From Genomes to targets, tools of the trade and big dreams

Martin Beaulieu

February 5, 2026

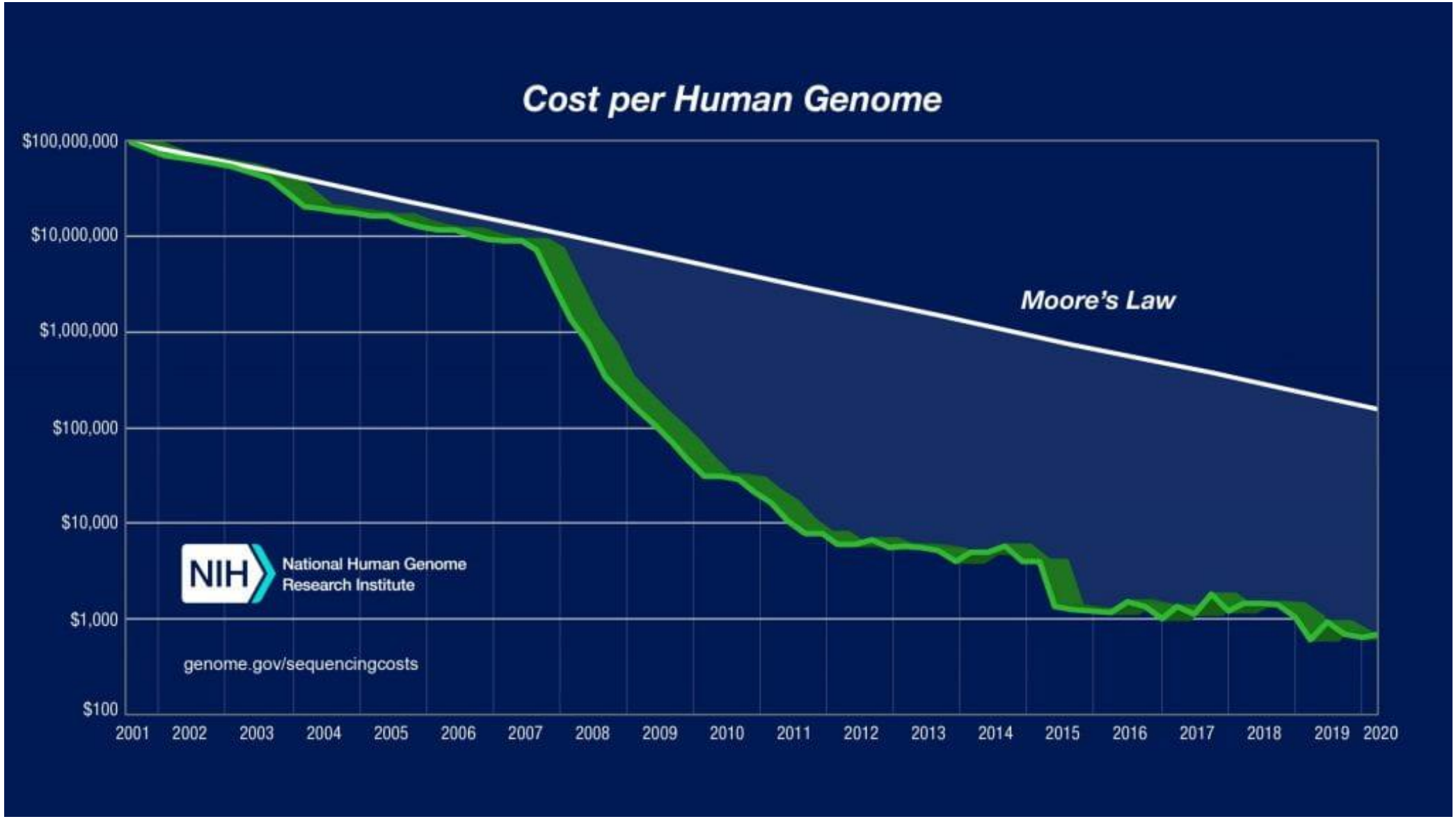
Lecture Plan

Part 1: What is it we measure, types of sequencing and meaning.

Part 2: Going from genome to drug development

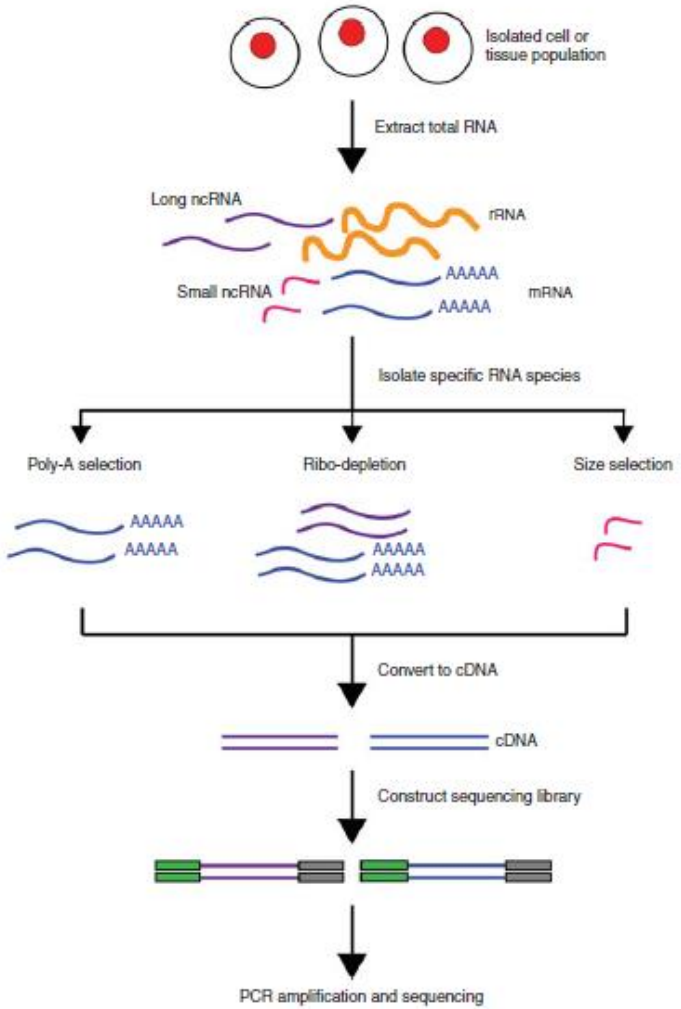
The Next generation Sequencing Revolution.

Hyperexponetial development!



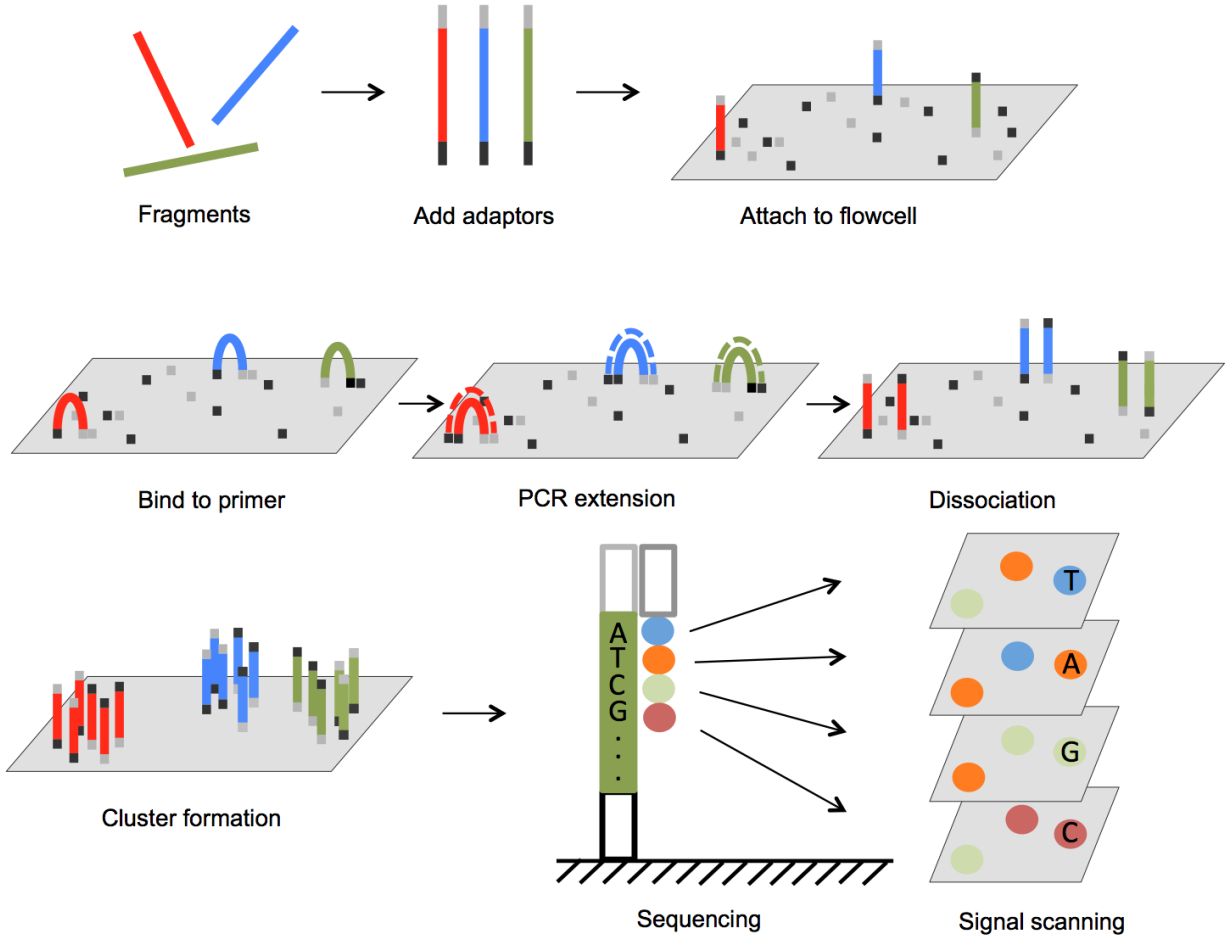
What is next generation sequencing?

Preparation



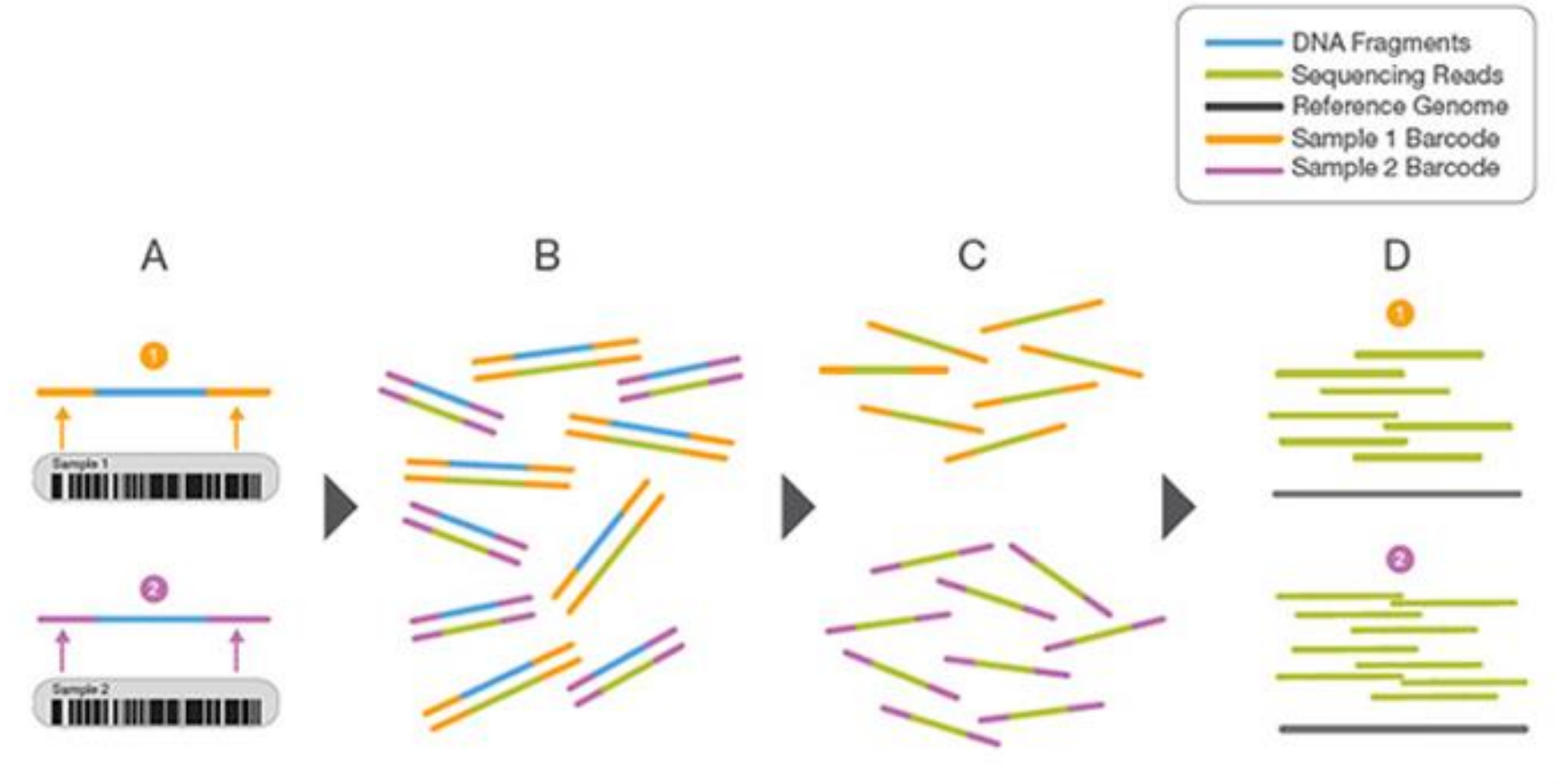
RNA Seq Blog.com

Sequencing



Illumina

Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

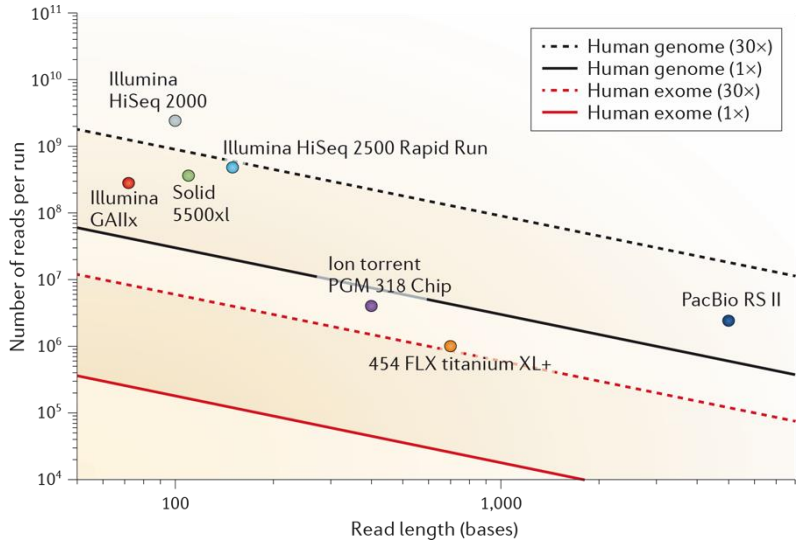
Illumina

Next generation sequencing metrics

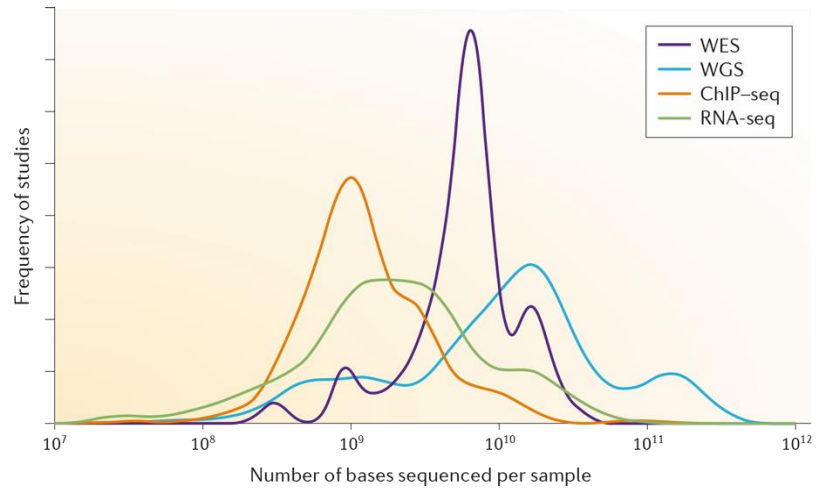
1. *Sequencing Depth (Read Depth): The average number of times each base is sequenced (e.g., 30x, 100x). Higher depth increases confidence in detecting rare variants or somatic mutations.*
2. *Coverage (Breadth): The percentage of the target genome or region that is covered by at least one read (e.g., 95% of the genome).*
3. *Base Quality Score (Q-Score): Measures the probability that a base call is incorrect. A Phred score of 30 (Q30) corresponds to a 1 in 1000 error rate (99.9% accuracy).*
4. *Uniformity of Coverage: Indicates how evenly reads are distributed across the target region. Poor uniformity results in some areas having too little depth (false negatives) and others too much (wasted resources).*
5. *On-Target Rate: For targeted panels, this measures the percentage of reads that map to the intended target region versus unintended, random background DNA.*
6. *Duplication Rate: The proportion of reads that are PCR duplicates, which can falsely inflate coverage and complicate variant calling.*
7. *GC-Bias: The tendency of some sequencing platforms to under- or over-represent regions with high or low GC content, affecting overall coverage uniformity.*

Next generation sequencing metrics use

Affected by technology



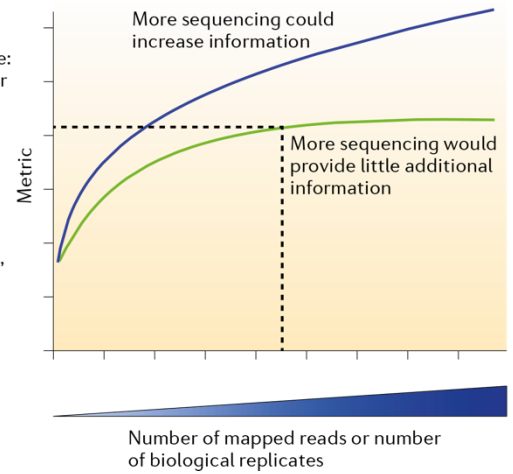
Different across use



Box 3 | **Staged sequencing for predicting sequencing requirements**

Possible metrics:

- General transcriptome coverage: percentage of genes covered over 90% at a given expression level
- Differential expression: number of differentially expressed genes
- Alternative isoform detection: percentage of split reads (that is, junction that spans reads)
- ChIP–seq peak detection: number of enriched loci



What you capture is what you get DNA.

Whole Genome,

Coding Genome

Specific interactions

Example, studying the chromatin

Capturing the chromatin

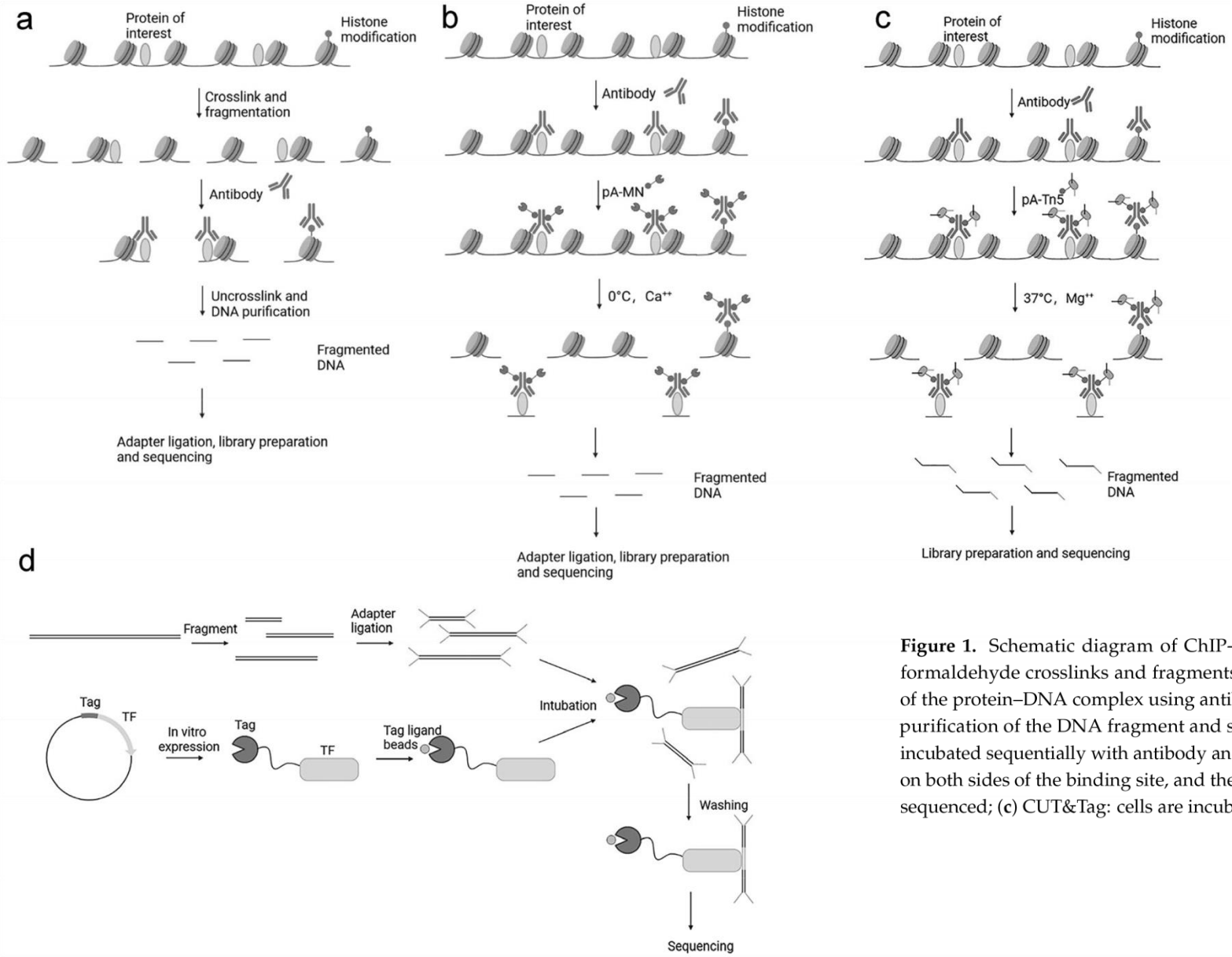


Figure 1. Schematic diagram of ChIP-seq, CUT&RUN, CUT&Tag, and DAP-seq. (**a**) ChIP-seq: formaldehyde crosslinks and fragments the protein and DNA, followed by immunoprecipitation of the protein–DNA complex using antibodies to the target protein, and finally uncrosslinking and purification of the DNA fragment and sequencing of the DNA fragment; (**b**) CUT&RUN: cells are incubated sequentially with antibody and pA-MN, and at 0 °C, Ca++ is added, MNase cleaves DNA on both sides of the binding site, and the fragmented DNA is adapter-ligated, library-prepared, and sequenced; (**c**) CUT&Tag: cells are incubated sequentially with antibody and pA-Tn5, and at 37 °C,

