DCEG Statistical Genetics Workshop 2023

Session 9: Functional genomics

Laboratory of Translational Genomics:

Ludmila Prokunina-Olsson, PhD

Kevin Brown, PhD

Jiyeon Choi, PhD

Laufey Amundadottir, PhD

Oscar Florez-Vargas, PhD

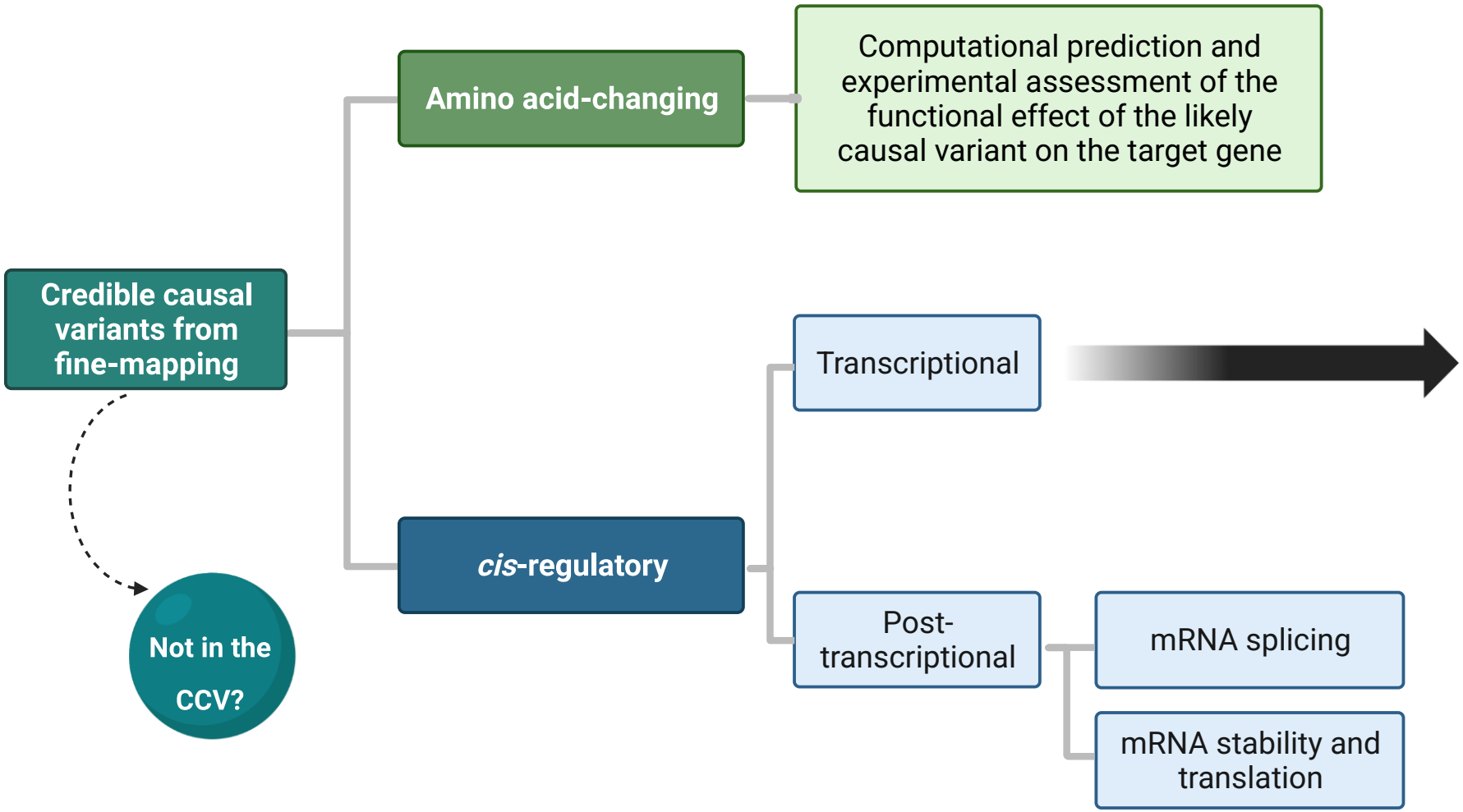
December 13, 2023

NIH NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics

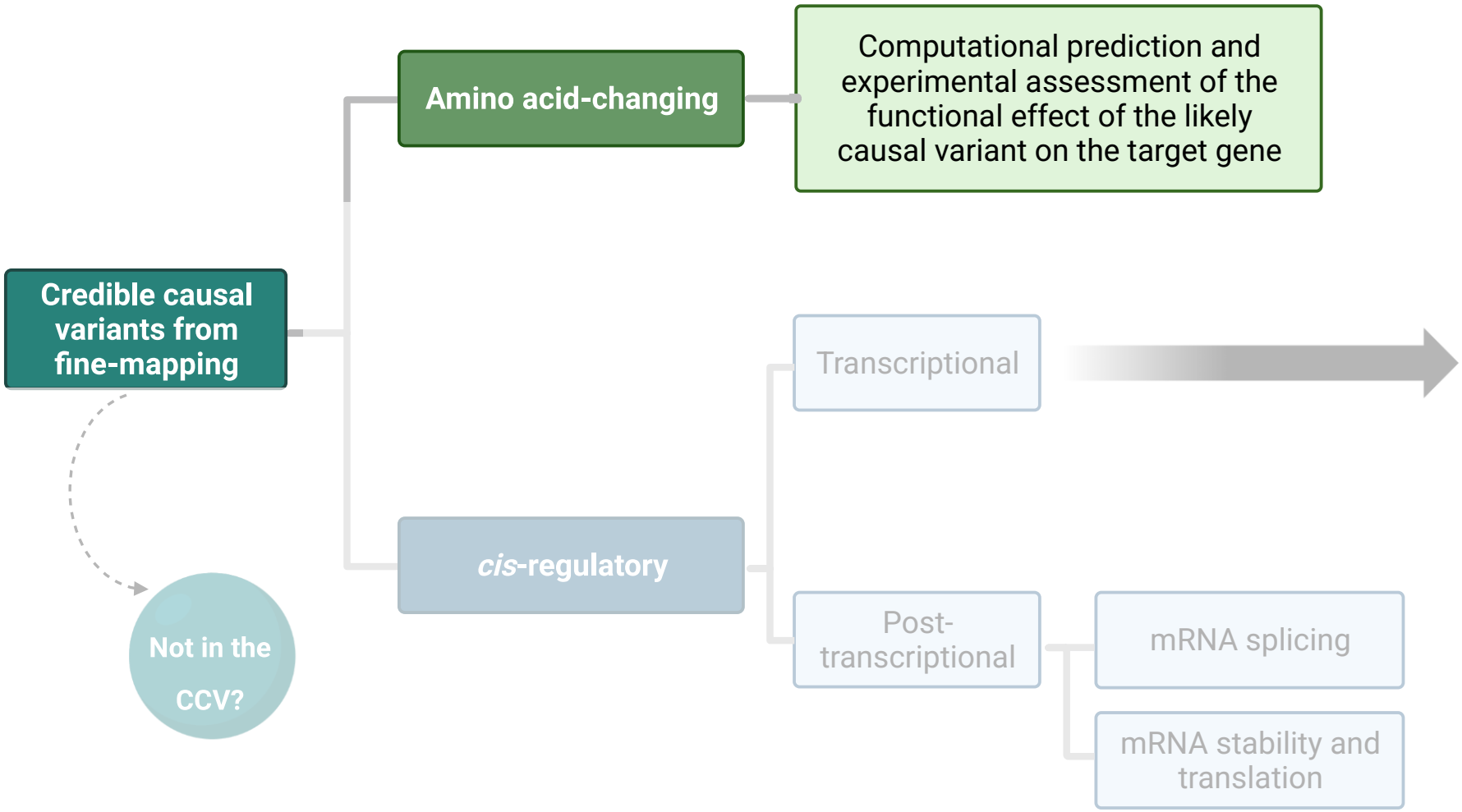
Post-GWAS questions:

- What are the “functional” variant(s) lead GWAS SNPs are tagging?
- What are the gene(s) through which underlying genetic variation influence risk?
- How does the functional variant influence gene function?
- What role does the gene play in cancer associated phenotypes?

Mechanisms of variant function



Mechanisms of variant function



Protein-coding variants: pathogenicity prediction

There are MANY tools to predict potential impact of non-synonymous protein-coding variants on protein function, based on:

- DNA and or protein sequence conservation
- Protein structure
- Other features (including SNPs and pathogenic variants)

Common variants from GWAS -> small effect sizes

- Various tools often give conflicting interpretations

Experimental validation required, many potential mechanisms by which to influence protein function

Note: protein-coding variants can also function via most of the other mechanisms by which non-coding variants can, including:

- be *cis*-regulatory!
- Regulate splicing

Nice review and list of tools (see Table 1):
https://www.frontiersin.org/articles/10.3389/fgene.2022.1010327/full

Relatively few fine-mapped variants are non-synonymous

Table 1 The genomic context in which a variant is found can be used as preliminary functional analysis

Classification	Approximate percentages[a]	Approximate numbers[a]
Intronic	40	1,047
Intergenic	32	838
Within non-coding sequence of a gene	10	262
Upstream	8	210
Downstream	4	105
Non-synonymous coding	3	79
3′ untranslated region	~1	26
Synonymous coding	~1	26
5′ untranslated region		
Regulatory region		
Nonsense-mediated decay transcript		
Unknown	~1	26
Splice site		
Gained stop codon		
Frameshift in a coding sequence		

The table broadly summarizes the genomic context of disease- and trait-associated SNPs annotated in the Catalog of Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) as of December 9th, 2010: 1,212 published genome-wide associations with $P < 5 \times 10^{-8}$ for 210 traits totaling 2,619 SNPs. Most of the SNPs are located in intergenic and intronic positions, but a small percentage are located upstream and downstream of genes, as well as in regulatory regions and splice sites. SNPs in these locations can be analyzed in more detail using more specific bioinformatics tools.
[a]Values are indicative and dependent on genomic boundaries used.

Freedman et al. (2011) Nat Genet. 2011 Jun;43(6):513-8.

Non-protein coding mechanisms

***cis*-regulation:**

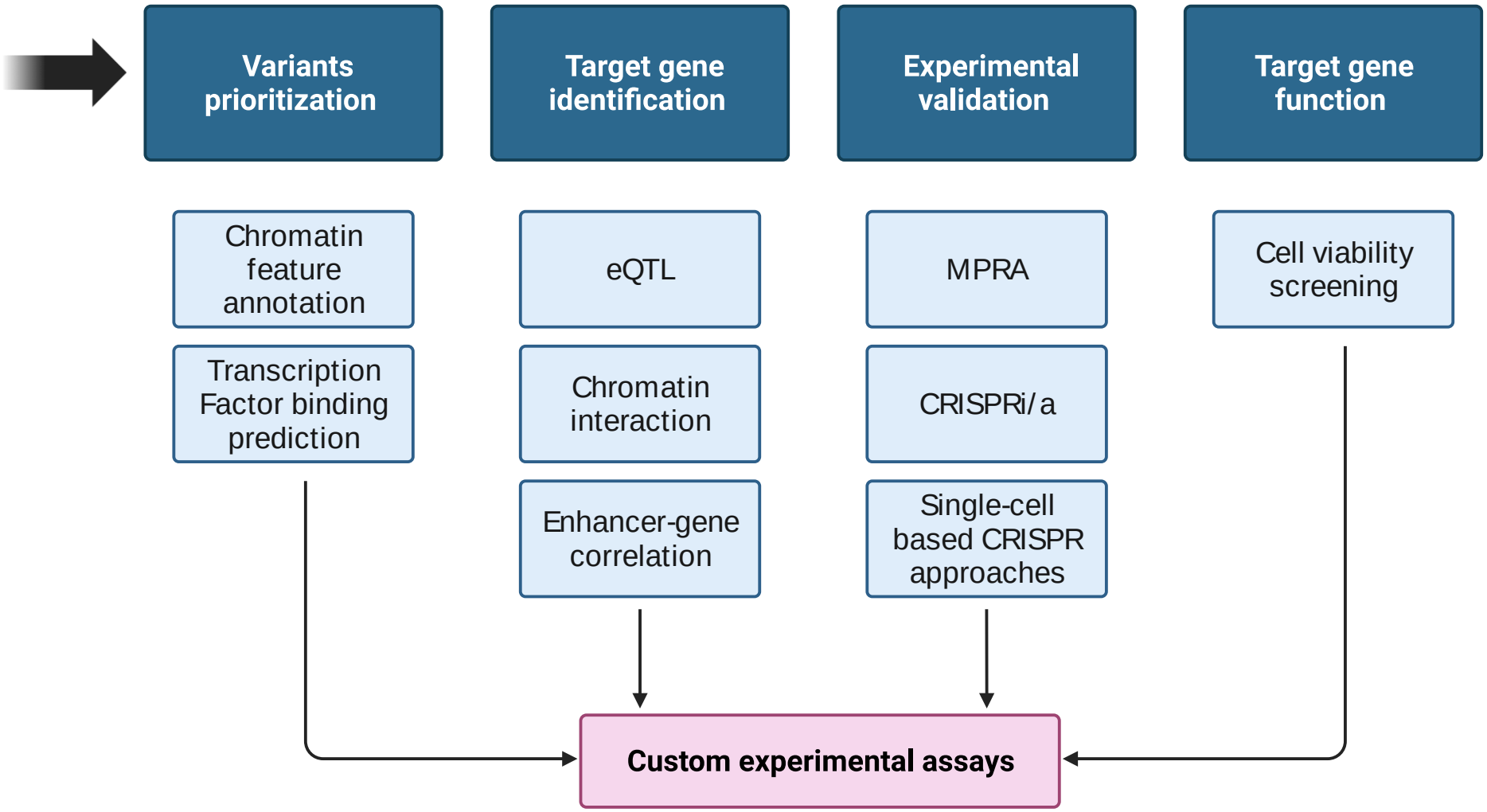
- Altered TF binding/regulatory activity in enhancer/promoter
  - Direct regulation of transcription
  - Regulation of chromatin accessibility
- Alteration of 3-Dimensional chromatin structure and localization
  - ->Changes to enhancer-promoter associations
- Altered CpG methylation

Post-transcriptional:

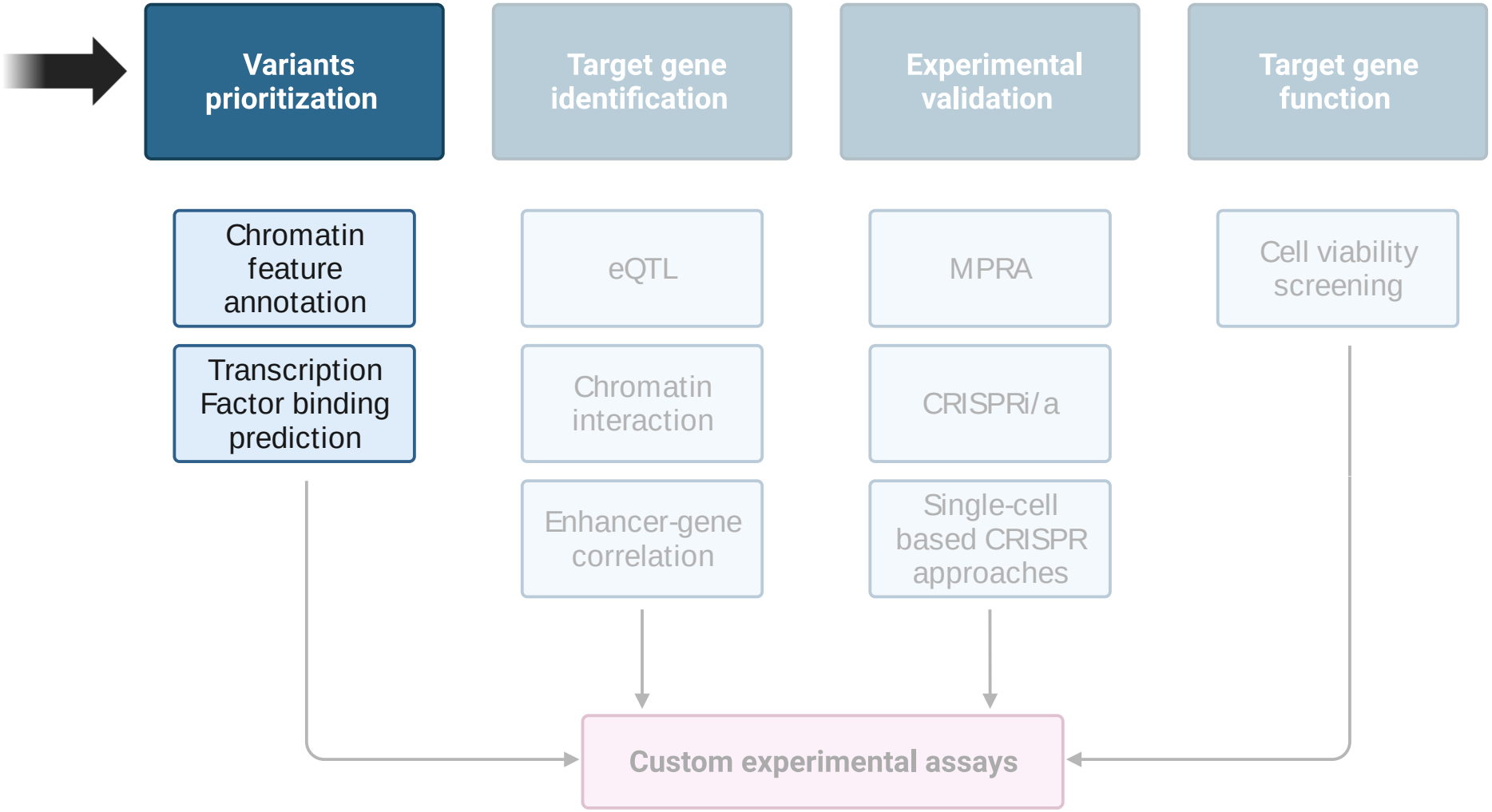
- Splicing (splice variants can be in both introns or exons)
- Post-transcriptional modifications, polyadenylation, etc.
- Transcript stability
- Translation efficiency
- Alternative codon usage, translational efficiency, and protein folding

Altered function of non protein-coding transcripts

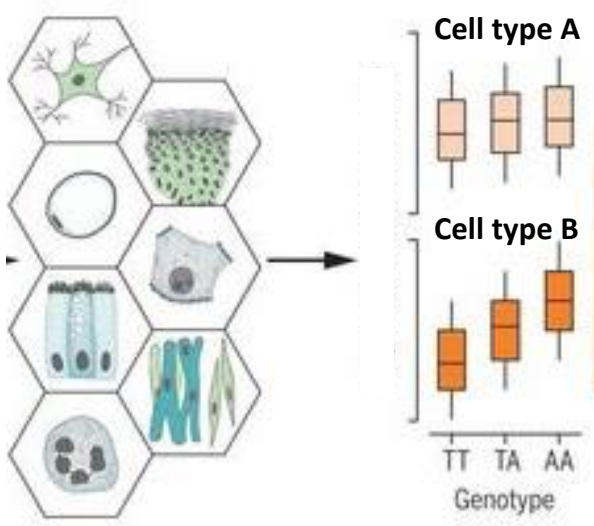
Identification of "causal" variants and target genes



Identification of "causal" variants and target genes

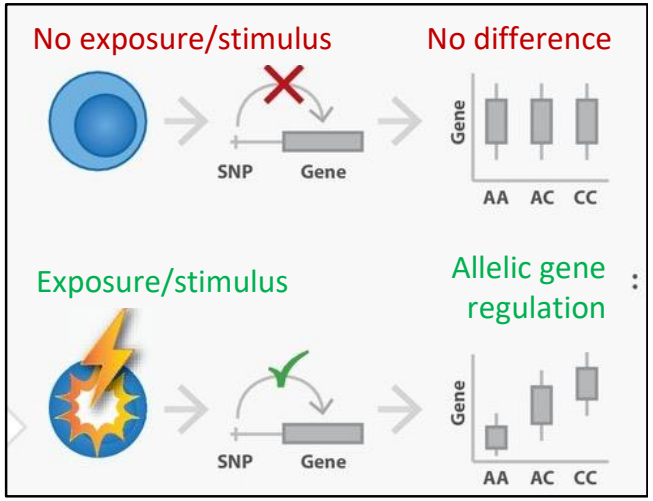


Complex function of cancer-associated variants



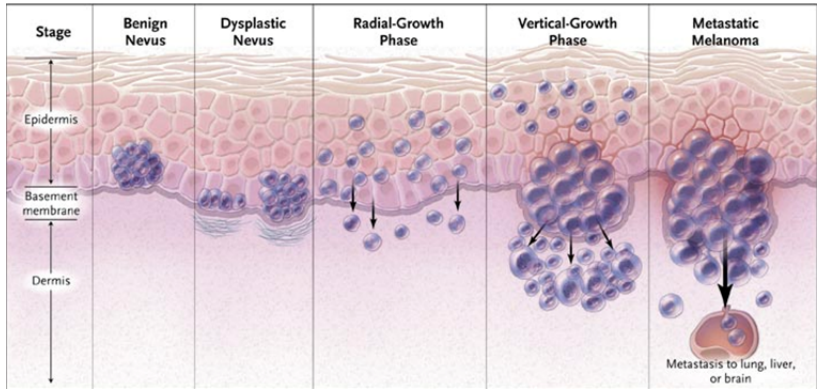
Modified from Kim-Hellmuth *et al. Science* 2020

- Cell-of-tumor-origin
- Primary vs. cancer cells
- Cells in TME



Modified from Wijst *et al. eLife* 2020

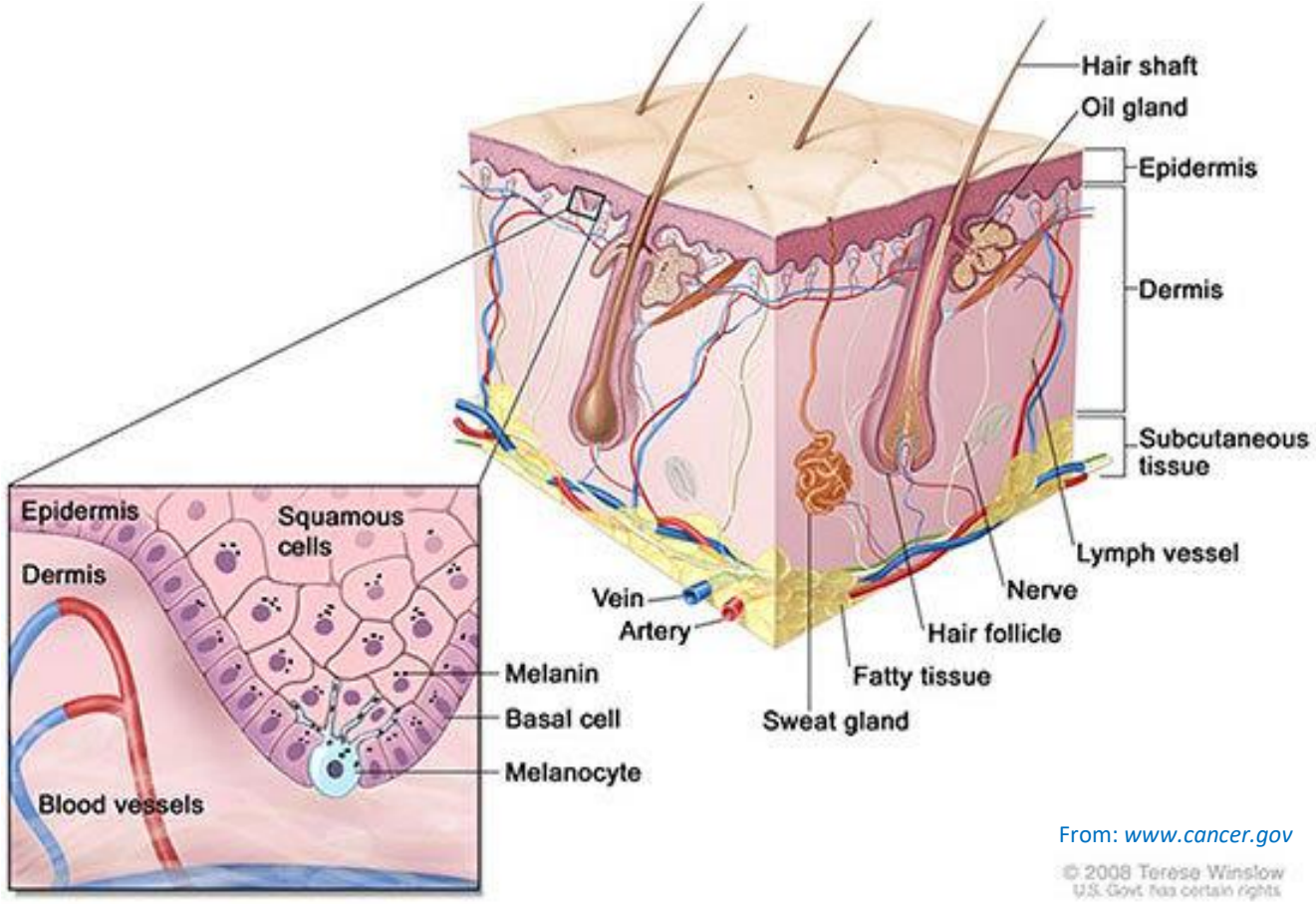
- Exposure/stimulus
- Age, sex, race/ethnicity/ancestry



