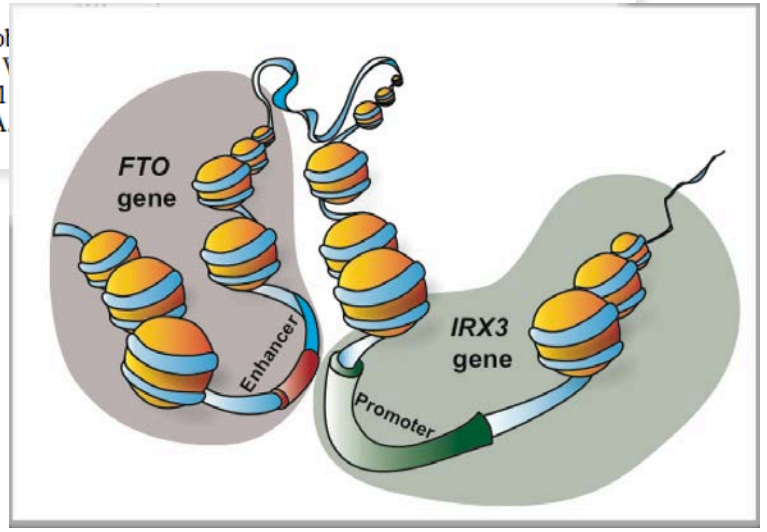
LETTER

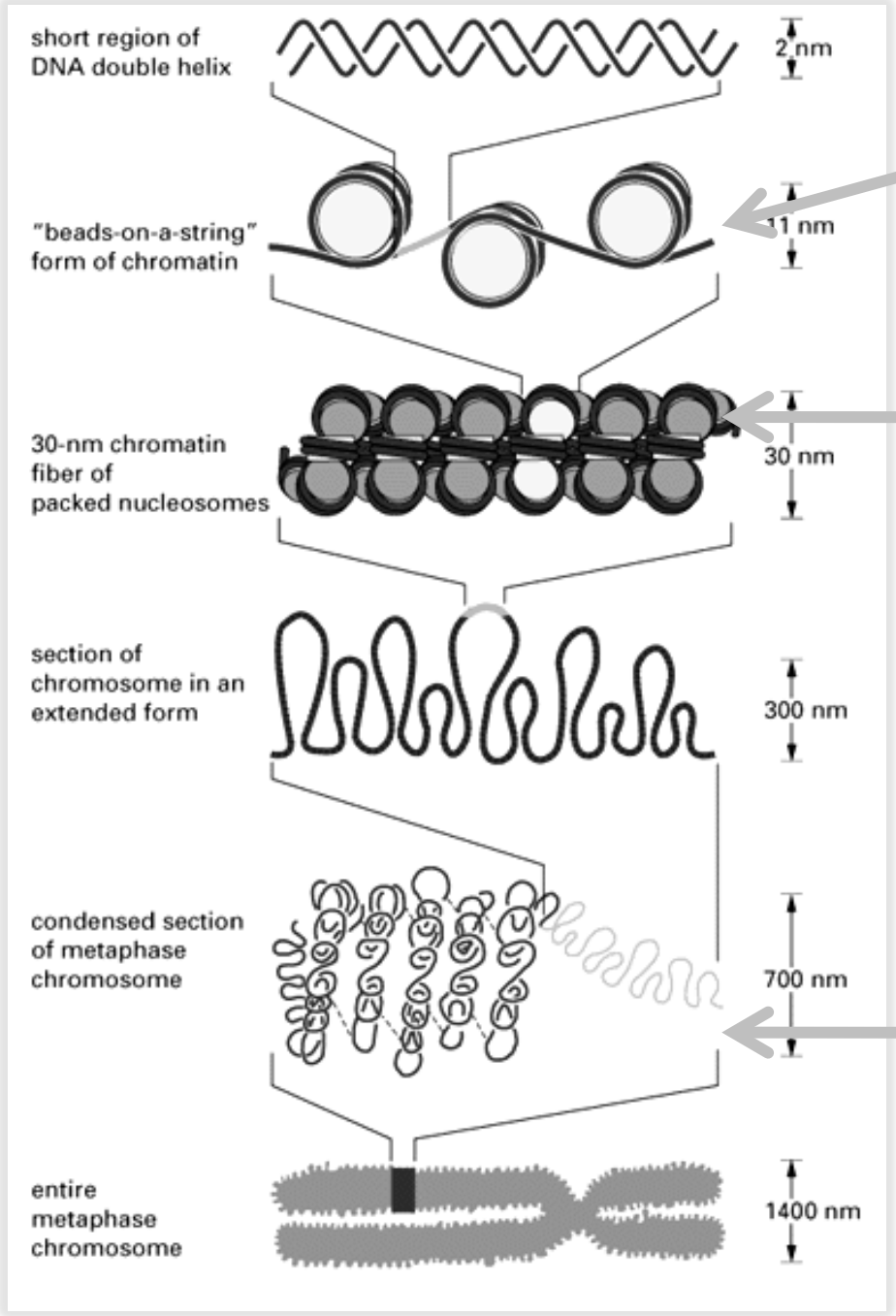
doi:10.1038/nature13138

Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*

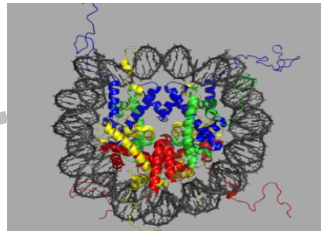
Scott Smemo[1]*, Juan J. Tena[2]*, Kyoung-Han Kim[3]*, Eric R. Gamazon[4], Nol
e[3], V
fael
lo A



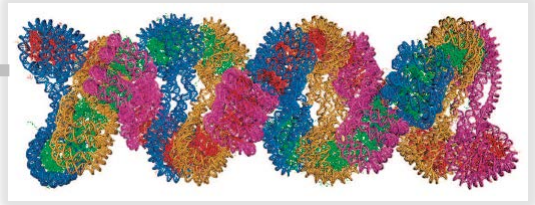




Felsenfeld & Groudine, Nature 2003

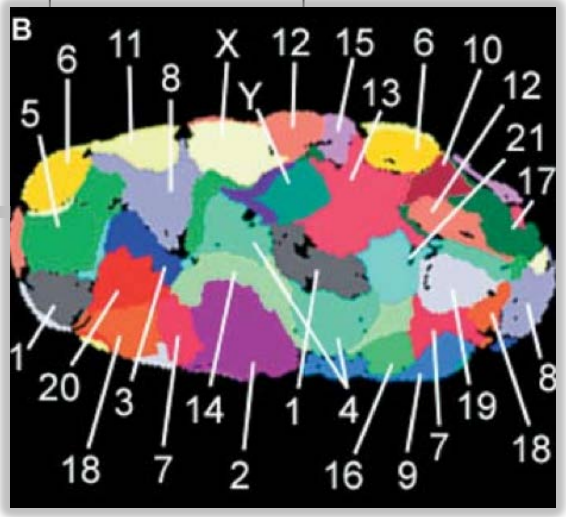


H2A H2B H3 and H4



Song et al. Science 2014

?

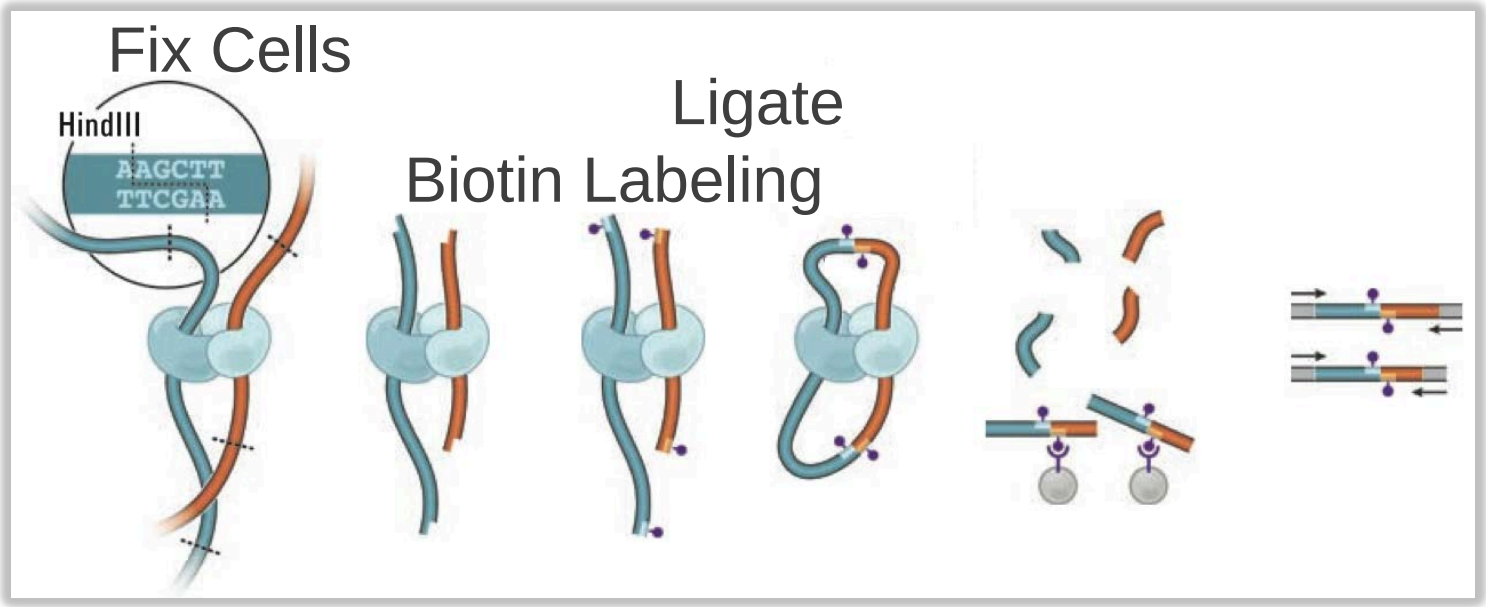


Bolzer et al., PLoS Biol. 2005

Opportunity to explore long-range chromatin interactions and regulation

- Genome-wide survey of long-range chromatin interactions in mammalian cells
- General features of chromatin organization and dynamics
- Local chromatin interactions reveal enhancer/promoter interactions
- Functional analysis of long-range regulatory elements