Mapping accessible genome

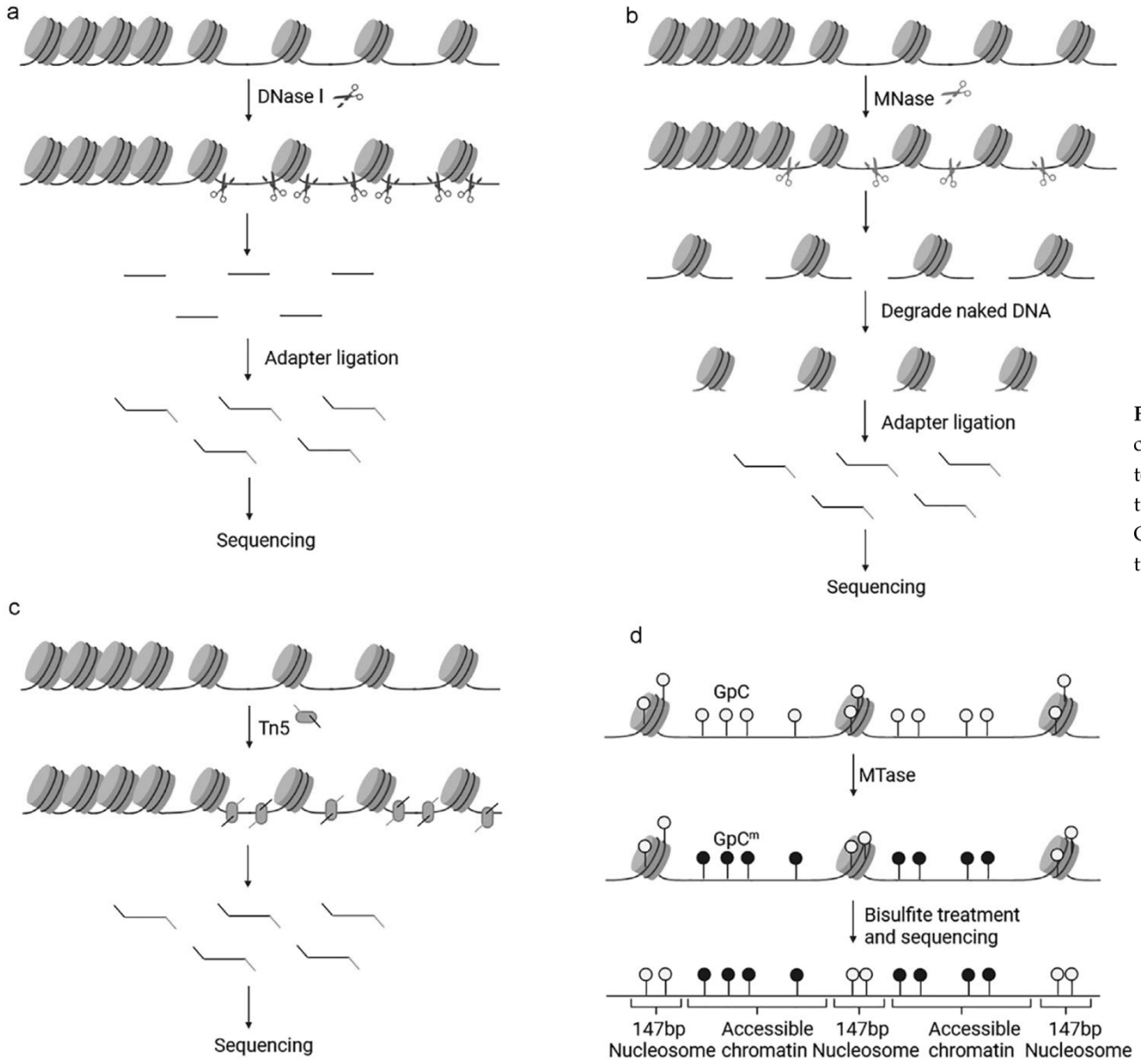
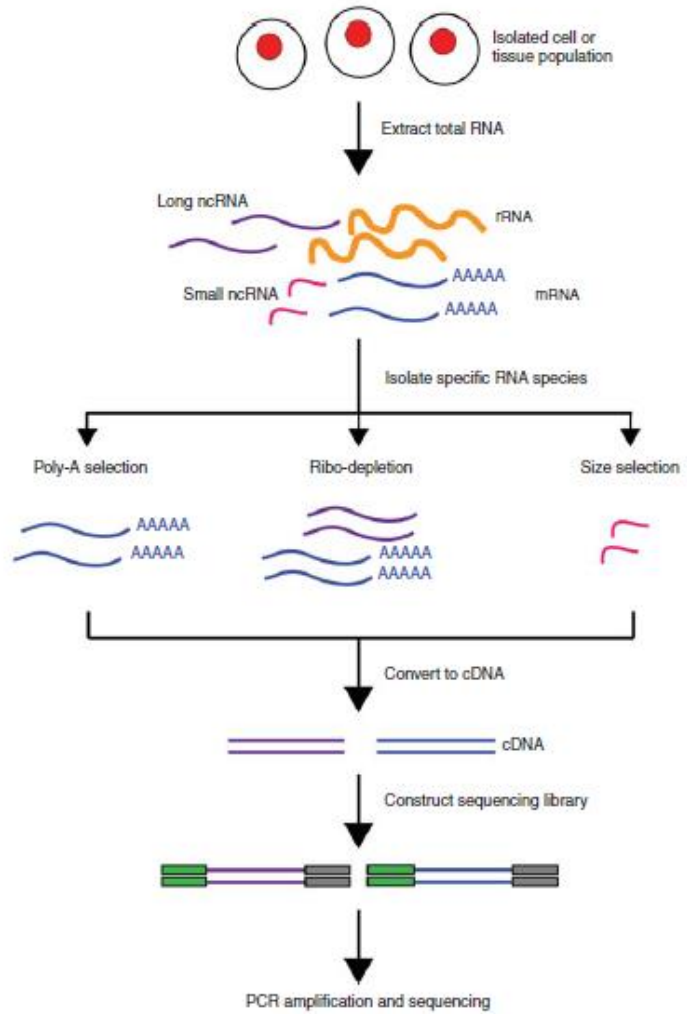


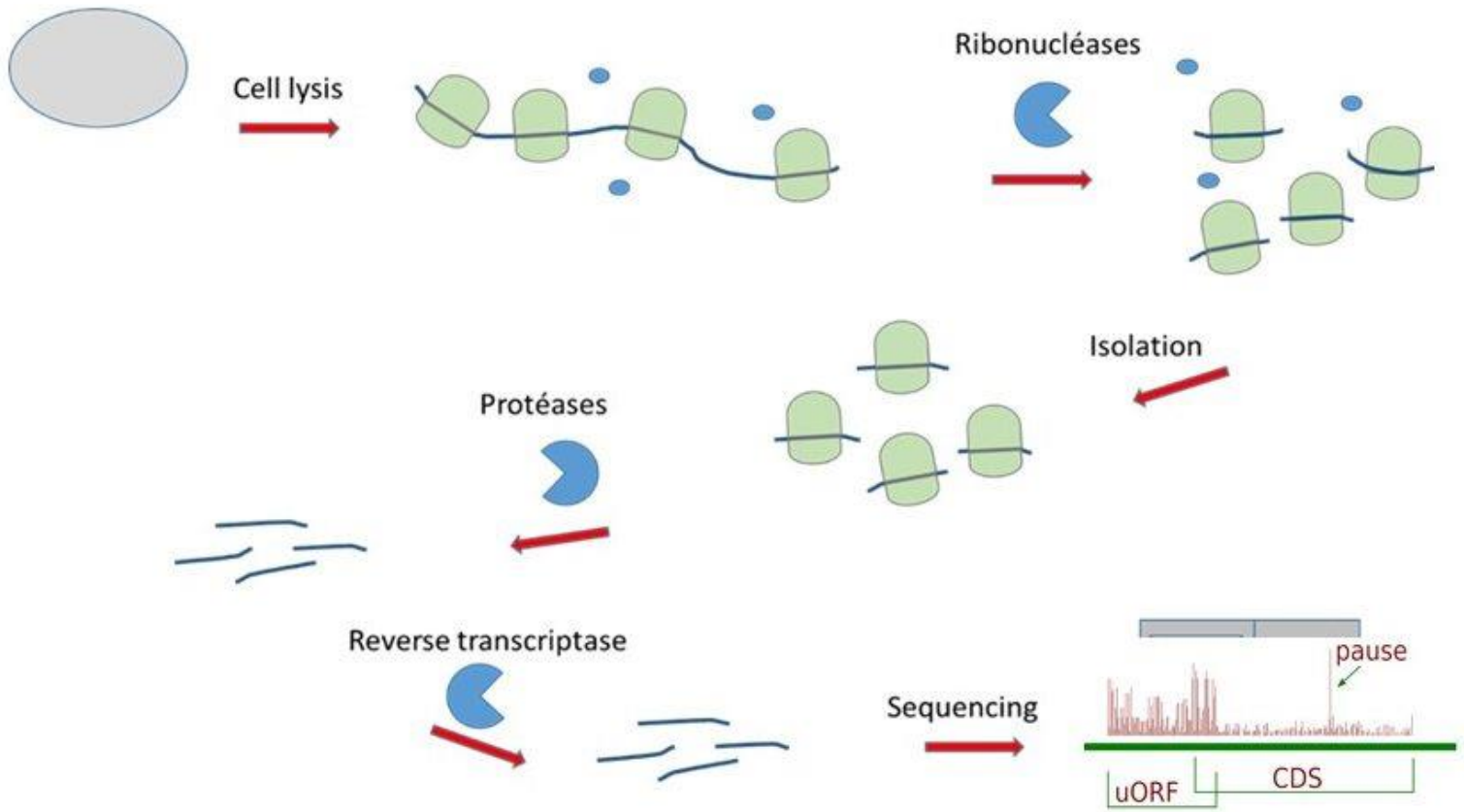
Figure 2. Schematic diagram of DNase-seq, MNase-seq, ATAC-seq, and Nome-seq. (**a**) DNase-seq cleaves accessible chromatin regions characterized by DHSs with DNase I; (**b**) MNase-seq uses MNase to cleave inter-nucleosomal DNA and to degrade the naked accessible DNA; (**c**) ATAC-seq uses Tn5 transposase to ligate sequencing adaptors while cutting DNA fragment; (**d**) Nome-seq uses M.CviPI GpC methyltransferase (MTase) to methylate GpC in accessible chromatin to GpCm, and bisulfite treatment can distinguish GpC from GpCm.

RNA Seq

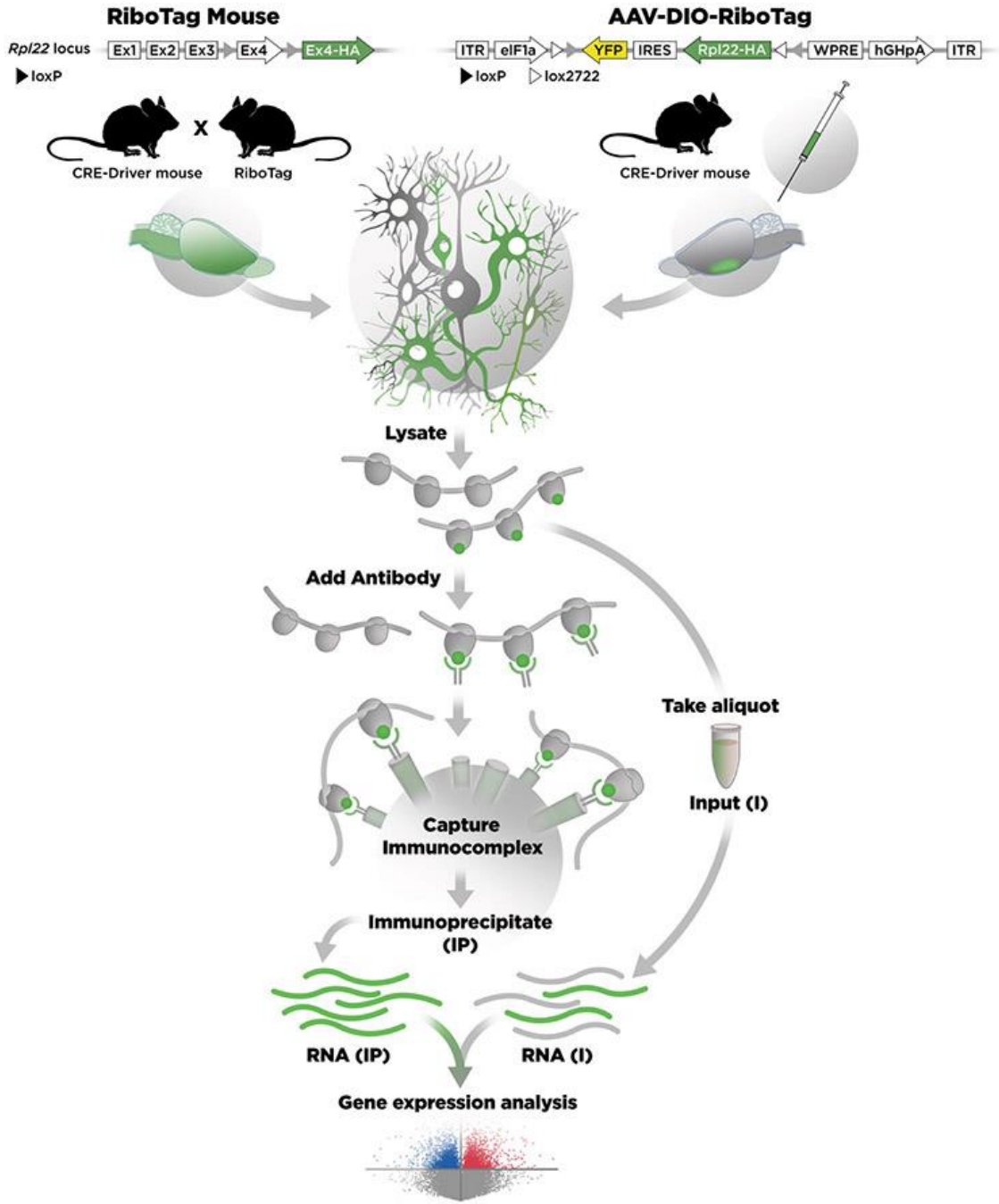
Bulk: Collect everything in the tissue. Sequence some. The more one covers the lesser the depth