Miller *et al, NEJM* 2006

- Initiation, progression, outcome
- Germline x somatic
- Specific driver events

Example: melanoma originates from melanocytes



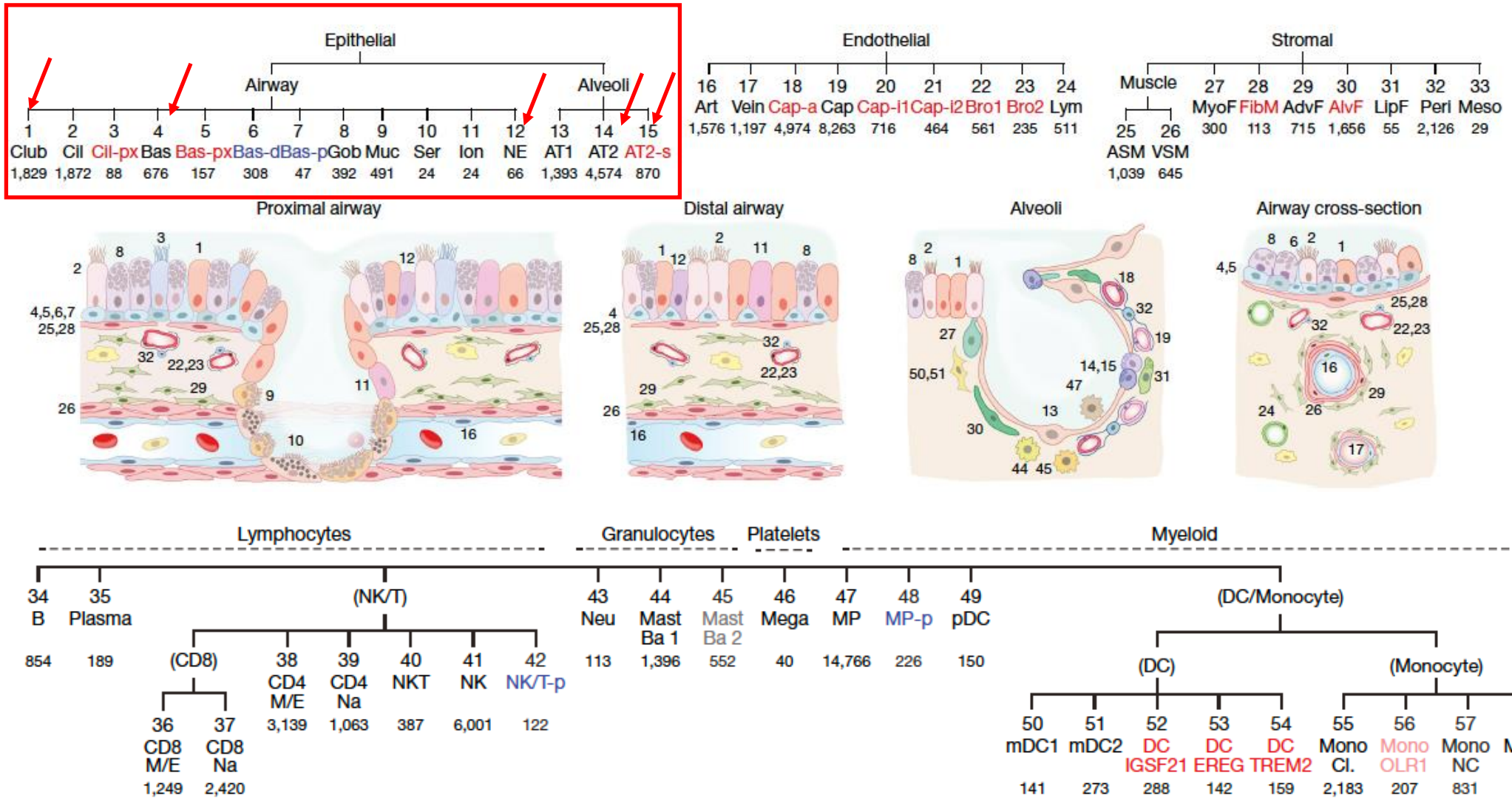
Melanocytes are a small fraction of the cells in a skin biopsy

Melanocyte-specific contributions to expression and allelic expression may not be well-reflected in bulk skin tissue for many genes

We can culture melanocytes and measure expression for QTLs!... But:

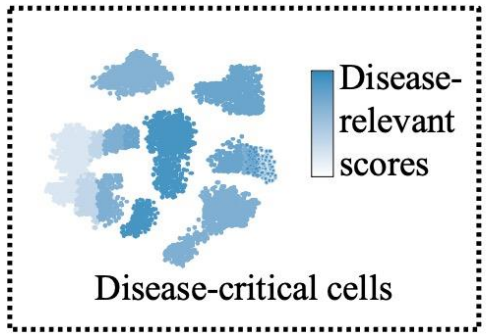
- Cultured cells -> artificial system
- Not possible for all cell types
- Cell of origin may not be known

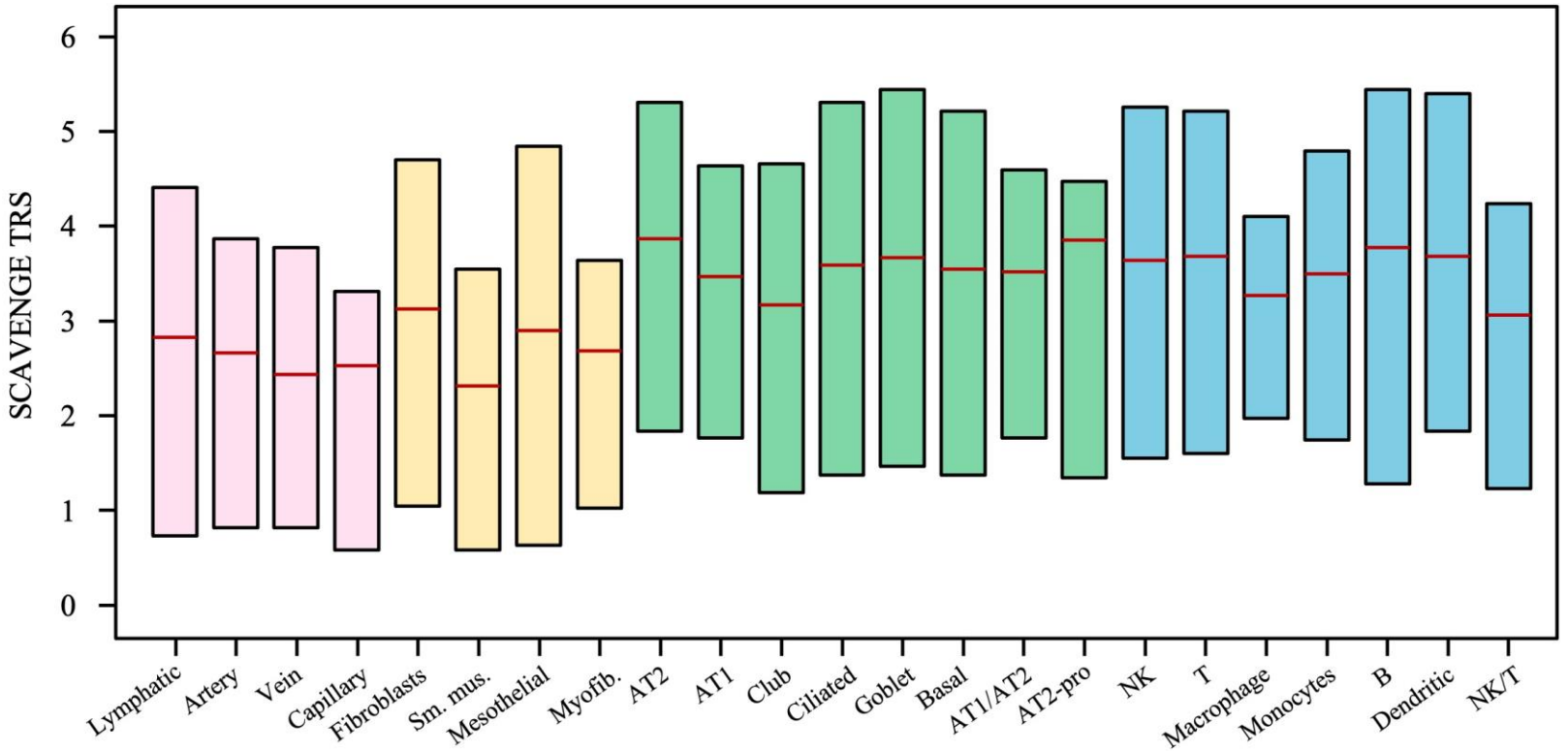
Cells of lung cancer origin are diverse



Travaglini *et al*. *Nature* 2020

Single-cell level trait relevance scores using single-cell ATACseq





Long et al, *bioRxiv* 2023.09.25.559336

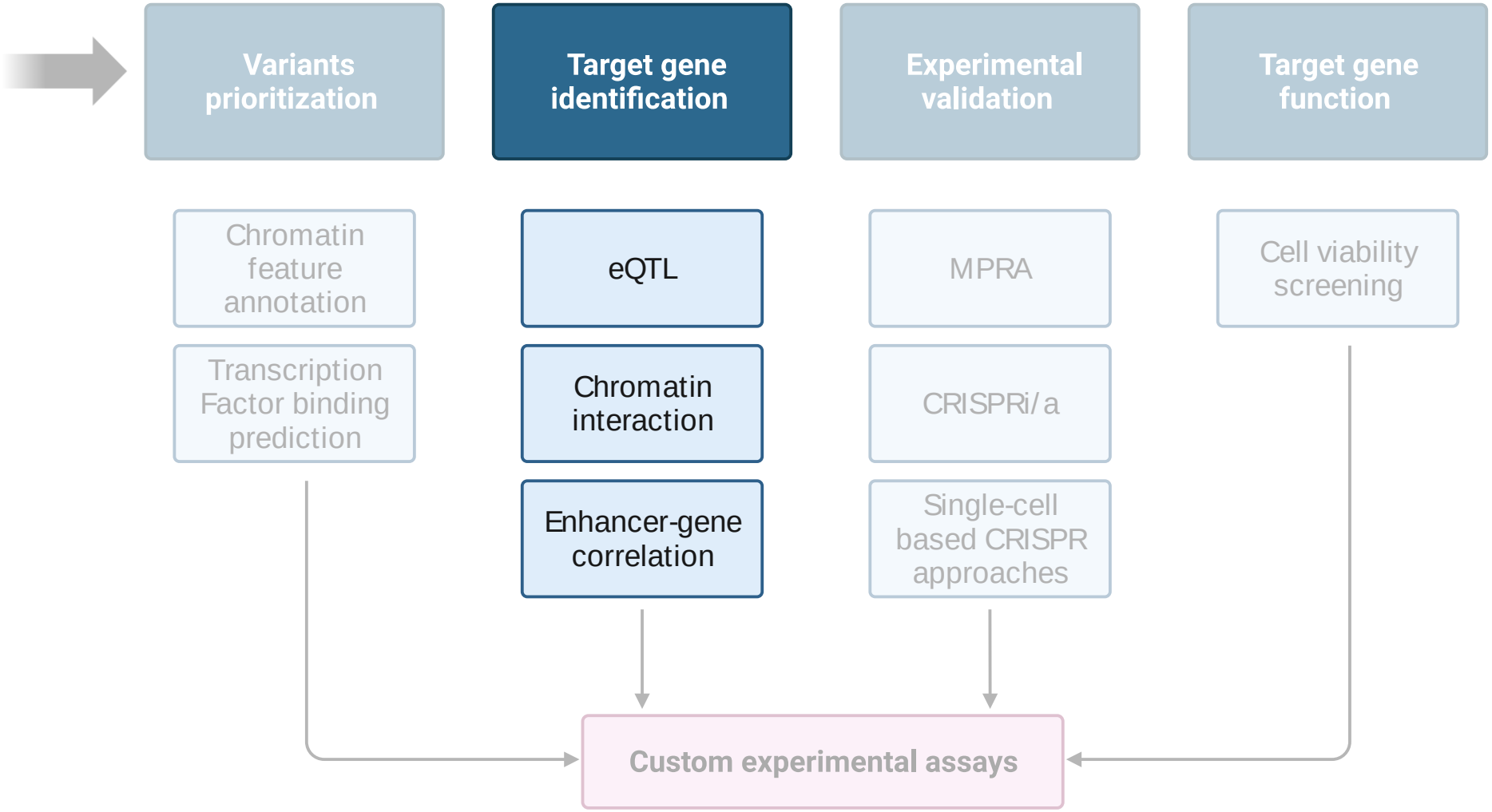
Variant prioritization tools

- SNP look-up tools with LD information
  - Ldlink (https://ldlink.nih.gov/)
  - Haploreg (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php)
  - Regulome DB (functional scores) (https://regulomedb.org/regulome-search/)
  - FORGEdb (functional scores) (https://forgedb.cancer.gov/)

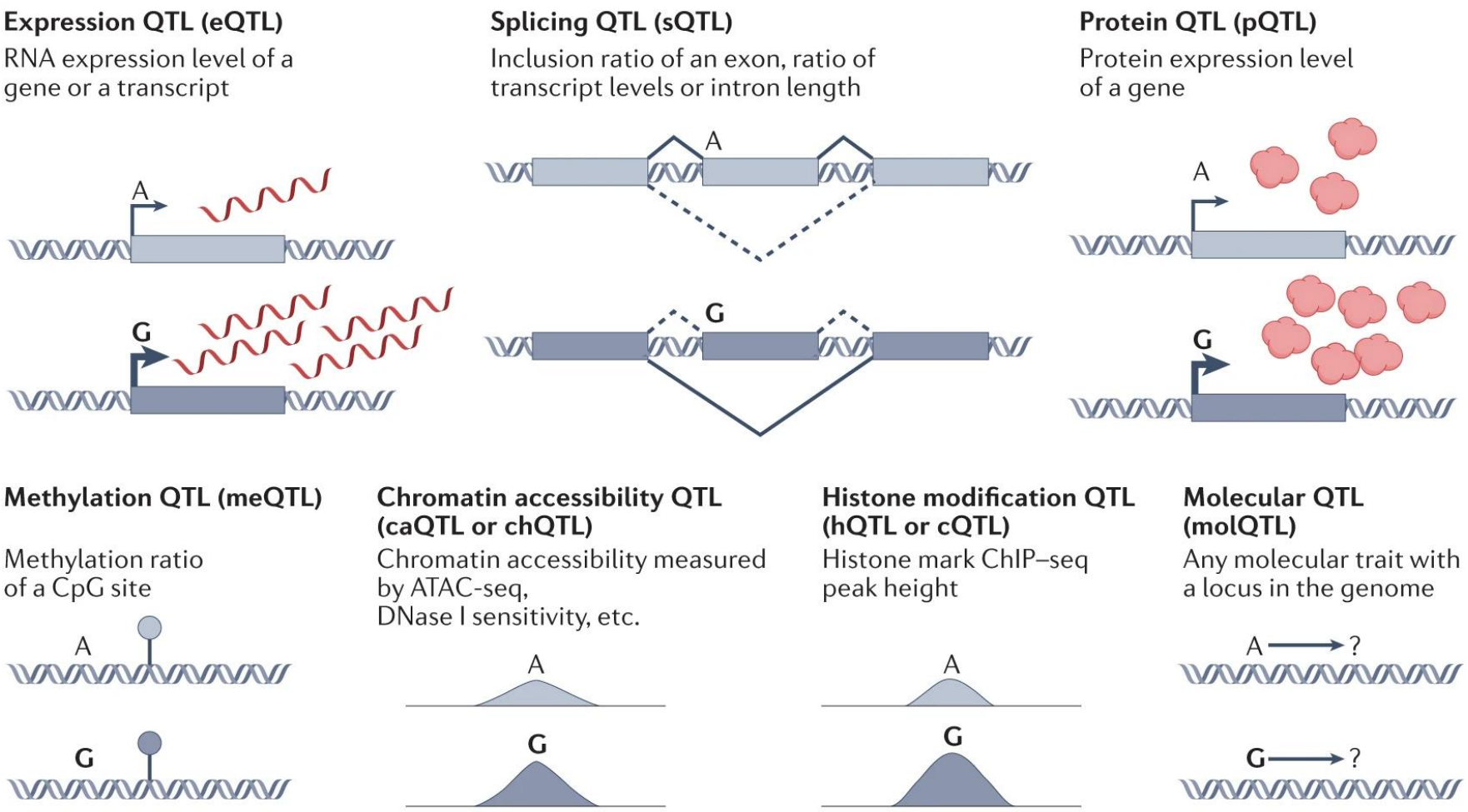
- Public epigenomic resources
  - ENCODE (https://www.encodeproject.org/)
    - http://encodec.encodeproject.org/
    - https://screen.encodeproject.org/
  - RoadMap Epigenomics Project
    - http://www.roadmapepigenomics.org/
  - Data from both are also housed in the UCSC Genome Browser
    - https://genome.ucsc.edu/

- Single cell-level resources
  - http://catlas.org/humanenhancer/#!/

Identification of "causal" variants and target genes



Types of molecular QTLs:



From Auget *et al.*, 2023, Nature Reviews Methods Primers; https://www.nature.com/articles/s43586-022-00188-6

Target gene identification tools

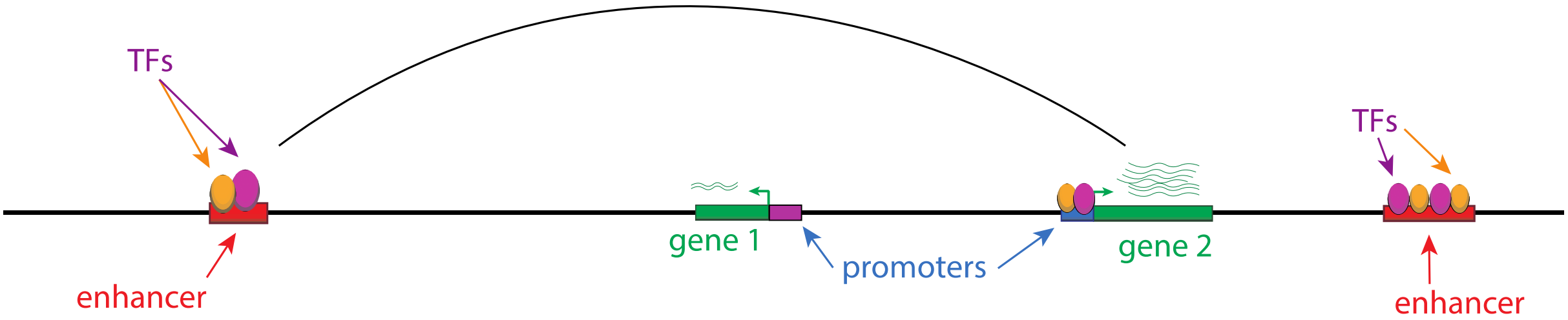
Some eQTL resources:

- GTEx portal (https://gtexportal.org/home/)
- eQTLgen (https://www.eqtlgen.org/)
- psychENCODE (http://resource.psychencode.org/)
- TCGA eQTL (http://gong_lab.hzau.edu.cn/PancanQTL/)
- Single cell eQTLs (https://eqtlgen.org/sc/)

QTLs + Colocalization:

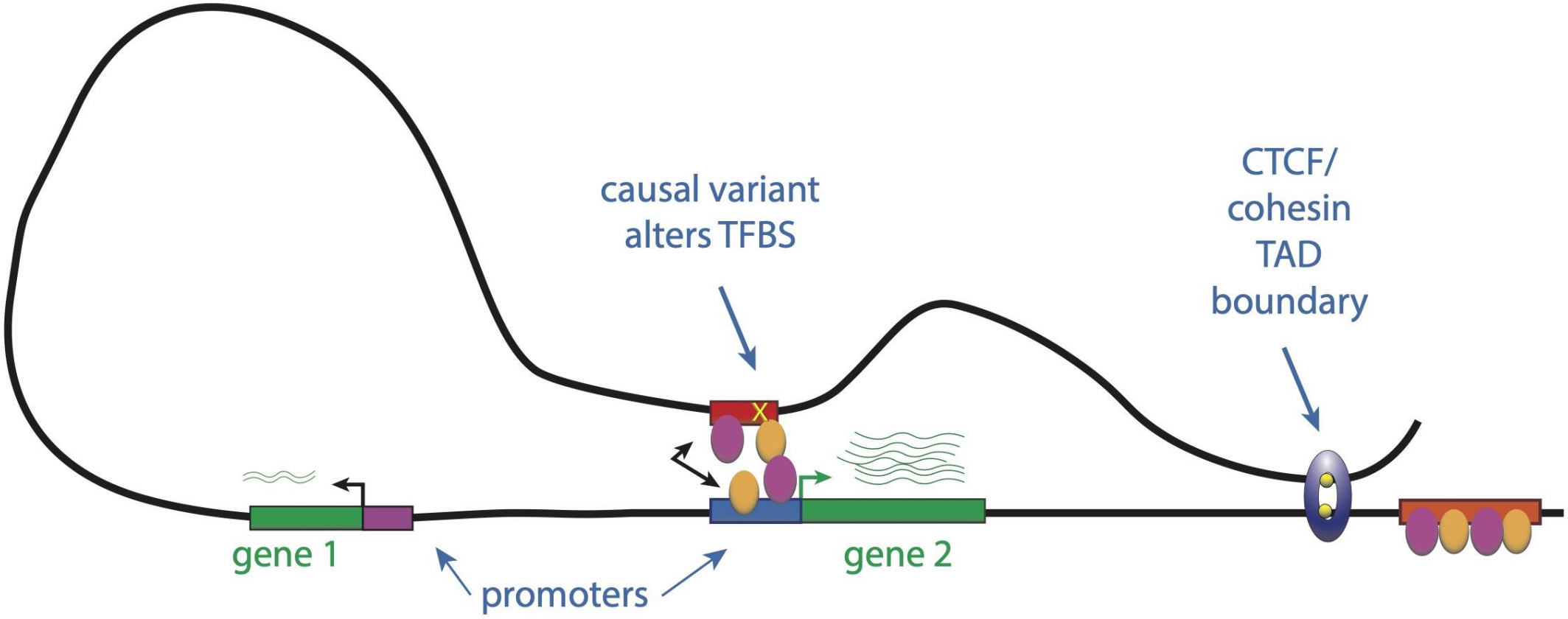
- ezQTL (https://analysistools.cancer.gov/ezqtl/)
- LocusFocus (https://locusfocus.research.sickkids.ca/)
- LocusCompare (http://locuscompare.com/)

Enhancer-gene chromatin interactions



The genome is not linear!