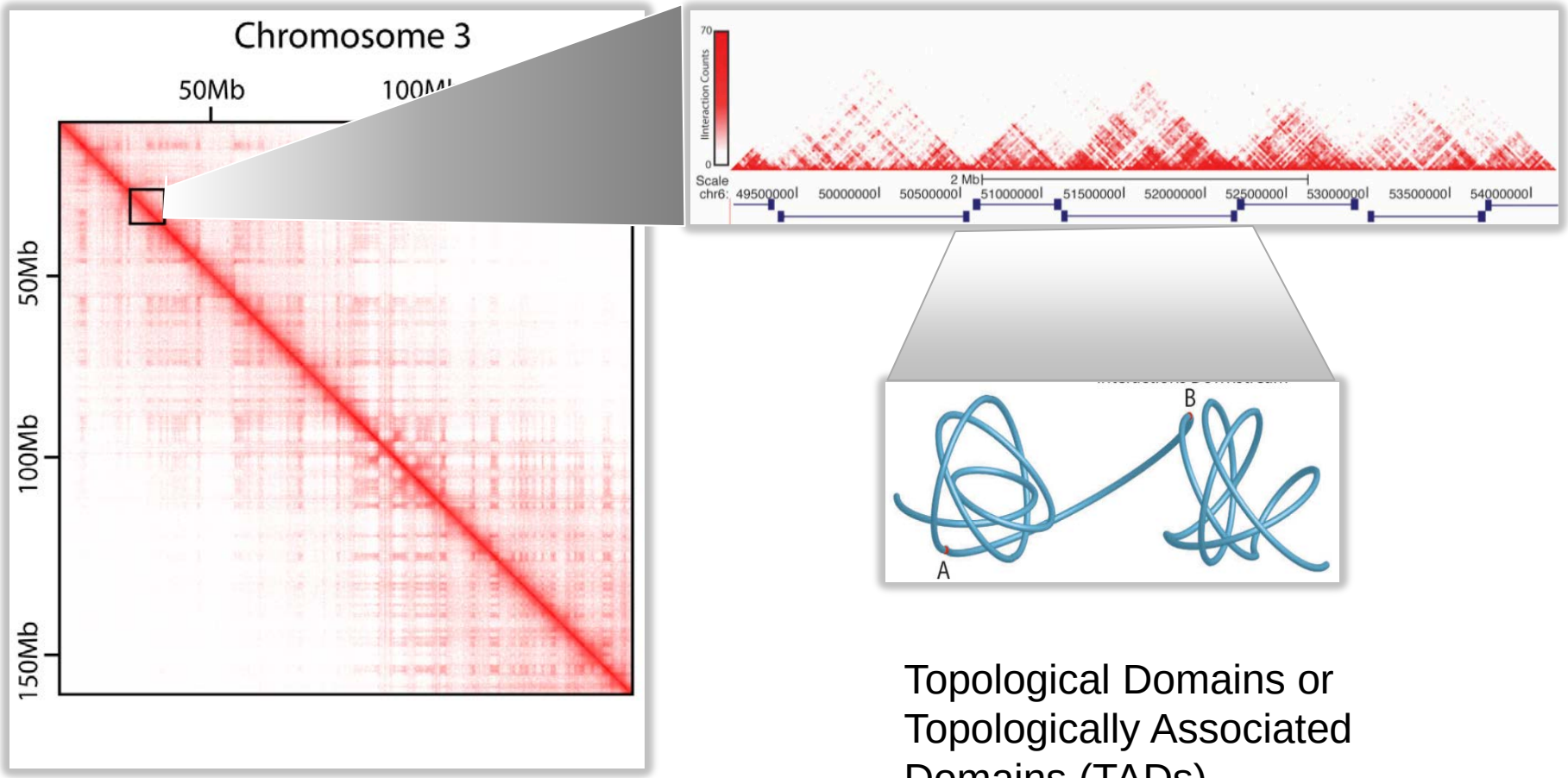
Hi-C: a method for genome-wide analysis of higher order chromatin structure



Cross Linking → Proximity Ligation → Sequencing

Lieberman-Aiden et al., Science 2009

Genome-wide analysis of higher order chromatin structure in human and mouse cells



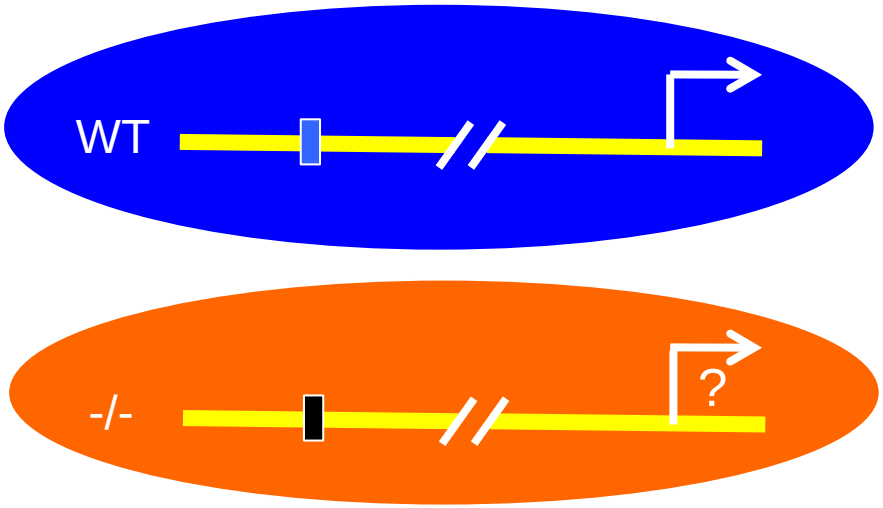
Higher Hi-C frequency = shorter spatial distance

Lower Hi-C frequency = longer spatial distance

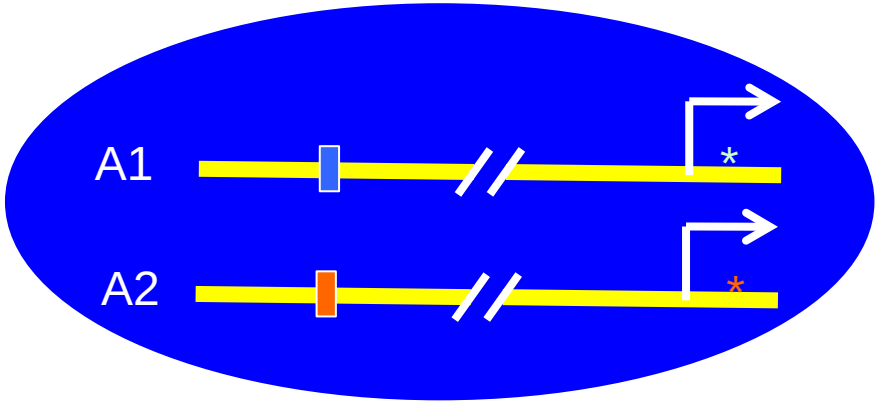
Topological Domains or Topologically Associated Domains (TADs)

Strategies for functional study of enhancers

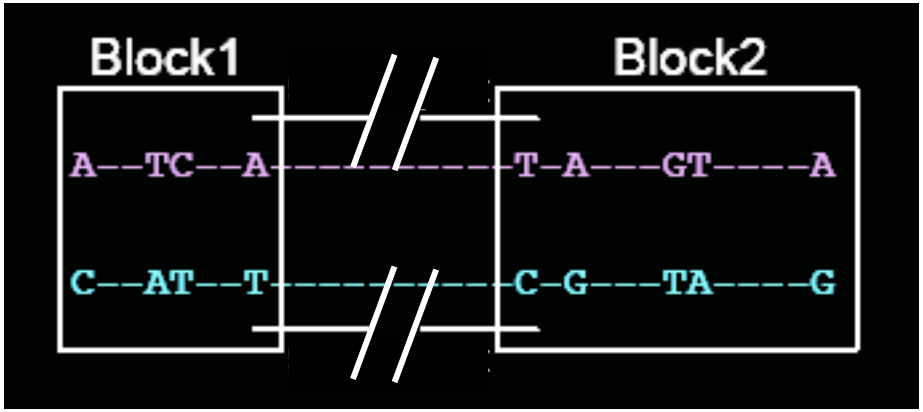
- Introduce mutations into each enhancer in their endogenous locus and test for changes in gene expression
  - Pros: most direct
  - Cons: low throughput; may not applicable to humans



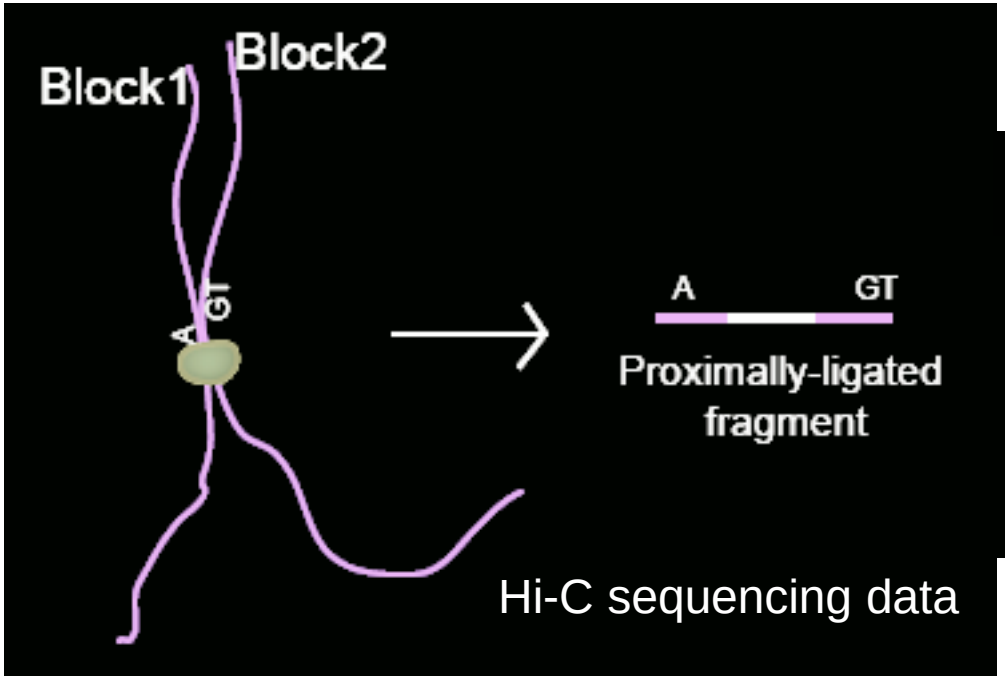
- Exploit the naturally occurring sequence variants (SNPs) between the two copies of DNA in each cell
  - Pros: global and genome-wide
  - Cons: need to know the haplotypes