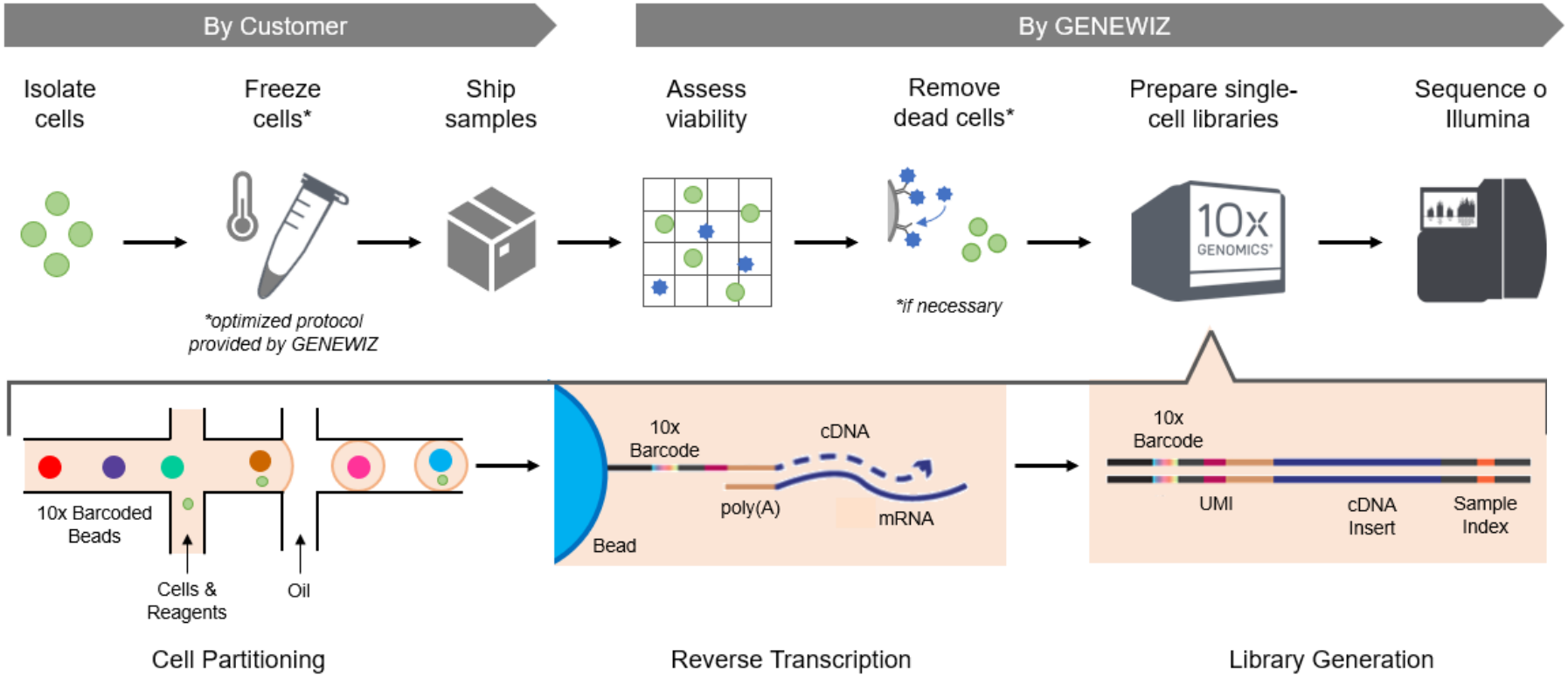
Mapping
Ribosomes
RIBO-Seq

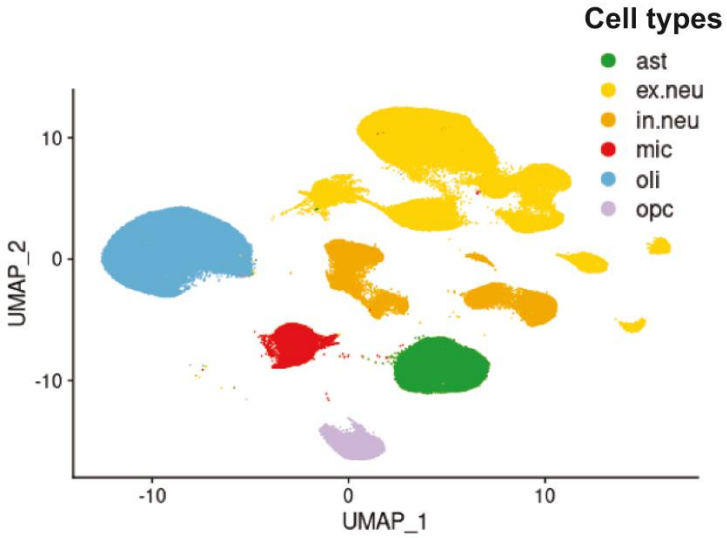


Mapping
Ribosomes
RIBO-Tag Seq



Single cells approach

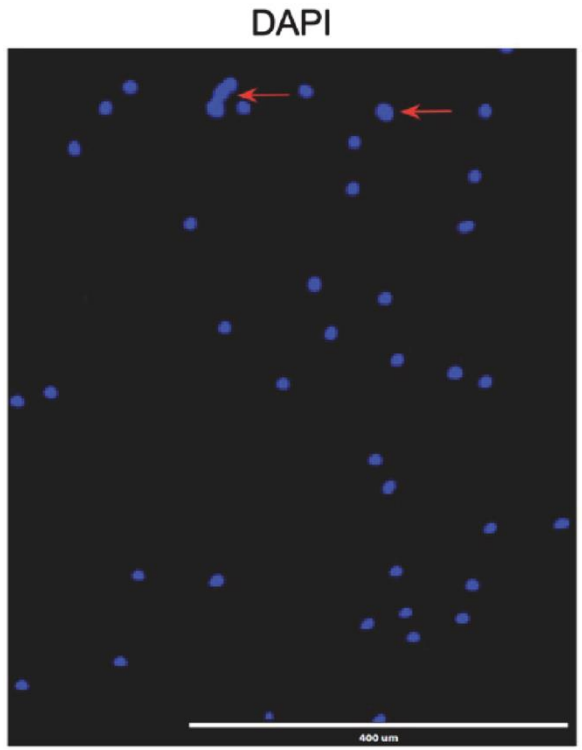




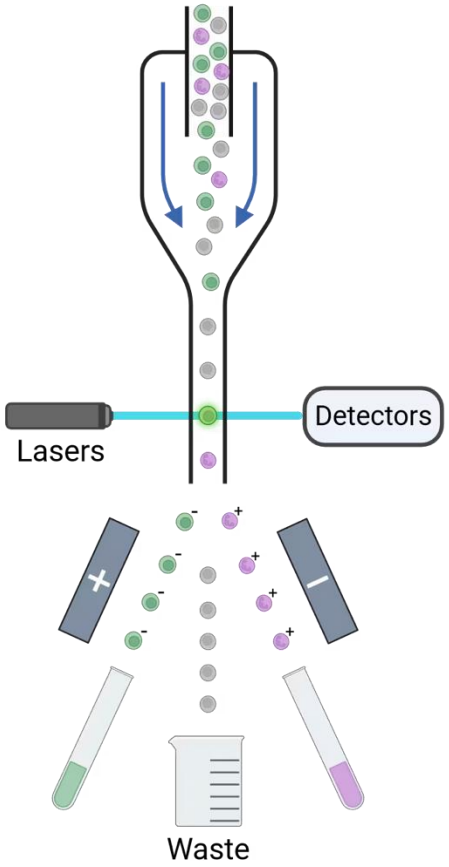
Single nucleus approach

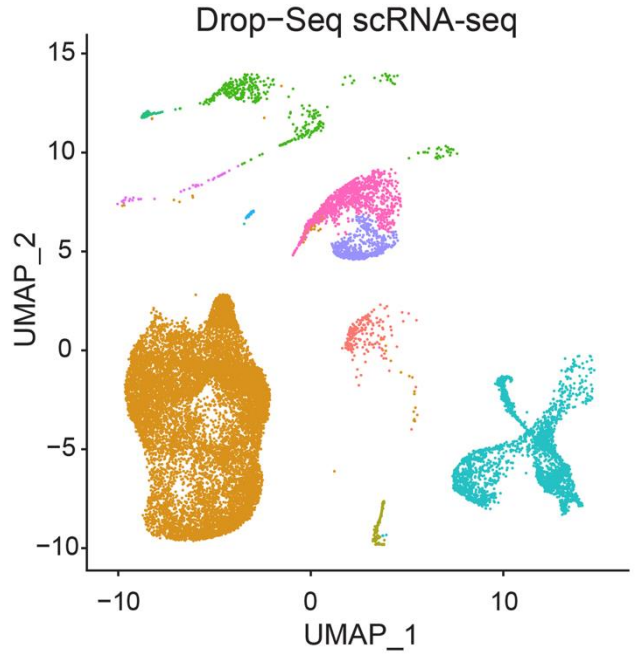
Lyse Cells

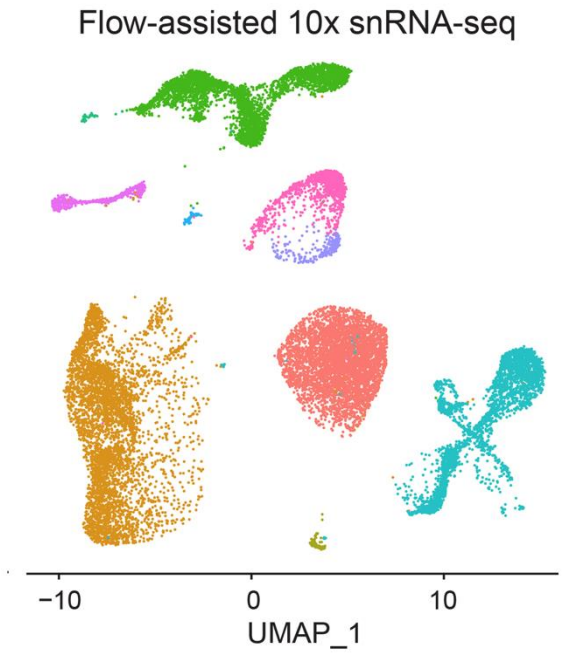
Stain Nuclei using DAPI



FACS Sort







Single cells and Single nucleus approaches

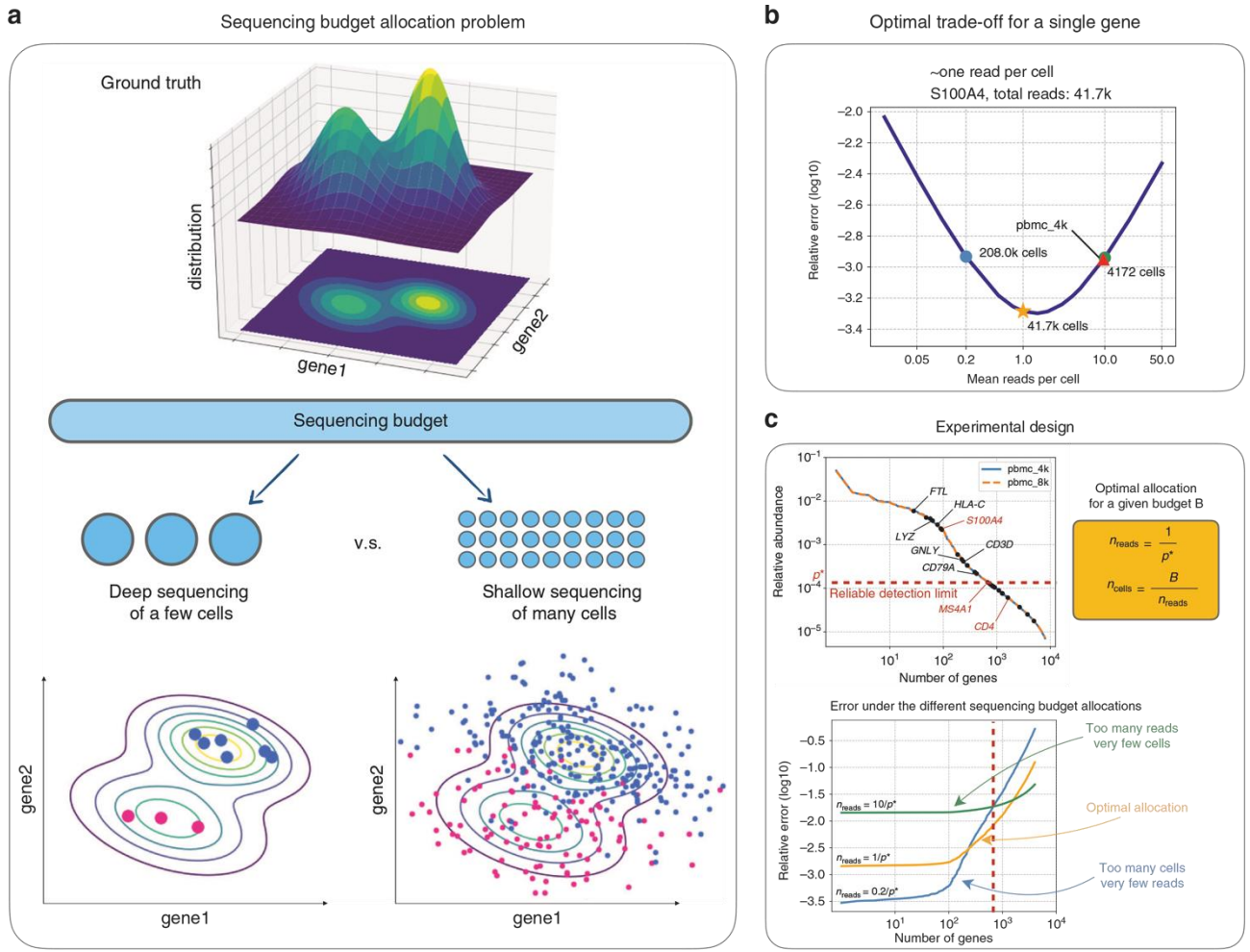


Fig. 1 Optimal sequencing budget allocation. a Description of the sequencing budget allocation problem. Consider estimating the underlying gene distribution (top) from the noisy read counts obtained via sequencing (bottom). With a fixed number of reads to be sequenced, deep sequencing of a few cells accurately estimates each individual cell but lacks coverage of the entire distribution (left), whereas a shallow sequencing of many cells covers the entire population but introduces a lot of noise (right). **b** Optimal tradeoff. The memory T-cell marker gene *S100A4* has 41.7k reads in the pbmc_4k dataset. For estimating the underlying gamma distribution $X_g \sim \mathrm{Gamma}(r_g, \theta_g)$, the relative error is plotted as a function of the sequencing depth, where the optimal error is obtained at a depth of one read per cell (orange star) and is two times smaller than that at the current depth of pbmc_4k (red triangle). **c** Experimental design. To determine the sequencing depth for an experiment, first the relative gene expression level can be obtained via pilot experiments or previous studies (top left). Then the researcher can select a set of genes of interest (i.e., some marker genes highlighted as black dots), of which the smallest relative expression level p^* (*MS4A1*) defines the reliable detection limit. Finally, the optimal sequencing depth is determined as $n^*_{\mathrm{reads}} = 1/p^*$ (top right). The errors under different tradeoffs are visualized as a function of the genes ordered from the most expressed to the least (bottom). The optimal sequencing budget allocation (orange) minimizes the worst-case error over all the genes of interest (left of the red dashed line), whereas both the deeper sequencing (green) and the shallower sequencing (blue) yield worse results.

NATURE COMMUNICATIONS | (2020)11:774 | https://doi.org/10.1038/s41467-020-14482-y | www.nature.com/naturecommunications

Spatial RNA Seq

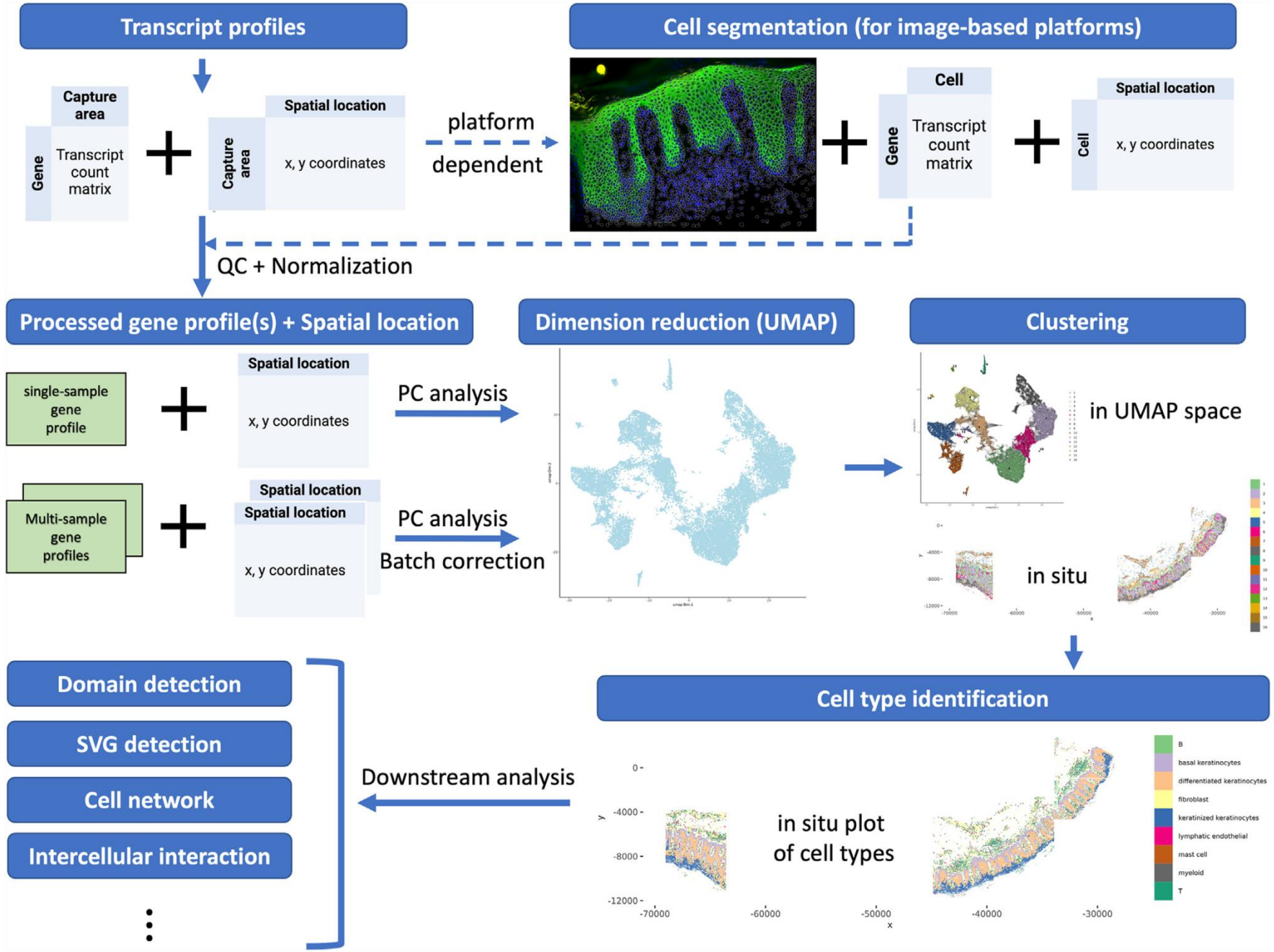
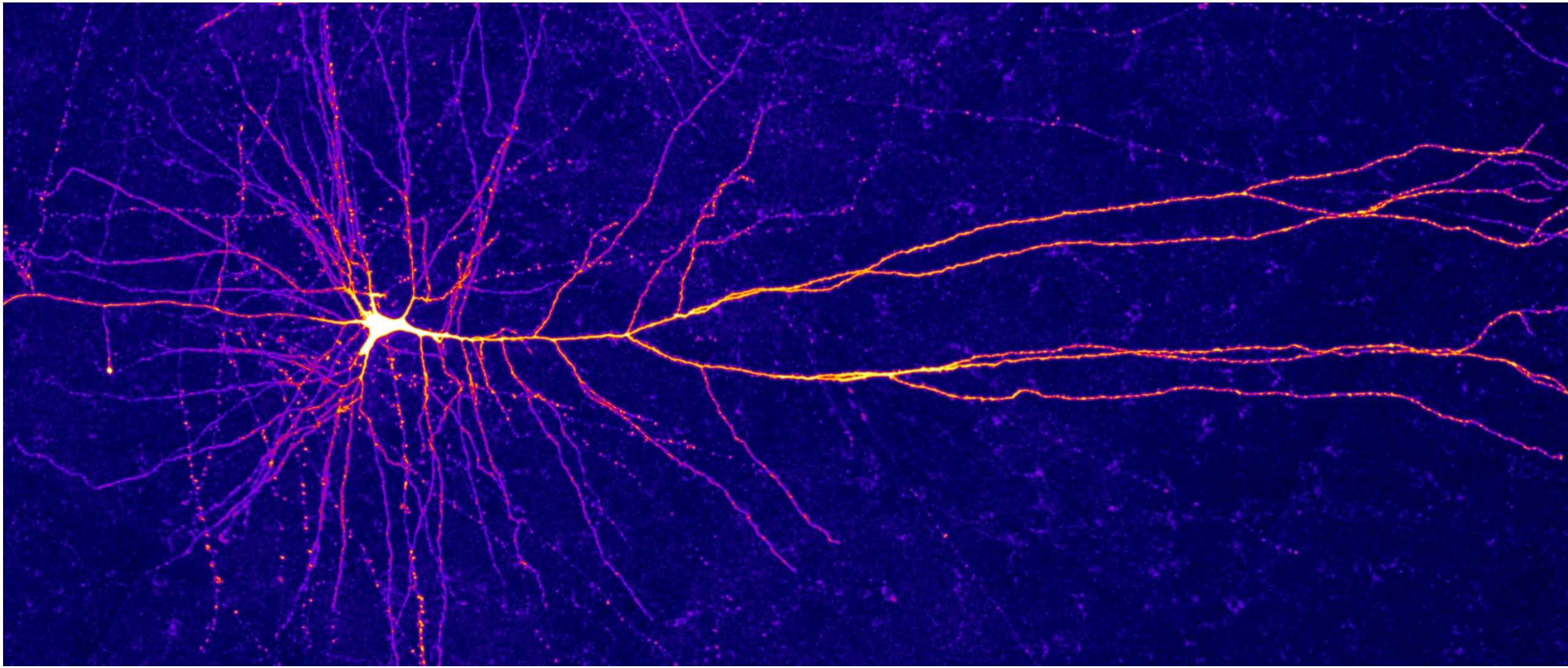


FIG 1. Workflow of spatial transcriptomic analysis. *QC*, Quality control.

Different input, different output



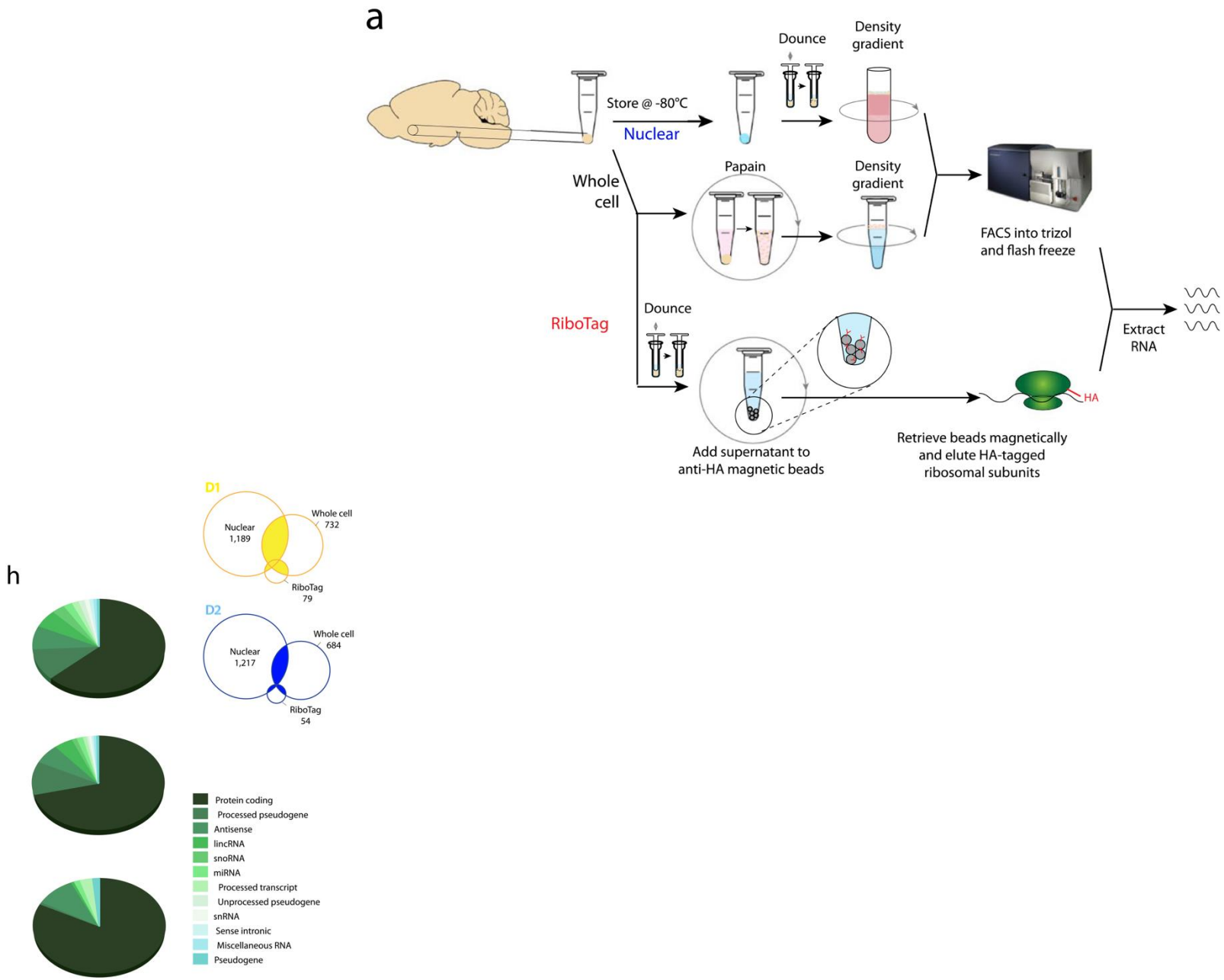


Figure 1. Library characterization demonstrates fewer differentially expressed transcripts and a predominance of protein coding genes in RiboTag compared to whole cell and nuclear RNAseq. (a) Method scheme for whole cell FACS, nuclear FACS, and RiboTag affinity purification showing key differences in steps involving sample preparation, cell dissociation, and retrieval of cellular fractions. (b) Density plot for all methods of ln([average FPKM] + 1) of all genes captured shows that RiboTag does not capture a number of genes. (c) Density plots for all methods of ln([average FPKM] + 1) of differentially expressed genes show similar distributions across method. (d) Overlap of D1- and D2-enriched differentially expressed genes across all methods (total D1 overlap = 134; total D2 overlap = 64). (e) Fold change of differentially expressed genes from most D1-enriched in yellow to most D2-enriched in blue, sorted by fold change in whole cell (black = low fold change, grey = not detected in the dataset). (f) Density plots for each method of ln([average FPKM] + 1) of D1 (dark line) and D2 (light line) differentially expressed genes in the respective cell types; medians are indicated with dashed lines.
(g) Mean-variance plots comparing ln(variance) to ln([average FPKM] + 1) of differentially expressed genes in pooled D1- and D2-MSNs show only slight differences across methods. (h) Gene biotype distributions for each method's differentially expressed genes show a decreasing proportion of protein coding genes from the RiboTag to the whole cell to the nuclear dataset.

Part 2: From Genomes to Targets

The Dream

The NIMH as a case study

The Problem

PSYCHIATRIC DRUG DISCOVERY

Revolution Stalled

Steven E. Hyman

Drug discovery is at a near standstill for treating psychiatric disorders such as schizophrenia, bipolar disorder, depression, and common forms of autism. Despite high prevalence and unmet medical need, major pharmaceutical companies are deemphasizing or exiting psychiatry, thus removing significant capacity from efforts to discover new medicines. In this Commentary, I develop a view of what has gone wrong scientifically and ask what can be done to address this parlous situation.

From an economic perspective, drugs for psychiatric disorders have historically been among the largest sources of revenue (Table 1) for the pharmaceutical industry. Given the high prevalence of psychiatric disorders (*1*), their massive effect on global disease burden (*2*), and limitations in the efficacy of current therapies, large markets for new and better therapies already exist, and current demographic and socioeconomic trends predict the development of expanded markets. Such projections are based on gains in global life expectancy, increasing attention to mental health and cognitive performance in the developed world, rapid unplanned urbanization in the developing world, and large numbers of individuals affected by conflict and postconflict situations worldwide.