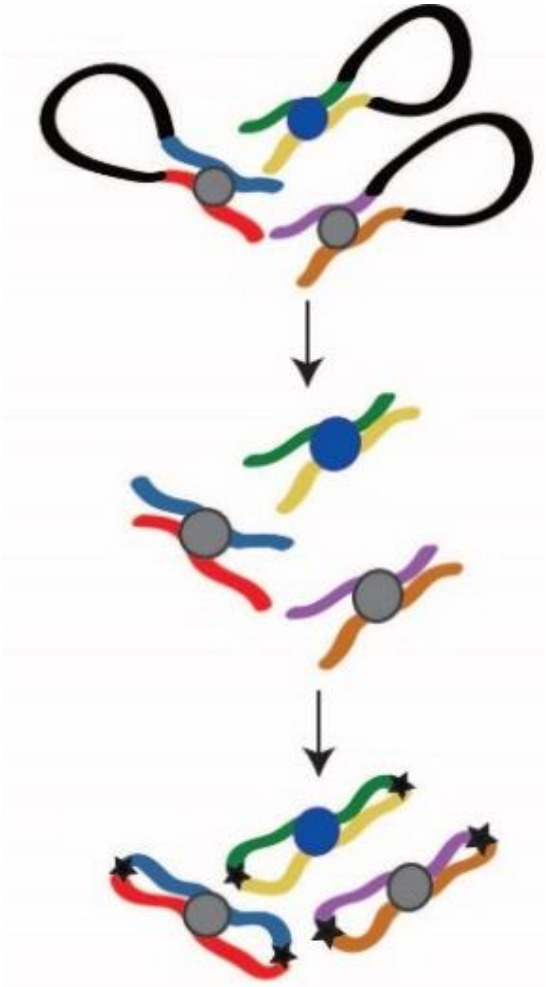
The non-linear genome



Features of regulatory enhancer-gene interactions:

- enhancer and gene in close physical proximity
- enhancer activity (chromatin openness, histone marks) correlated with gene expression

Chromatin conformation-based methods to identify enhancer-gene interactions



DNA crosslinking

Restriction enzyme digestion

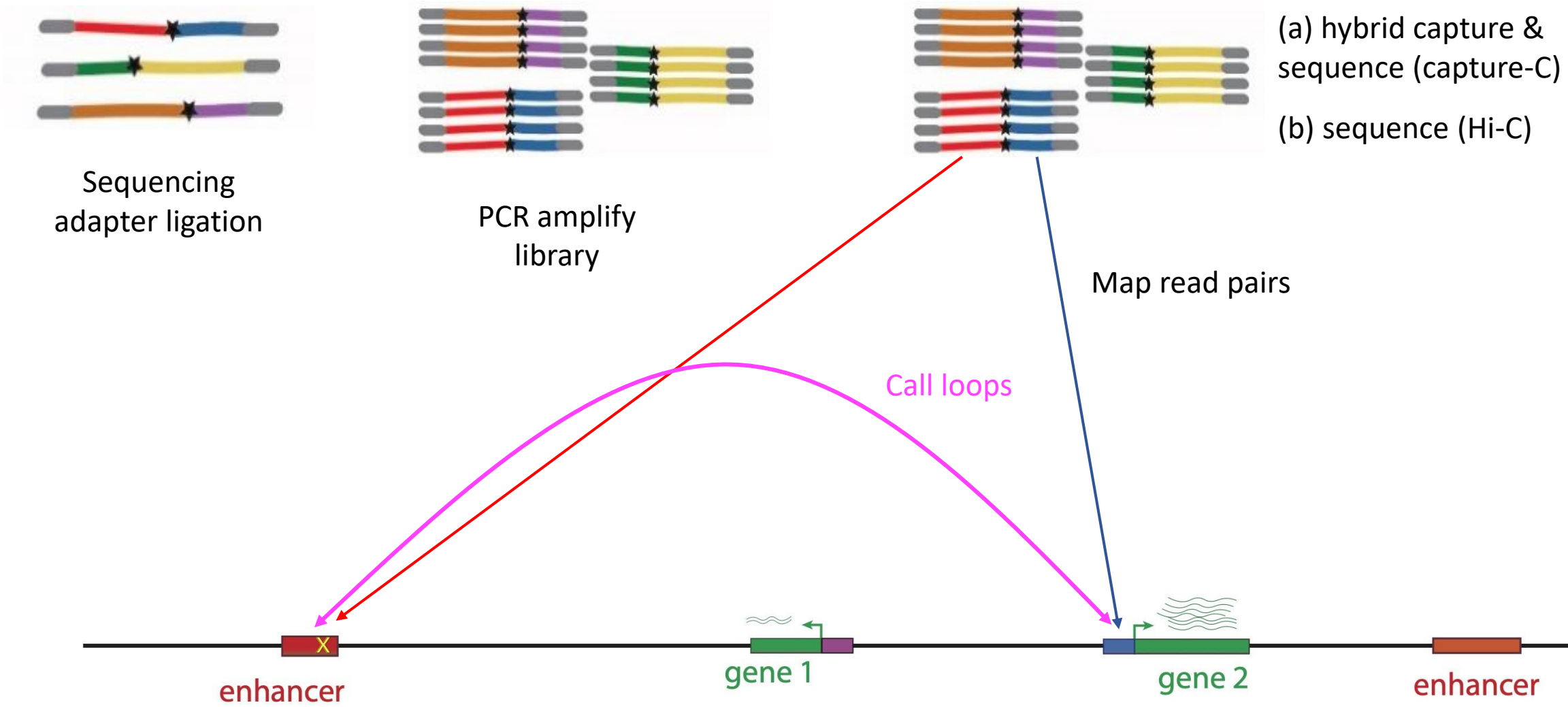
Biotin labeling and ligation

Reverse crosslink, protein digestion, shearing

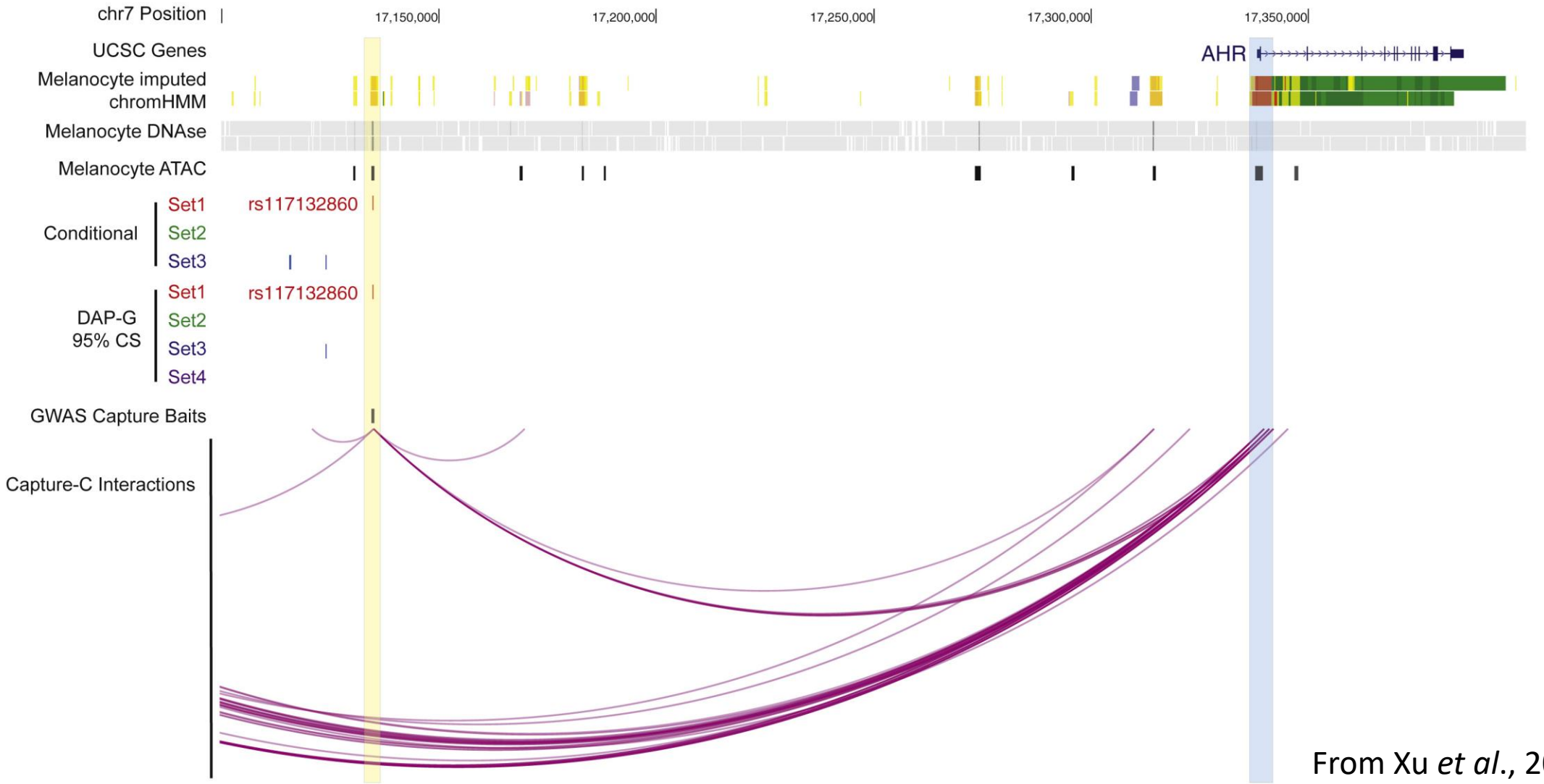
Capture

Sequencing adapter ligation

Sequencing-based chromatin conformation assay (HiC/Capture-C)

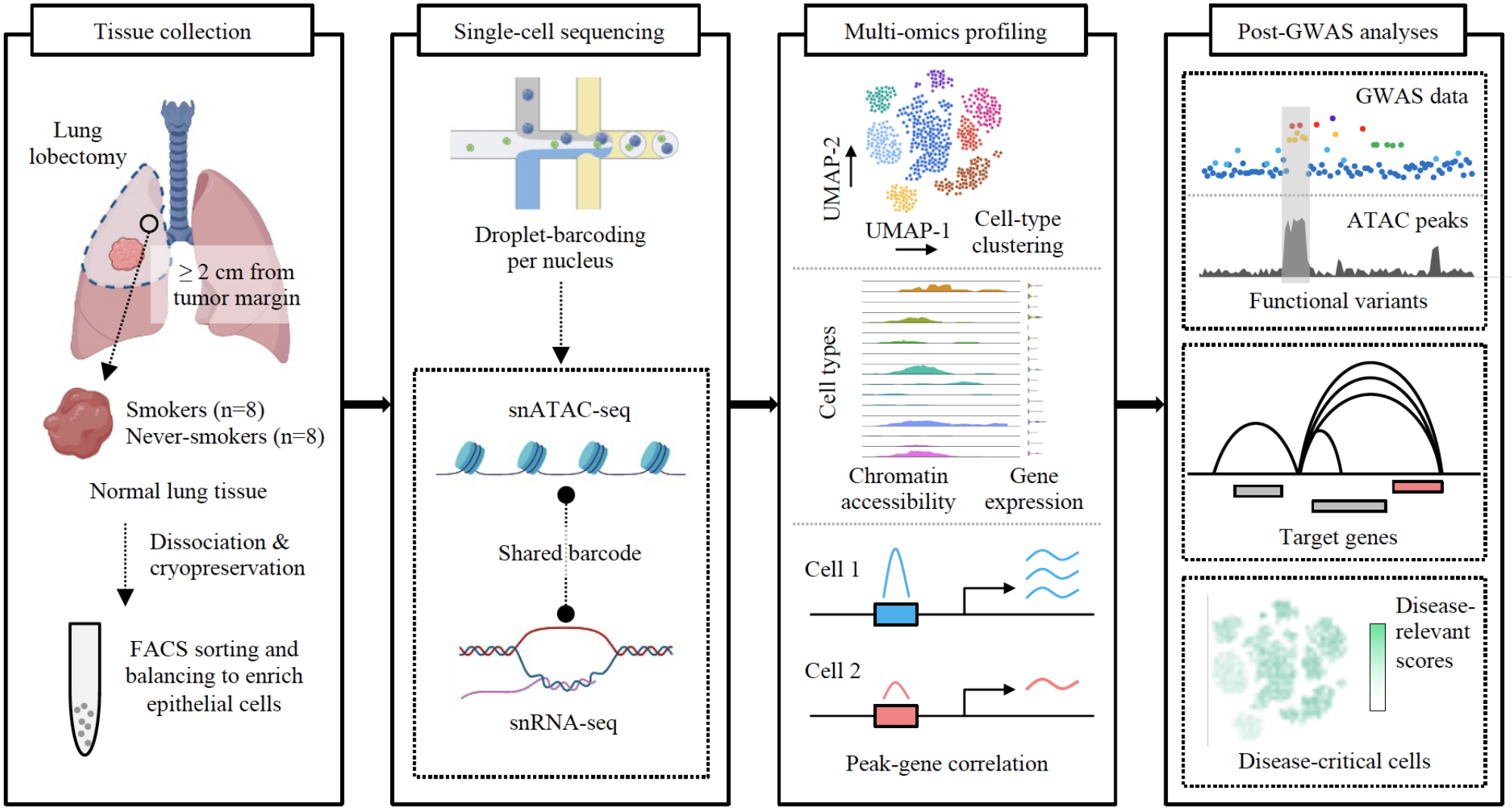


Example: capture-C in melanocytes identifies *AHR* as a target for a functional melanoma risk variant

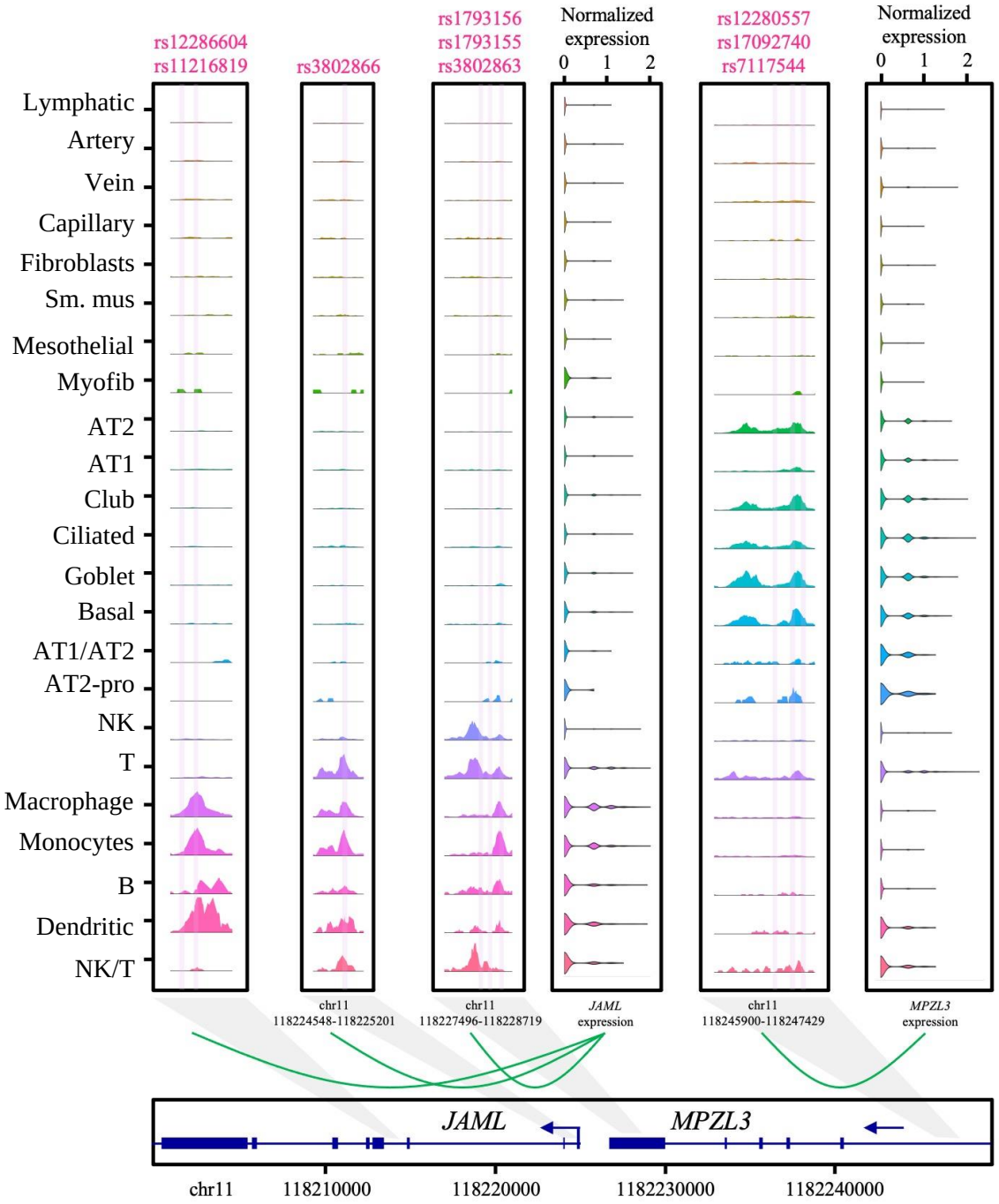


From Xu *et al*., 2021, *AJHG*

Single-cell ATACseq/RNAseq for variant-gene connection



Cell type-specific target genes in a single locus



Long et al, *bioRxiv* 2023.09.25.559336

Public resources for chromatin interaction and enhancer-gene correlation

Bulk cell or tissue data:

3D Genome Browser (http://3dgenome.fsm.northwestern.edu/)

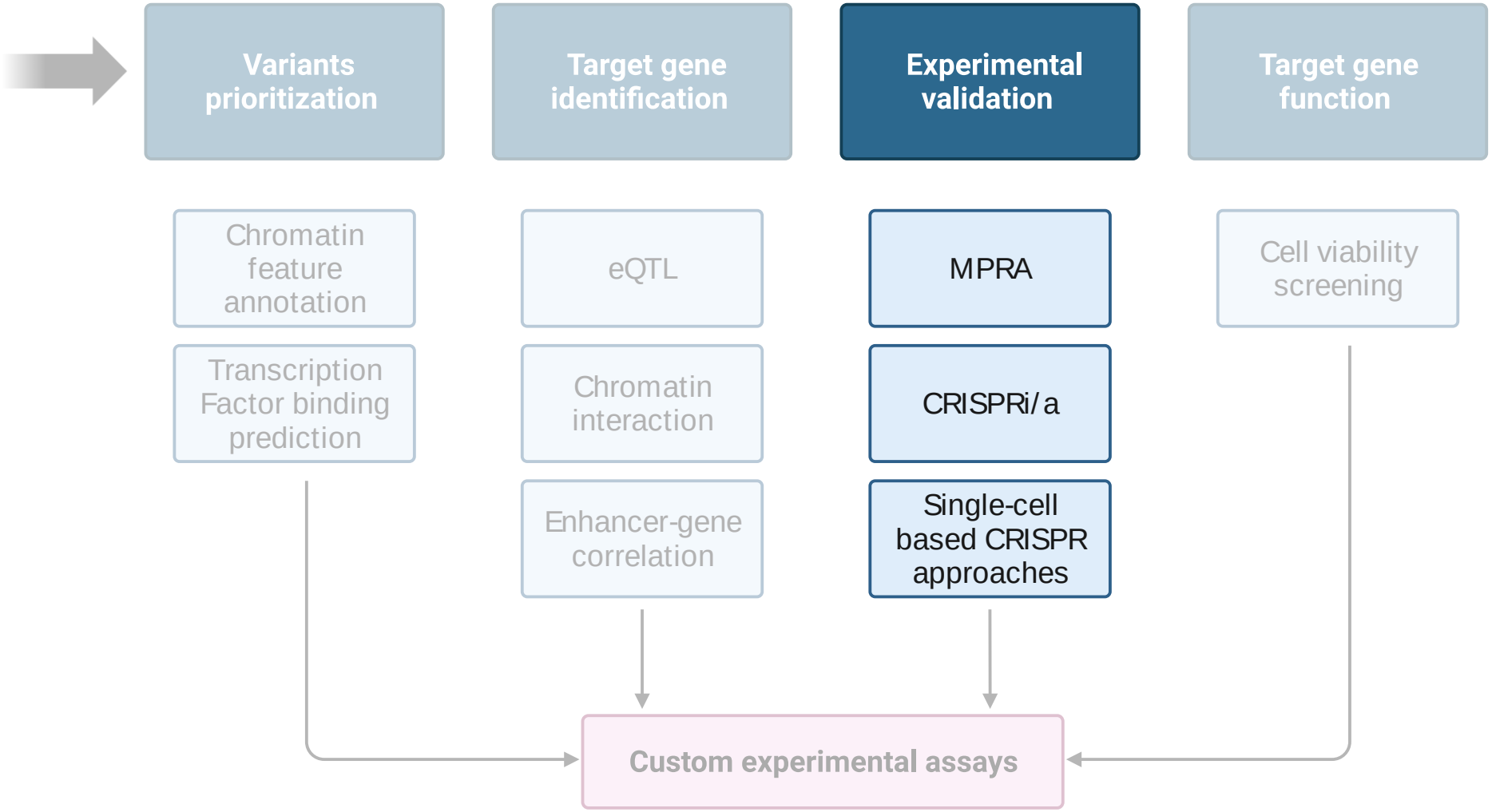
- Incorporates chromatin conformation data from many cell types
- Can assess interactions for specific genetic variants

Single-cell data:

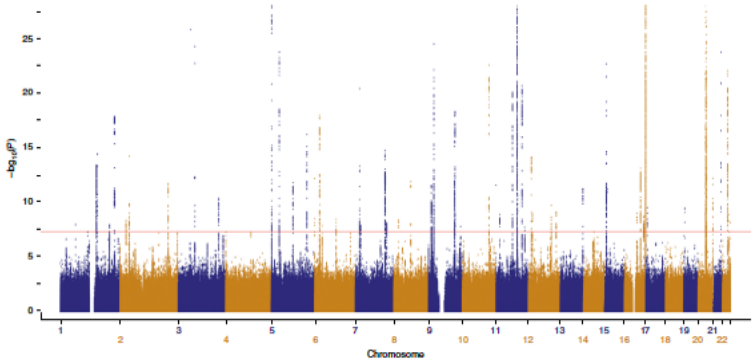
Activity-by-Contact (ABC) Browser (https://www.engreitzlab.org/resources)