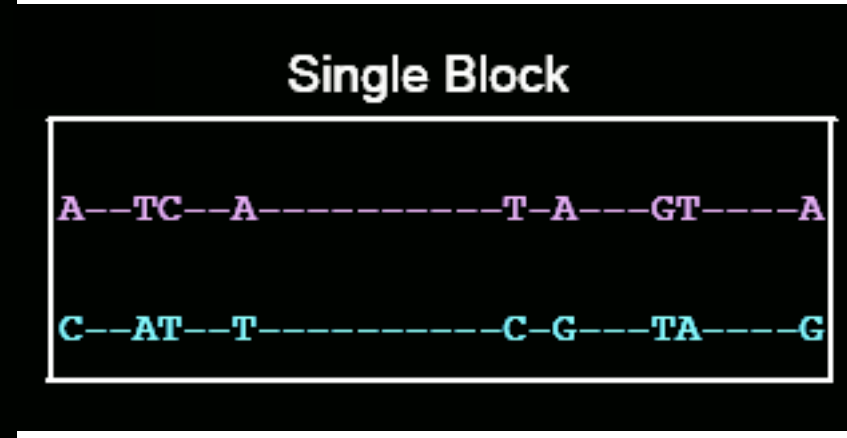


Hi-C data can inform on haplotypes- Haplo-seq



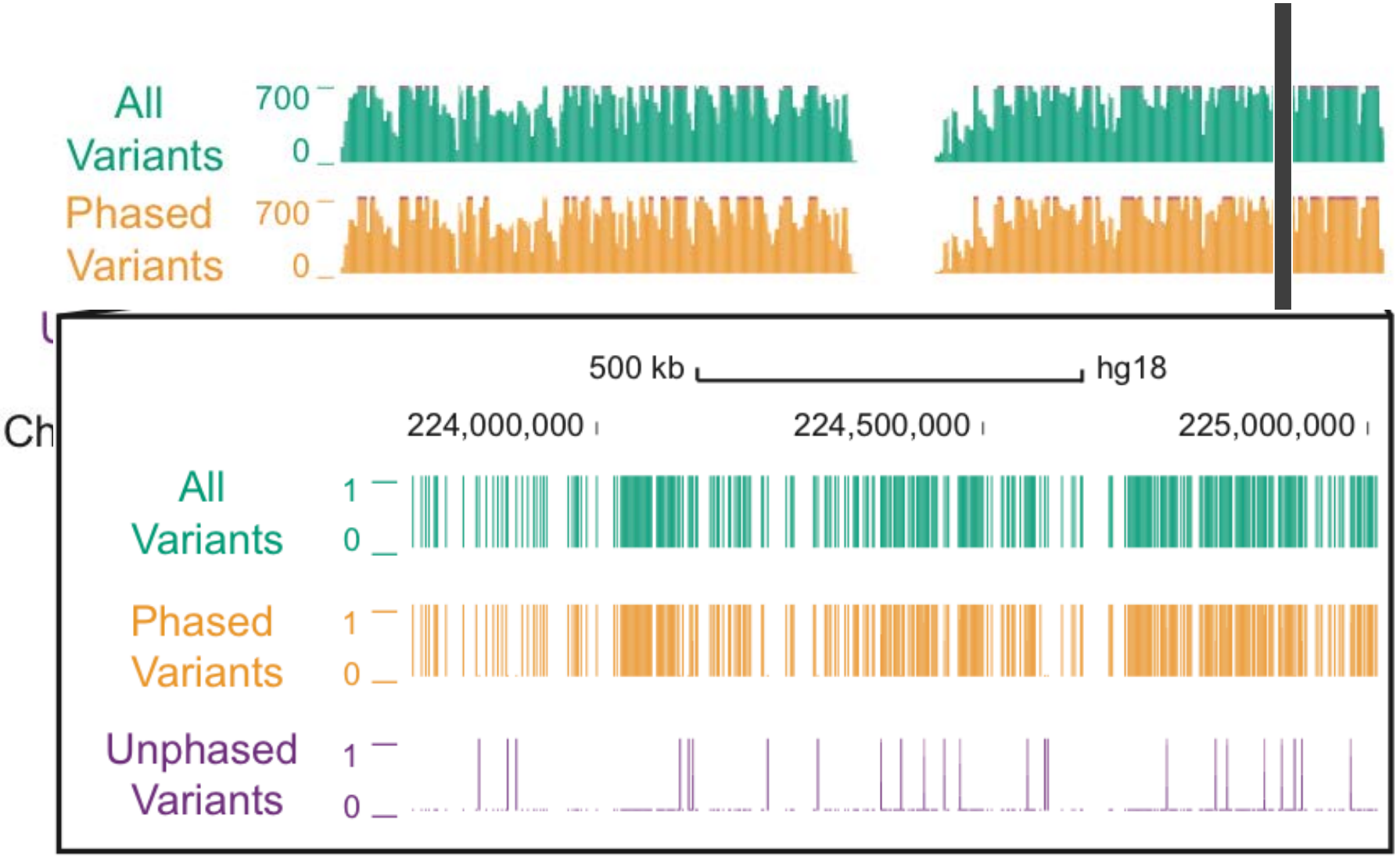
Conventional Whole Genome Shotgun sequencing





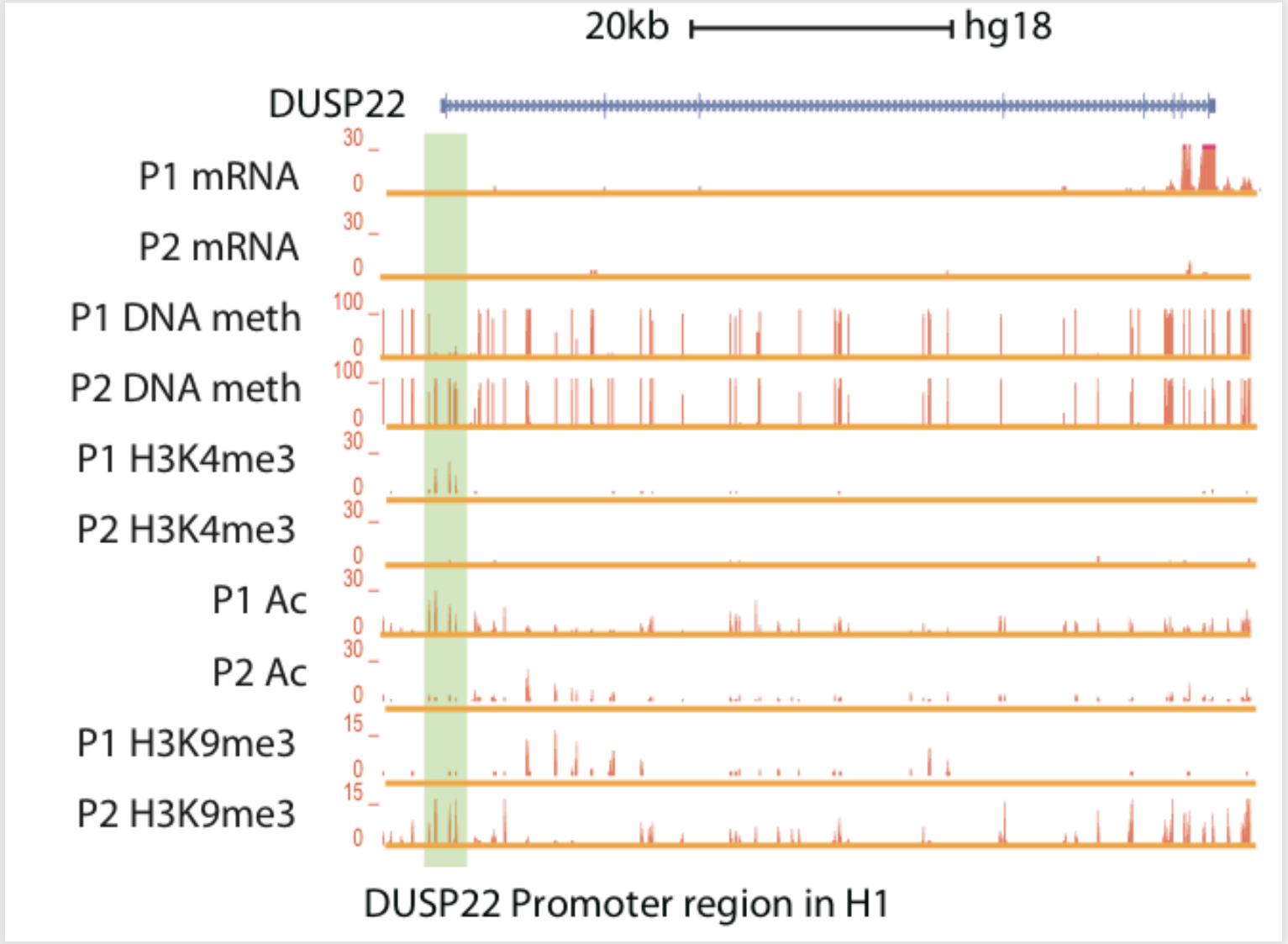
Selvaraj et al. Nat Bioltech 2013

Complete haplotypes in H1 hESC using HaploSeq



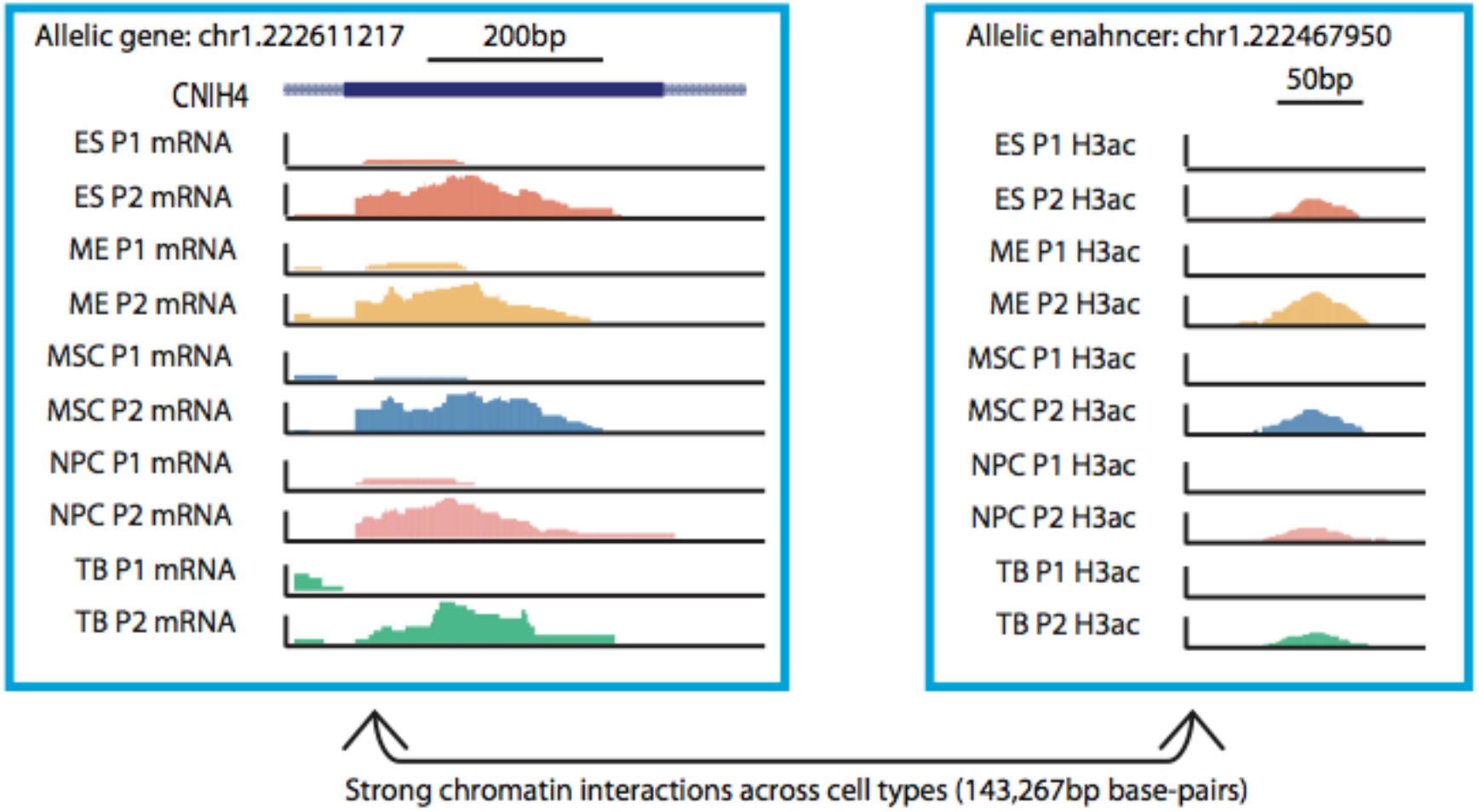
Bing Ren Laboratory (unpublished)