Advances continue to be made in modes of cognitive psychotherapy (*3*) and in device-based psychiatric treatments (*4*); but despite the growing market opportunities, major pharmaceutical companies recently announced substantial cutbacks or complete discontinuation of efforts to discover new drugs for psychiatric disorders (*5*, *6*). This exodus creates a dangerous gap in the public health ecosystem, because safe and effective drugs can be readily deployed in both primary care and specialist settings and, when generic, can be highly cost-effective. Although large companies may continue to in-license new drug candidates, their complete or near complete exits from psychiatric research deplete expertise and financial resources from therapeutics discovery. Here, I describe how we arrived at this crossroads and how we might get back on a productive path of discovery.

Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
E-mail: seh@harvard.edu

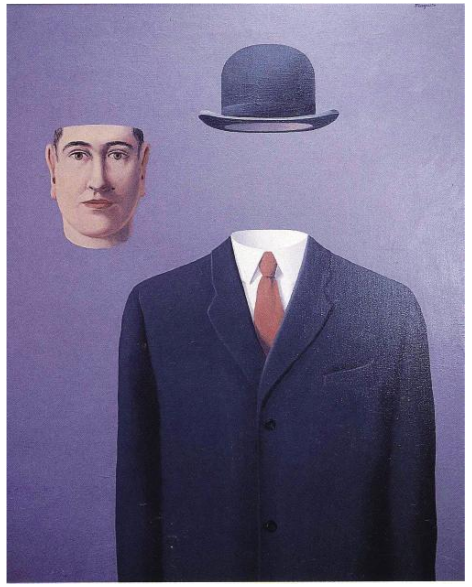


Fig. 1. Terrible thing to waste. Pharma has removed psychiatric diseases from its body of drug discovery projects. Geneticists might help reverse the trend. [*The Pilgrim* by René Magritte. 1966]

BRILLIANT PROMISE UNFULFILLED

John Cade discovered the sedating effects of lithium in 1949, first in guineas pigs and soon thereafter in manic patients. In the remarkably short period that followed—less than a decade—prototypes of the other major classes of psychiatric drugs were identified: the antipsychotic drugs (beginning with chlorpromazine), antidepressants [the tricyclic drug imipramine and the monoamine oxidase inhibitor (MAOI) iproniazid], and benzodiazepines (chlordiazepoxide). Each of these discoveries had an important component of serendipity but also motivated path-breaking research on neurotransmitter release, receptors, and transporters (*7*).

What has happened—or rather not happened—in the intervening half-century was as unexpected as the initial spate of discoveries. Many antidepressant drugs have been developed since the 1950s, but none has improved on the efficacy of imipramine or the first MAOIs, leaving many patients with modest benefits or none at all (*8*, *9*). Antipsychotic drugs achieved a peak in efficacy—never equaled (*10*) and still not understood—with the discovery of clozapine in the mid-1960s. Although valproic acid and other drugs developed as anticonvulsants were shown in the early 1980s to be mood stabilizers, lithium remains a mainstay of treatment for bipolar disorder, despite its serious toxicities. There are still no broadly useful pharmacological treatments for the core symptoms of autism—social deficits, language delay, narrowed interests, and repetitive behaviors—or for the disabling negative (deficit) and cognitive symptoms of schizophrenia (*11*, *12*). The molecular targets of all of today's approved psychiatric drugs are the same as the targets of their pre-1960 prototypes (Table 2), and their mechanisms of action are not understood beyond a few initial molecular events (*13*). Indeed, the critical molecular target (or targets) for lithium have not been established with certainty (*13*).

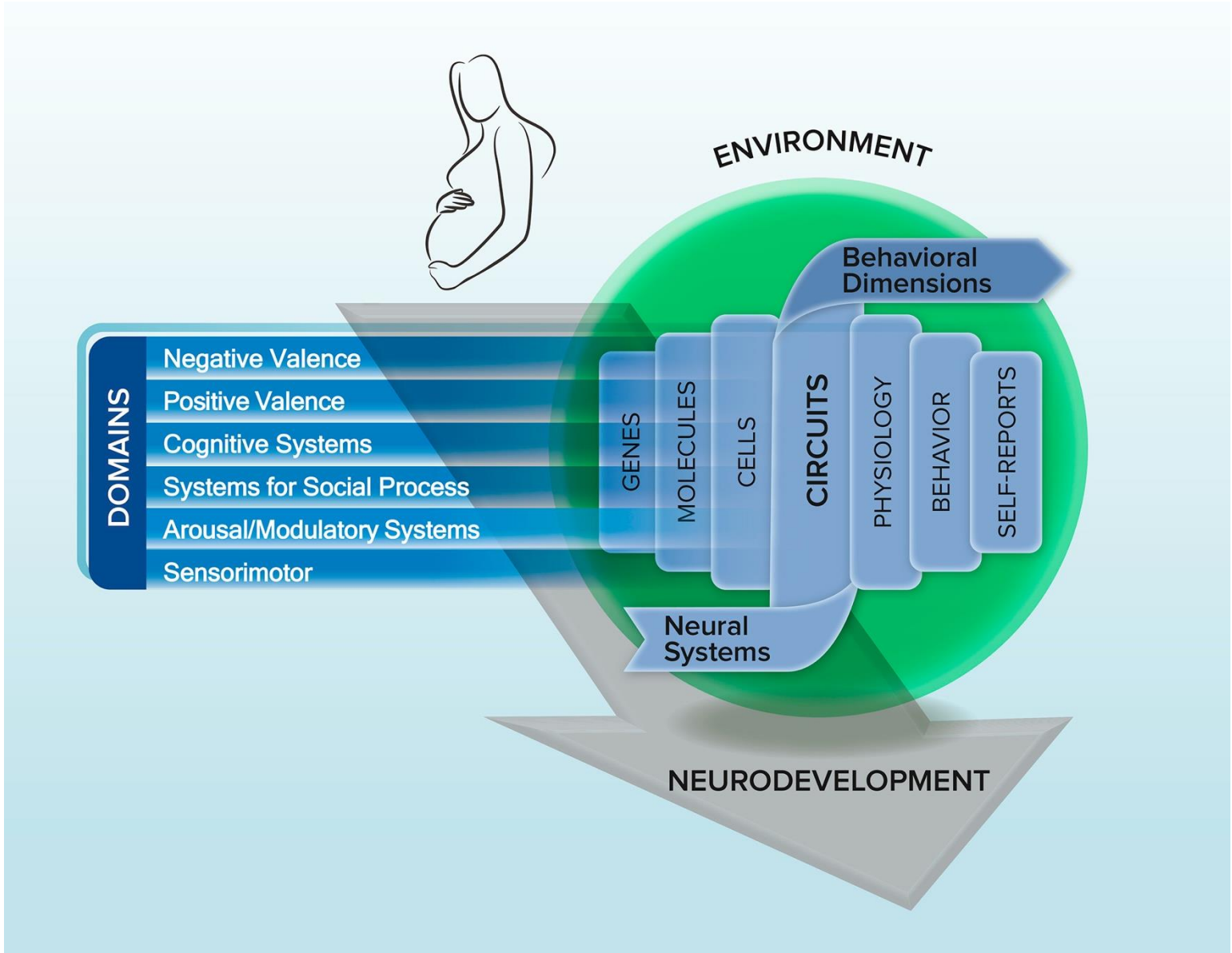
There has been some progress, of course. Important advances have been made in the safety and tolerability of antidepressants. More recently, clinical observation and small clinical trials with ketamine have suggested that the *N*-methyl-D-aspartate (NMDA) glutamate receptor channel represents a possible new target for antidepressants associated with more rapid onset of therapeutic effects (*14*). A second generation of antipsychotic drugs was developed on the

The Dream

The NIMH as a case study

The Solution?

-GWAS validated hypothesis
-Genome wide & Computational approaches
-Integration across dimensions
-Scrap legacy models



Sanilow et al., 2019 Neuron

The Dream

The NIMH as a case study

No new Therapeutic Principle from rational approaches

Current development centered around rediscovering old “socially controversial” drugs.

The Result so far: THC, Psychadelics, Ketamine…

Who said only AI can hallucinate?

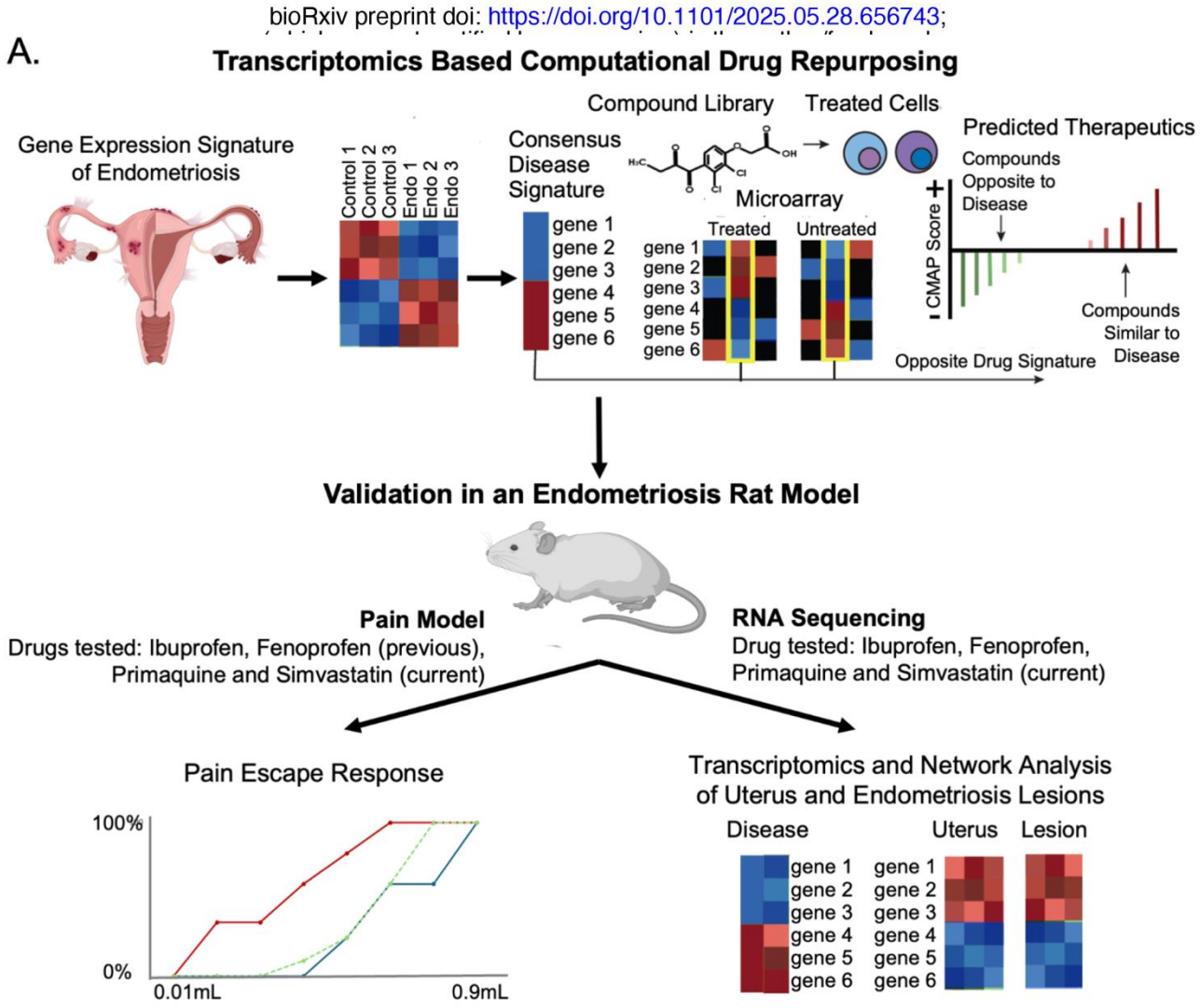
So, what can we do with genomic information?
What can we learn to guide
-Personalized medicine
-Toxicology
-Target identification

Here are a few existing studies

So, what can we do with genomic information?

What can we learn to guide
-Personalized medicine
-Toxicology
-Target identification
-Drug repurposing based on genome wide impact?

Here are a few existing studies



Interesting design but limited in scope

Published in final edited form as:
Nat Med. 2017 September ; 23(9): 1102–1111. doi:10.1038/nm.4386.

Sex-Specific Transcriptional Signatures in Human Depression

Benoit Labonté[1], **Olivia Engmann**[1], **Immanuel Purushothaman**[1], **Caroline Menard**[1], **Junshi Wang**[2], **Chunfeng Tan**[3], **Joseph R Scarpa**[1,4], **Gregory Moy**[1], **Yong-Hwee E Loh**[1], **Michael Cahill**[1], **Zachary S Lorsch**[1], **Peter J. Hamilton**[1], **Erin S Calipari**[1], **Georgia E. Hodes**[1], **Orna Issler**[1], **Hope Kronman**[1], **Madeline Pfau**[1], **Aleksander Obradovic**[1], **Yan Dong**[2], **Rachel Neve**[5], **Scott Russo**[1], **Andrew Kazarskis**[4], **Carol Tamminga**[3], **Naguib Mechawar**[6,7], **Gustavo Turecki**[6,7], **Bin Zhang**[4,*], **Li Shen**[1,*], and **Eric J Nestler**[1,*]

Sex as a variable

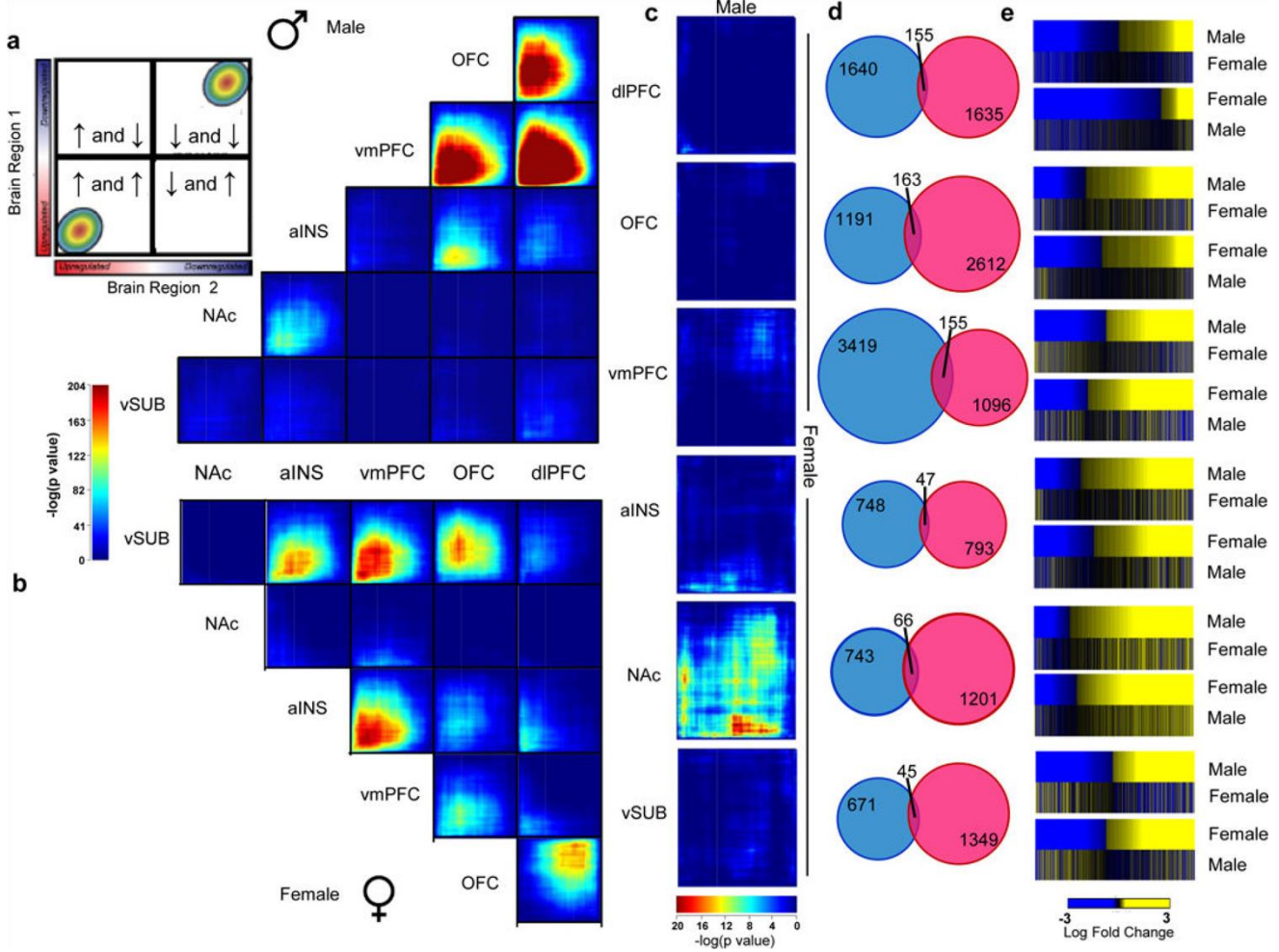


Figure 1.

Differential expression profiles in humans with MDD reveal distinct sex-specific transcriptional signatures across brain regions. **a**,**b**, Rank-rank hypergeometric overlap (RRHO) maps comparing region to region transcriptional profiles in **a**, males and **b**, females with MDD. The upper left panel in **a** displays the overlap relationship across brain regions. The color bar between **a** and **b** represents degree of significance. **c**, RRHO maps directly comparing male and female transcriptional profiles across brain regions. Degree of significance is depicted in the color bar below the RRHO maps. **d**, Venn diagrams displaying low overlap between genes differentially expressed ($p<0.05$) in males (blue) and females (pink) across brain regions. **e**, Heatmaps comparing transcriptional changes (log fold change; below the heatmaps) in males and females with MDD compared to controls across brain regions.

Published in final edited form as:
Nat Med. 2017 September ; 23(9): 1102–1111. doi:10.1038/nm.4386.

Sex-Specific Transcriptional Signatures in Human Depression

Benoit Labonté[1], **Olivia Engmann**[1], **Immanuel Purushothaman**[1], **Caroline Menard**[1], **Junshi Wang**[2], **Chunfeng Tan**[3], **Joseph R Scarpa**[1,4], **Gregory Moy**[1], **Yong-Hwee E Loh**[1], **Michael Cahill**[1], **Zachary S Lorsch**[1], **Peter J. Hamilton**[1], **Erin S Calipari**[1], **Georgia E. Hodes**[1], **Orna Issler**[1], **Hope Kronman**[1], **Madeline Pfau**[1], **Aleksander Obradovic**[1], **Yan Dong**[2], **Rachel Neve**[5], **Scott Russo**[1], **Andrew Kazarskis**[4], **Carol Tamminga**[3], **Naguib Mechawar**[6,7], **Gustavo Turecki**[6,7], **Bin Zhang**[4,*], **Li Shen**[1,*], and **Eric J Nestler**[1,*]

Gene expression signatures are different across sex

Different gene expression signature have the same phenotype

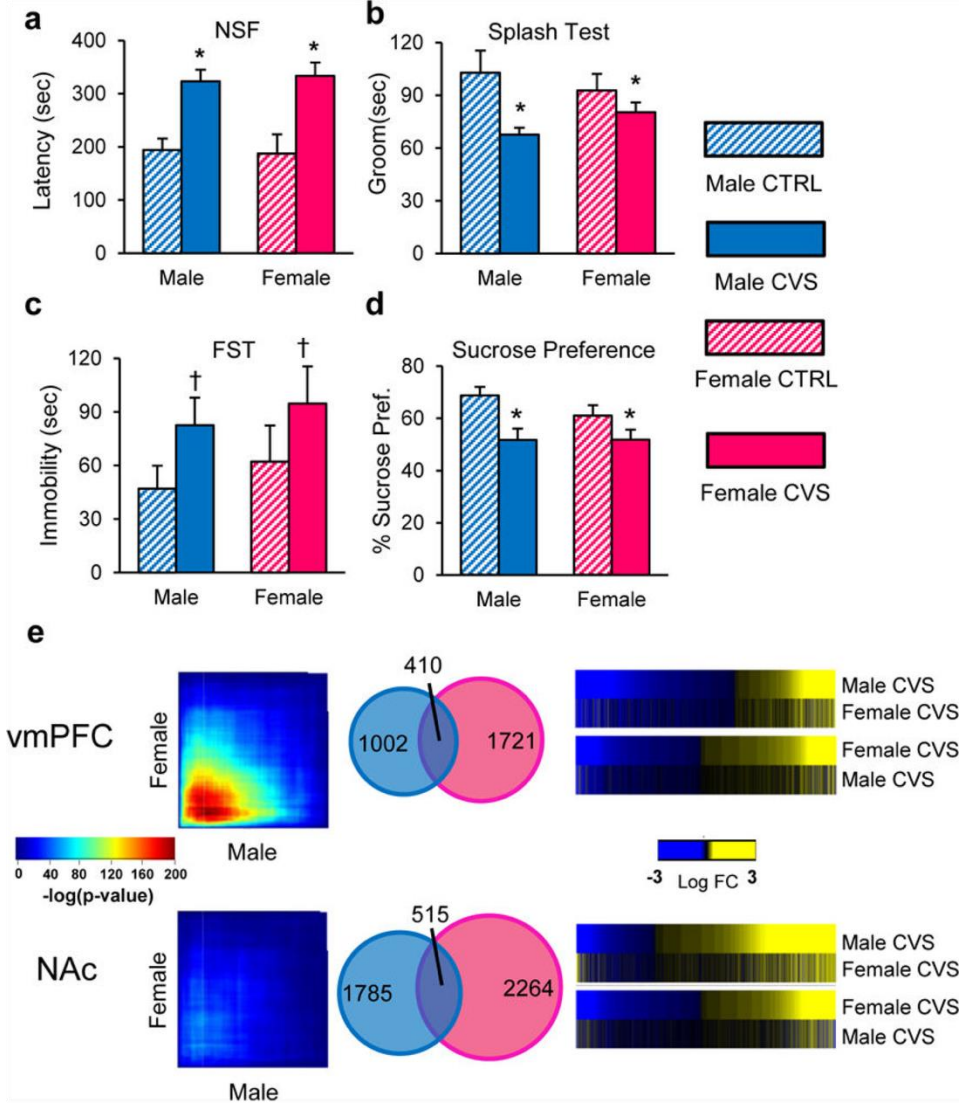


Figure 2.