- Incorporates both single-cell level enhancer-gene correlation and chromatin contact data

Identification of "causal" variants and target genes



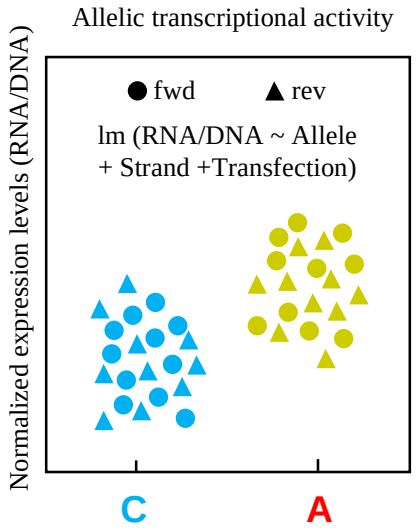
Massively Parallel Reporter Assays (MPRA) can test allelic transcriptional activity GWAS variants

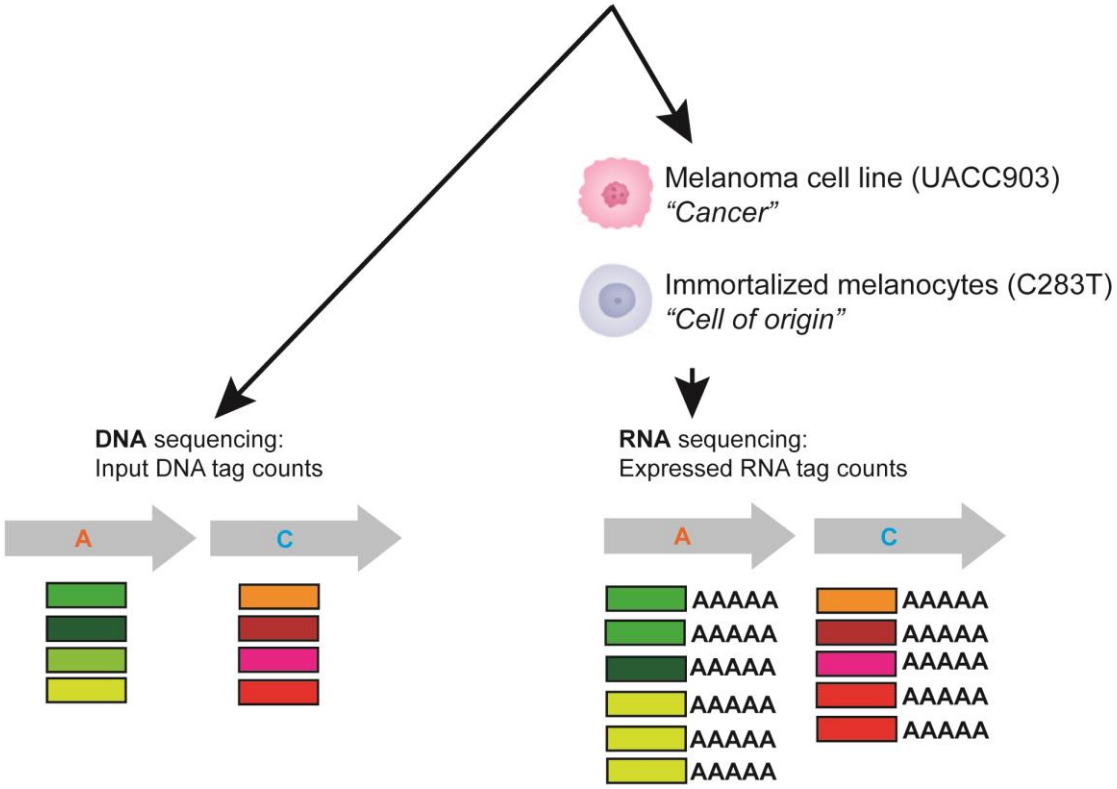


Melanoma GWAS
(Landi *et al.*, 2020 *Nature Genetics*)
36,760 cases/ 375,188 controls

54 loci
$R^2 > 0.8$
LLR < 1:1000

1,992 variants
~244,000 oligos





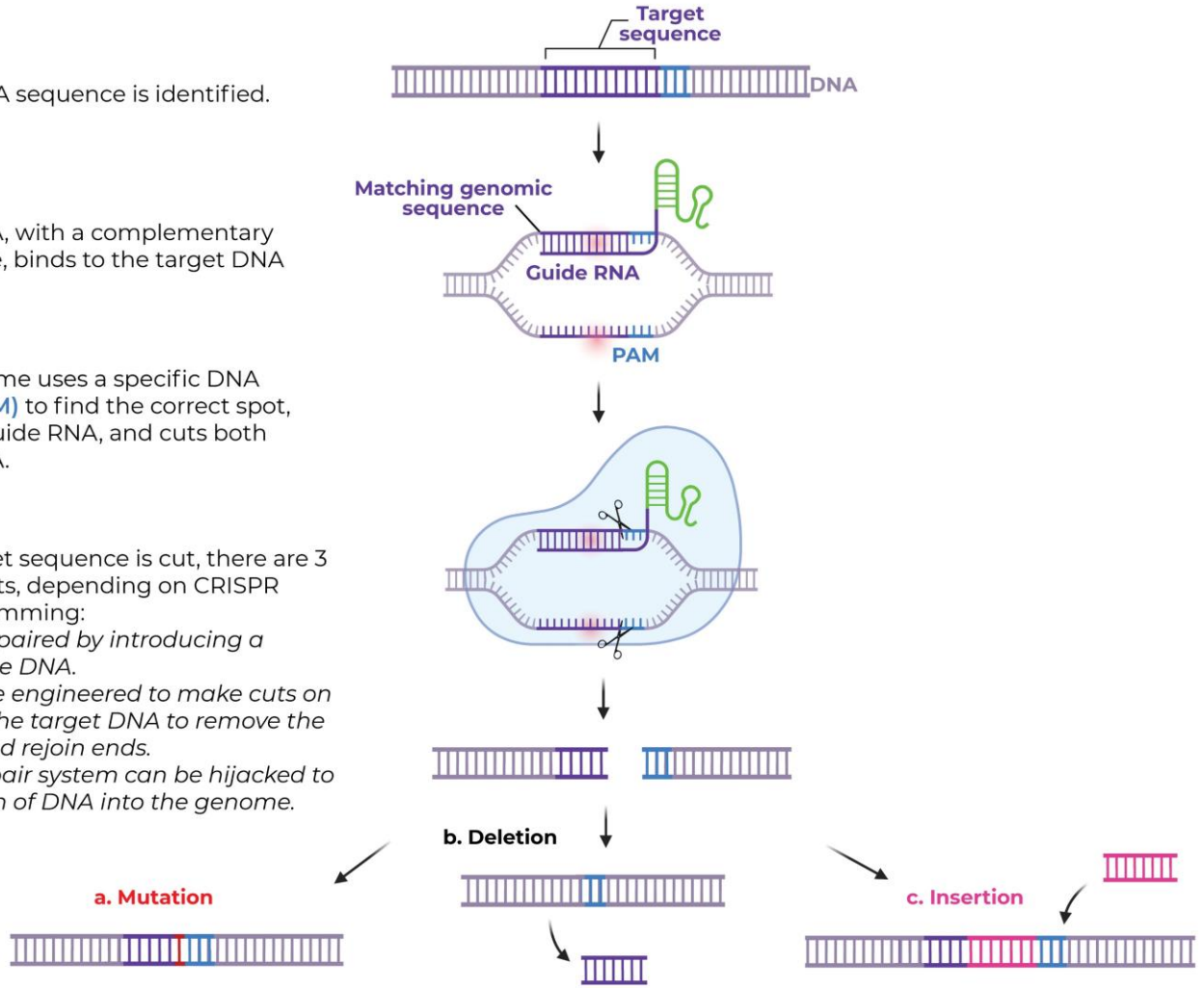
Long and Yin…Choi, *AJHG,* 2022

MPRA-based resources

- MPRA
  - https://www.varianteffect.org/resources (region-specific, limited)

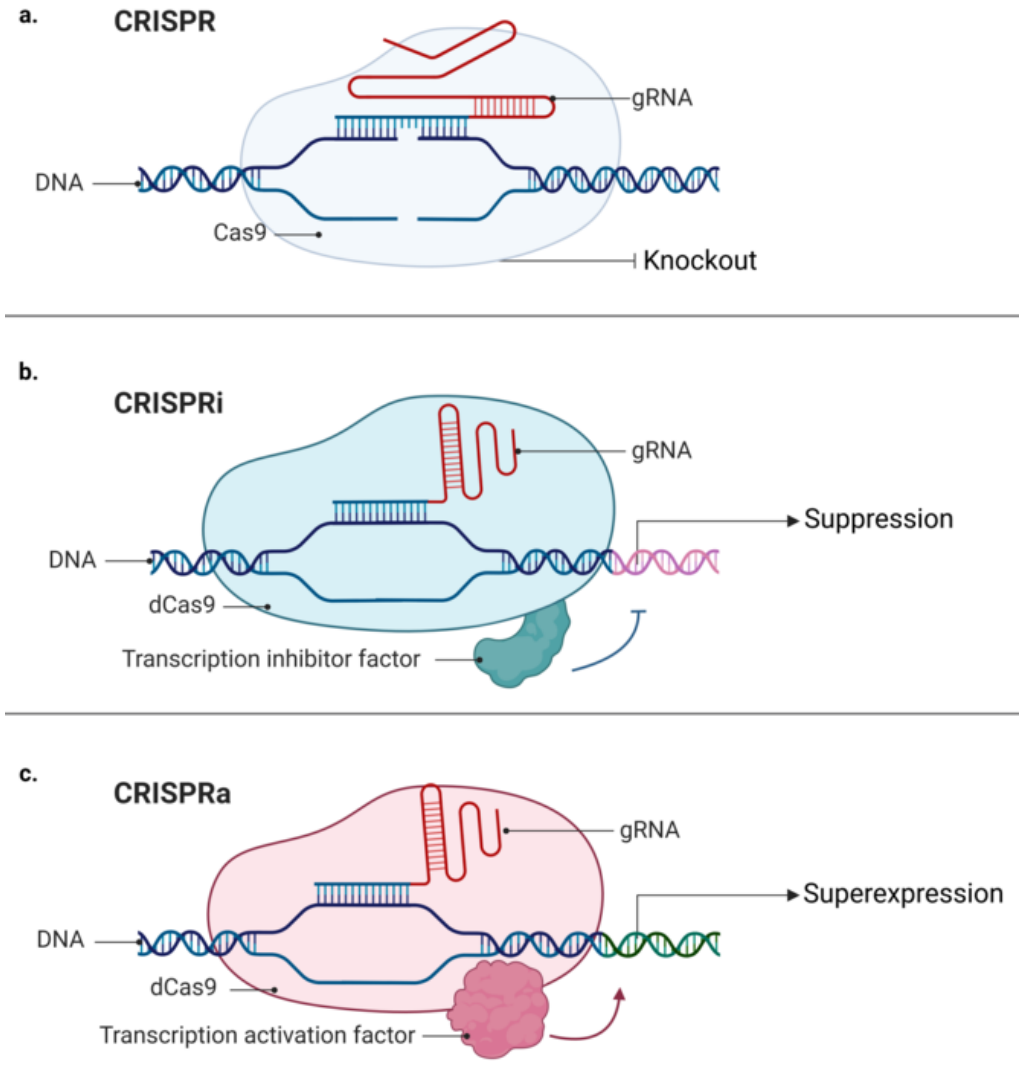
CRISPR as a tool for discovery and validation

1. The target DNA sequence is identified.
2. The guide RNA, with a complementary DNA sequence, binds to the target DNA sequence.
3. The Cas9 enzyme uses a specific DNA sequence (PAM) to find the correct spot, binds to the guide RNA, and cuts both strands of DNA.
4. Once the target sequence is cut, there are 3 potential results, depending on CRISPR system programming:
   a. The cut is repaired by introducing a mutation in the DNA.
   b. Enzymes are engineered to make cuts on either side of the target DNA to remove the target DNA and rejoin ends.
   c. The DNA repair system can be hijacked to insert a section of DNA into the genome.



https://www.umassmed.edu/rti/biology/crispr-cas9/

CRISPR, CRISPR-inhibition, CRISPR-activation



Knock-out:

- SNP/regulatory region
- Gene (introduce truncating mutations)

Inhibits:

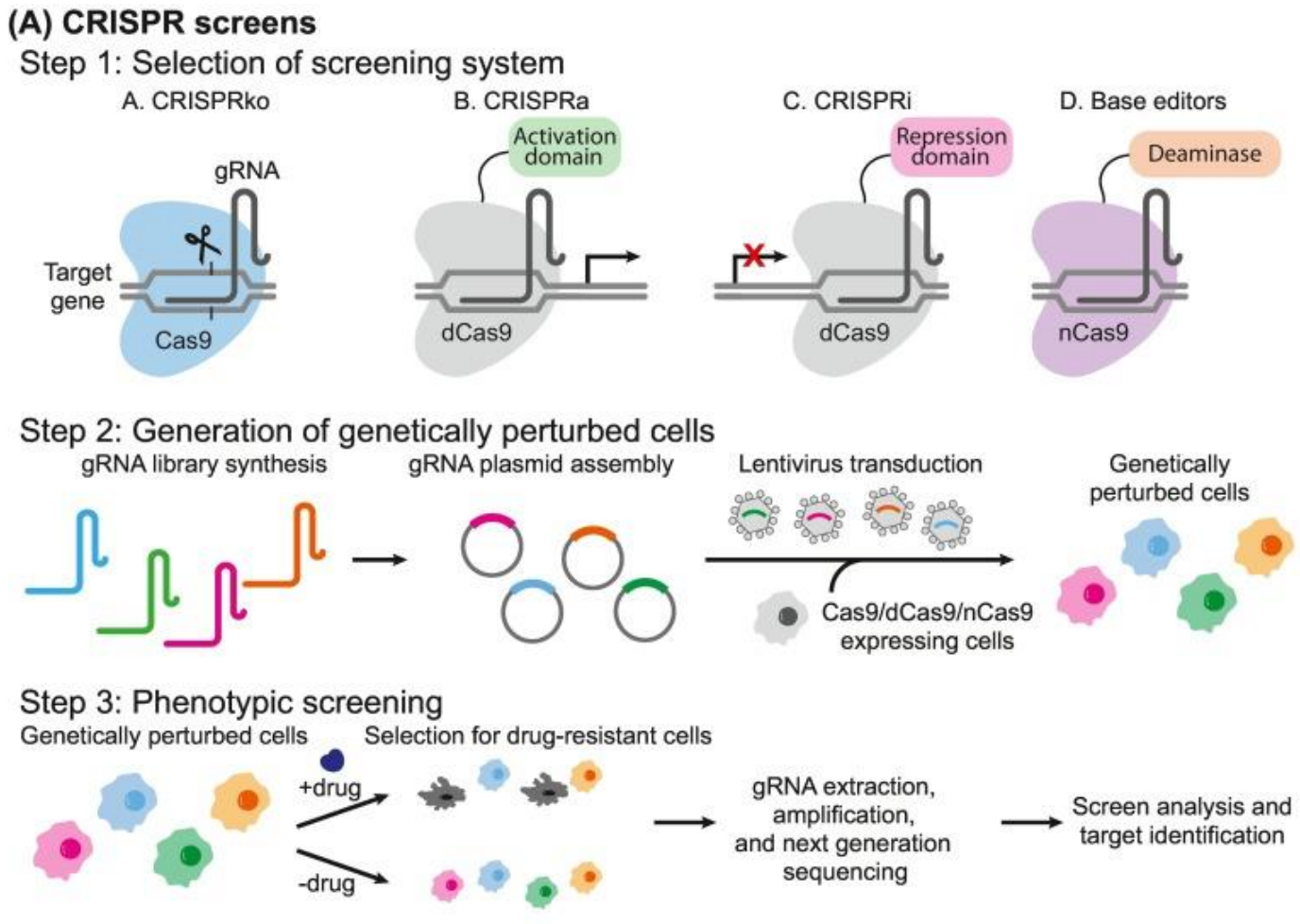
- SNP/regulatory activity
- Gene (targeting gene promoters)

Activates:

- SNP/regulatory activity
- Gene (targeting gene promoters)

From Almeida *et al*., 2022, *Genes*

Discovery: CRISPR screens



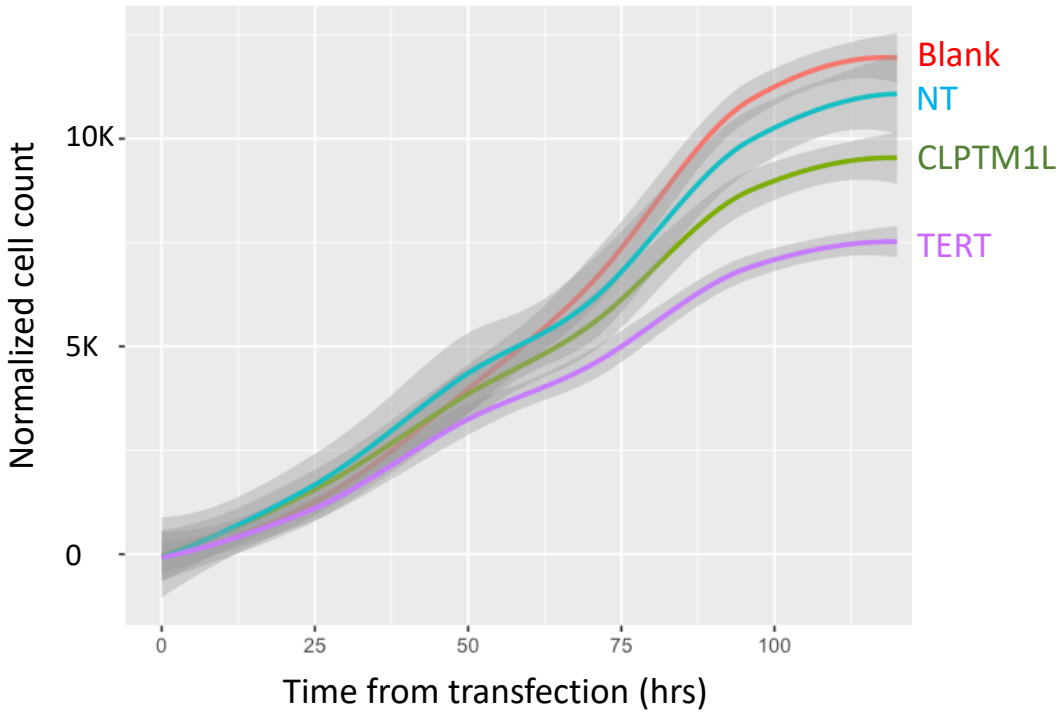
Modell et al., Trends Pharmacol Sci 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9726229/pdf/nihms-1854548.pdf

CRISPRi Screen (variant based) example – chr5p15.33

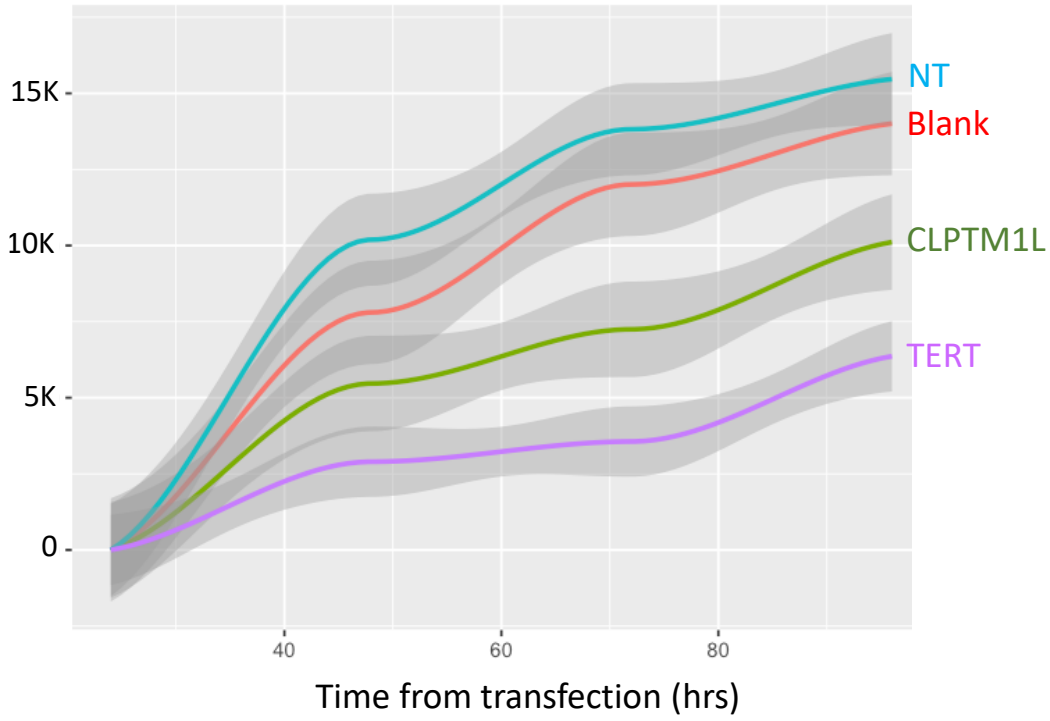
Hypothesis: functional variants underlying chr5p15.33 signals mediate risk via noncoding regulatory effects on *TERT* and/or *CLPTM1L* expression leading to changes in cellular growth

Approach: CRISPR and MPRA screens to target all known chr5p15.33 signals simultaneously

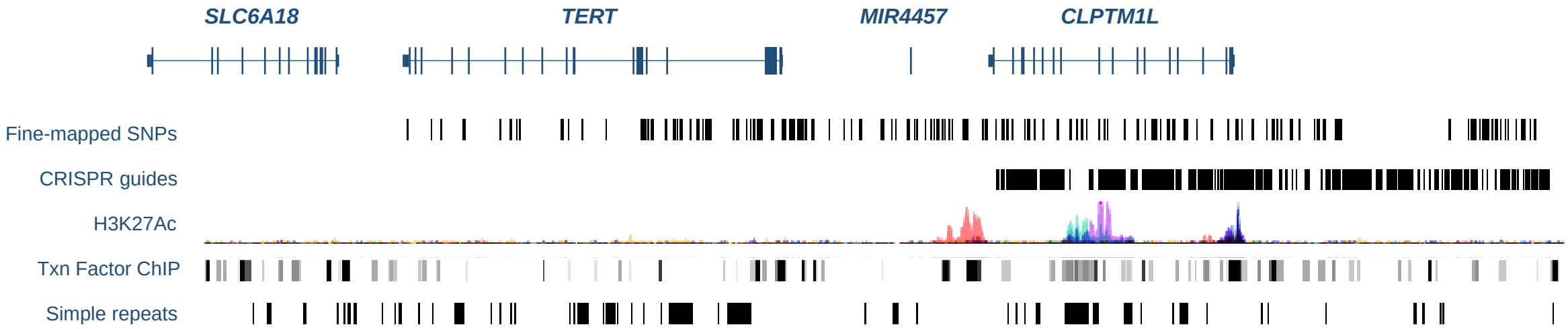
PANC-1 cells

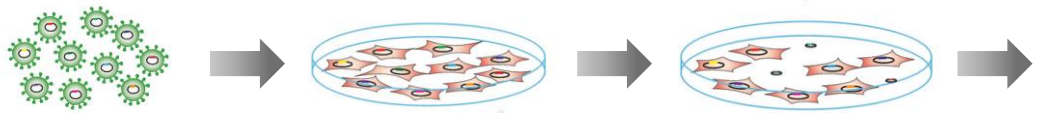


MiaPaCa-2 cells



Tiled CRISPRi screen across 10 GWAS signals at chr5p15.33





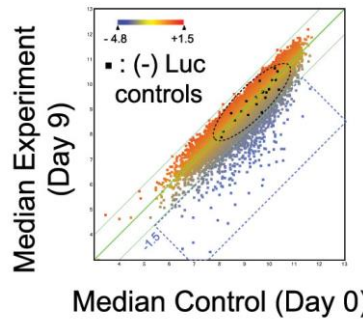
Packaged Pooled Lentiviral sgRNA Library