Allele-specific transcription, chromatin state and DNA methylation in H1 cells



Bing Ren Laboratory (unpublished)

Allele-specific transcription is correlated with allelic chromatin state at enhancers

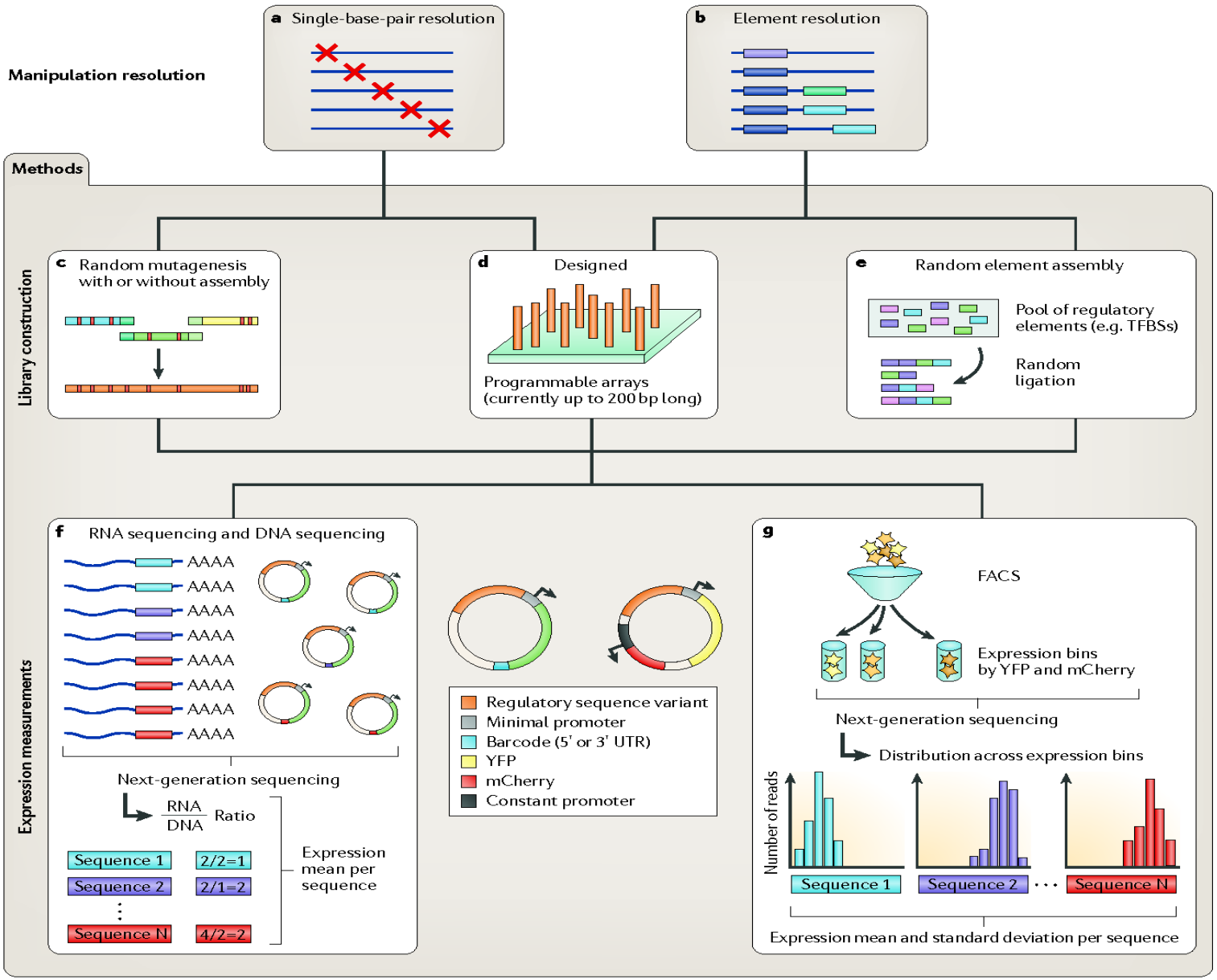


Bing Ren Laboratory (unpublished)

Functional Genomics @ Scale

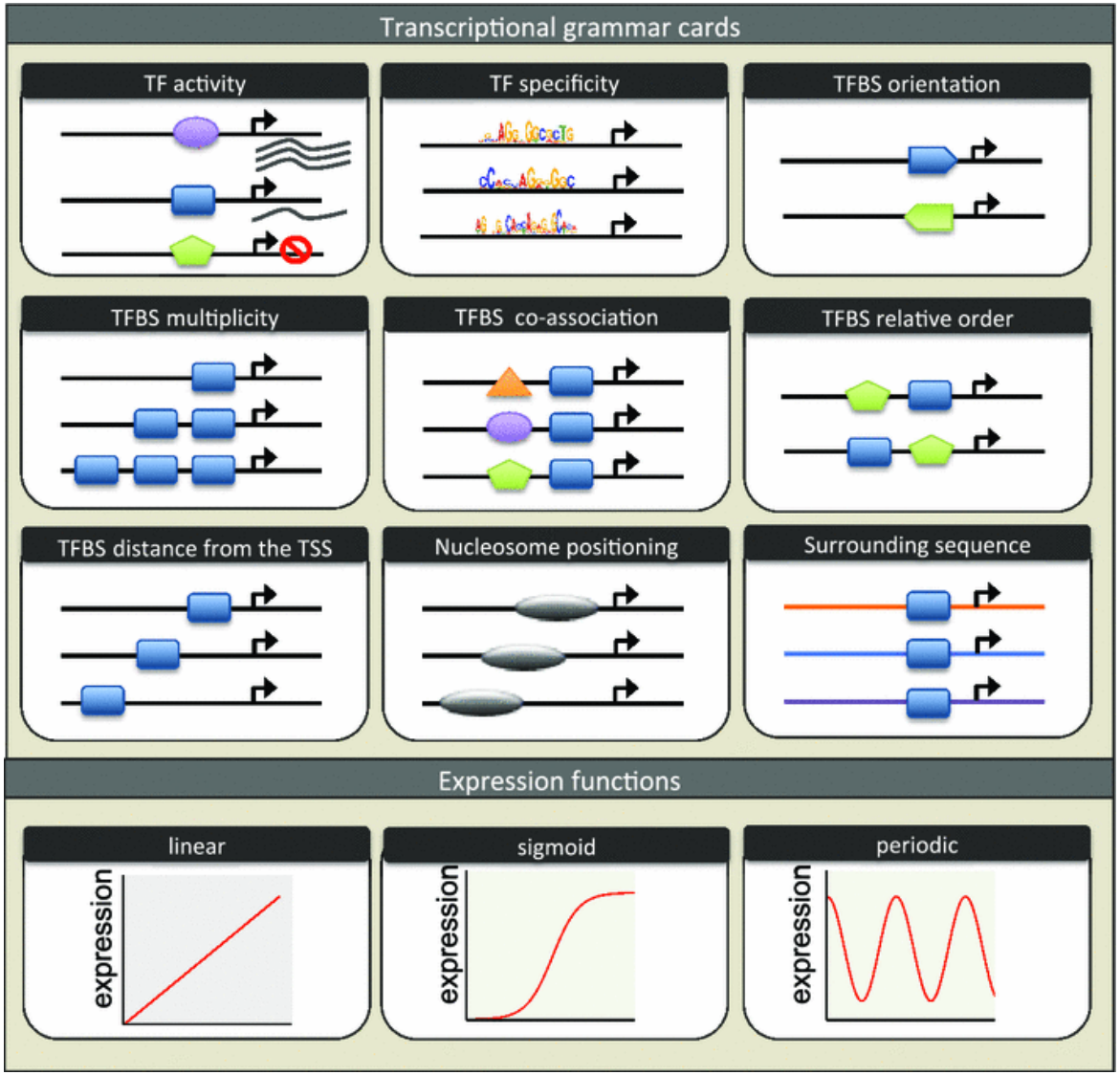
- Resources for Interpretation of variants
- Functional validation of variants

Dissection of regulatory sequences using massively parallel reporter assays



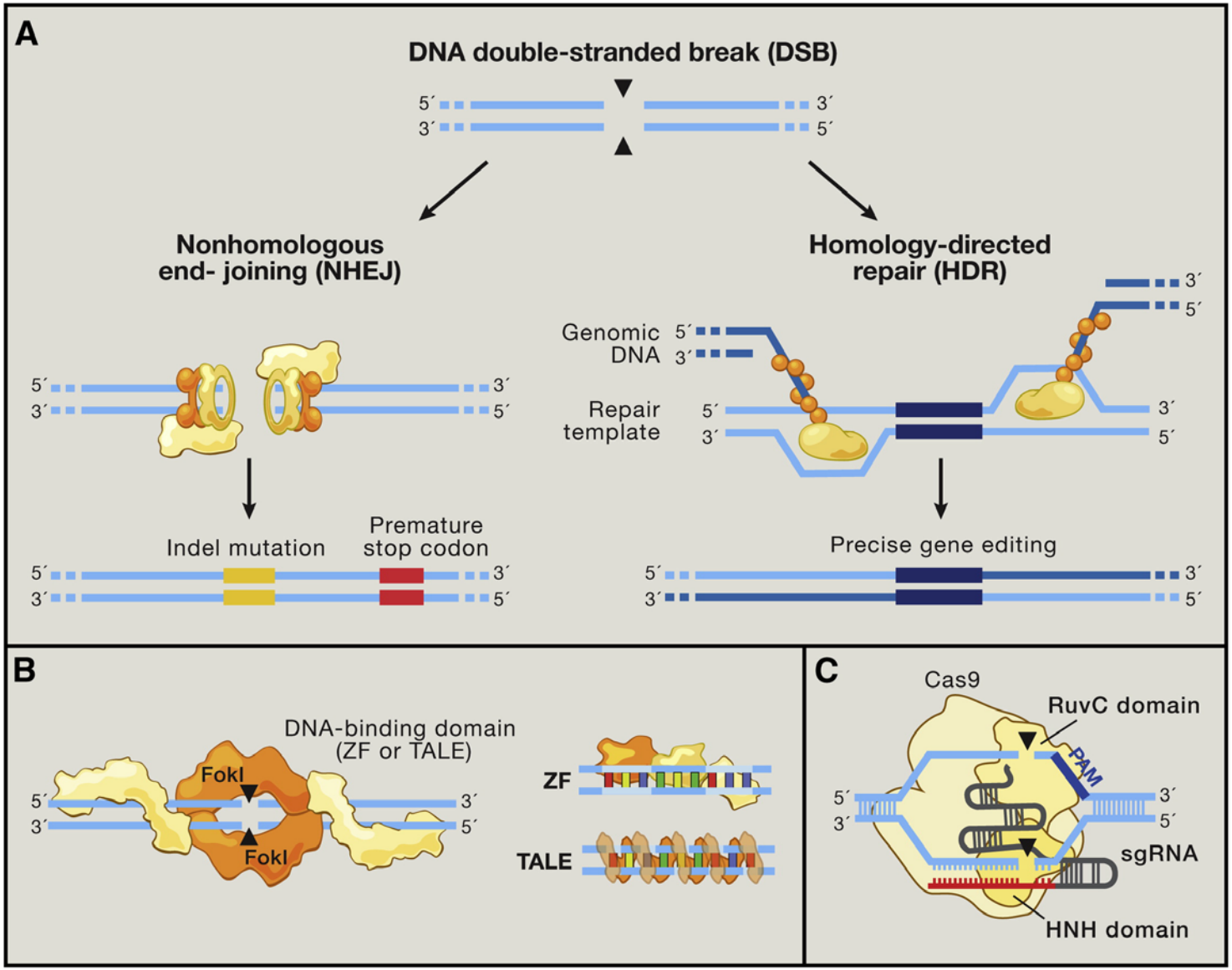
Levo and Segal (2014) NATURE REVIEWS | **GENETICS**