Chronic variable stress (CVS) induces an equivalent depressive-like phenotype in male and female mice despite the induction of largely distinct transcriptional profiles. **a**, Quantification of latency to eat in the novelty suppressed feeding (NSF) test, **b** time spent grooming in the splash test, **c** time swimming in the forced swim test (FST) and **d** sucrose preference in male (blue) and female (pink) mice. Bars, mean ± sem; * $p<0.05$; † $p<0.1$. **e**, RRHO maps comparing male and female stressed mice in the vmPFC and NAc. Degree of significance is depicted in the color bar in between the RRHO maps. Venn diagrams displaying overlap between genes differentially expressed ($p<0.05$) in male (blue) and female (pink) stressed mice in both brain regions. Heatmaps comparing transcriptional

Cell

The coding genome
Modulating drug response
and population heterogeneity

Pharmacogenomics of GPCR Drug Targets

Alexander S. Hauser,[1,2,*] Sreenivas Chavali,[1] Ikuo Masuho,[3] Leonie J. Jahn,[4] Kirill A. Martemyanov,[3] David E. Gloriam,[2] and M. Madan Babu[1,5,*]

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK
[2]Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark
[3]Department of Neuroscience, The Scripps Research Institute Florida, Jupiter, FL 33458, USA
[4]The Novo Nordisk Foundation Center for Biosustainability, Technical University Denmark, Kemitorvet 2800 Kgs. Lyngby, Denmark
[5]Lead Contact
*Correspondence: alexander.hauser@sund.ku.dk (A.S.H.), madanm@mrc-lmb.cam.ac.uk (M.M.B.)
https://doi.org/10.1016/j.cell.2017.11.033

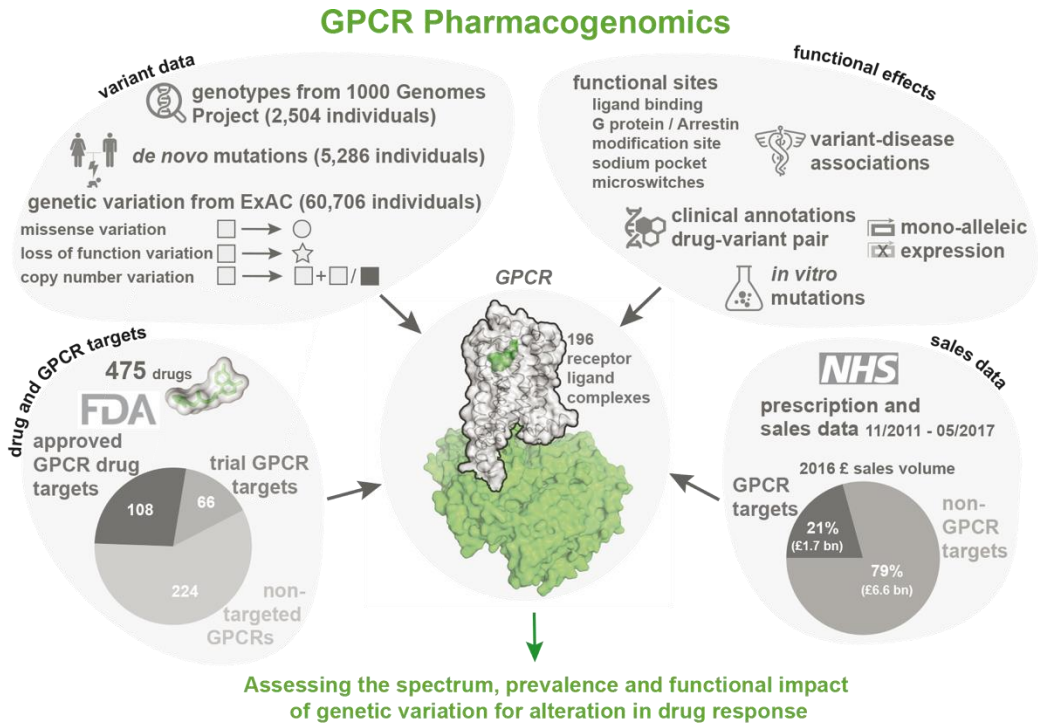


Figure 1. Pharmacogenomic Landscape of GPCR Drug Targets
Schematic highlighting the different types of data analyzed in this study, ranging from data on drug targets, variants, functional effects, sequences, structures to information on prescription, and sales of drugs in the UK.

SUMMARY

Natural genetic variation in the human genome is a cause of individual differences in responses to medications and is an underappreciated burden on public health. Although 108 G-protein-coupled receptors (GPCRs) are the targets of 475 (~34%) Food and Drug Administration (FDA)-approved drugs and account for a global sales volume of over 180 billion US dollars annually, the prevalence of genetic variation among GPCRs targeted by drugs is unknown. By analyzing data from 68,496 individuals, we find that GPCRs targeted by drugs show genetic variation within functional regions such as drug- and effector-binding sites in the human population. We experimentally show that certain variants of μ-opioid and Cholecystokinin-A receptors could lead to altered or adverse drug response. By analyzing UK National Health Service drug prescription and sales data, we suggest that characterizing GPCR variants could increase prescription precision, improving patients' quality of life, and relieve the economic and societal burden due to variable drug responsiveness.

Cell

Pharmacogenomics of GPCR Drug Targets

Alexander S. Hauser,[1,2,*] Sreenivas Chavali,[1] Ikuo Masuho,[3] Leonie J. Jahn,[4] Kirill A. Martemyanov,[3] David E. Gloriam,[2] and M. Madan Babu[1,5,*]

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK
[2]Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark
[3]Department of Neuroscience, The Scripps Research Institute Florida, Jupiter, FL 33458, USA
[4]The Novo Nordisk Foundation Center for Biosustainability, Technical University Denmark, Kemitorvet 2800 Kgs. Lyngby, Denmark
[5]Lead Contact
*Correspondence: alexander.hauser@sund.ku.dk (A.S.H.), madanm@mrc-lmb.cam.ac.uk (M.M.B.)
https://doi.org/10.1016/j.cell.2017.11.033

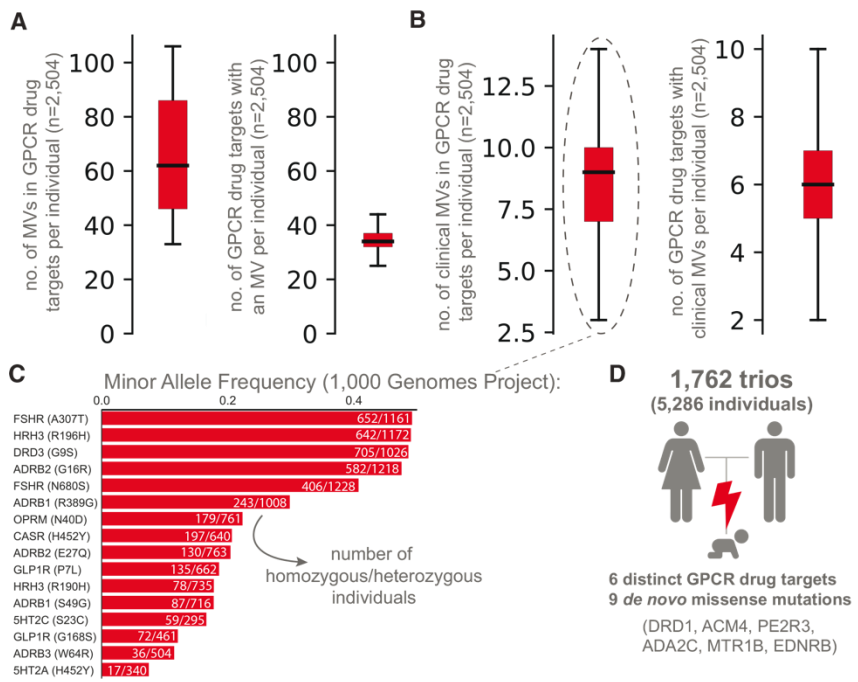


Figure 2. Distribution of Individuals Harboring Missense Variation in GPCR Drug Targets

(A and B) Estimates based on genotype data from 2,504 individual genomes was made per individual on (A) number of missense variants in GPCR drug targets (left) and the number of GPCR drug targets harboring a missense variation (right) and (B) number of clinically known variants that alter efficacy of drug response or toxicity in GPCR drug targets (left) and the number of affected GPCR drug targets with clinically known mutations (right).

(C) Allele frequencies of variants, known to alter drug response in 2,504 individuals (number of homozygous/heterozygous carriers) (Table S2).

(D) Analysis of 1,762 studied trios (father-mother-offspring) revealed a total of 9 *de novo* missense mutations in 6 GPCR drug targets.

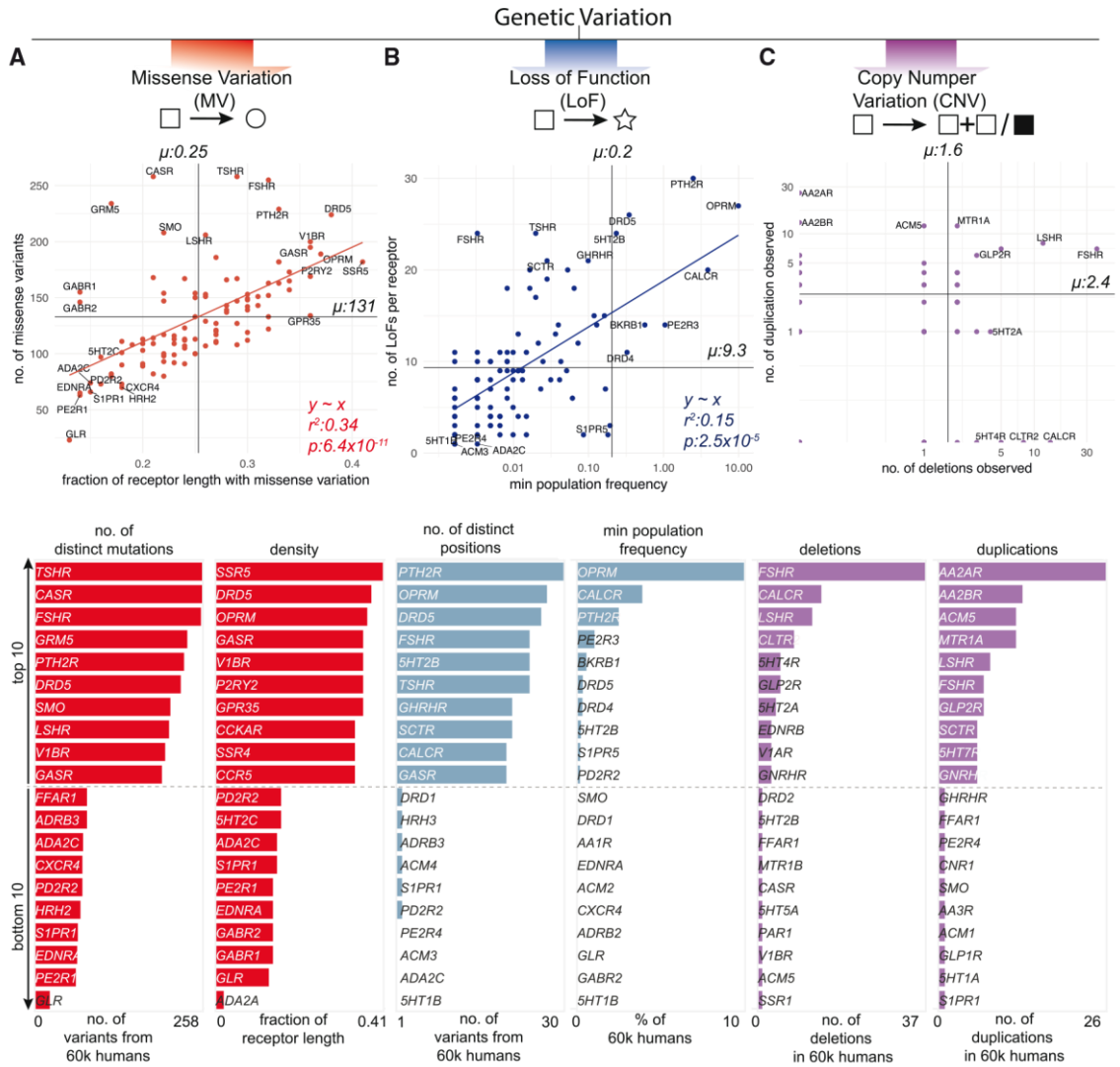
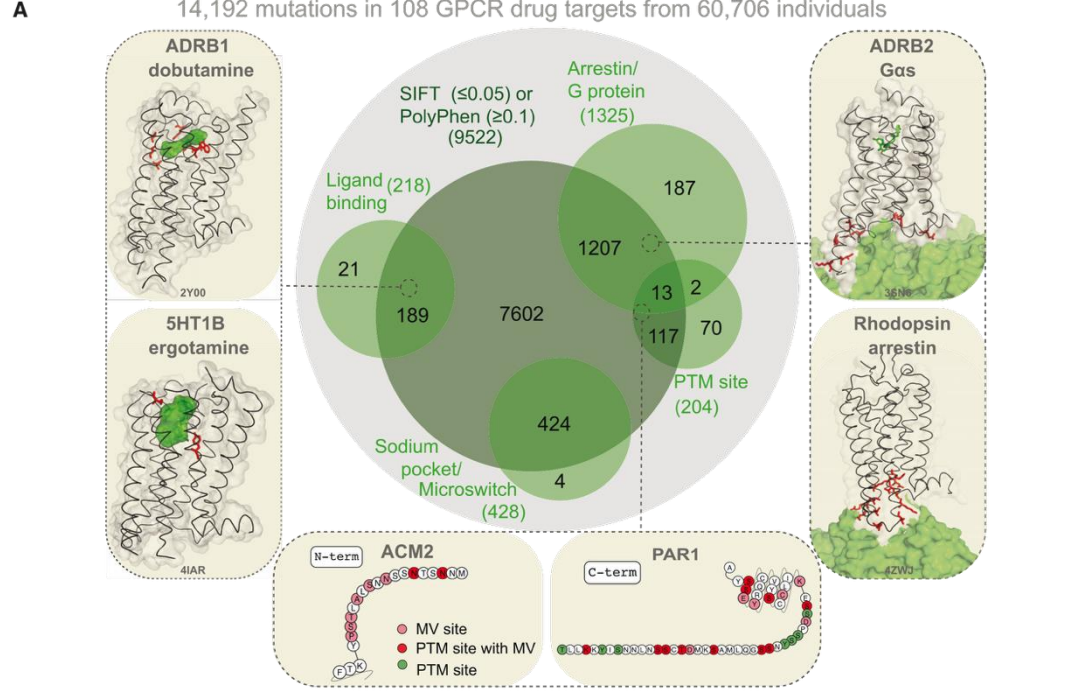
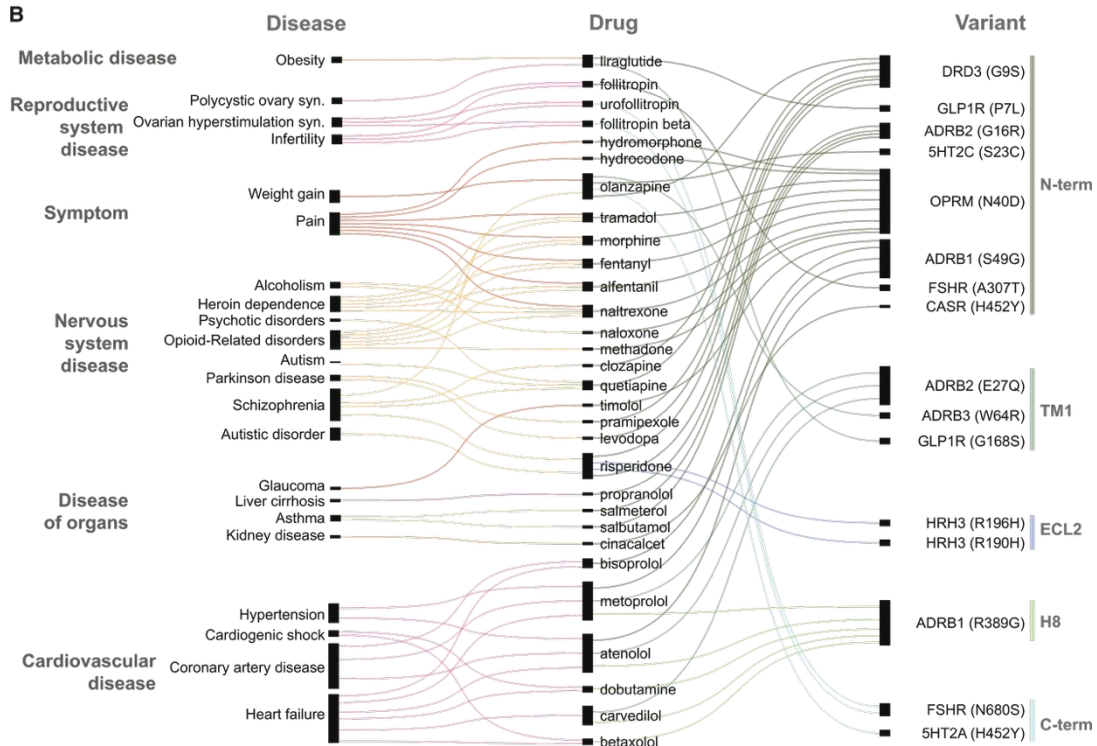


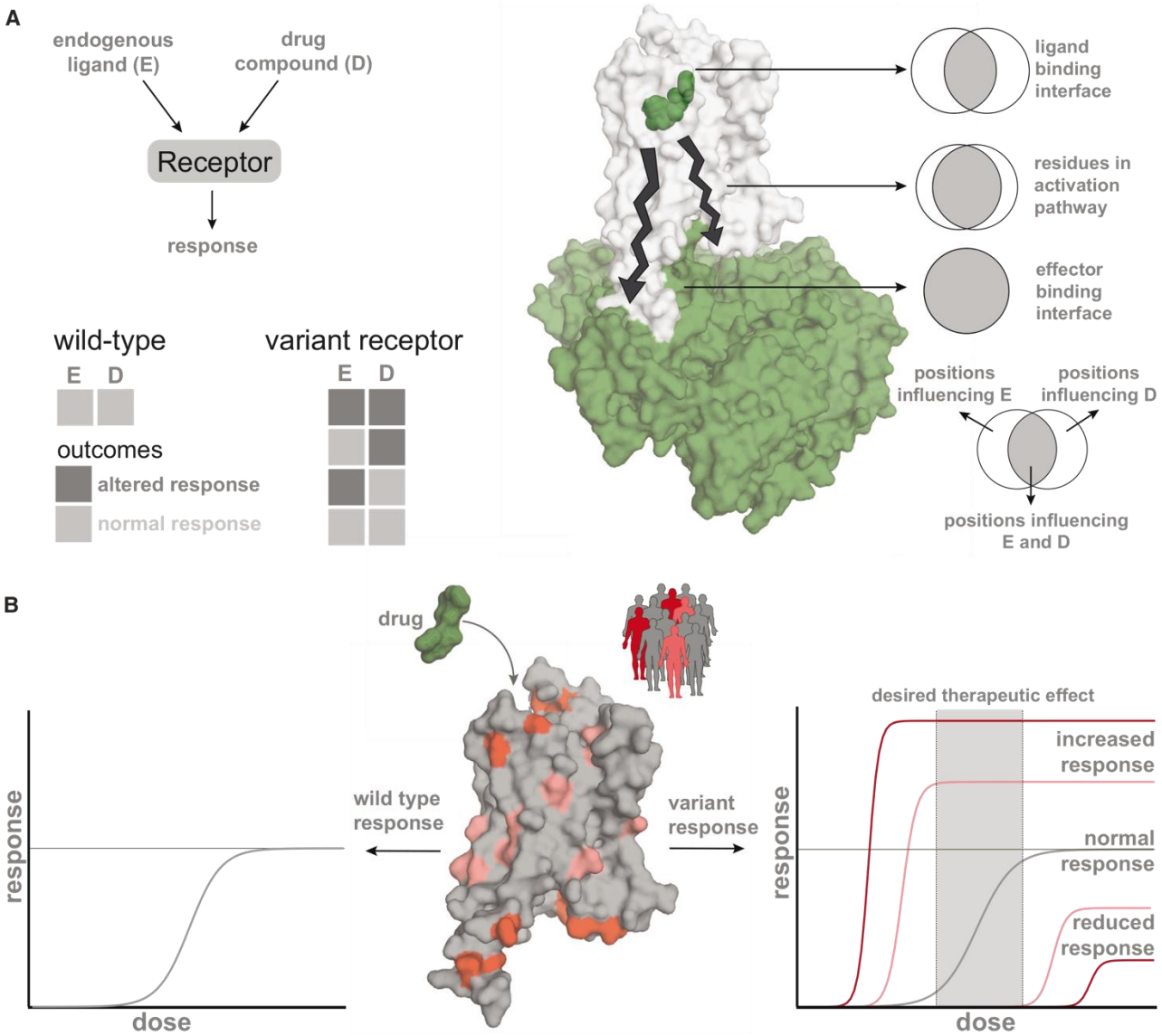
Figure 3. Genetic Variation Landscape of GPCR Drug Targets

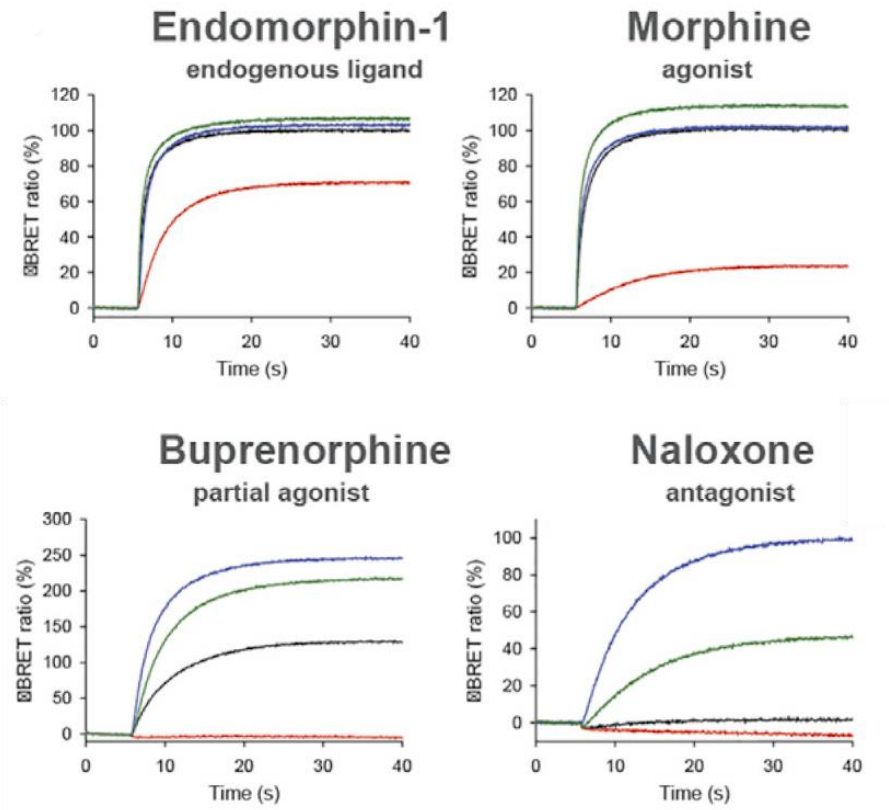
(A–C) Scatterplots of (A) missense variation (red), (B) loss-of-function mutations (blue), and (C) copy-number variation (purple) for GPCR drug targets. Each mutation type shows the number of observed variants (separated into deletions and duplications for CNVs) for a given GPCR drug target. Missense variation density was obtained by normalizing number of missense mutations to the receptor sequence length. Loss-of-function mutations are presented as the minimum percentage of individuals harboring at least one copy of a protein-truncating variant (STAR Methods). Correlations and mean values (μ) are shown for MVs and LoFs. Mean values (μ) for the distributions are provided. Genetic variation landscapes of GPCR drug targets that are in clinical trials are provided in Table S3 and S4). Lower half of the figure shows the distribution of top 10 (upper panels) and bottom 10 ranking GPCR drug targets (lower panels).