Transduced cells

Selection

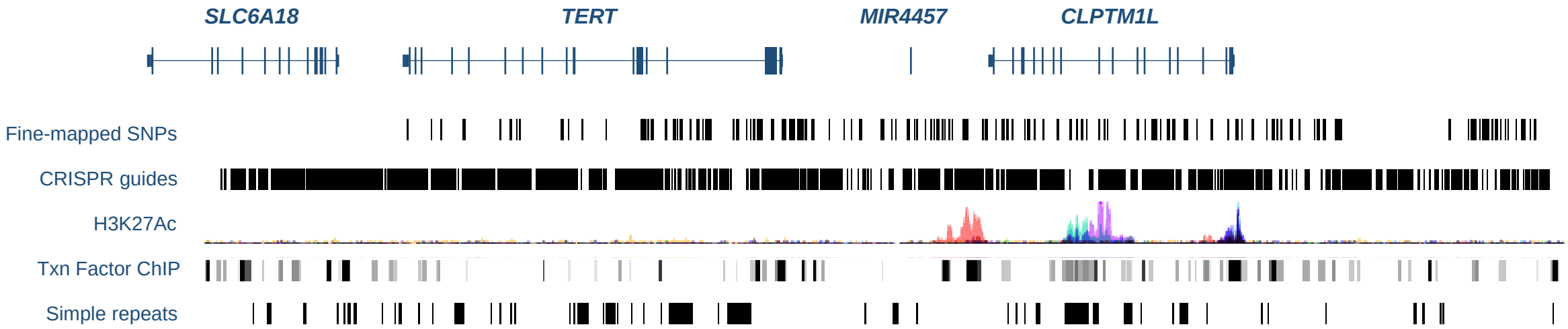
Harvest at 0, 7, 14, 21 days

Amplification & Next-Gen sgRNA sequencing

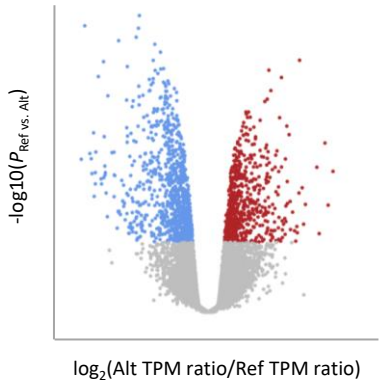


Analysis of sgRNA depletion (CRISPRi) & enrichment (CRISPRa)

Massively Parallel Reporter Assay (MPRA) screen across 5p15.33



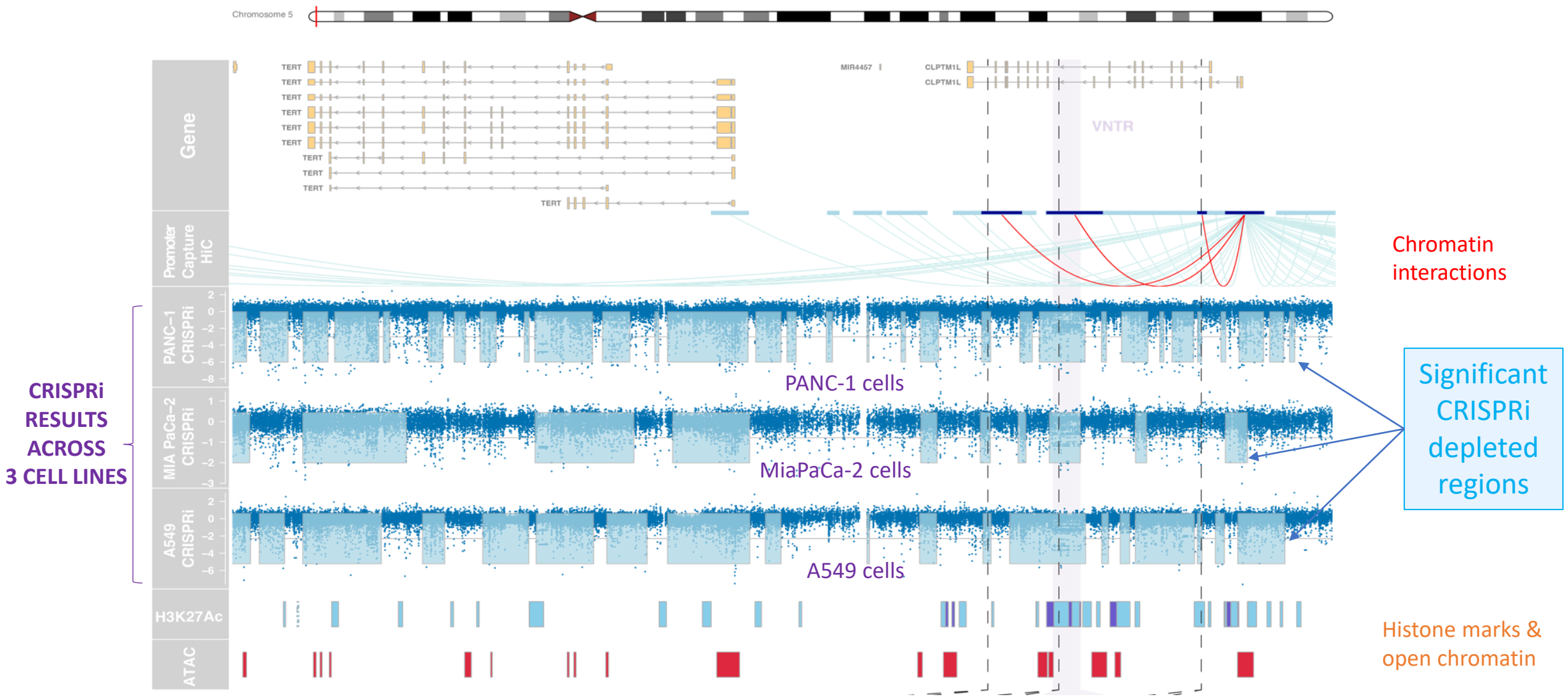
MPRA



Assess allele specific gene regulatory effects across all signals

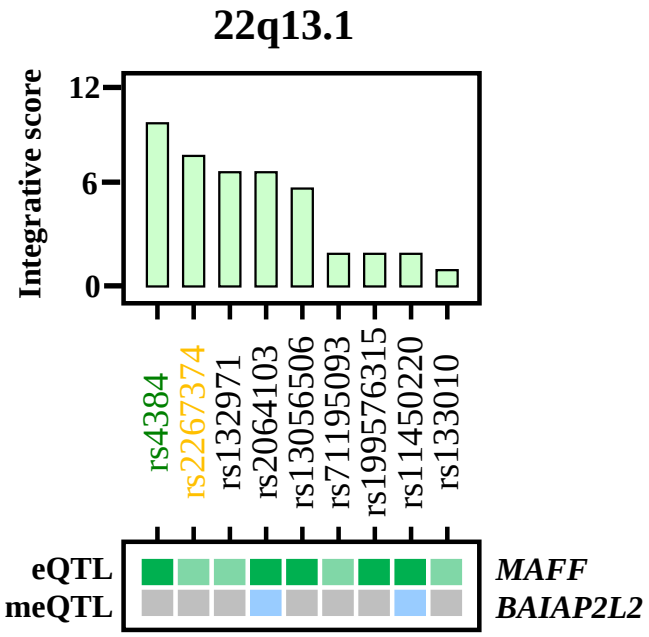
- 175 fine-mapped variants
- ~60,000 reporter constructs
- 2-3 cell lines each used for pancreatic cancer, melanoma, bladder and lung cancer

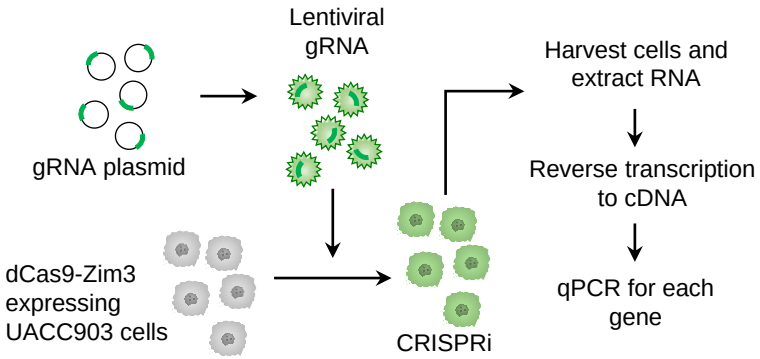
Chr5p15.33 - CRISPRi screen results in pancreatic and lung cancer cells

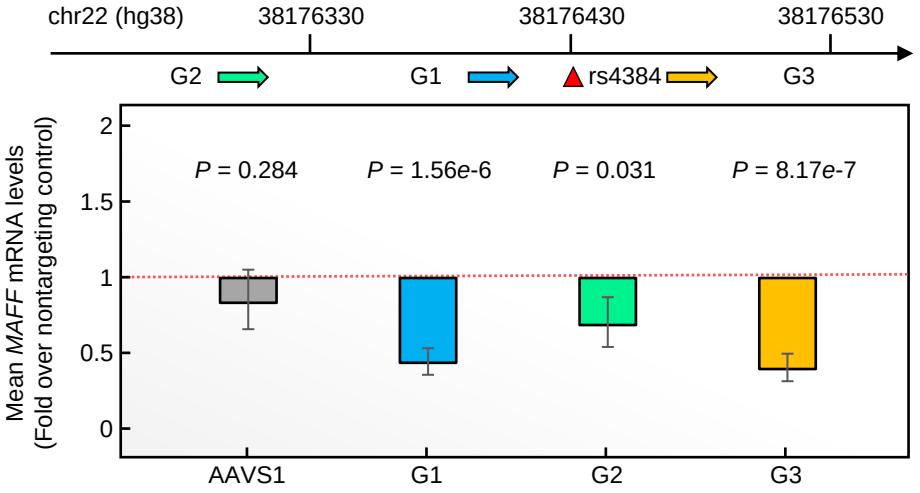


➔Next steps: individual CRISPRi validation for 10-30 SNPs to assess effects on *TERT* and *CLPTM1L* expression

Validation: CRISPRi can validate variant gene connection

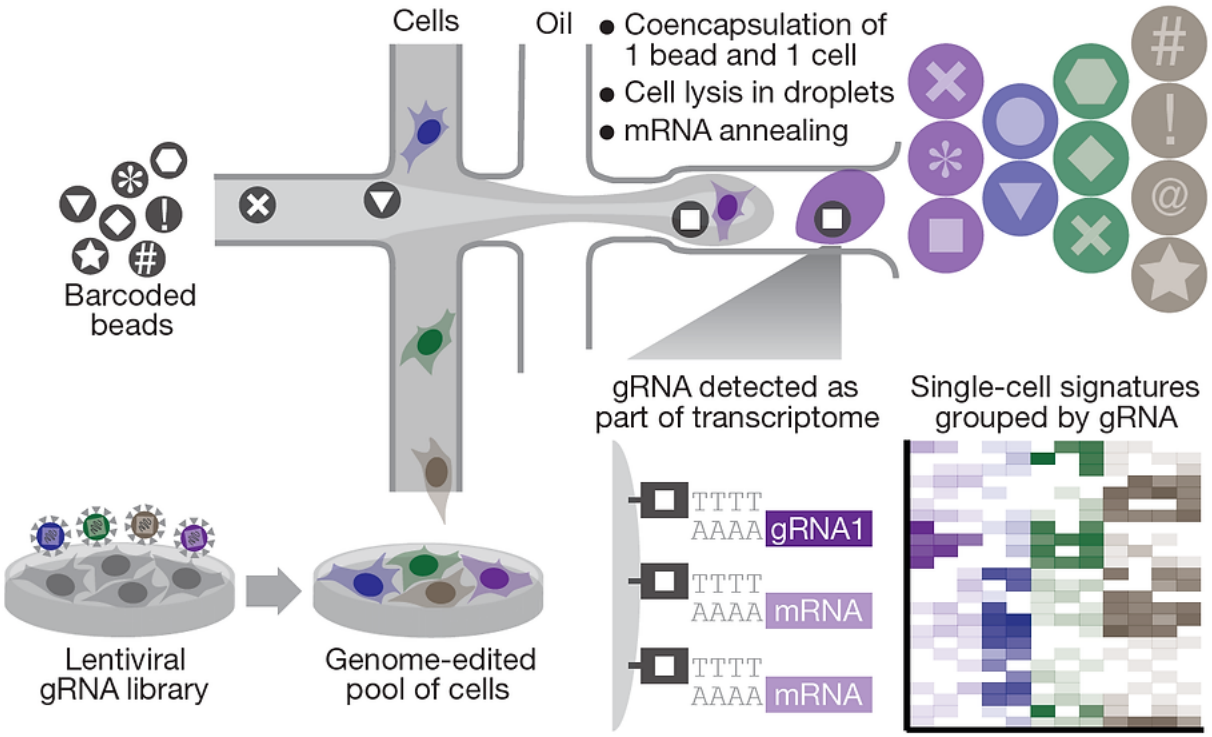






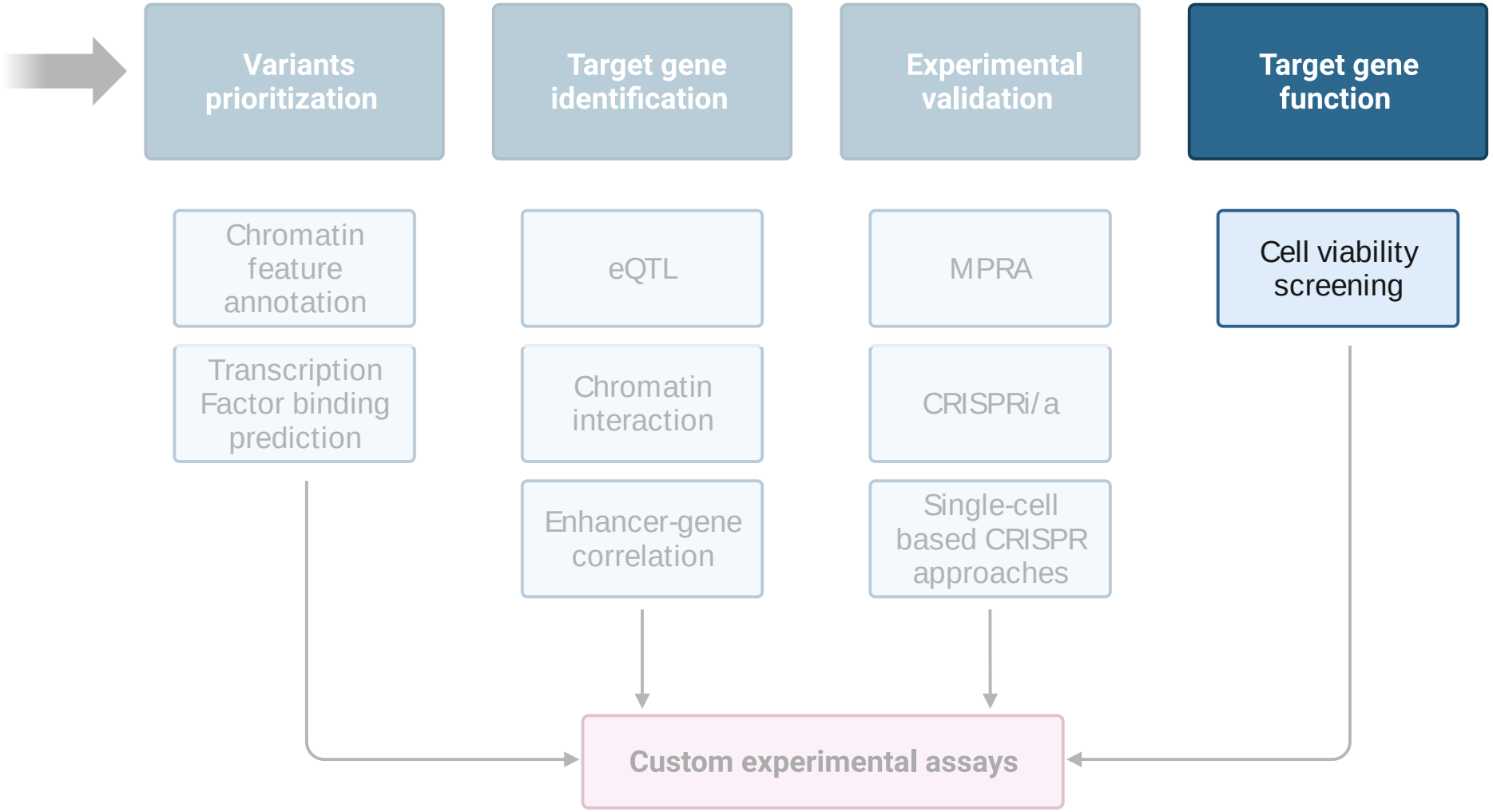
Long and Yin…Choi, *AJHG,* 2022

CROP-seq (CRISPR screening with scRNAseq readout)



https://www.bocklab.org/resources/crop-seq

Identification of "causal" variants and target genes



Target Gene Function

High-throughput (typically CRISPR) methods:

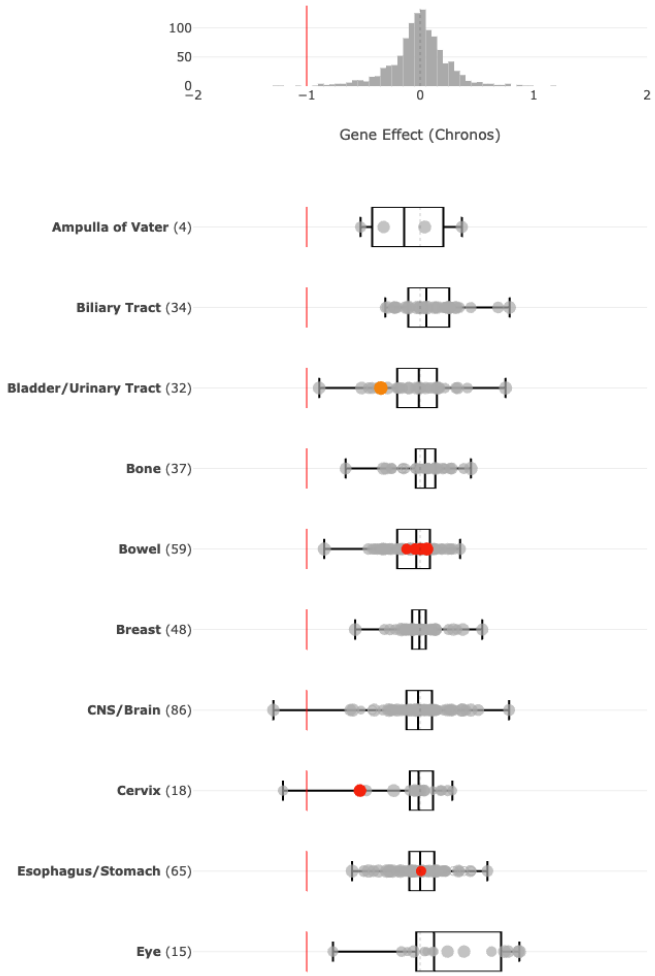
- Publicly-available screen data for effect on cell growth/viability
  - DepMap: Cancer Dependency Map Project (https://depmap.org/portal/)
- Custom pooled CRISPR-based screens -> need selectable phenotype

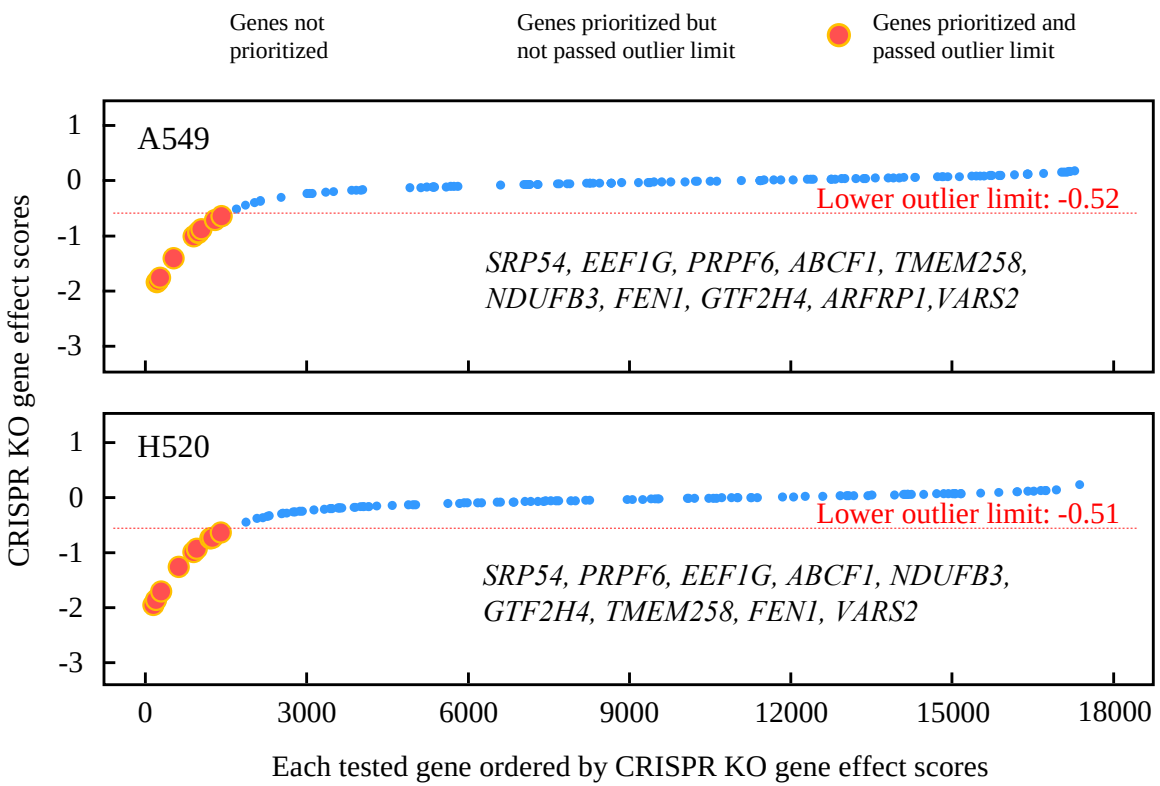
Individual gene assessment:

- Typically over-expression/knockdown and assess cellular phenotypes
- There is no one-size-fits-all assay
- Cell type and cell context matters!