Understanding the Grammar of Gene Expression Regulation



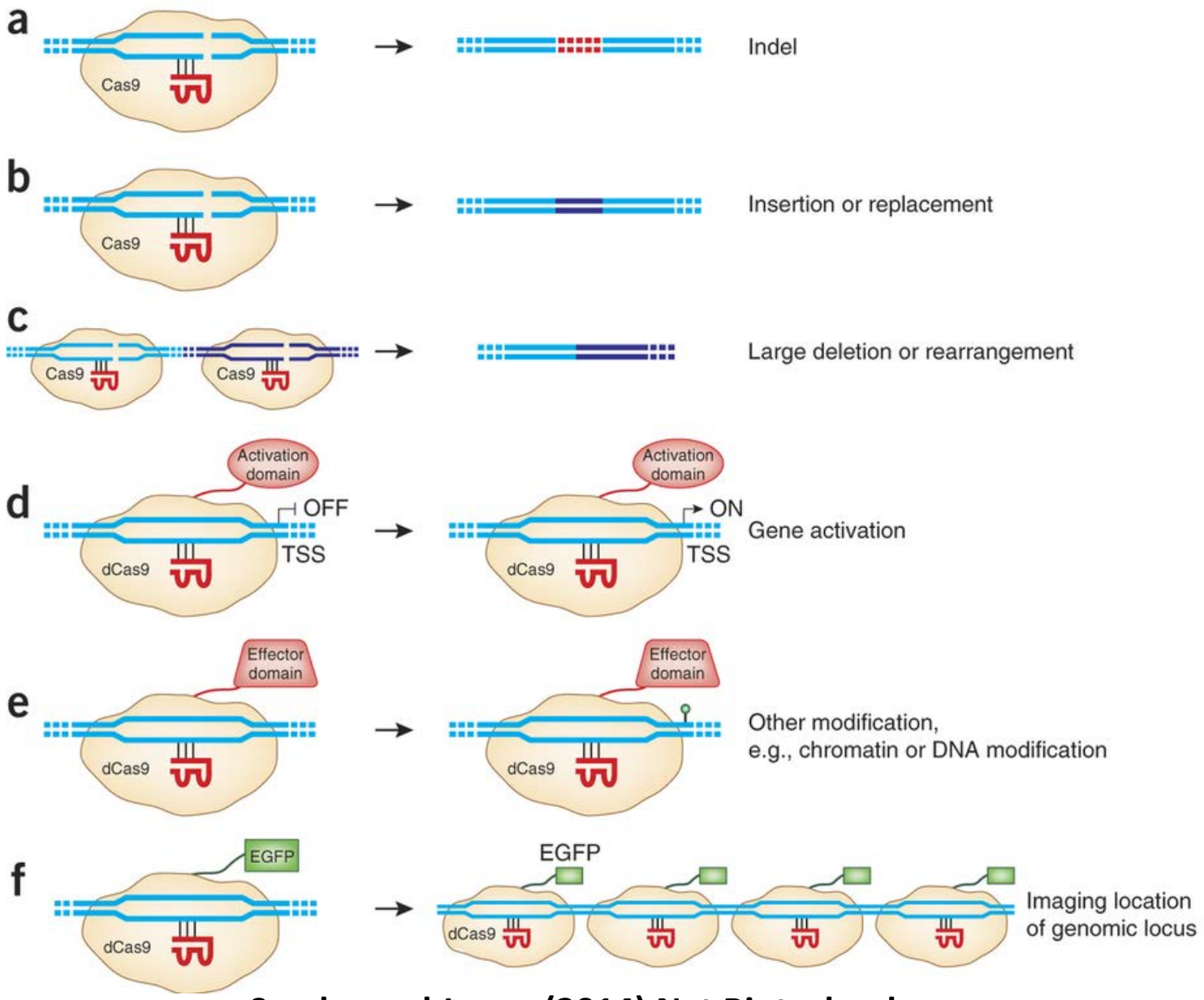
Weingarten-Gabbay and Segal (2014) Hum Genet

Powerful New Genome Editing Approaches



Hsu et al. Cell 2014 157, 1262-1278

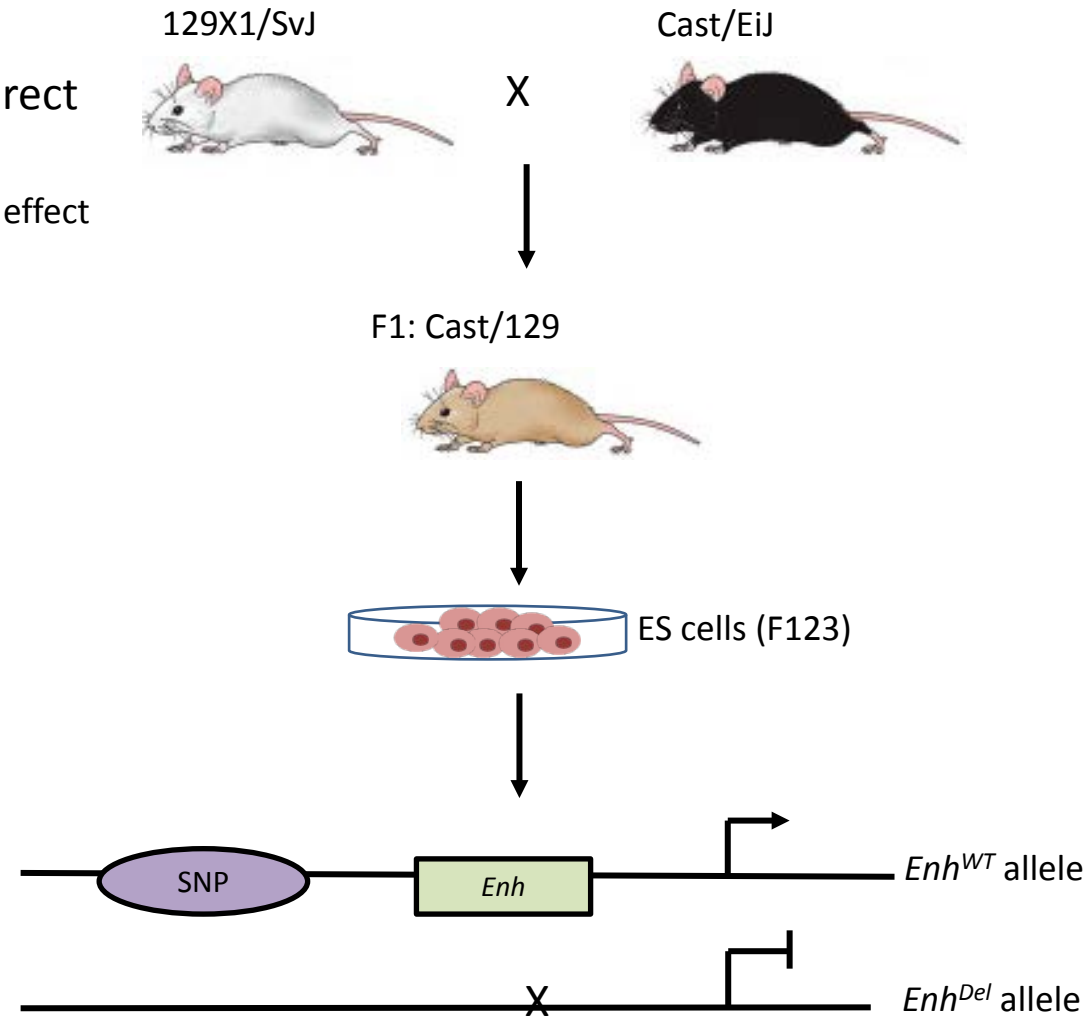
CRISPR/Cas9 Genome Editing Applications



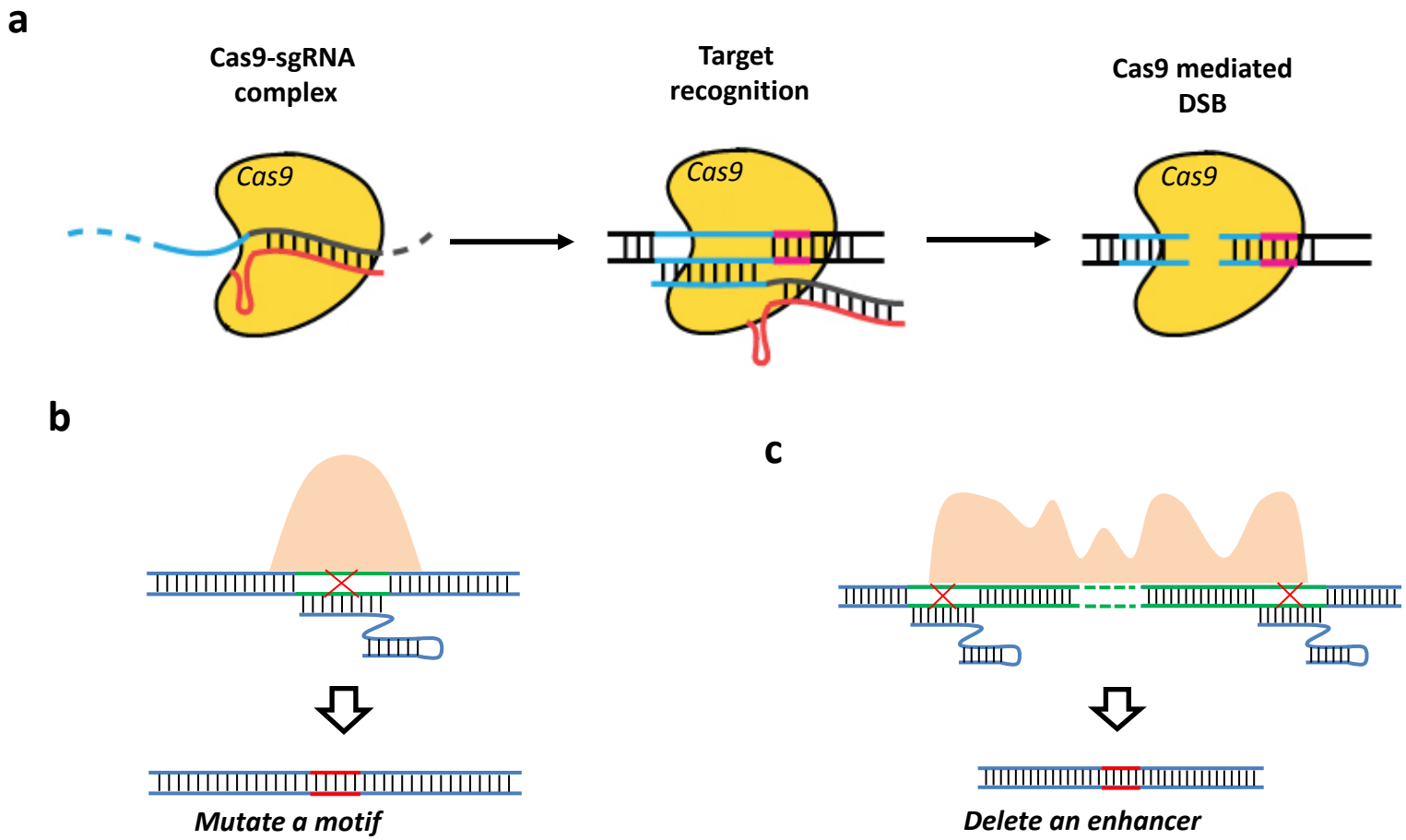
Sander and Joung (2014) Nat Biotechnol.