See also Figures S1, S2, and S3.

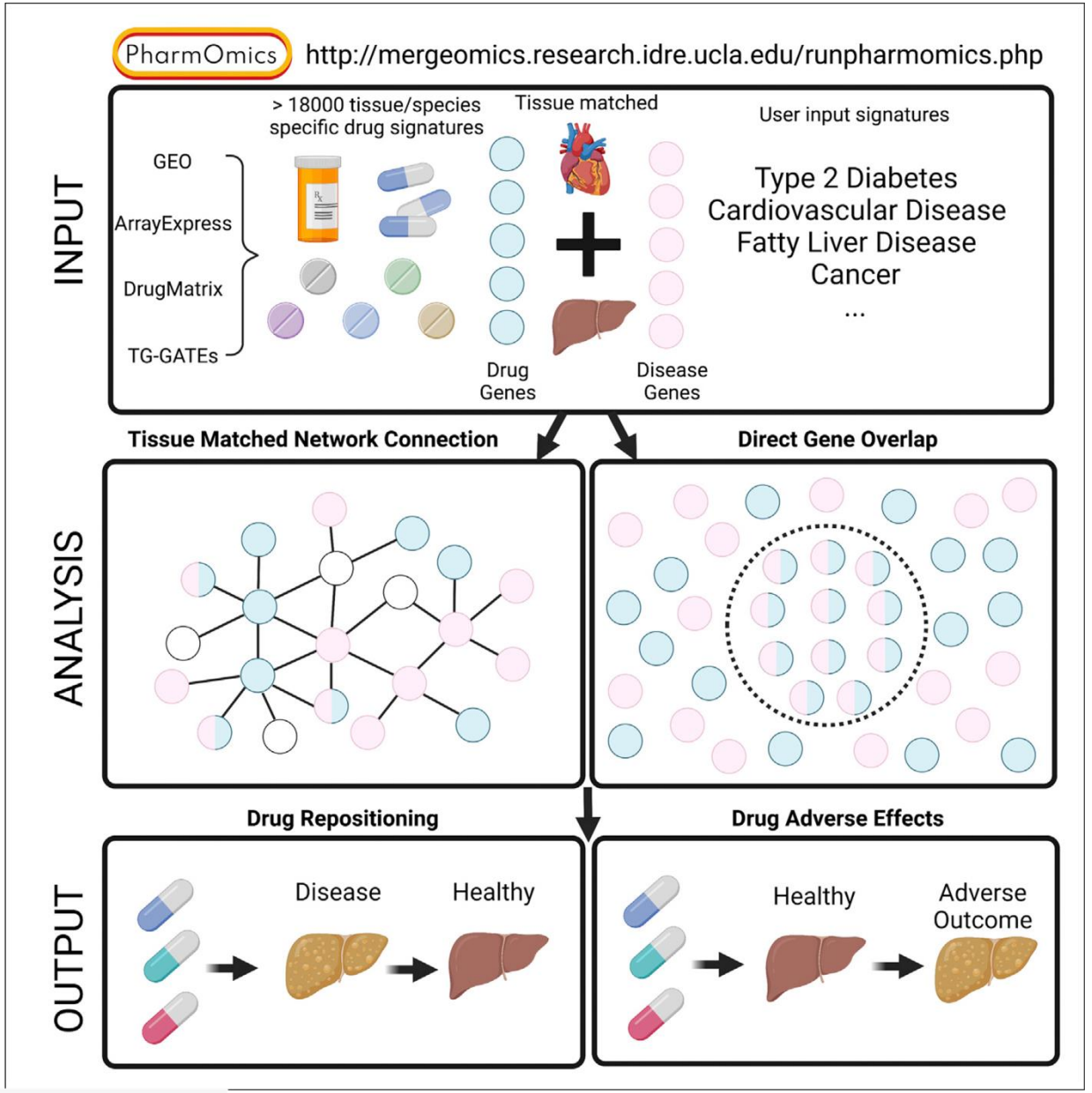

A
14,192 mutations in 108 GPCR drug targets from 60,706 individuals
ADRB1
dobutamine
2Y00
5HT1B
ergotamine
4IAR
SIFT (≤0.05) or
PolyPhen (≥0.1)
(9522)
Arrestin/
G protein
(1325)
Ligand
binding
(218)
PTM site
(204)
Sodium
pocket/
Microswitch
(428)
21
189
7602
1207
187
13
2
117
70
424
4
ADRB2
Gαs
3SN6
Rhodopsin
arrestin
4ZWJ
N-term
ACM2
C-term
PAR1
MV site
PTM site with MV
PTM site


B
Disease
Drug
Variant
Metabolic disease
Obesity
Reproductive system disease
Polycystic ovary syn.
Ovarian hyperstimulation syn.
Infertility
Symptom
Weight gain
Pain
Nervous system disease
Alcoholism
Heroin dependence
Psychotic disorders
Opioid-Related disorders
Autism
Parkinson disease
Schizophrenia
Autistic disorder
Disease of organs
Glaucoma
Liver cirrhosis
Asthma
Kidney disease
Cardiovascular disease
Hypertension
Cardiogenic shock
Coronary artery disease
Heart failure
liraglutide
follitropin
urofollitropin
follitropin beta
hydromorphone
hydrocodone
olanzapine
tramadol
morphine
fentanyl
alfentanil
naltrexone
naloxone
methadone
clozapine
quetiapine
timolol
pramipexole
levodopa
risperidone
propranolol
salmeterol
salbutamol
cinacalcet
bisoprolol
metoprolol
atenolol
dobutamine
carvedilol
betaxolol
DRD3 (G9S)
GLP1R (P7L)
ADRB2 (G16R)
5HT2C (S23C)
OPRM (N40D)
ADRB1 (S49G)
FSHR (A307T)
CASR (H452Y)
N-term
ADRB2 (E27Q)
ADRB3 (W64R)
GLP1R (G168S)
TM1
HRH3 (R196H)
HRH3 (R190H)
ECL2
ADRB1 (R389G)
H8
FSHR (N680S)
5HT2A (H452Y)
C-term


A
endogenous ligand (E)
drug compound (D)
Receptor
response
wild-type
E D
variant receptor
E D
outcomes
altered response
normal response
ligand binding interface
residues in activation pathway
effector binding interface
positions influencing E
positions influencing D
positions influencing E and D
B
drug
wild type response
variant response
response
dose
desired therapeutic effect
increased response
normal response
reduced response
response
dose


Endomorphin-1
endogenous ligand
Morphine
agonist
Buprenorphine
partial agonist
Naloxone
antagonist
ΔBRET ratio (%)
Time (s)

Predicting drug efficacy and toxicity from drug induced changes in expression profile



Chen et al., iScience *25*, 104052
April 15, 2022
https://doi.org/10.1016/j.isci.2022.104052

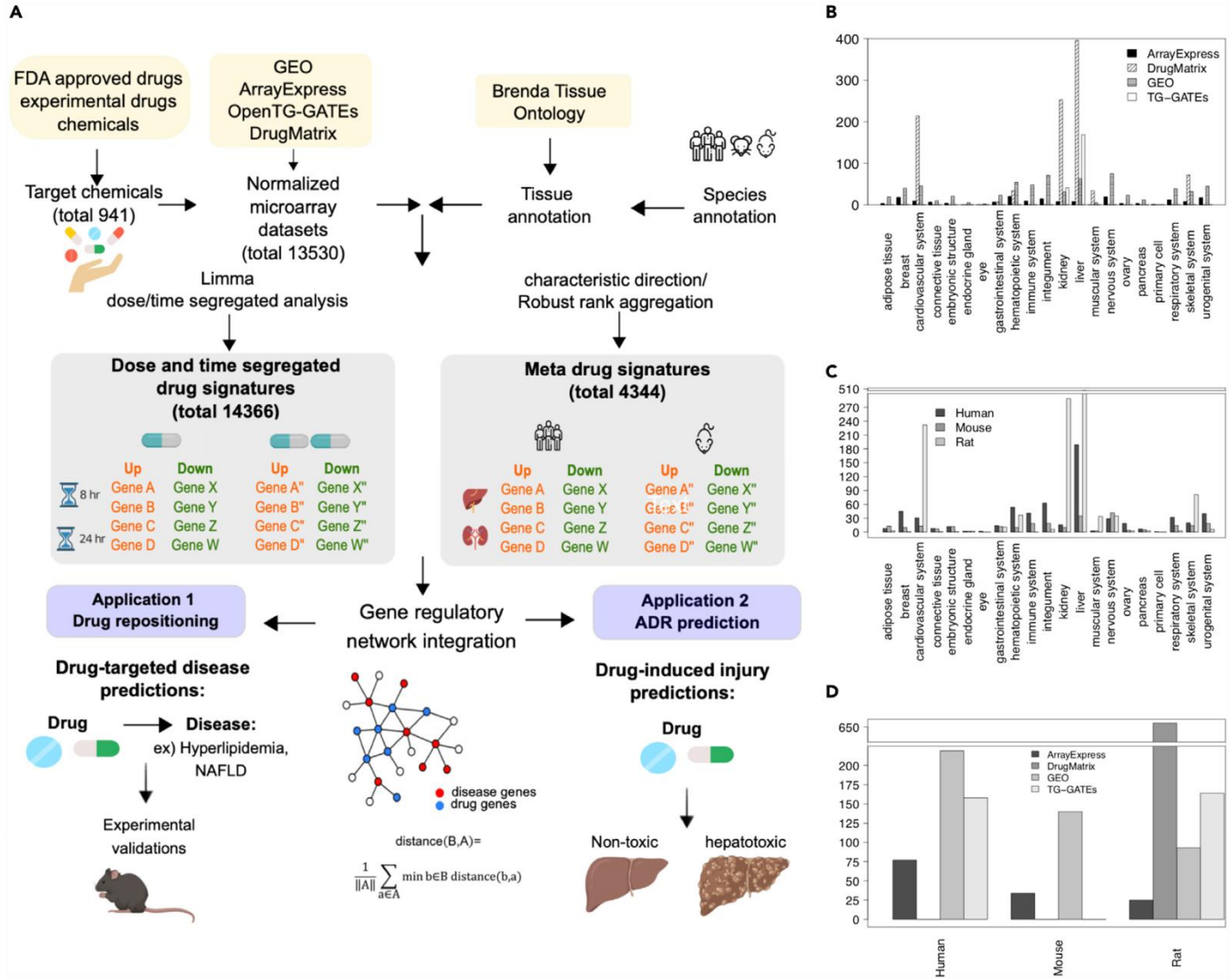


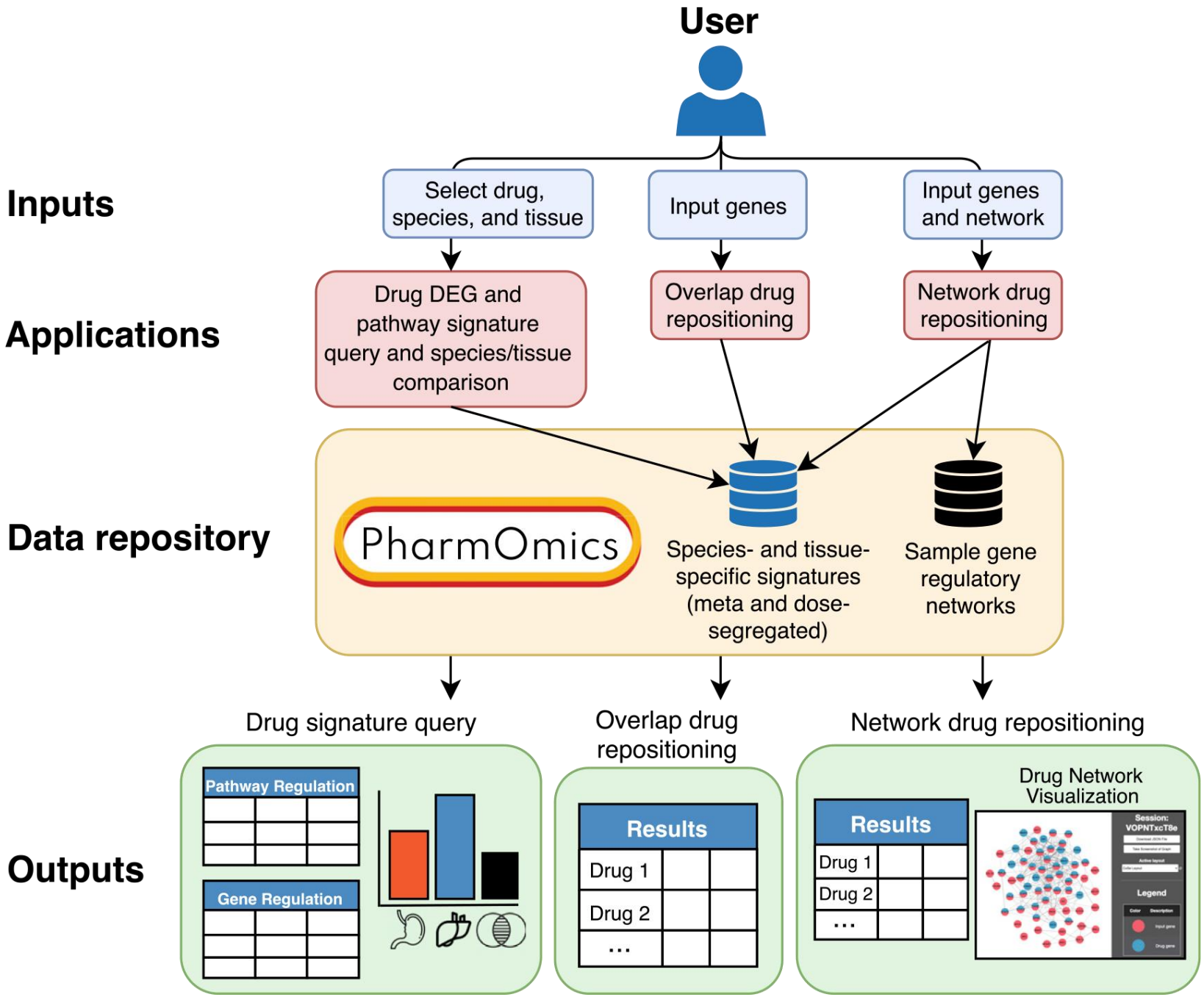
Figure 1. PharmOmics data processing pipeline and database summary

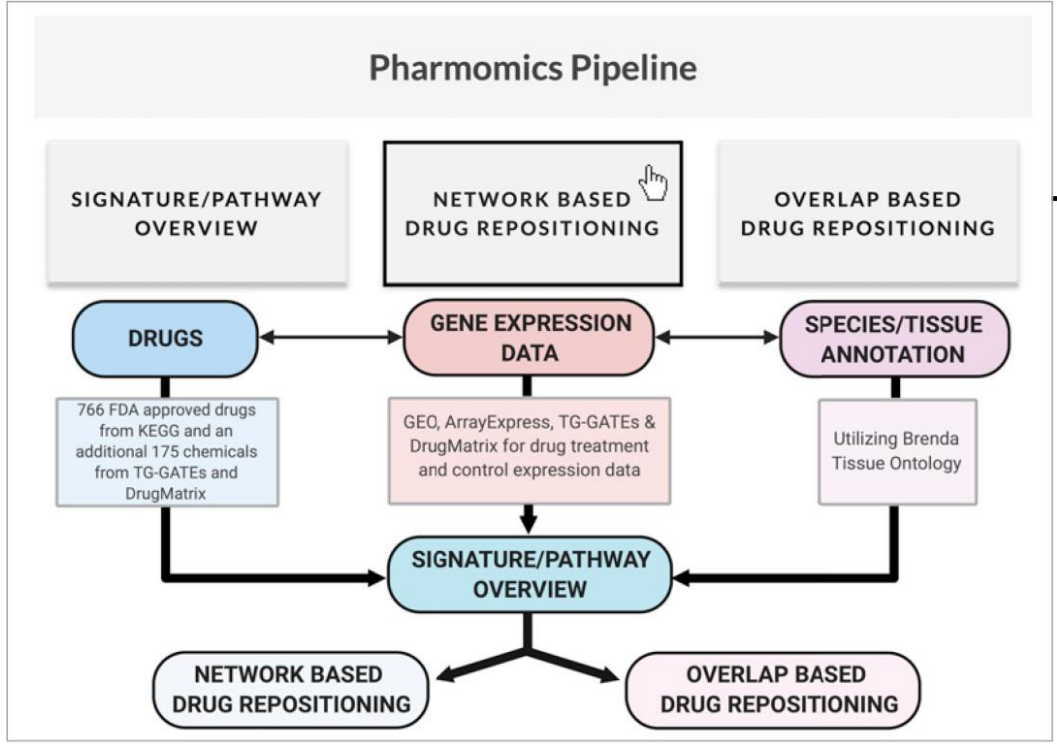
(A) FDA-approved drugs were searched against GEO, ArrayExpress, TG-GATEs, and DrugMatrix data repositories. Additional experimental drugs and chemicals from TG-GATEs and DrugMatrix were also included. Datasets were first annotated with tissue and species information, followed by retrieval of dose-/time-segregated signatures using LIMMA (Ritchie et al., 2015) or meta-analysis drug signatures using GeoDE (Clark et al., 2014) and Robust Rank Aggregation (Kolde et al., 2012). These signatures were used to conduct drug repositioning analysis and hepatotoxicity prediction based on either direct gene overlaps or a gene-network-based approach.

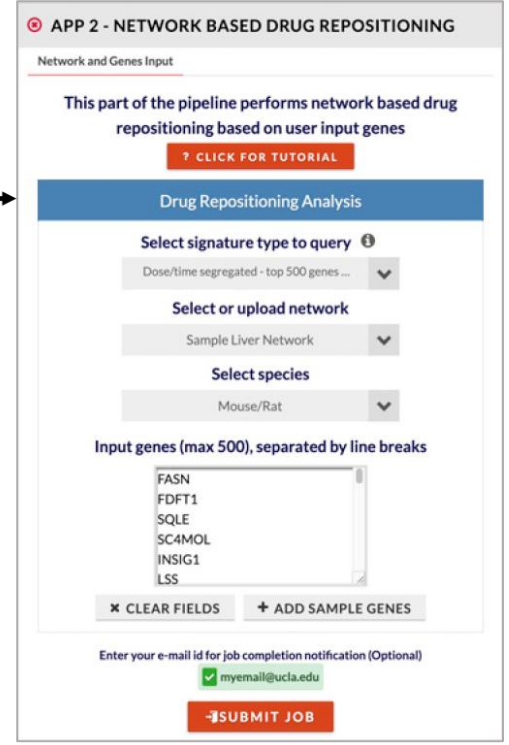
(B) Summary of available datasets based on data sources and tissues. Y axis indicates unique dataset counts, and X axis indicates tissue and data resources.

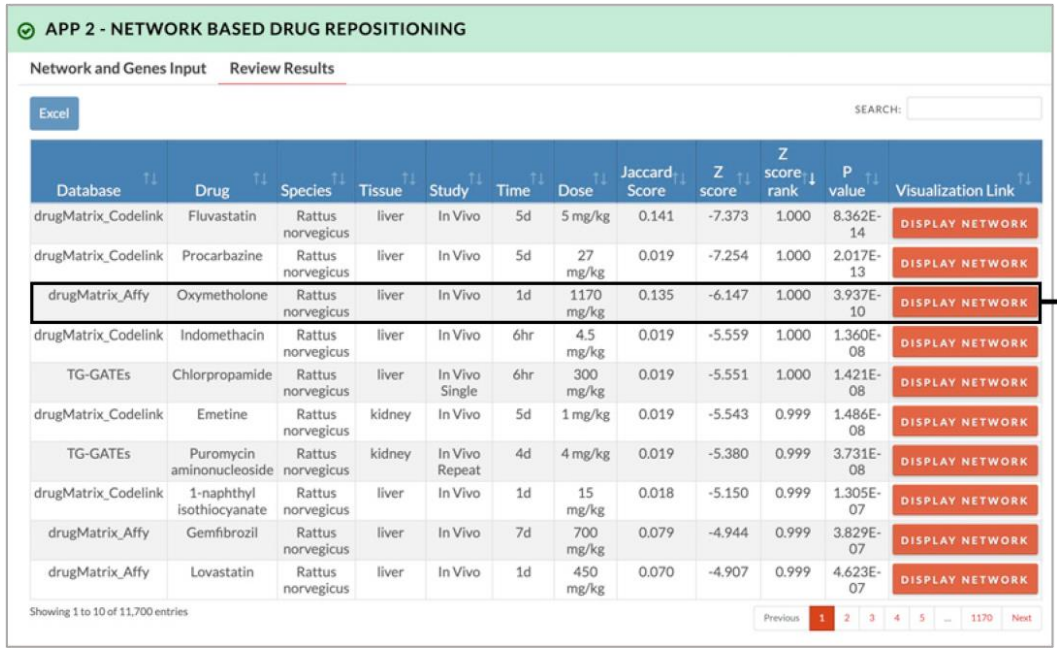
(C) Summary of available datasets based on tissues and species. Y axis indicates unique dataset counts, and X axis indicates tissue and species.