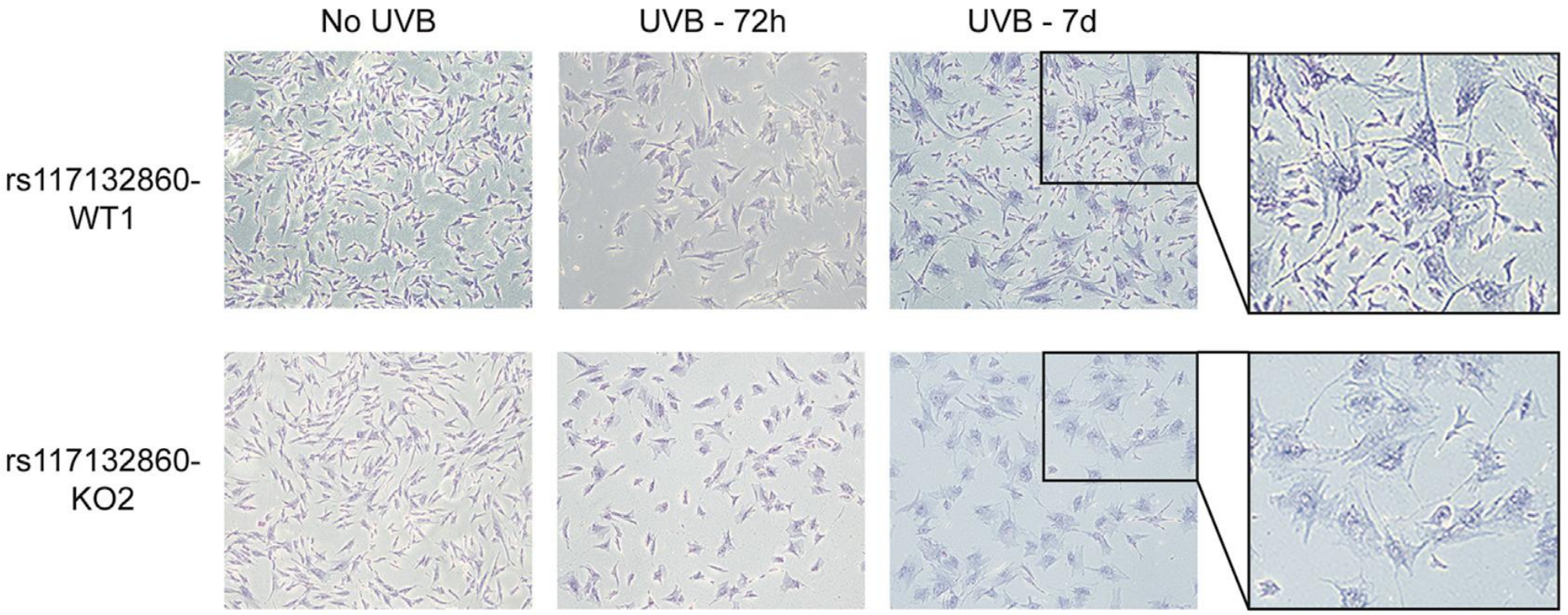
Publicly available CRISPR viability screen data

- DepMap cell viability screening (https://depmap.org/portal/)



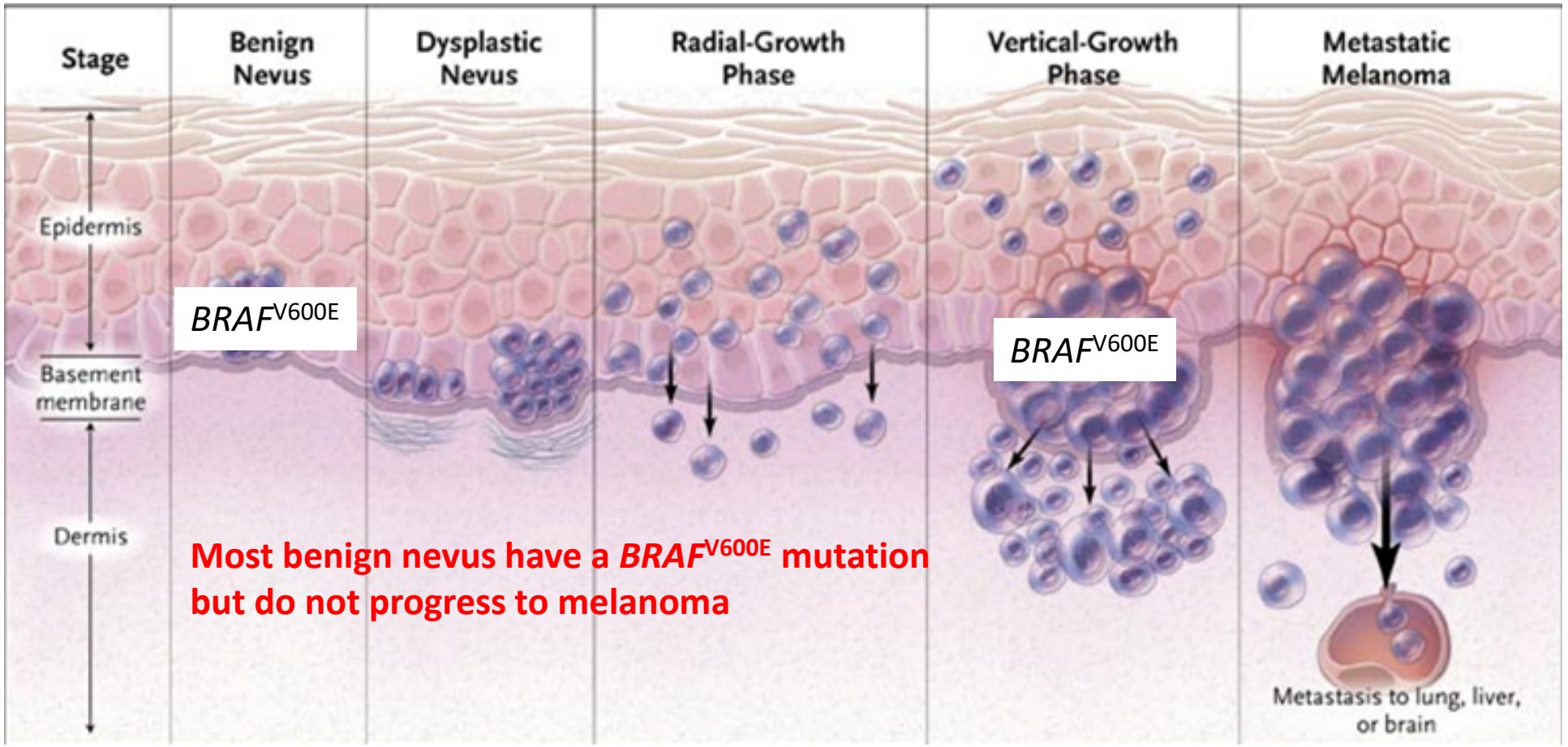


UV exposure-specific function of *AHR* melanoma-risk variant



Xu et al., 2021, *AJHG*

BRAF^{V600E}-induced senescence is a barrier to melanoma progression

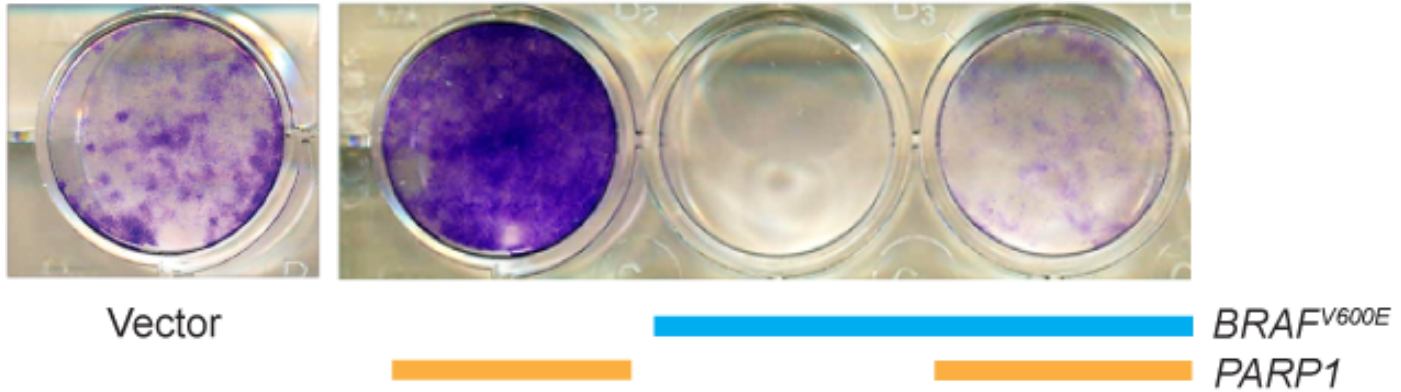


Miller *et al, NEJM* 2006

PARP1 function: promotes melanocyte growth, bypass senescence, and transforms immortalized melanocytes

✓Does PARP1 help bypass *BRAF*^{V600E}-induced senescence in primary melanocytes?

Primary melanocytes growing attached to culture dish
More purple: more cell growth

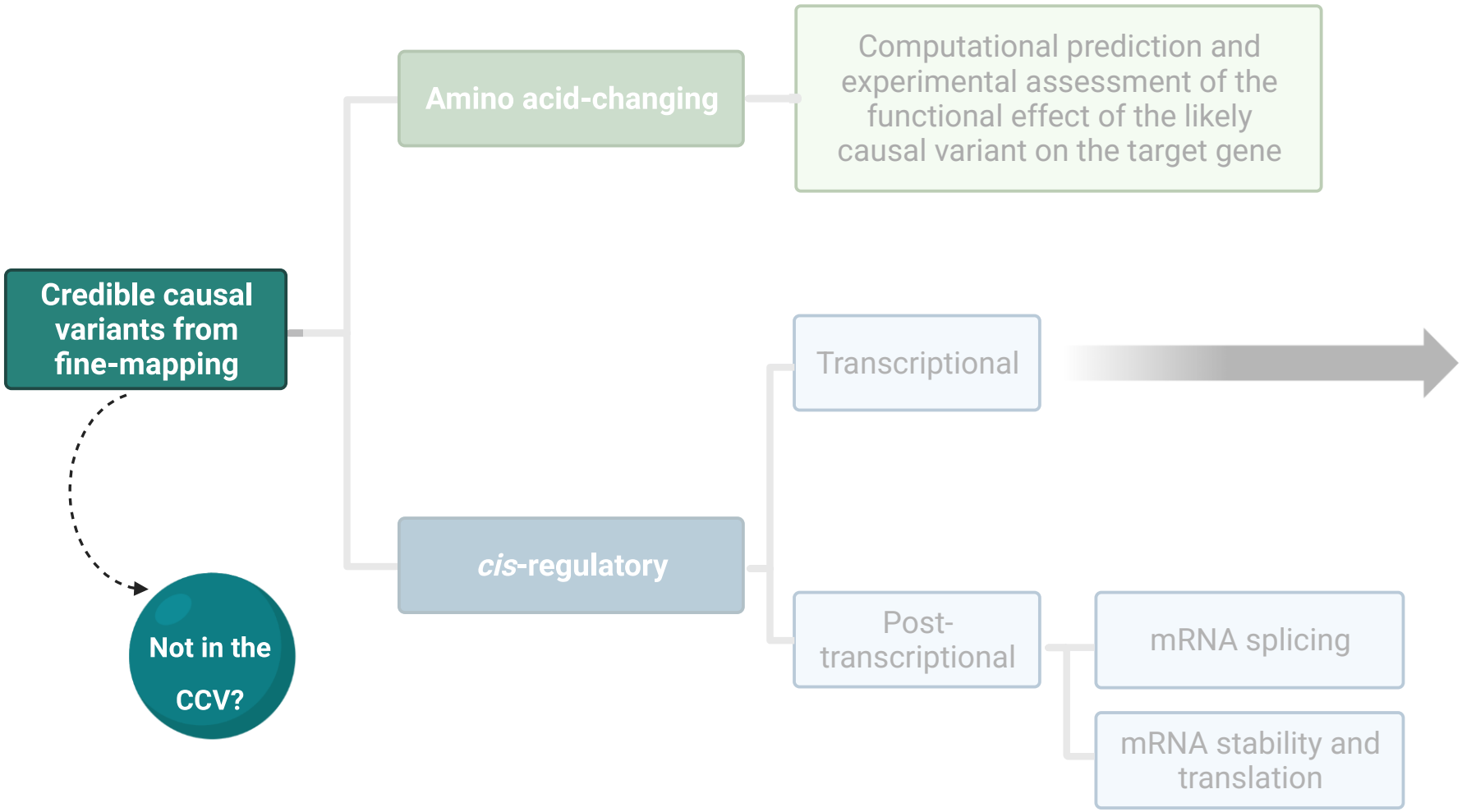


✓Does PARP1 promote malignant transformation of immortalized melanocytes?

Soft agar assay
(tumorous potential by growth in suspension, unattached to any matrix
More dots of colonies: more transformation)

Mai Xu (LTG)

Mechanisms of variant function



From statistics to function

Genetic variants directly genotyped

Genetic variants imputed based on various panels and tools

Which of the available variant(s) can statistically account for the signal?

Predicted functionality of the candidate variants

Assumption:
We know all the variants in the region, they can be scored or well imputed in relevant sets

Examples

Assumption:
statistical association =functionality=causality;
Single vs. multiple variants (haplotypes) that might need to work together

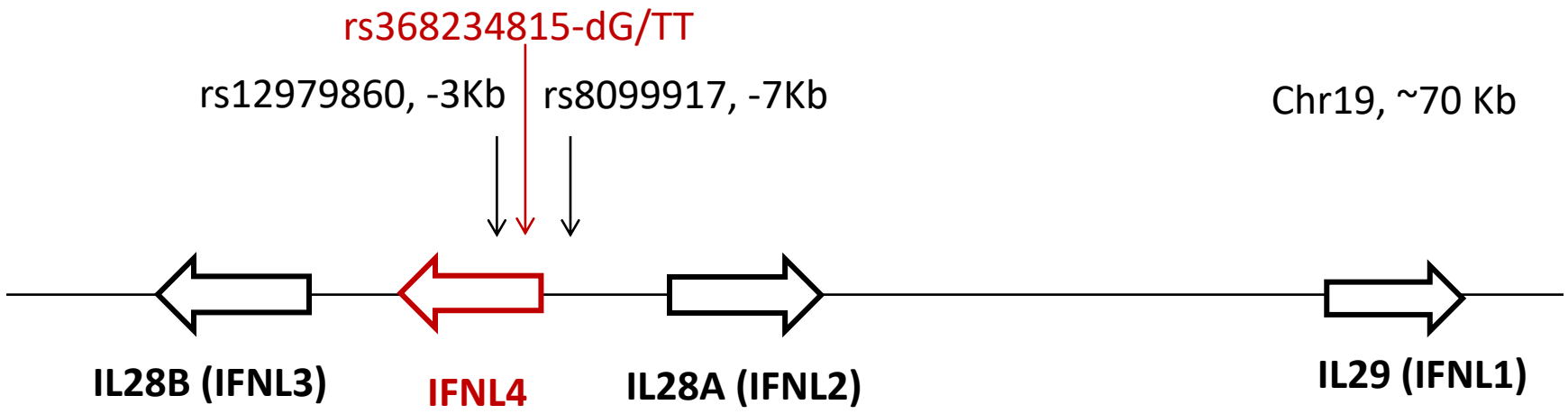
Examples

What is functional? Almost anything at least in some conditions. But is some functionality somewhere = causality for a specific phenotype?

- Level:
- mRNA – eQTLs, stability, modifications, splicing, polyadenylation
- DNA – binding sites for various regulatory factors, methylation
- Chromatin – open/closed regions, long-range interactions
- Protein – pQTLs, stability, localization, modifications, etc

- Derived phenotypes – anything that can be quantified
- Somatic mutagenesis and mutation signatures; immune cell ratios, cell viability, proliferation, senescence or stemness scores, relative telomere length, hormone ratios, etc

Not all variants are SNPs and we still don't know all of them. Example 1: A small polymorphic indel



2009: GWAS for spontaneous and treatment-induced clearance of hepatitis C virus (HCV). One of the strongest GWAS signals ever reported, ORs ~3-10

70 Kb region encoding type-III interferons IL28A, IL28B, IL29, ~95% similarity

A number of efforts to functionally characterize this GWAS signal

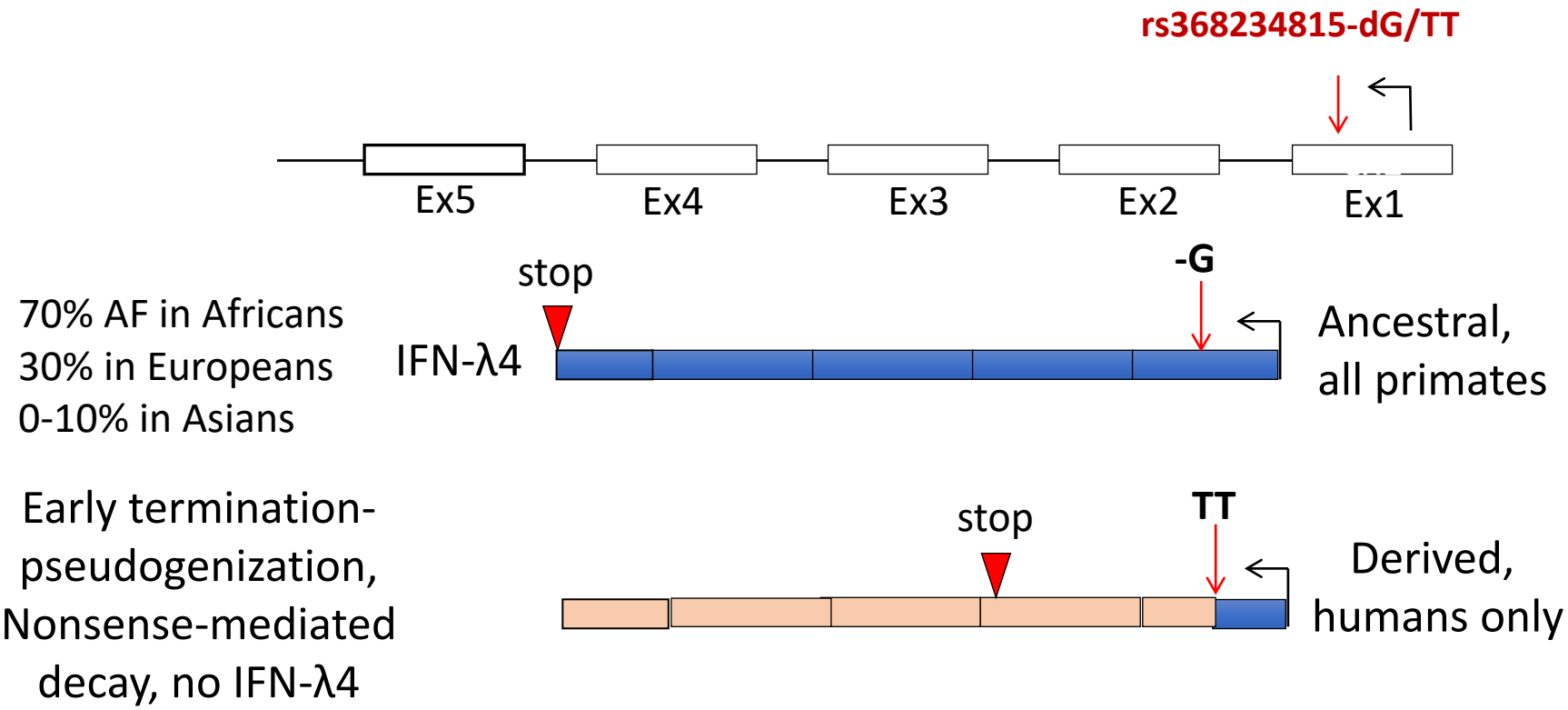
2013: We discovered a novel gene, Interferon Lambda 4 (IFNL4)

GWAS SNP - within first intron of IFNL4, only 367 bp from the novel variant

rs368234815-dG/TT in first exon, explains and improves GWAS signal, function - frame-shift variant

Ge et al, Nature, 2009; Thomas et al, Nature, 2009
Kotenko et al, Sheppard et al, Nat Immun, 2002
Prokunina-Olsson et al, Nat Gen, 2013

Molecular phenotype of the GWAS signal: creating or eliminating IFNL4 protein

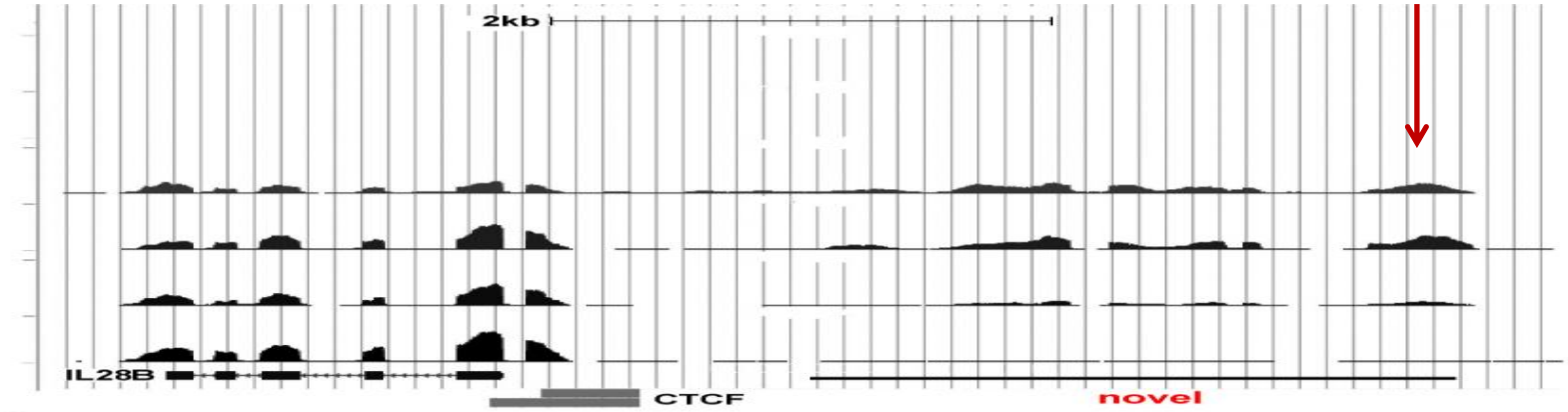


From USCS: Homo sapiens interferon, lambda 4 (gene/pseudogene) (IFNL4)

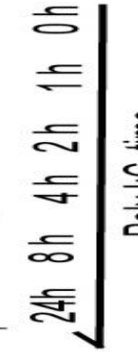
RefSeq Summary (NR_074079): This gene is a polymorphic pseudogene which, in some humans, encodes the interferon (IFN) lambda 4 protein. Humans are polymorphic for the dinucleotide TT/deltaG allele. Compared to the ancestral state in non-human primates, the TT allele produces a frameshift in the coding region of this gene which is predicted to induce nonsense-mediated mRNA decay.