Validate the cis-regulatory functions of enhancers

- Enhancer knockout provide direct evidence
  - Test the transcription enhancing effect
  - Test if the effect is in *cis*.

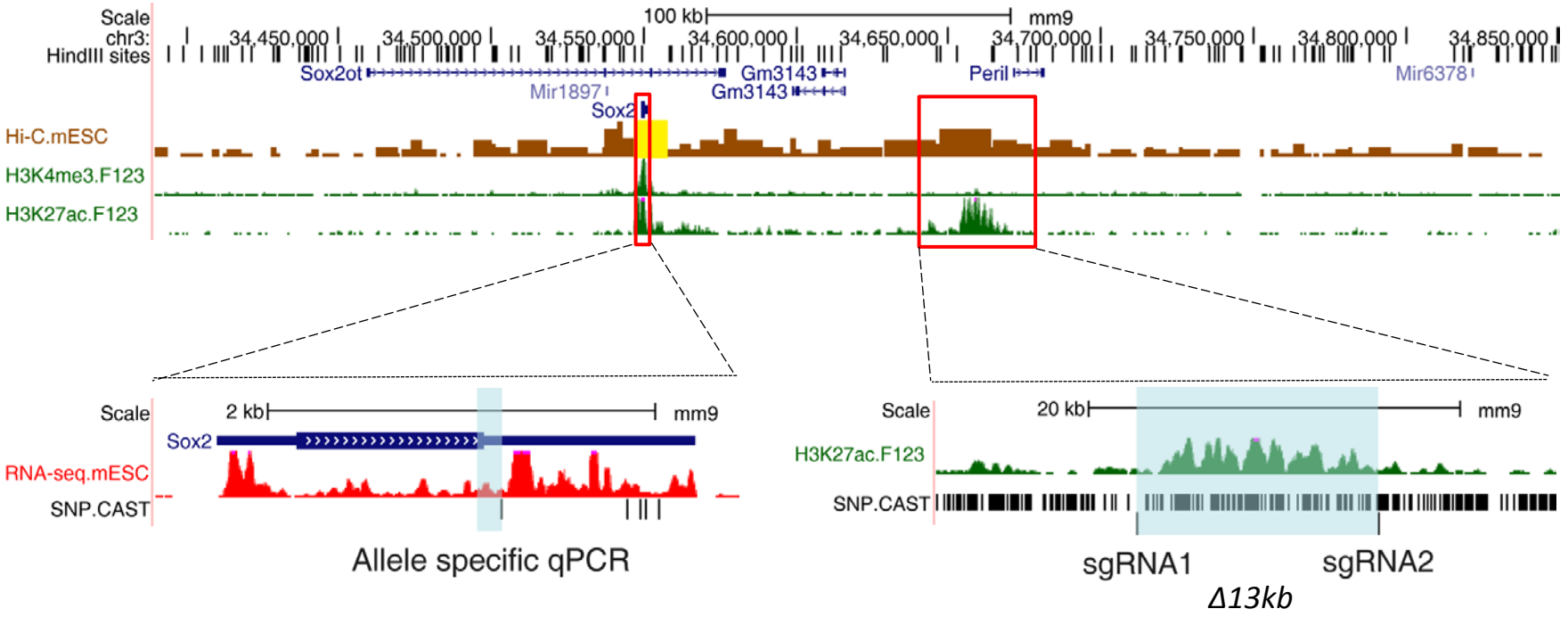


Using CRISPR/Cas9 to mutate enhancers



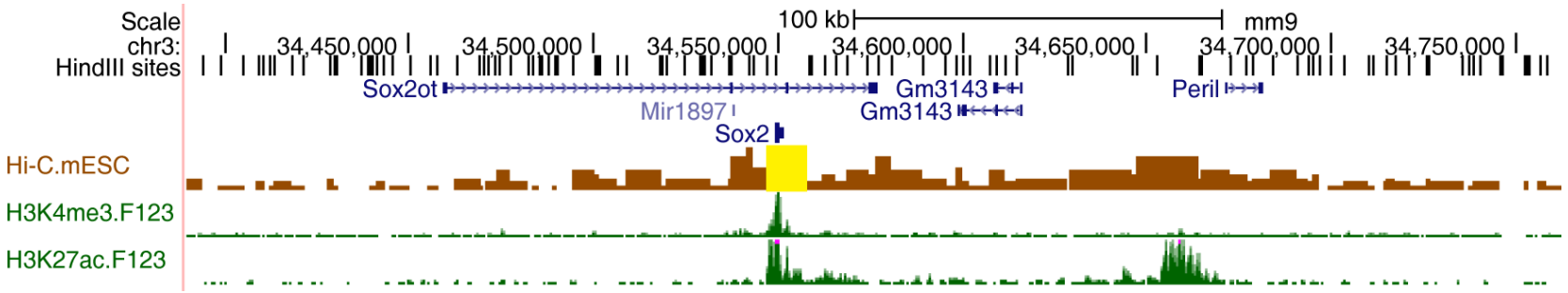
Bing Ren Laboratory

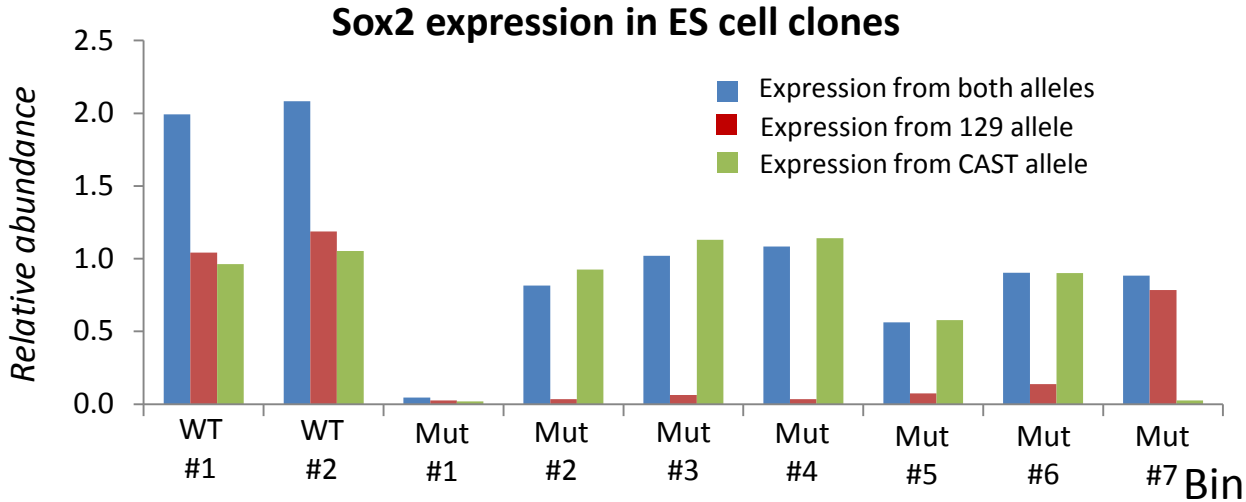
Validate Sox2 enhancer function using CRISPR/Cas9



Bing Ren Laboratory (unpublished)

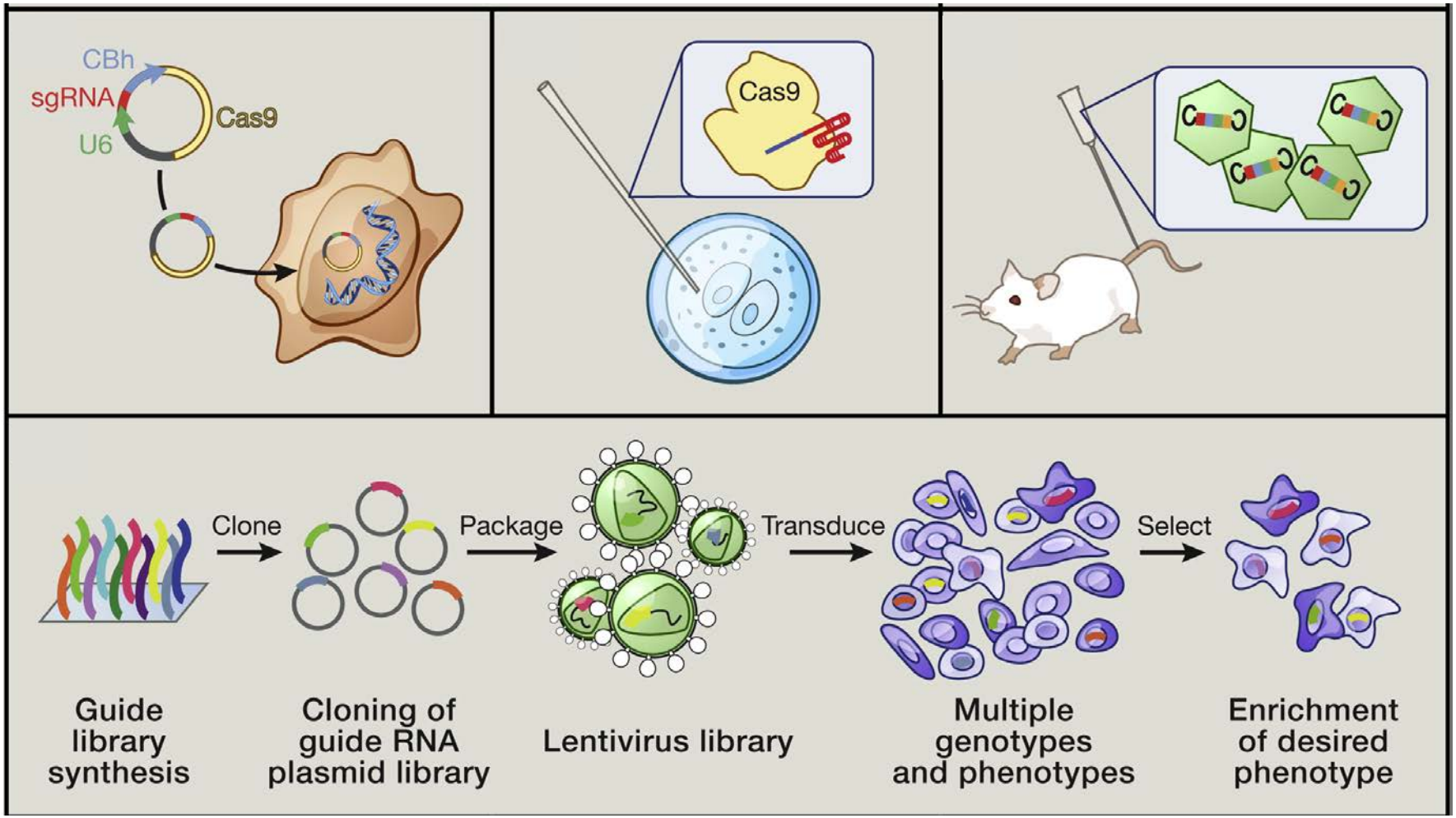
Sox2 expression is completely driven by a distal enhancer





Bing Ren Laboratory

Cas9 editing tools can be used in a variety of contexts to assess the function of sequence variants



Hsu et al. Cell 2014 157, 1262-1278

Received 29 April; accepted 10 June; published online 22 June 2014; doi:10.1038/nbt.2951

nature biotechnology

Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing

Dirk Heckl[1,5], Monika S Kowalczyk[2,6], David Yudovich[1,6], Roger Belizaire[1,3], Rishi V Puram[1], Marie E McConkey[1], Anne Thielke[2], Jon C Aster[3], Aviv Regev[2,4] & Benjamin L Ebert[1,2]

Genome sequencing studies have shown that human malignancies often bear mutations in four or more driver genes[1], but it is difficult to recapitulate this degree of genetic complexity in mouse models using conventional breeding. Here we use the CRISPR-Cas9 system of genome editing[2–4] to overcome this limitation. By delivering combinations of small guide RNAs (sgRNAs) and Cas9 with a lentiviral vector, we modified up to five genes in a single mouse hematopoietic stem cell (HSC), leading to clonal outgrowth and myeloid malignancy. We thereby generated models of acute myeloid leukemia (AML) with cooperating mutations in genes encoding epigenetic modifiers, transcription factors and mediators of cytokine signaling, recapitulating the combinations of

cell populations for myeloid malignancies, is complicated by the difficulty of using common nonviral gene transfer methods in these cells.