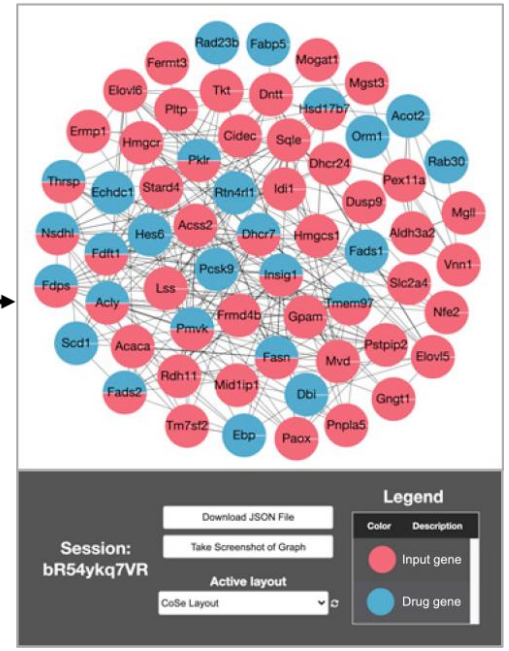
(D) Summary of available datasets based on data sources and species. Y axis indicates unique dataset counts, and X axis indicates data resources and species.


User
Inputs
Applications
Data repository
Outputs
Select drug, species, and tissue
Input genes
Input genes and network
Drug DEG and pathway signature query and species/tissue comparison
Overlap drug repositioning
Network drug repositioning
PharmOmics
Species- and tissue-specific signatures (meta and dose-segregated)
Sample gene regulatory networks
Drug signature query
Overlap drug repositioning
Network drug repositioning
Pathway Regulation
Gene Regulation
Results
Drug 1
Drug 2
...
Drug Network Visualization
Results
Drug 1
Drug 2
...
Session: VOPNTxcT8e
Legend

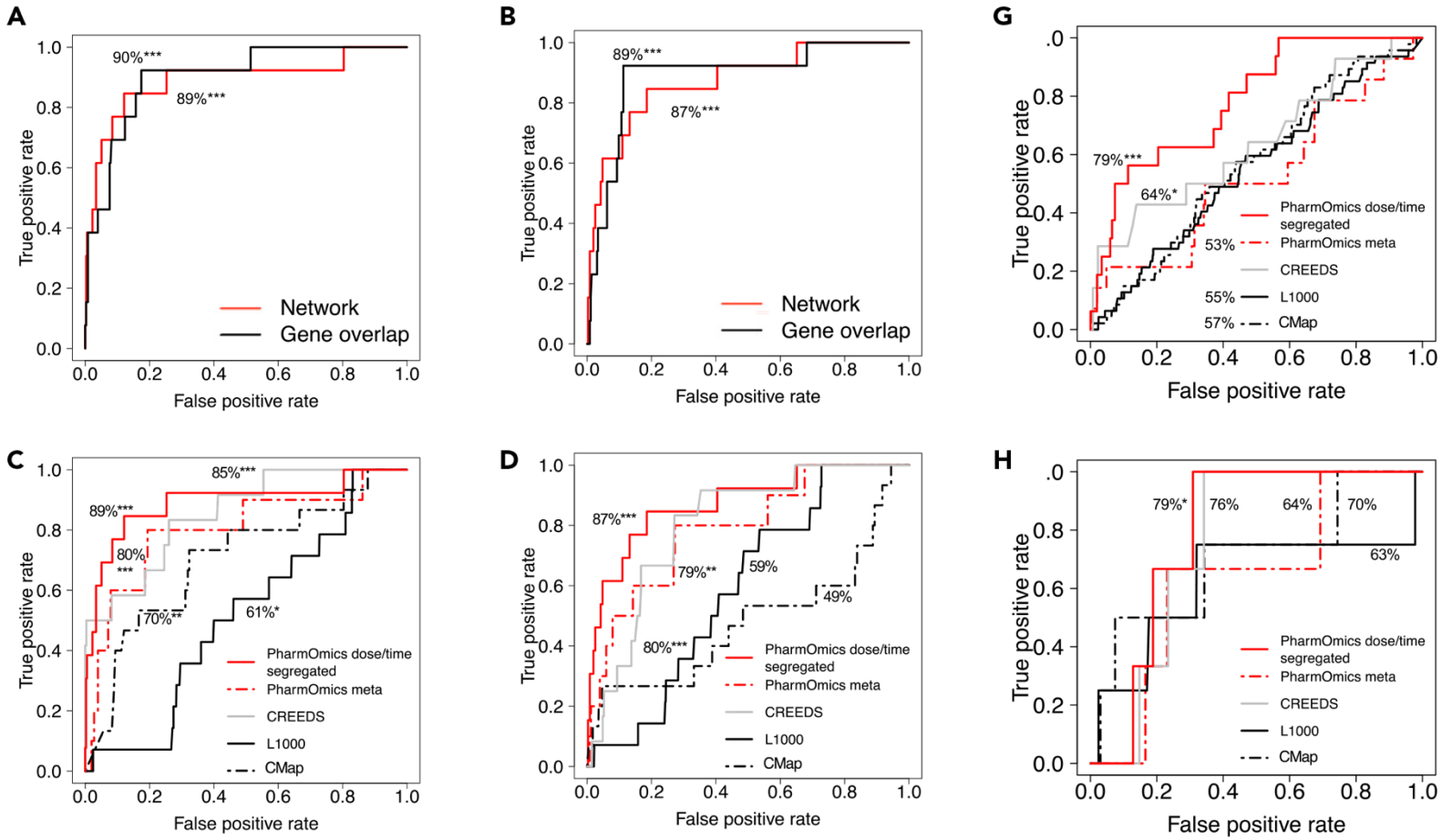

Pharmomics Pipeline
SIGNATURE/PATHWAY OVERVIEW
NETWORK BASED DRUG REPOSITIONING
OVERLAP BASED DRUG REPOSITIONING
DRUGS
GENE EXPRESSION DATA
SPECIES/TISSUE ANNOTATION
766 FDA approved drugs from KEGG and an additional 175 chemicals from TG-GATEs and DrugMatrix
GEO, ArrayExpress, TG-GATEs & DrugMatrix for drug treatment and control expression data
Utilizing Brenda Tissue Ontology
SIGNATURE/PATHWAY OVERVIEW
NETWORK BASED DRUG REPOSITIONING
OVERLAP BASED DRUG REPOSITIONING


APP 2 - NETWORK BASED DRUG REPOSITIONING
Network and Genes Input
This part of the pipeline performs network based drug repositioning based on user input genes
? CLICK FOR TUTORIAL
Drug Repositioning Analysis
Select signature type to query
Dose/time segregated - top 500 genes ...
Select or upload network
Sample Liver Network
Select species
Mouse/Rat
Input genes (max 500), separated by line breaks
FASN
FDFT1
SQLE
SC4MOL
INSIG1
LSS
✖ CLEAR FIELDS
+ ADD SAMPLE GENES
Enter your e-mail id for job completion notification (Optional)
myemail@ucla.edu
SUBMIT JOB


APP 2 - NETWORK BASED DRUG REPOSITIONING
Network and Genes Input
Review Results
Excel
SEARCH:
Database | Drug | Species | Tissue | Study | Time | Dose | Jaccard Score | Z score | Z score rank | P value | Visualization Link
drugMatrix_Codelink | Fluvastatin | Rattus norvegicus | liver | In Vivo | 5d | 5 mg/kg | 0.141 | -7.373 | 1.000 | 8.362E-14 | DISPLAY NETWORK
drugMatrix_Codelink | Procarbazine | Rattus norvegicus | liver | In Vivo | 5d | 27 mg/kg | 0.019 | -7.254 | 1.000 | 2.017E-13 | DISPLAY NETWORK
drugMatrix_Affy | Oxymetholone | Rattus norvegicus | liver | In Vivo | 1d | 1170 mg/kg | 0.135 | -6.147 | 1.000 | 3.937E-10 | DISPLAY NETWORK
drugMatrix_Codelink | Indomethacin | Rattus norvegicus | liver | In Vivo | 6hr | 4.5 mg/kg | 0.019 | -5.559 | 1.000 | 1.360E-08 | DISPLAY NETWORK
TG-GATEs | Chlorpropamide | Rattus norvegicus | liver | In Vivo Single | 6hr | 300 mg/kg | 0.019 | -5.551 | 1.000 | 1.421E-08 | DISPLAY NETWORK
drugMatrix_Codelink | Emetine | Rattus norvegicus | kidney | In Vivo | 5d | 1 mg/kg | 0.019 | -5.543 | 0.999 | 1.486E-08 | DISPLAY NETWORK
TG-GATEs | Puromycin aminonucleoside | Rattus norvegicus | kidney | In Vivo Repeat | 4d | 4 mg/kg | 0.019 | -5.380 | 0.999 | 3.731E-08 | DISPLAY NETWORK
drugMatrix_Codelink | 1-naphthyl isothiocyanate | Rattus norvegicus | liver | In Vivo | 1d | 15 mg/kg | 0.018 | -5.150 | 0.999 | 1.305E-07 | DISPLAY NETWORK
drugMatrix_Affy | Gemfibrozil | Rattus norvegicus | liver | In Vivo | 7d | 700 mg/kg | 0.079 | -4.944 | 0.999 | 3.829E-07 | DISPLAY NETWORK
drugMatrix_Affy | Lovastatin | Rattus norvegicus | liver | In Vivo | 1d | 450 mg/kg | 0.070 | -4.907 | 0.999 | 4.623E-07 | DISPLAY NETWORK
Showing 1 to 10 of 11,700 entries
Previous 1 2 3 4 5 ... 1170 Next


Rad23b
Fabp5
Mogat1
Fermt3
Elovl6
Tkt
Dntt
Mgst3
Pitp
Hsd17b7
Acot2
Ermp1
Hmgcr
Cidec
Sqle
Orm1
Pkir
Dhcr24
Rab30
Thrsp
Echdc1
Stard4
Rtn4rl1
Idi1
Pex11a
Dusp9
Mgll
Nsdhl
Acss2
Hes6
Dhcr7
Hmgcs1
Aldh3a2
Fdft1
Fads1
Pcsk9
Insig1
Vnn1
Fdps
Lss
Slc2a4
Acly
Tmem97
Frmd4b
Gpam
Nfe2
Pmvk
Scd1
Acaca
Fasn
Mvd
Pstpip2
Elovl5
Rdh11
Mid1ip1
Dbi
Fads2
Gngt1
Tm7sf2
Ebp
Paox
Pnpla5
Session: bR54ykq7VR
Download JSON File
Take Screenshot of Graph
Active layout
CoSe Layout
Legend
Color
Description
Input gene
Drug gene

Test with known drugs



Test potential drug repurposing

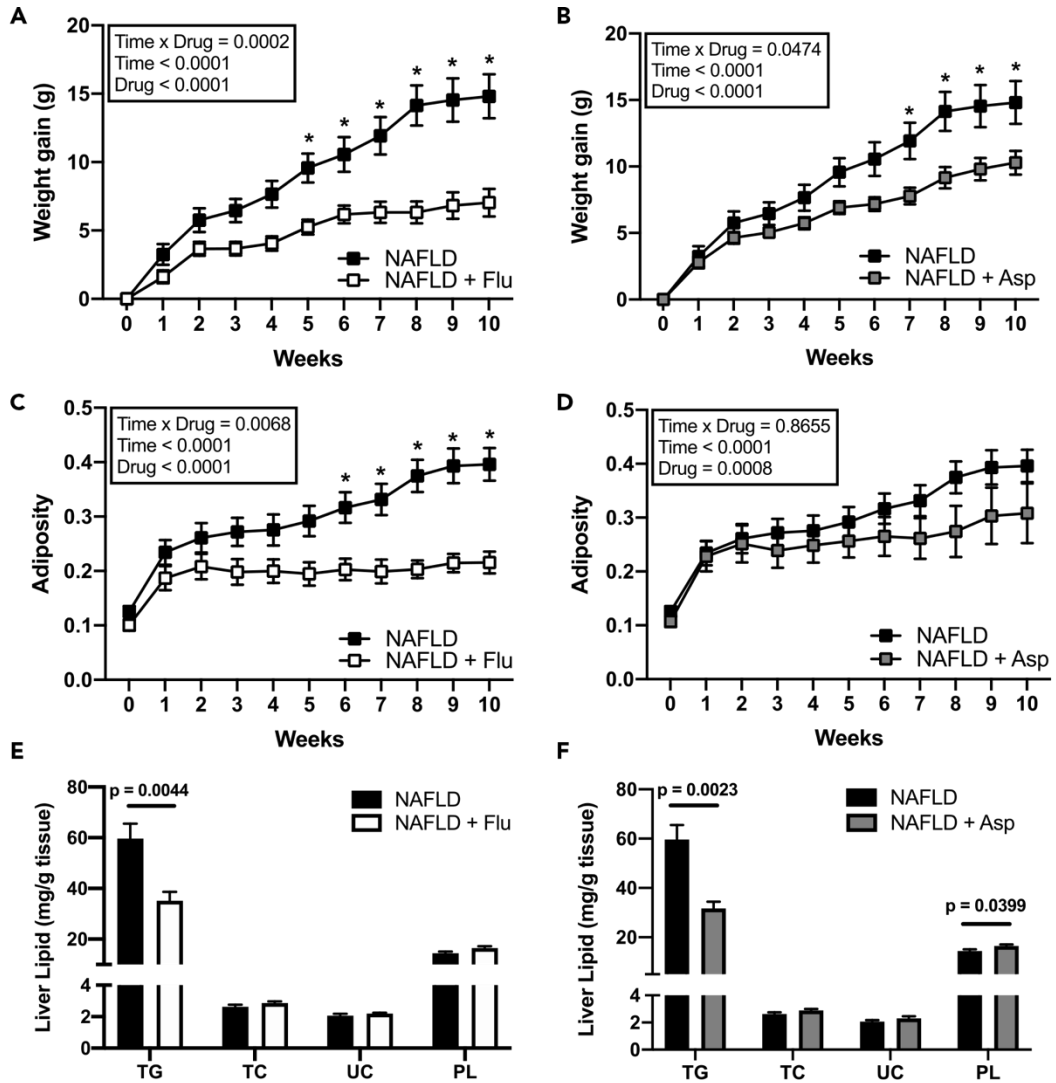


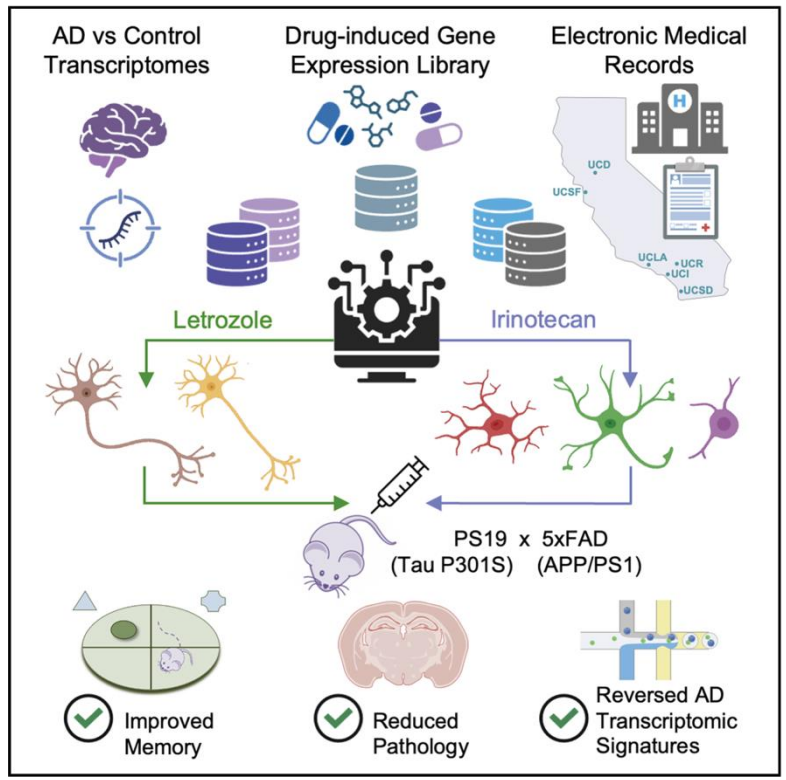
Figure 4. ***In vivo*** **validation of top predicted drugs fluvastatin and aspirin on preventing NAFLD phenotypes in a diet-induced NAFLD mouse model**

Cell

Article

Cell-type-directed network-correcting combination therapy for Alzheimer's disease

Graphical abstract



Authors

Yaqiao Li, Carlota Pereda Serras, Jessica Blumenfeld, ..., Michael J. Keiser, Yadong Huang, Marina Sirota

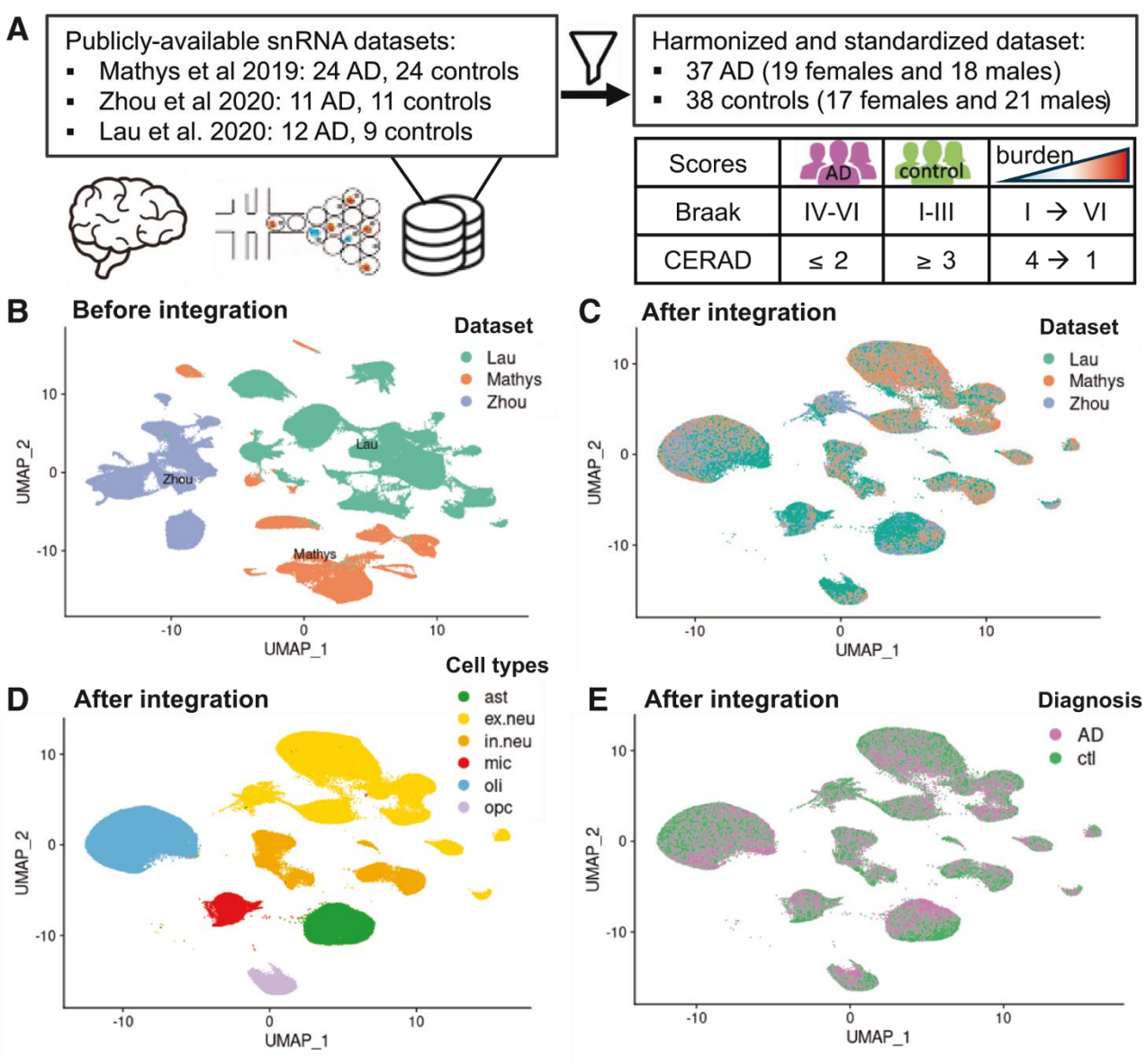
Correspondence

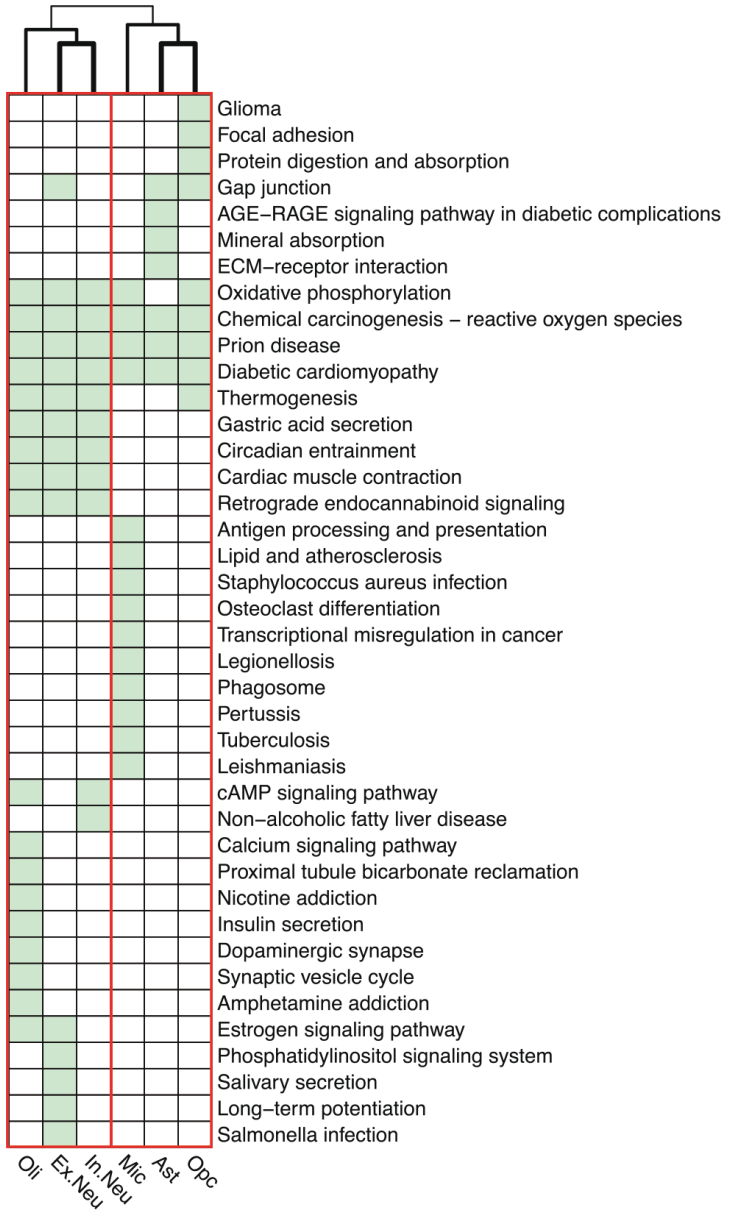
yaqiao.li@ucsf.edu (Y.L.), yadong.huang@gladstone.ucsf.edu (Y.H.), marina.sirota@ucsf.edu (M.S.)