Prokunina-Olsson et al, Nat Gen, 2013

Variant and gene discovery through functional studies

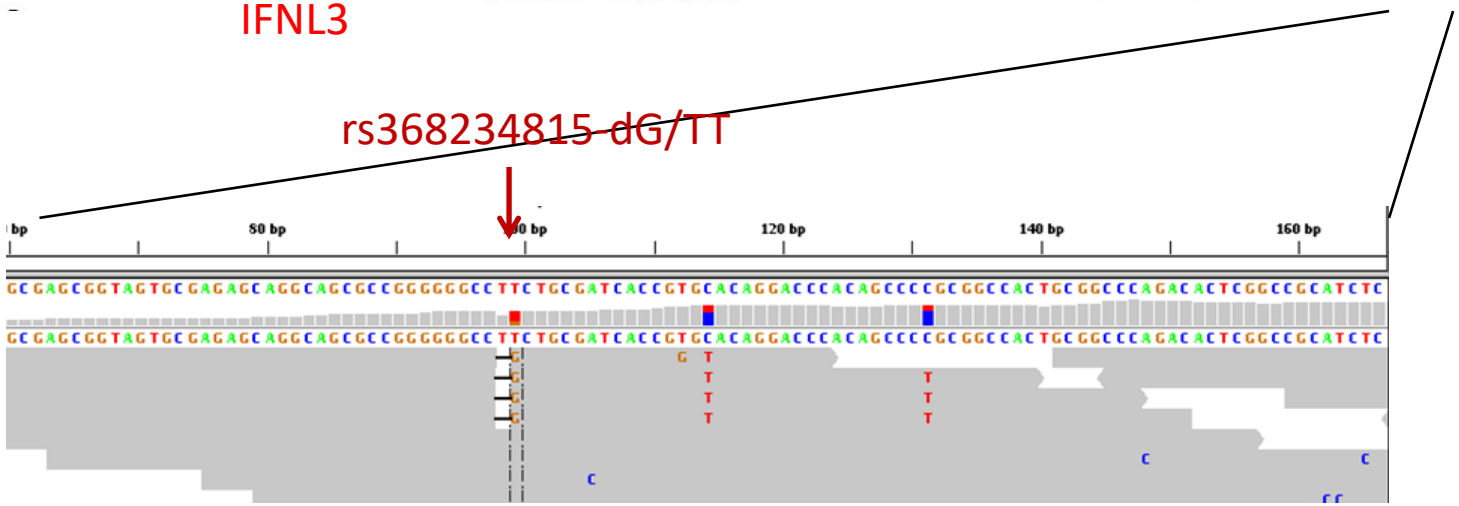


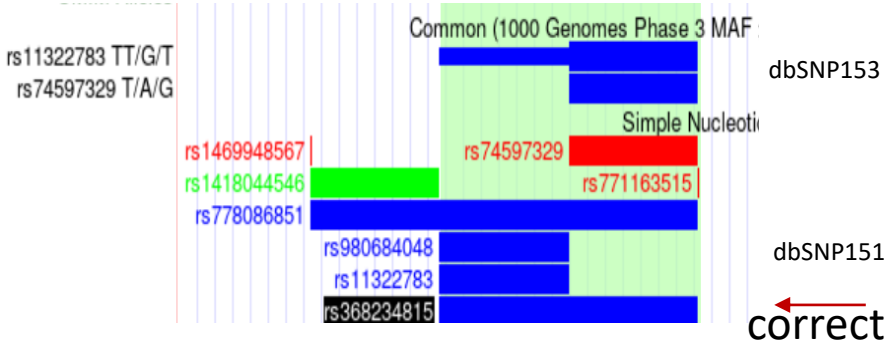
IFNL3



RNA-sequencing in time series,
In primary human
hepatocytes from a single donor.
In vitro treated with polyIC to
mimic infection with RNA-viruses

rs368234815 dG/TT





Currently: in 1KG rs368234815-TT/-G is represented by
rs74597329: T/G (and A) – no Forgedb annotation, HaploReg 4
rs11322783 – TT/T (and G) - no Forgedb annotation

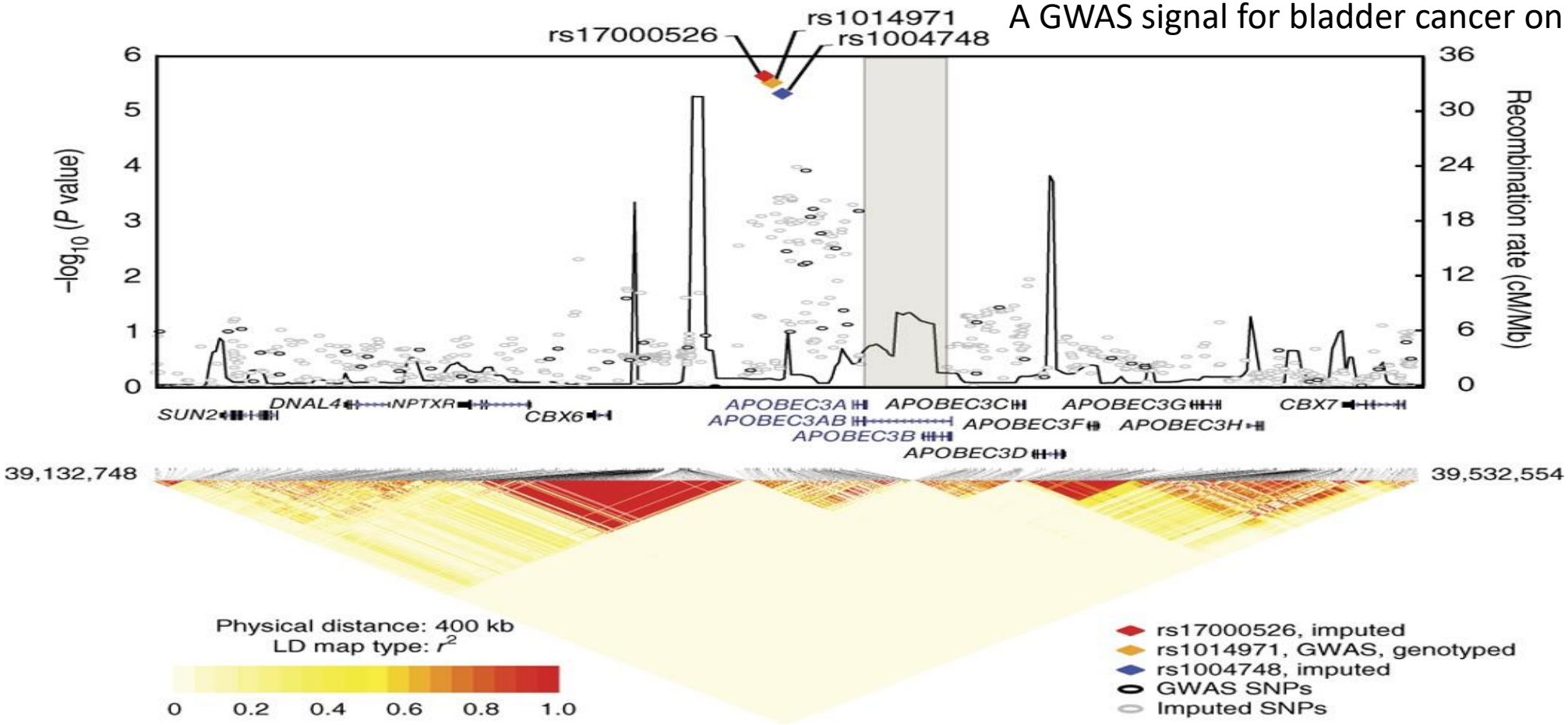
Prokunina-Olsson et al, Nat Gen, 2013

Summary on the Example 1

- Even with extremely strong GWAS signals (ORs=3-10) functional interpretations are not obvious
- The genome data and annotations maybe incomplete/misleading or confusing
- Functional analysis = discovery of the causal variant, novel gene and mechanism
- Without knowing this variant, the functional studies would be very different and not informative
- Even after 10 years after discovery of IFNL4 and the causal variant, genome annotation (dbSNP153 is still confusing and unhelpful) and negatively affecting downstream analyses
- The initial GWAS SNP, rs12979860 (now intron of IFNL4) is not “the causal” variant but a good proxy for it (TT/dG). Rs12979860 = “IL28B marker” is a patented marker for predicting outcomes of HCV clearance on several therapies
- Prediction markers are not necessarily the causal ones – matter of timing (priority), convenience of genotyping, commerce. Hard to change once established.

Not all variants are SNPs and we still don't know all of them. Example 2: Polymorphic indels

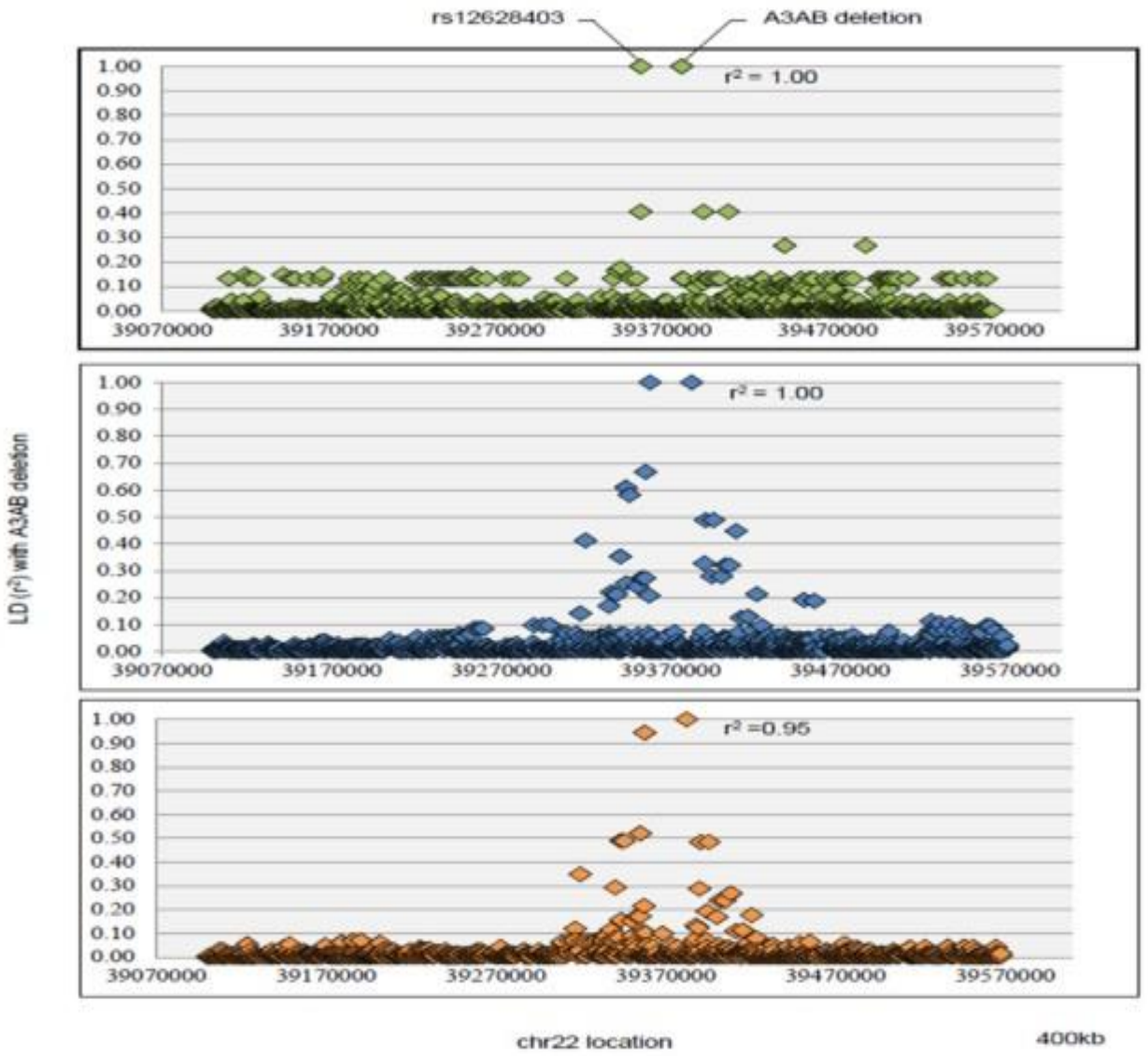
A GWAS signal for bladder cancer on chr 22q13.1



Germline A3AB deletion (~30 kb) eliminating APOBEC3B and merging coding part of APOBEC3A with 3'UTR of APOBEC3B – a super-active DNA-editing fusion APOBEC3A enzyme

Middlebrooks, Banday et al, Nat Gen, 2016

Intronic SNP rs12628403 is the only proxy for the A3AB deletion



LD between rs12628403 and A3AB deletion:

In Europeans, r^2 ~ 0.92-1.00, 7%
East Asians, r^2 ~ 0.95-1.00, 34%
Africans, no proxy, CNV has a frequency of 4.2% while rs12628403 is monomorphic

If both unavailable - cannot be imputed and have to be directly genotyped

Genotype landscape of this region is affected by the A3AB deletion

Very different functional consequences that need to be considered

Molecular phenotype: quantitative trait forgenome-wide somatic mutations caused by APOBEC3B enzymes

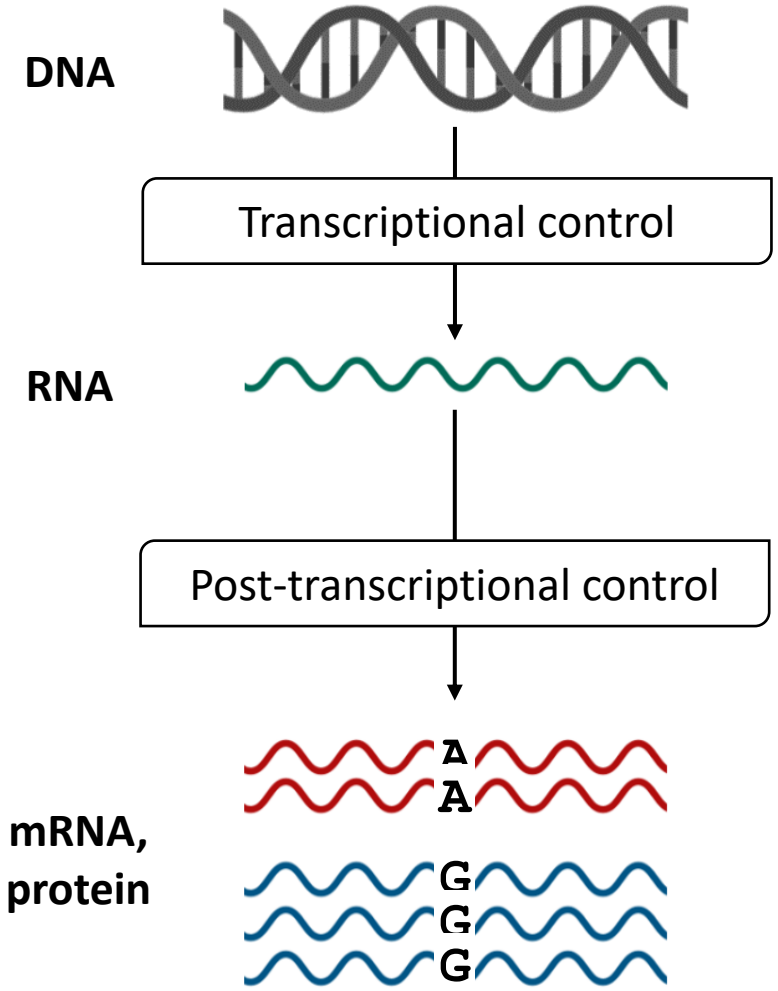
Middlebrooks, Banday et al, Nat Gen, 2016

Example 3. It make take more than one variant …

- We tend to assign all the blame on a single variant that statistically explains the association signal, other variants are ignored
- Functionally – the effects of several variants might be required to achieve a nuanced regulation through different but complementary mechanisms
- Risk of severe COVID-19 and OAS1 region

Banday, Stanifer, Florez-Vargas et al, Nat Gen, 2022

Multiple approaches to explore the molecular contribution of rs10774671-A and rs1131454-A to the risk of severe COVID-19

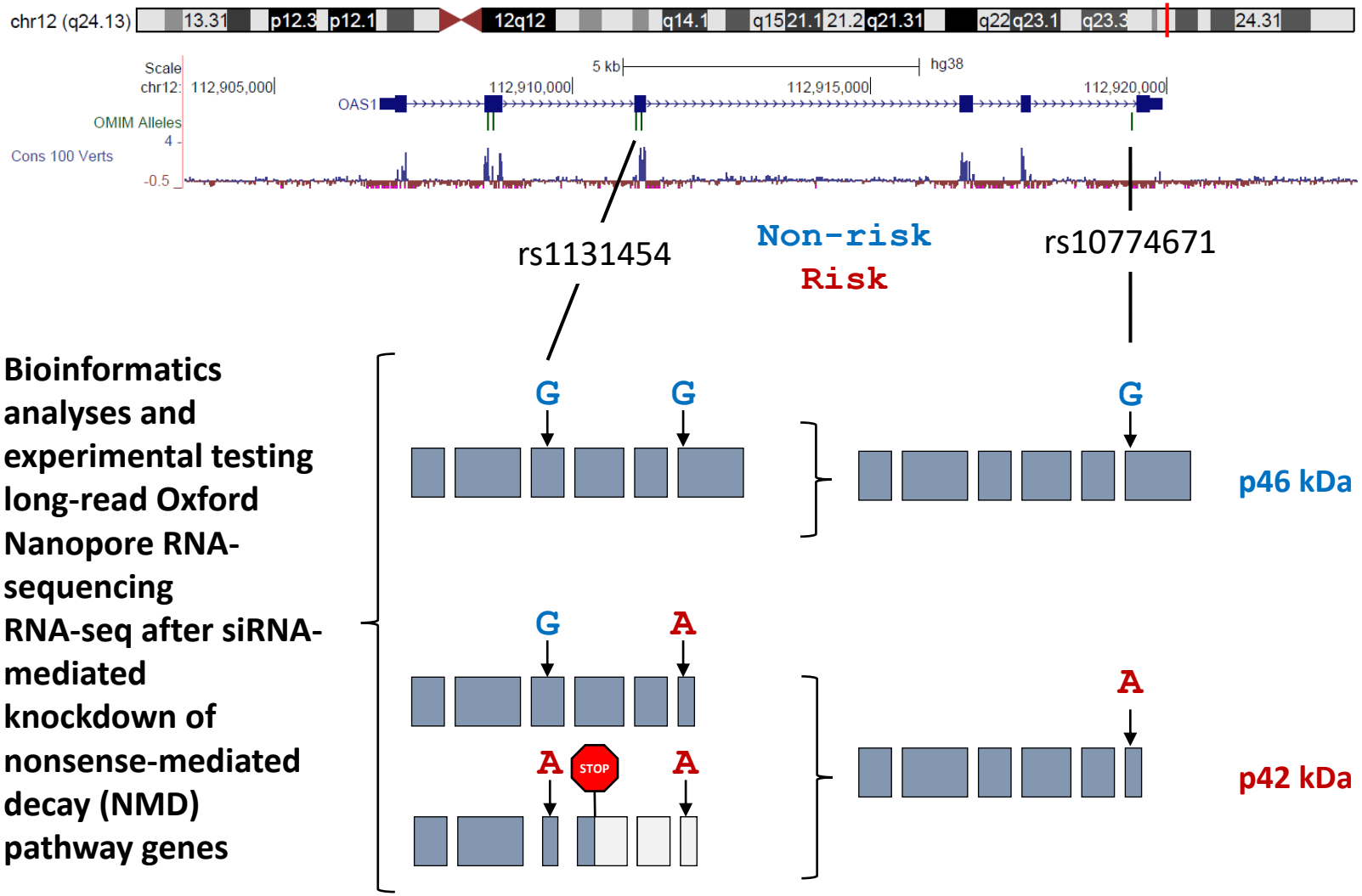


- **Bioinformatic analyses of multiple datasets:**
- **ATAC-seq:** chromatin accessibility analysis
- **ChIP-seq:** DNA-protein interactions
- **Hi-C:** chromatin interactions
- **Conclusion: these mechanisms are not important here**

- ***In silico* and *in vitro*:**
- rs10774671 – affects splicing of OAS1
- rs1131454 creates a putative exonic splicing enhancer/silencer of exon 3
- **Bionformatic analyses, ExonTrap and RT-PCR:**
  - ✓ rs1131454-G allele creates a long exon 3 isoform
  - ✓ rs1131454-A allele creates short exon 3 isoform
- **Conclusion: these mechanisms are important for nonsense-mediated decay of mRNA carrying the risk alleles**

Banday, Stanifer, Florez-Vargas et al, Nat Gen, 2022

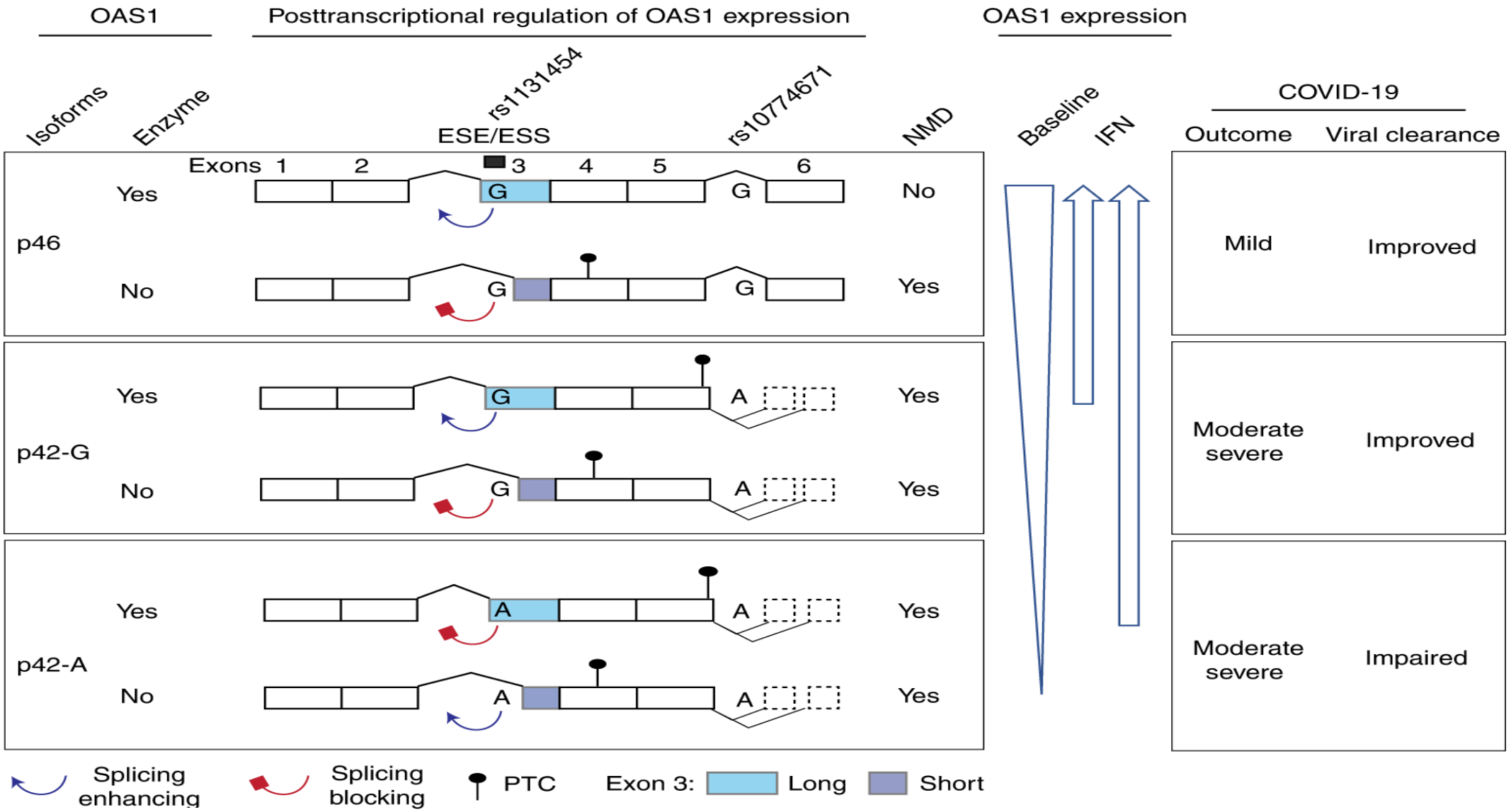
Through different mechanisms risk alleles of both SNPs contribute to nonsense-mediated decay (NMD) resulting in mRNA degradation, decreased protein expression and impaired sensing and clearance of SARS-CoV-19