To perform genome editing with high efficiency in primary HSPCs, and to track the engineered cells *in vivo*, we generated a modular lentiviral sgRNA:Cas9 vector for modeling of myeloid malignancies by genome editing in primary HSPCs *in vivo* (**Fig. 1a** and **Supplementary Fig. 1**). This lentiviral vector simultaneously delivers the *Streptococcus pyogenes cas9* gene, a chimeric sgRNA and a fluorescent marker, similar to our recently developed system[2,3,13,14]. This enables the targeting of any genomic locus in a broad range of cell types, and consequent nonhomologous end-joining (NHEJ)-mediated gene disruption, by a one-step exchange of the target site (spacer).

Current favorite example: the challenge of understanding non-coding variants

VOLUME 46 | NUMBER 7 | JULY 2014 NATURE GENETICS

A molecular basis for classic blond hair color in Europeans

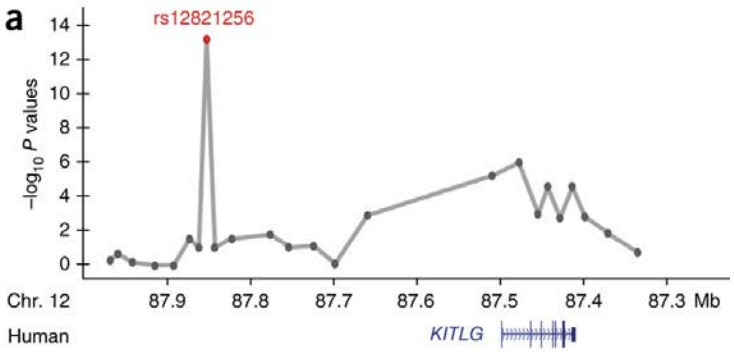
Catherine A Guenther[1,2], Bosiljka Tasic[2,3,5], Liqun Luo[2,3], Mary A Bedell[4] & David M Kingsley[1,2]

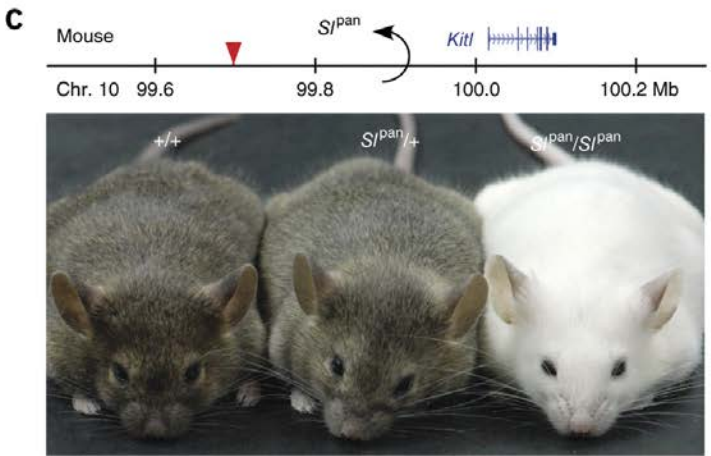
Hair color differences are among the most obvious examples of phenotypic variation in humans. Although genome-wide association studies (GWAS) have implicated multiple loci in human pigment variation, the causative base-pair changes are still largely unknown[1]. Here we dissect a regulatory region of the *KITLG* gene (encoding KIT ligand) that is significantly associated with common blond hair color in northern Europeans[2]. Functional tests demonstrate that the region contains a regulatory enhancer that drives expression in developing hair follicles. This enhancer contains a common SNP (rs12821256) that alters a binding site for the lymphoid enhancer-binding factor 1 (LEF1) transcription factor, reducing LEF1 responsiveness and enhancer activity in cultured human keratinocytes. Mice carrying ancestral or derived variants of the human *KITLG* enhancer exhibit significant differences in hair pigmentation, confirming that altered regulation of an essential growth factor contributes to the classic blond hair phenotype found in northern Europeans.

improve genetic predictions compared to common linked markers[13] and facilitate comparison of traits and mutations among both past and present populations[14].

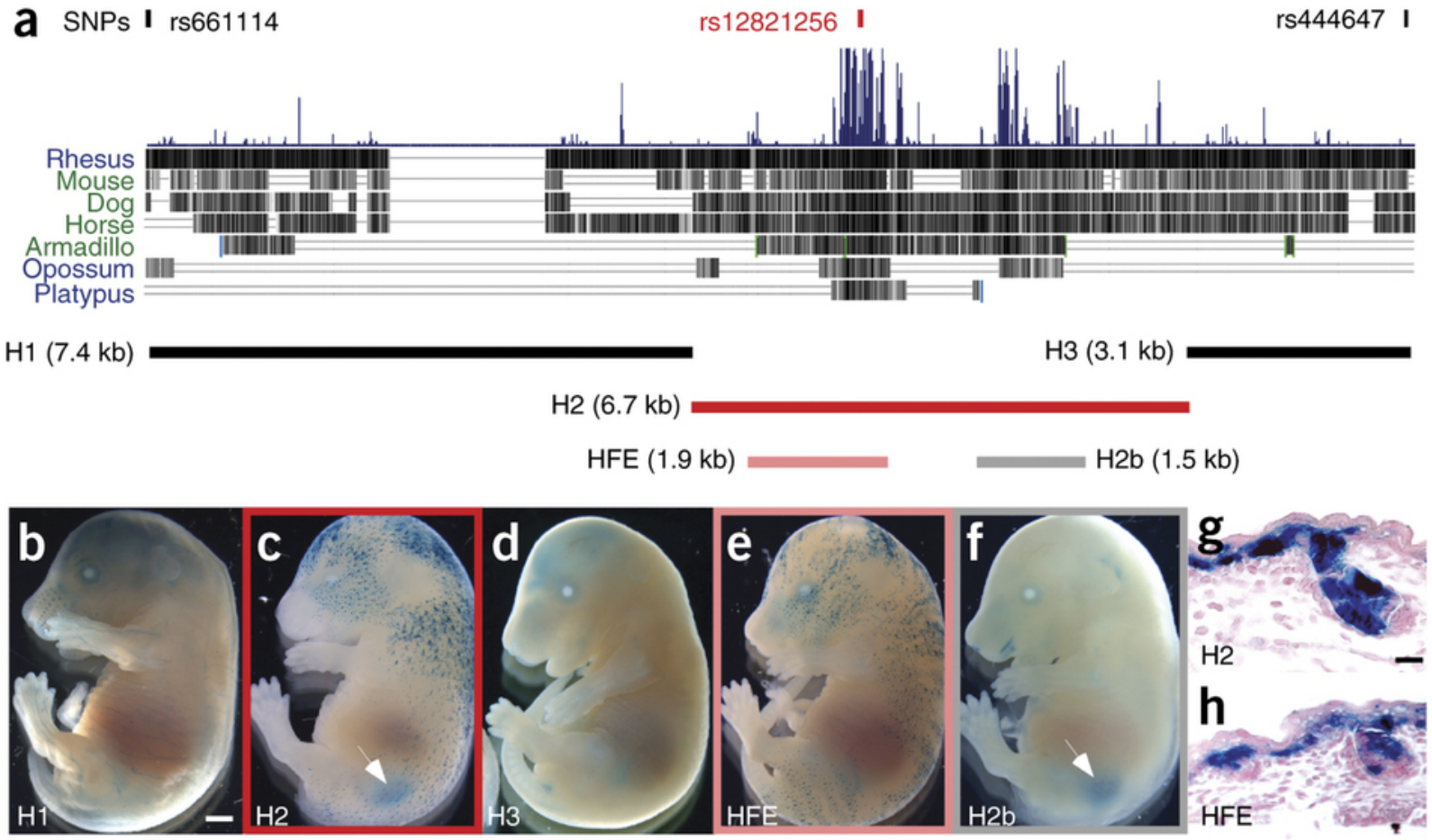
Human *KITLG* (mouse *Kitl*) encodes a secreted ligand for the KIT receptor tyrosine kinase and has an essential role in the development, migration and differentiation of many different cell types in the body, including melanocytes, blood cells and germ cells[15]. Null mutations affecting *Kitl* or *Kit* are lethal in mice, and hypomorphic alleles cause white hair, mast cell defects, anemia and sterility[16–18]. A noncoding SNP (rs12821256) located in a large intergenic region over 350 kb upstream of the *KITLG* transcription start site is significantly associated with blond hair color in Iceland and The Netherlands[2] (**Fig. 1a**). This SNP shows relatively large odds ratios of 1.9–2.4 per allele associated with blond versus brown hair in northern Europeans (multiplicative model[2]). Together with variants in other genes, rs12821256 helps explain 3–6% of the variance in categorical hair color scores[2] and is now one of several markers used for predictive testing of human hair color[19]. The blond-associated A>G substitution at this position is prevalent in northern European populations but virtually absent