In brief

A multi-cell-type drug discovery strategy targeting dysregulated gene networks in neurons and glia identified letrozole and irinotecan as a combination therapy that significantly improved memory and reduced pathology in preclinical Alzheimer's models.

Single nucleus Seq what can be done with public datasets

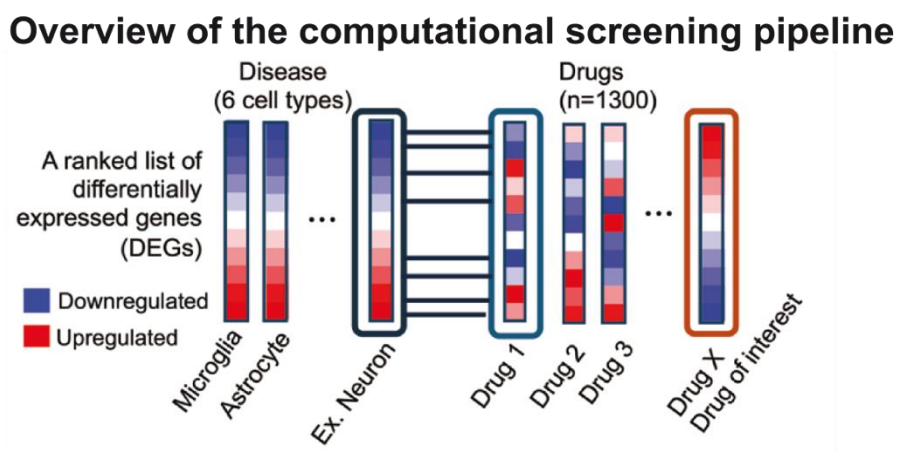


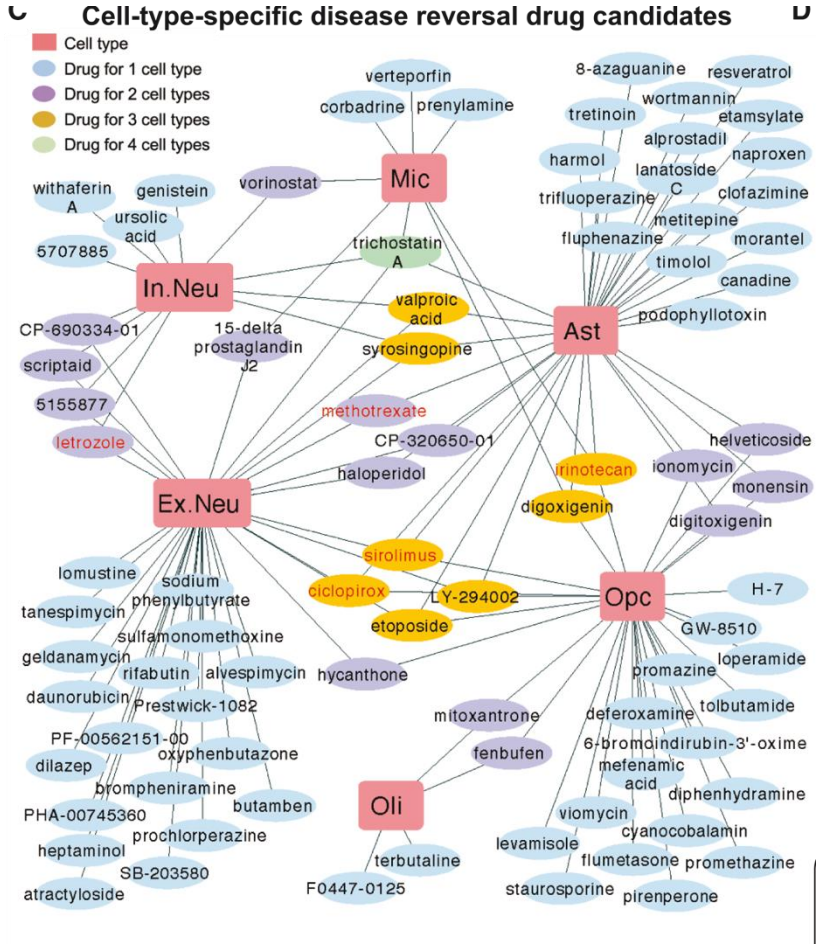


Used : Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031. –

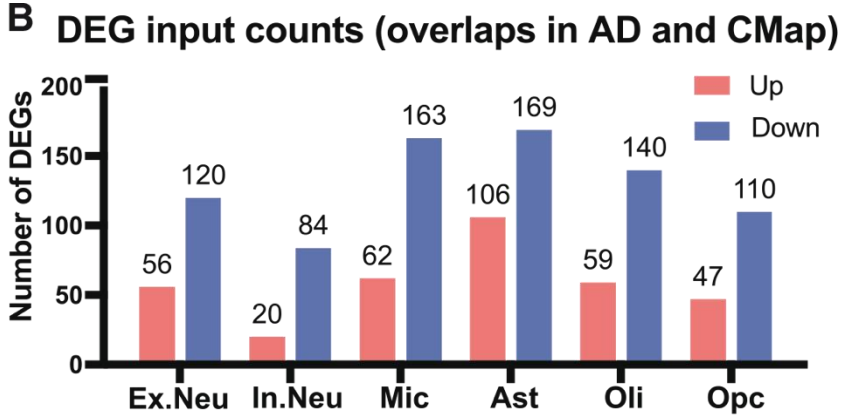
Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0. -

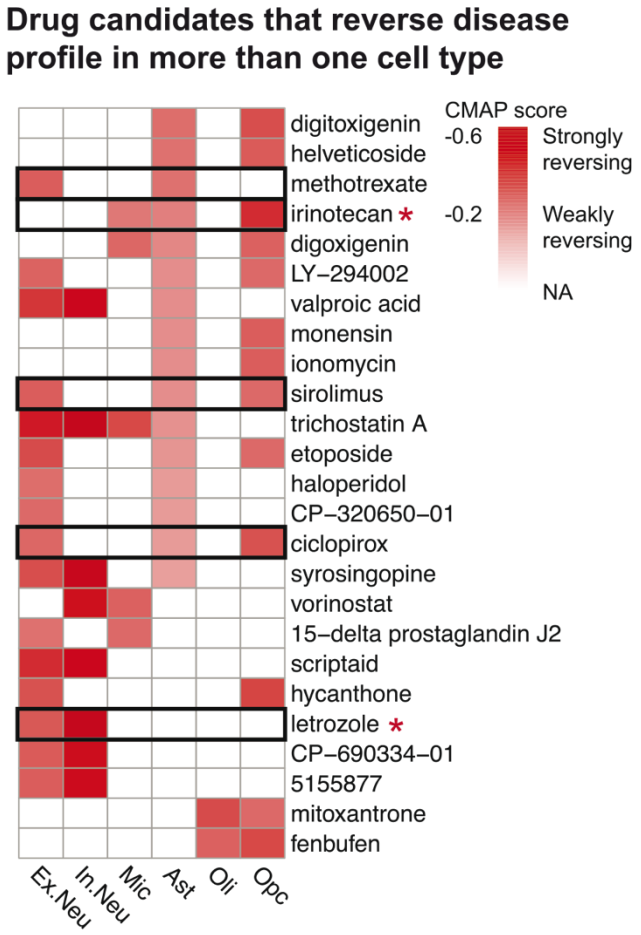
A





B





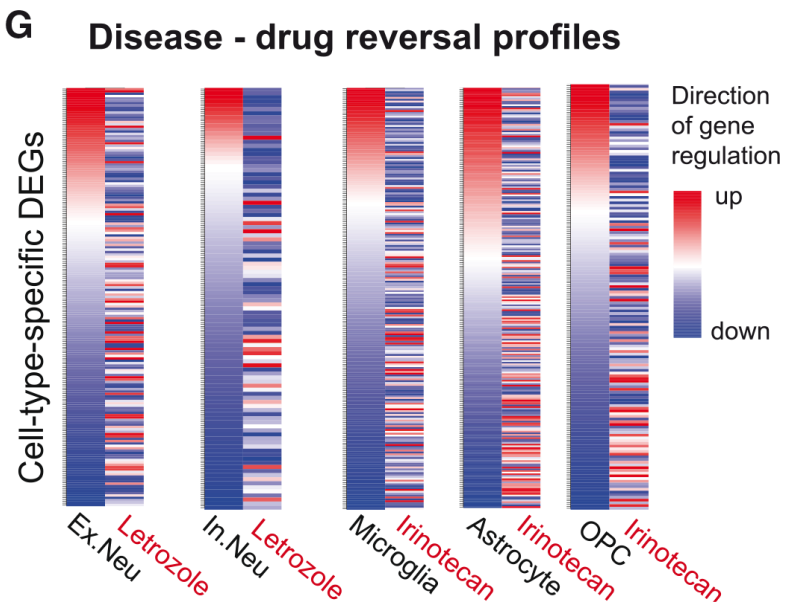
Using cell-type-specific AD profiles from our integrated analysis, we screened for network-correcting drug candidates by querying each profile against the **Connectivity Map (CMap),** a drug expression database generated with human cancer cell lines, via a computational pipeline that matches gene expression profiles of diseases and existing drugs

Post Hoc human efficacy study

Reversal of disease signature

E **AD outcome measures after drug exposure from UC-wide Electronic Medical Record**

	relative risk scores	P-value	total # of patients (exposed to drug)	% AD in drug group	original indications
Letrozole	0.466	1.27E-37	10841	1.01	breast cancer
Irinotecan	0.195	9.83E-23	2227	0.35	cancer
Methotrexate	0.650	4.10E-12	27547	1.22	cancer, autoimmune
Ciclopirox	0.927	0.0306	47642	2.44	Skin infection, dermatitis
Sirolimus	0.509	0.0209	4155	0.34	cancer, immuno-suppression



An so they go showing their two drugs ”work” in an animal model…

Future frontiers

- Improving upon these classic computational approaches using machine learning
- Improve understanding of larger scale genome organization (not addressed by Alpha-Genome)
- Generate artificial genome or optimized RNAs for therapy

nature machine intelligence

Review article

https://doi.org/10.1038/s42256-025-01007-9

Transformers and genome language models

Received: 29 January 2024

Accepted: 31 January 2025

Published online: 13 March 2025

Check for updates

Micaela E. Consens[1,2,3], Cameron Dufault[1], Michael Wainberg[2,4,5,6,7], Duncan Forster[2,8,9], Mehran Karimzadeh[2,10,11,12], Hani Goodarzi[10,11,12], Fabian J. Theis[13,14,15,16], Alan Moses[1,17] & Bo Wang[1,2,3,18]

Large language models based on the transformer deep learning architecture have revolutionized natural language processing. Motivated by the analogy between human language and the genome's biological code, researchers have begun to develop genome language models (gLMs) based on transformers and related architectures. This Review explores the use of transformers and language models in genomics. We survey open questions in genomics amenable to the use of gLMs, and motivate the use of gLMs and the transformer architecture for these problems. We discuss the potential of gLMs for modelling the genome using unsupervised pretraining tasks, specifically focusing on the power of zero- and few-shot learning. We explore the strengths and limitations of the transformer architecture, as well as the strengths and limitations of current gLMs more broadly. Additionally, we contemplate the future of genomic modelling beyond the transformer architecture, based on current trends in research. This Review serves as a guide for computational biologists and computer scientists interested in transformers and language models for genomic data.

Further Reading

Further Reading

Science

RESEARCH ARTICLES

Cite as: H. Zhang *et al.*, *Science* 10.1126/science.adr8470 (2025).

Deep generative models design mRNA sequences with enhanced translational capacity and stability

He Zhang[1], Hailong Liu[2], Yushan Xu[2]†, Haoran Huang[2], Yiming Liu[2], Jia Wang[2], Yan Qin[2], Haiyan Wang[2], Lili Ma[2], Zhiyuan Xun[2], Xuzhuang Hou[2], Timothy K. Lu[4,5,6]*, Jicong Cao[1]*

[1]Raina Biosciences, Cambridge, MA, USA. [2]Raina Biosciences, Beijing, China. [3]Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, MA, USA. [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Synthetic Biology Center, Massachusetts Institute of Technology, Cambridge, MA, USA.

†Present address: Broad Institute of MIT and Harvard, Cambridge, MA, USA.

***Corresponding author. Email: caojicong@rainabio.com (J.C.); tim@lugroup.org (T.K.L.)**

Despite the success of mRNA COVID-19 vaccines, extending this modality to more diseases necessitates substantial enhancements. We present GEMORNA, a generative RNA model that utilizes Transformer architectures tailored for mRNA coding sequences (CDSs) and untranslated regions (UTRs), to design novel mRNAs with enhanced expression and stability. GEMORNA-designed full-length mRNAs exhibited up to a 41-fold increase in firefly luciferase expression compared to an optimized benchmark in vitro. GEMORNA-generated therapeutic mRNAs achieved up to a 15-fold enhancement in human erythropoietin (EPO) expression and substantially elicited antibody titers of COVID vaccine in mice. Additionally, GEMORNA's versatility extends to circular RNA, substantially enhancing circular EPO expression and boosting anti-tumor cytotoxicity in CAR-T cells. These advancements highlight deep generative AI's vast potential for mRNA therapeutics.

Messenger RNA (mRNA) vaccines have proven effective in preventing COVID-19 (*1*, *2*). There are numerous efforts to extend mRNA therapeutics to other indications (*3*, *4*), but stronger and longer-lasting protein expression is necessary for these applications to be successful (*5*). Designing optimal mRNA sequences, including the coding sequences (CDSs) and untranslated regions (UTRs), is of paramount importance to realizing the broad potential of mRNA therapeutics via enhancing their translational capacity (*6*), but remains challenging due to the extensive potential mRNA sequence space (Fig. 1A). Within this extremely large space, mRNAs exhibit wide differences in expression-related properties. Moreover, the presence of multiple optimization metrics with complex interdependencies, along with hidden objectives linked to implicit cellular mechanisms, adds to the complexity of this problem.

In past decades, numerous efforts have sought to address the challenges of mRNA sequence optimization. Previous studies have demonstrated that optimizing nucleotide usage, such as GC content and U percentage, or codon usage, for example via the codon adaptation index (CAI), can enhance mRNA translation efficiency (*7*–*9*). However, these methods oversimplify the problem by prioritizing a single optimization objective, focusing primarily on local optimization of individual nucleotide and codons without considering sequence and structural context. To incorporate richer contextual information, deep learning-based models have been proposed for CDS optimization (*10*, *11*), both utilizing Long Short-Term Memory (LSTM) networks. However, these models face limitations due to LSTM's insufficient capacity in processing long gene sequence and also inefficiencies arising from non-parallelized training, which constrain the training data size and hindering model generalizability (*12*). A recent study incorporated structural features into the optimization objective, enabled global optimization via dynamic programming (*13*), and demonstrated a 128-fold increase in antibody response to COVID-19 mRNA vaccines compared to a codon-optimized baseline. However, this algorithm has not been adapted to mRNA sequences with chemical modifications (*13*, *14*), resulting in mismatches and reduced effectiveness in designing sequences for therapeutic mRNAs (*6*, *15*).

It has been challenging to design 5′ UTRs de novo since the mechanisms governing 5′ UTR regulation of mRNA translation are not fully understood. Recently, Zeng *et al.* developed a 5′ UTR design approach based on minimizing 5′ UTR secondary structure that demonstrated enhanced translational efficiency in cell-based assays (*16*). However, other factors that may influence initiation efficiency, such as mean ribosome load (MRL) and UTR-CDS interaction, were not considered (*17*). Several groups have described machine learning-based methods for 5′ UTR design (*18*, *19*), which initially trained predictive models on labeled data and then incorporated them into genetic algorithms for 5′ UTR sequence evolution. These methods depend heavily on the reliability of the predictive models, and the evolutionary algorithms can be hindered by local optima.

We developed generative models to design mRNA sequences with improved translational capacity. Drawing

Further Reading

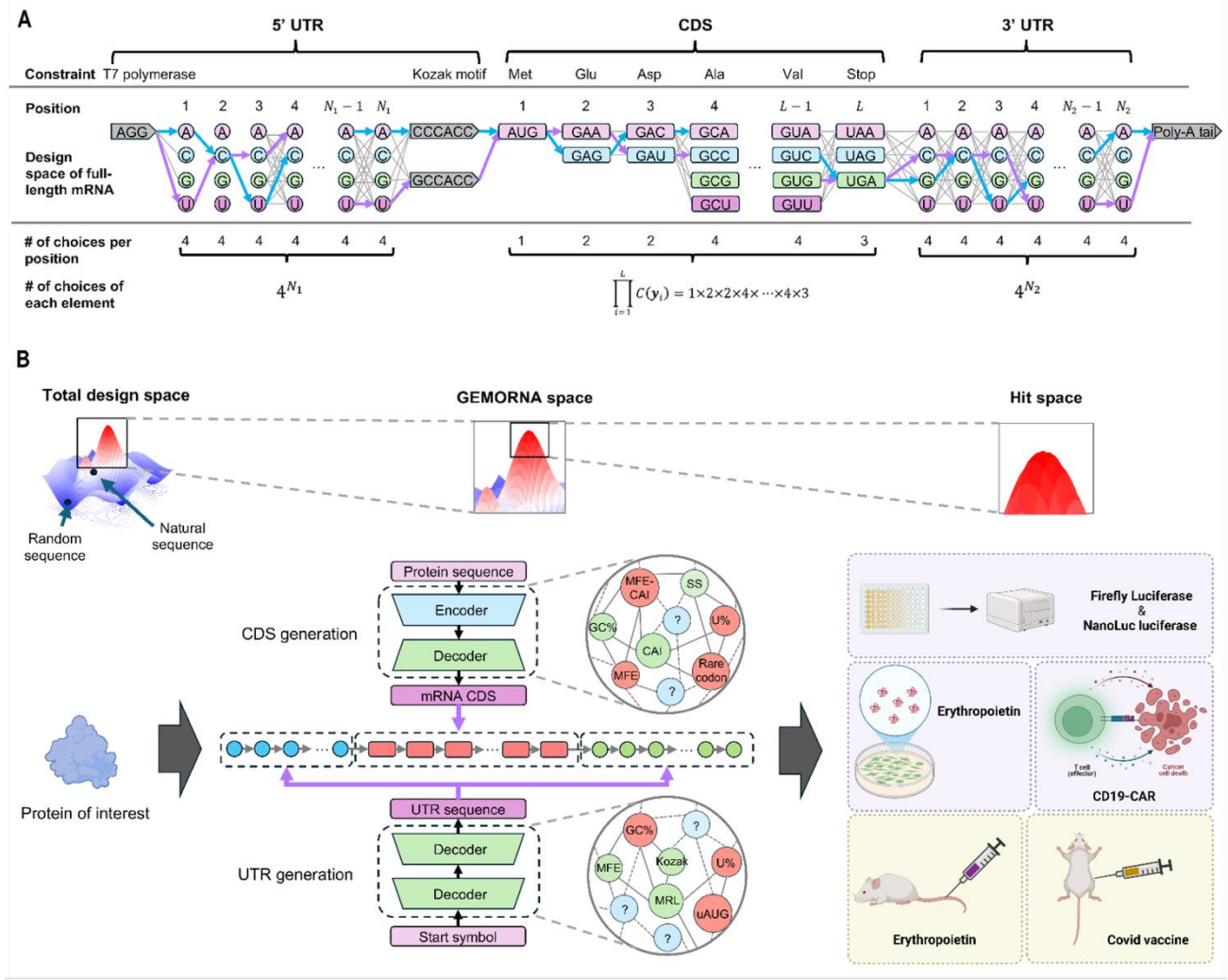


Fig. 1. Overview of designing full-length therapeutic mRNAs with GEMORNA. (**A**) Schematic representation of all possible mRNAs encoding firefly luciferase, including the 5′ UTR, CDS, and 3′ UTR sequences, illustrating the vast potential mRNA design space. Arrows between adjacent nucleotides (or codons) indicate possible choices at each position, with each path in the graph represents a unique full-length mRNA sequence with distinct UTR and CDS configurations. The highlighted blue path represents the natural alpha-globin UTRs with a codon-optimized CDS, while the purple path corresponds to GEMORNA's design. Note that the UTR's length (N_1 and N_2) can vary, whereas the CDS length (L) is constrained by the required protein length. $C(y_i)$ denotes the number of possible codon choice for a given amino acid y_i at each position. (**B**) GEMORNA-driven full-length mRNA design pipeline. The vast total design space is narrowed down by GEMORNA into a high-potency generation space, further validated through in vitro and in vivo experiments to reach hit space. To achieve this, GEMORNA models are trained to capture key features influencing mRNA expression and stability, ensuring that the GEMORNA mRNAs fall within the high-potency generation space.