- Bioinformatics analyses and experimental testing
- long-read Oxford Nanopore RNA-sequencing
- RNA-seq after siRNA-mediated knockdown of nonsense-mediated decay (NMD) pathway genes

Banday, Stanifer, Florez-Vargas et al, Nat Gen, 2022

Two SNPs on the same haplotype contribute to COVID-19 outcomes



Banday, Stanifer, Florez-Vargas et al, Nat Gen, 2022

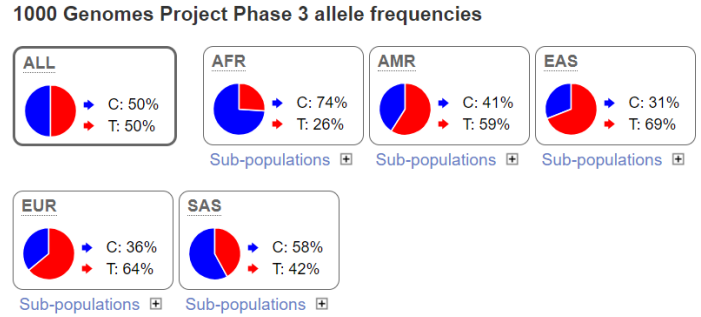
Example 4. Fake SNPs and synthetic associations

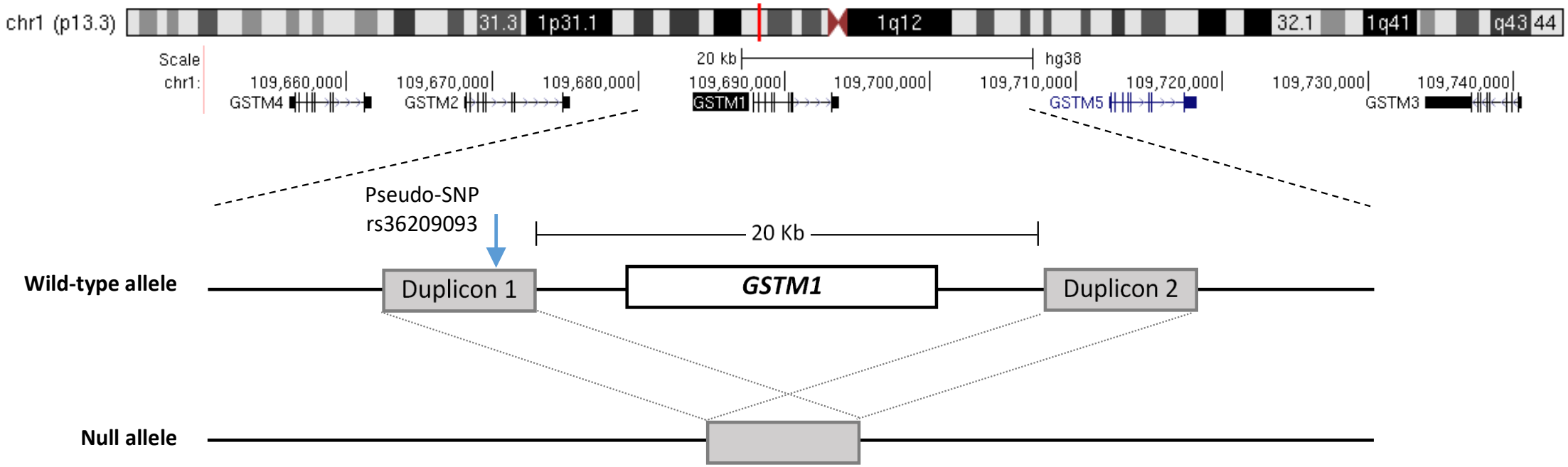
rs36209093

Bladder cancer – p = 3.21E-18, in our meta-analysis, Koutros et al, Eur Urology, 2023

Other GWAS-significant associations in GWAS catalog:

Metabolite measurements; HDL, urinary metabolites, free and total cholesterol, cholesterol ratio, GSTM1 levels

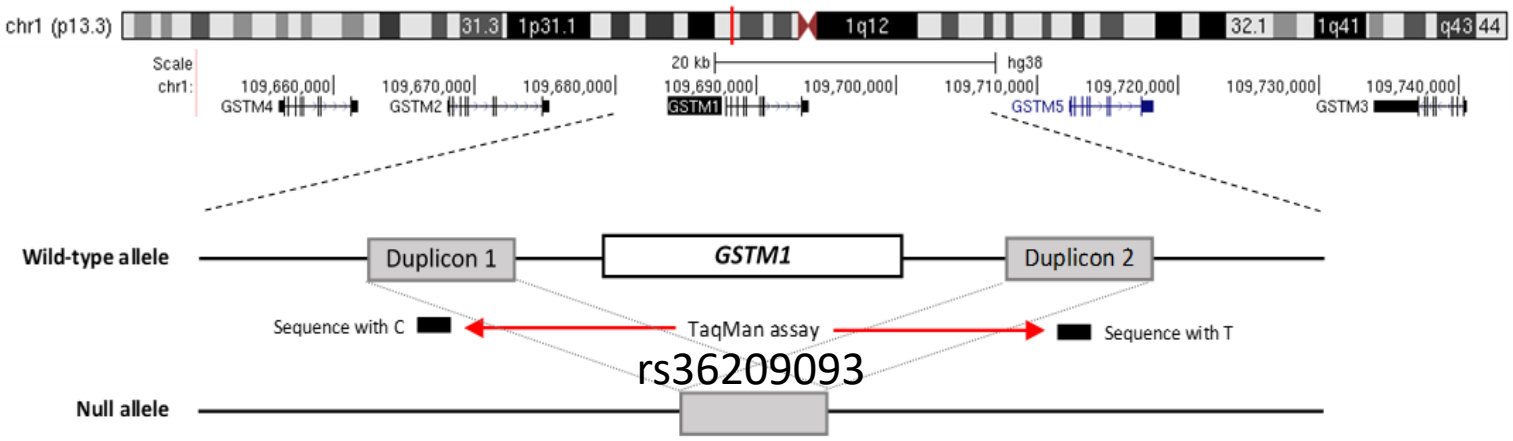


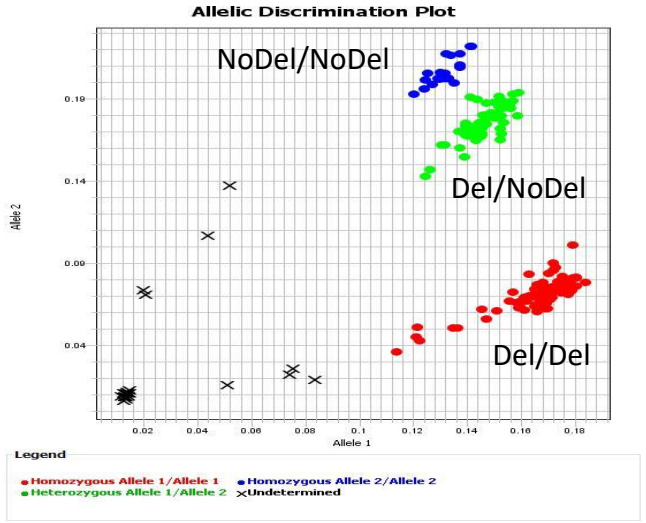


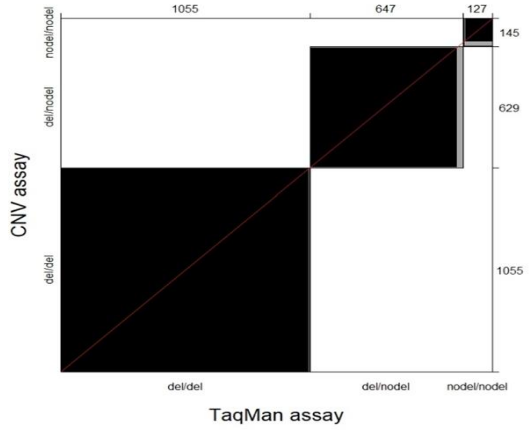
Koutros et al, Eur Urology, 2023

Florez-Vargas et al, in preparation

Example 4. rs36209093 is a fake SNP but is still useful as a proxy







Custom TaqMan assay, 98.1% concordance rate in 1102 samples previously genotyped with a CNV assay.

Koutros et al, Eur Uro, 2023 and Florez-Vargas in preparation

Our custom TaqMan genotyping assay for rs36209093

Dupl 1	C
Dupl 1	C
Dupl 2	T
Dupl 2	T

2C:2T
No del/
No del

Dupl 1	C
Dupl 1	C
Dupl 2	T
Dupl 2	del

No del/
Del
2C:1T

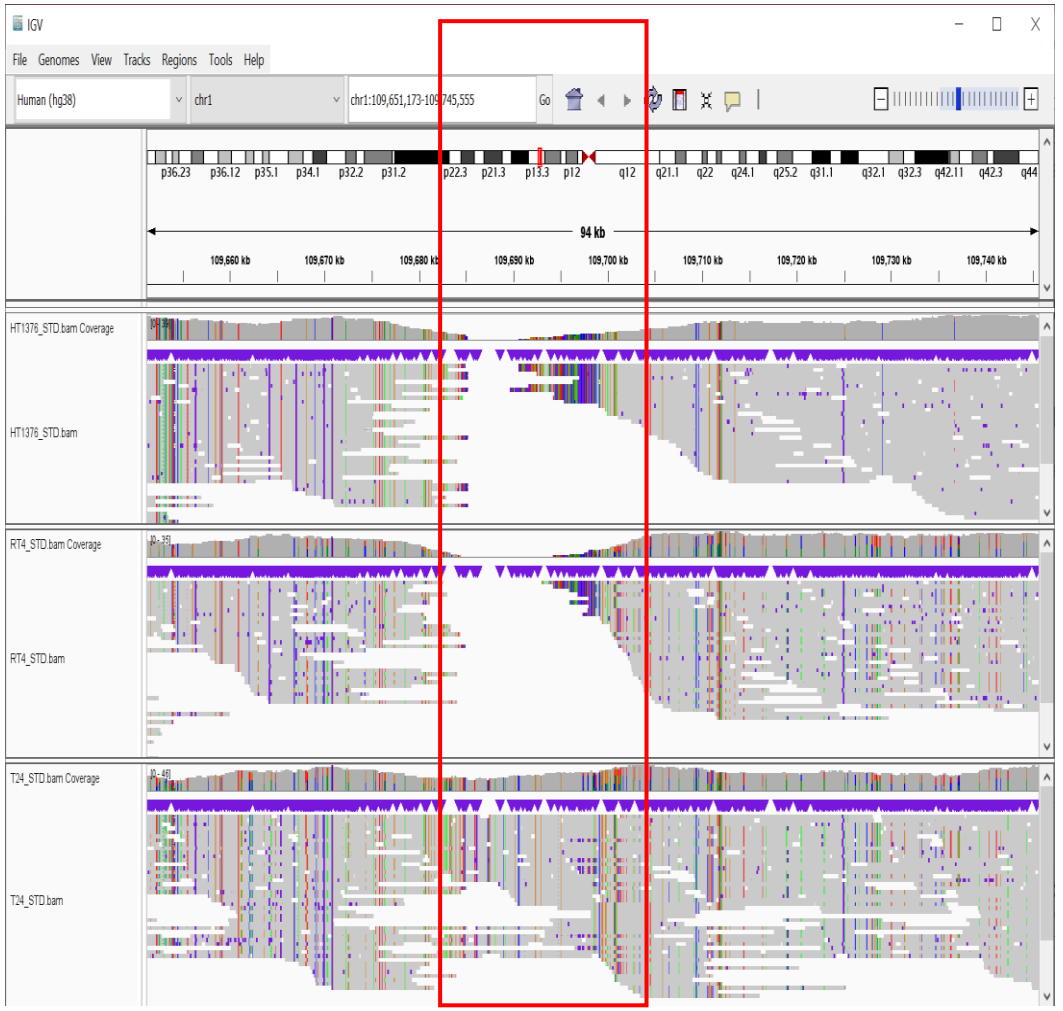
Dupl 1	C
Dupl 1	C
Dupl 2	del
Dupl 2	del

Del/
Del
2C:0T

Our experience with exploring the GSTM1 GWAS signal

Targeted PacBio long-read sequencing (~100Kb) in 1KG samples – poor coverage, low confidence about borders

Whole-genome long-read PacBio sequencing in several cell lines – looks better, still hard to define borders



Human Pangenomes – a new reference resource

Article

A draft human pangenome reference

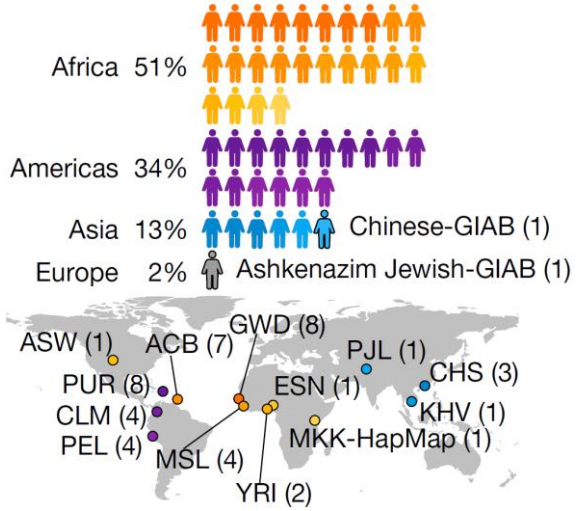
https://doi.org/10.1038/s41586-023-05896-x
Received: 9 July 2022
Accepted: 28 February 2023
Published online: 10 May 2023
Open access
Check for updates

Wen-Wei Liao[1,2,3,60], Mobin Asri[4,60], Jana Ebler[5,6,60], Daniel Doerr[5,6], Marina Haukness[4], Glenn Hickey[4], Shuangjia Lu[1,2], Julian K. Lucas[4], Jean Monlong[4], Haley J. Abel[7], Silvia Buonaiuto[8], Xian H. Chang[4], Haoyu Cheng[9,10], Justin Chu[9], Vincenza Colonna[8,11], Jordan M. Eizenga[4], Xiaowen Feng[9,10], Christian Fischer[11], Robert S. Fulton[12,13], Shilpa Garg[14], Cristian Groza[15], Andrea Guarracino[11,16], William T. Harvey[17], Simon Heumos[18,19], Kerstin Howe[20], Miten Jain[21], Tsung-Yu Lu[22], Charles Markello[4], Fergal J. Martin[23], Matthew W. Mitchell[24], Katherine M. Munson[17], Moses Njagi Mwaniki[25], Adam M. Novak[4], Hugh E. Olsen[4], Trevor Pesout[4], David Porubsky[17], Pjotr Prins[11], Jonas A. Sibbesen[26], Jouni Sirén[4], Chad Tomlinson[12], Flavia Villani[11], Mitchell R. Vollger[17,27], Lucinda L. Antonacci-Fulton[12], Gunjan Baid[28], Carl A. Baker[17], Anastasiya Belyaeva[28], Konstantinos Billis[23], Andrew Carroll[28], Pi-Chuan Chang[28], Sarah Cody[12], Daniel E. Cook[28], Robert M. Cook-Deegan[29], Omar E. Cornejo[30], Mark Diekhans[4], Peter Ebert[5,6,31], Susan Fairley[23], Olivier Fedrigo[32], Adam L. Felsenfeld[33], Giulio Formenti[32], Adam Frankish[23], Yan Gao[34], Nanibaa' A. Garrison[35,36,37], Carlos Garcia Giron[23], Richard E. Green[38,39], Leanne Haggerty[23], Kendra Hoekzema[17], Thibaut Hourlier[23], Hanlee P. Ji[40], Eimear E. Kenny[41], Barbara A. Koenig[42], Alexey Kolesnikov[28], Jan O. Korbel[23,43], Jennifer Kordosky[12], Sergey Koren[44], HoJoon Lee[40], Alexandra P. Lewis[17], Hugo Magalhães[5,6], Santiago Marco-Sola[45,46], Pierre Marijon[5,6], Ann McCartney[44], Jennifer McDaniel[47], Jacquelyn Mountcastle[32], Maria Nattestad[28], Sergey Nurk[44], Nathan D. Olson[47], Alice B. Popejoy[48], Daniela Puiu[49], Mikko Rautiainen[44], Allison A. Regier[12], Arang Rhie[44], Samuel Sacco[30], Ashley D. Sanders[50], Valerie A. Schneider[51], Baergen I. Schultz[33], Kishwar Shafin[28], Michael W. Smith[33], Heidi J. Sofia[53], Ahmad N. Abou Tayoun[52,53], Françoise Thibaud-Nissen[51], Francesca Floriana Tricomi[23], Justin Wagner[47], Brian Walenz[44], Jonathan M. D. Wood[20], Aleksey V. Zimin[49,54], Guillaume Bourque[55,56,57], Mark J. P. Chaisson[22], Paul Flicek[23], Adam M. Phillippy[44], Justin M. Zook[47], Evan E. Eichler[17,58], David Haussler[4,58], Ting Wang[12,13], Erich D. Jarvis[32,58,59], Karen H. Miga[4], Erik Garrison[11]✉, Tobias Marschall[5,6]✉, Ira M. Hall[1,2]✉, Heng Li[9,10]✉ & Benedict Paten[4]✉

Here the Human Pangenome Reference Consortium presents a first draft of the human pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals[1]. These assemblies cover more than 99% of the expected sequence in each genome and are more than 99% accurate at the structural and base pair levels. Based on alignments of the assemblies, we generate a draft pangenome that captures known variants and haplotypes and reveals new alleles at structurally complex loci. We also add 119 million base pairs of euchromatic polymorphic sequences and 1,115 gene duplications relative to the existing reference GRCh38. Roughly 90 million of the additional base pairs are derived from structural variation. Using our draft pangenome to analyse short-read data reduced small variant discovery errors by 34% and increased the number of structural variants detected per haplotype by 104% compared with GRCh38-based workflows, which enabled the typing of the vast majority of structural variant alleles per sample.

The human reference genome has formed the backbone of human genomics since its initial draft release more than 20 years ago[2]. The primary sequences are a mosaic representation of individual haplotypes containing one representative scaffold sequence for each chromosome. There are 210 Mb of gap or unknown (151 Mb) or computationally simulated sequences (59 Mb) within the current GRCh38 release, constituting 6.7% of the primary chromosome scaffolds. Missing reference sequences create an observational bias, or streetlamp effect, which limits studies to be within the boundaries of the reference. Recently, the Telomere-to-Telomere (T2T) consortium finished the first complete sequence of a haploid human genome, T2T-CHM13, which provides a contiguous representation of each autosome and of chromosome X, with the exception of some ribosomal DNA arrays that remain to be fully resolved[3]. Using T2T-CHM13 directly improves genomic analyses; for example, discovering 3.7 million additional single-nucleotide polymorphisms (SNPs) in regions non-syntenic to GRCh38 and better representing the true copy number variants (CNVs) of samples from the 1000 Genomes Project (1KG) compared with GRCh38 (refs. 1,4).

A list of affiliations appears at the end of the paper.



https://humanpangenome.org/

N = 47 individuals
98% are non-Europeans

with the goal of increasing to 350 individuals by mid-2024

- The Human Pangenome Reference Consortium (HPRC) presents a first draft of the human pangenome reference.
- **47 phased, diploid assemblies**.
- The data is based on Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing.
- HPRC add **119 million** base pairs of euchromatic polymorphic sequences and 1,115 gene duplications relative to the existing reference GRCh38. Roughly **90 million of the additional base pairs are derived from structural variation**.

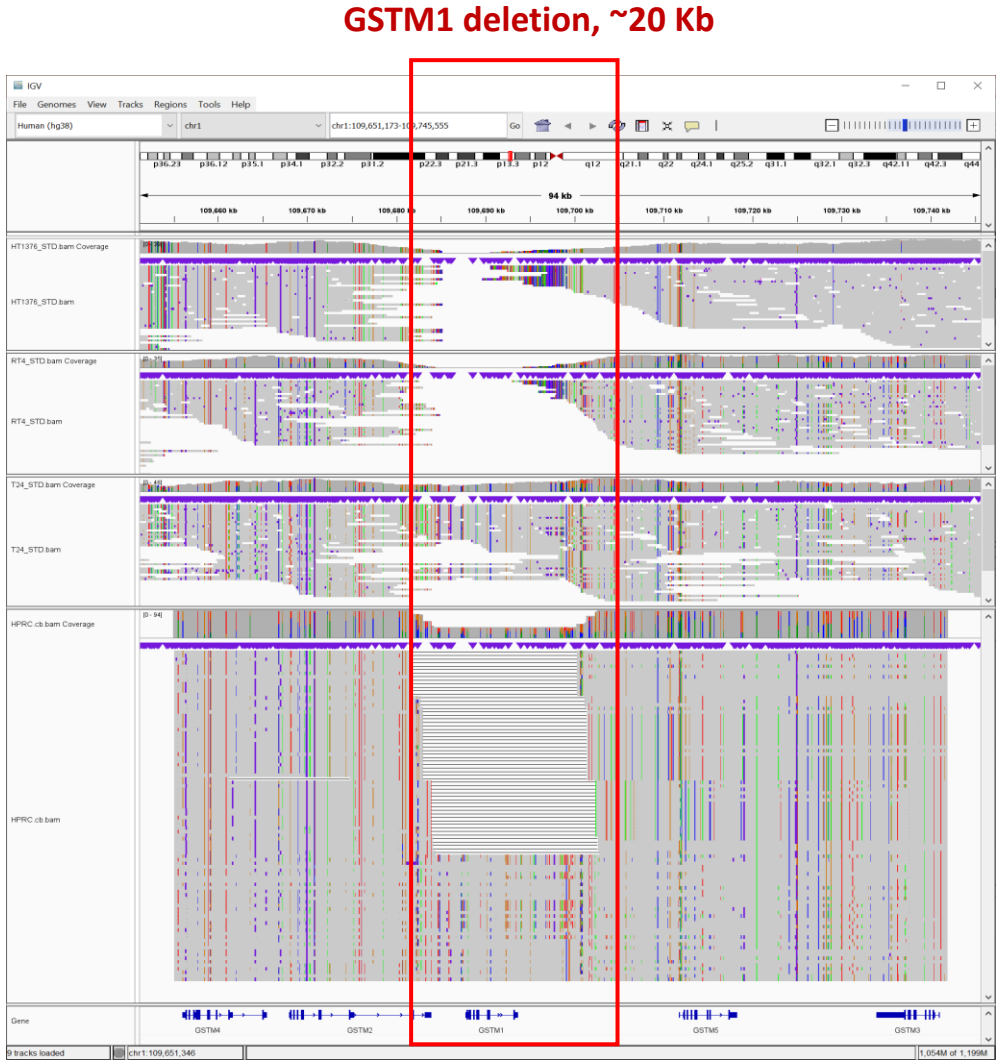
Practical session with Oscar Florez-Vargas: Explore the GSTM1 region based on the HPRC assemblies

Liao et al., Nature 2023

Better results based on long-read genome assemblies

Targeted PacBio long-read sequencing (~100Kb) in 1KG samples – poor coverage, low confidence about borders

Whole-genome long-read PacBio sequencing in several cell lines – looks better, still hard to define borders



Assemblies
1 read = 1 chromosome