Variant interpretation: population/mouse genetics

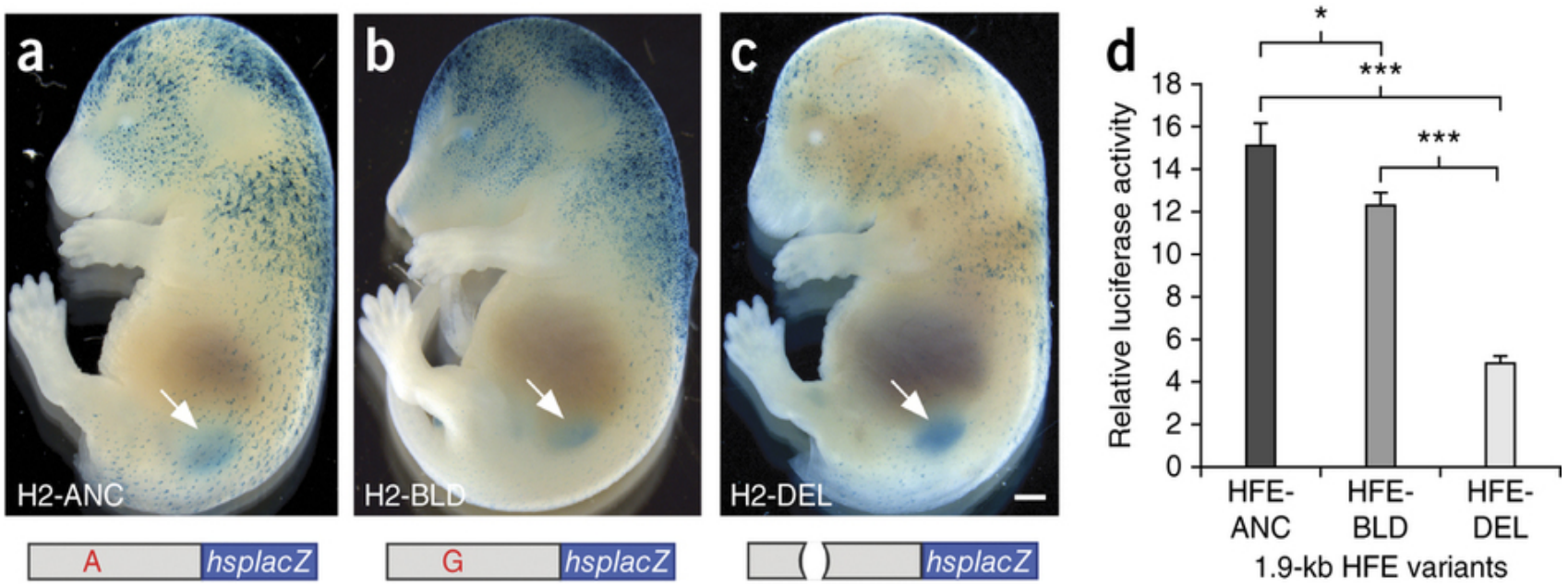




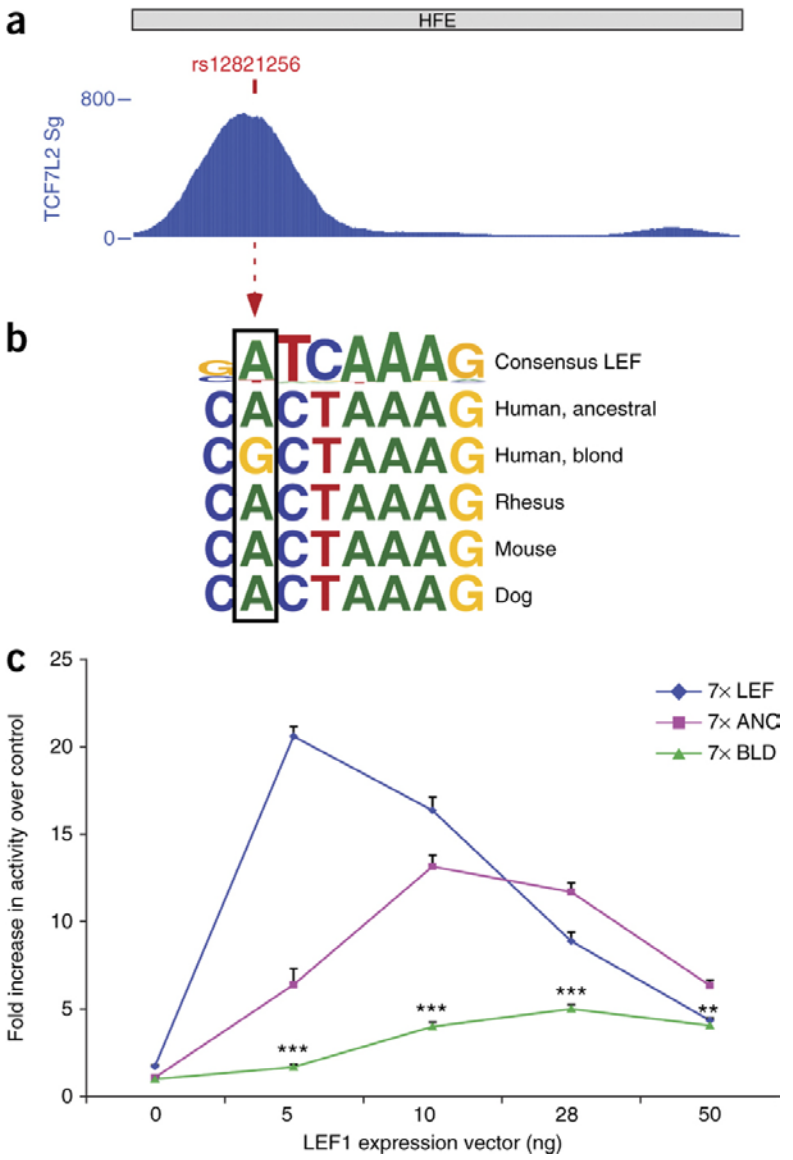
Variant interpretation: sequence conservation



Variant interpretation: in vivo functional assay

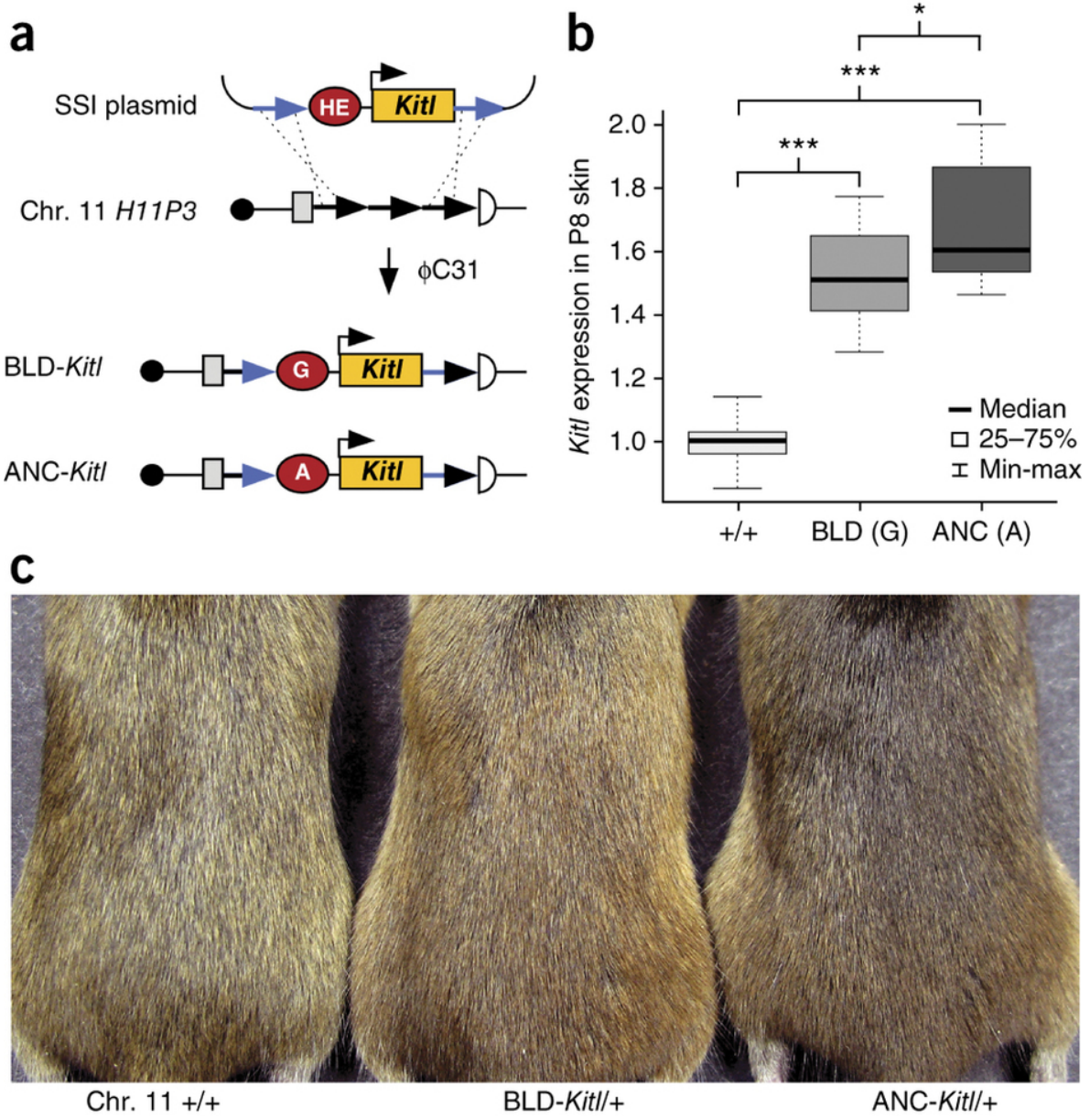


Variant interpretation: functional assay in cell culture



(A)
(G)

Variant interpretation: functional assay -transgenics



The study Kingsley highlights why it is still so difficult to identify the causal basis of human trait associations:

- The associated SNP (rs12821256) maps more than 350 kb from KITLG,
- acts at a specific anatomical site whose active enhancers have not yet been characterized in large-scale studies of human chromatin marks,
- alters a sequence that does not perfectly match a LEF1 consensus binding site and
- only causes an approximately 20% reduction in the activity of a previously unrecognized hair follicle enhancer.
- BUT the study also illustrate how these difficulties can now be overcome using:
- information from human population surveys,
- large-scale genome annotation projects and
- transcription factor interaction databases in combination with
- detailed functional tests of enhancer activity in cell lines and in mice.

Breakout Session (Gerstein/Myers)
Integrating functional genomics with DNA sequence variants

- **1) What is function in genomics & how do we use it to determine the effect of variants?**
  - What are the different aspects of function and why is it hard to study? For instance, molecular (or biochemical) function vs cellular role vs organismal phenotype.
  - What are the problems in defining function? Is it meaningful to localize a function to a single place on the genome so it can be affected by a single variant? How should one think about the functional effect of large block variants?
  - Is it possible to quantitatively systematize some aspects of function so that they can be precisely related and correlated with genomic variants? In particular, what are the paradigms available to inter-relate function with variants (eg QTLs & allelic effects and phenotypes resulting from a single disruption)?

- **2) How do we inter-relate function & variants on a large scale?**
  - Is this best done by individual investigators pooling together individual results into a database or is it best done by large-scale, highly standardized experiments? What is the role of special big data database architectures for aggregating the knowledge of many functional assays?
  - Is it more effective to follow up on the many disease-associated variants uncovered by sequencing in great detail rather than doing broad genome-wide functional characterization beforehand?
  - Are there ways for new high-throughput technologies and computational approaches to significantly help with this endeavor?
  - How do we prioritize those experiments and assays that provide more functional information compared to others? Is there a particular way of assessing the information in particular experiments?

- **3) How do we validate functional effects of variants in genomics?**
  - Is it possible to validate thousands (or millions) of assertions about the genome with one or two small-scale validation experiments?
  - Is it possible to do validation at a very large scale? Is medium-scale validation possible and useful? How to think about the cost of this?
  - How do we incorporate the results of validation into quantitative error estimates for the functional assertions being made?