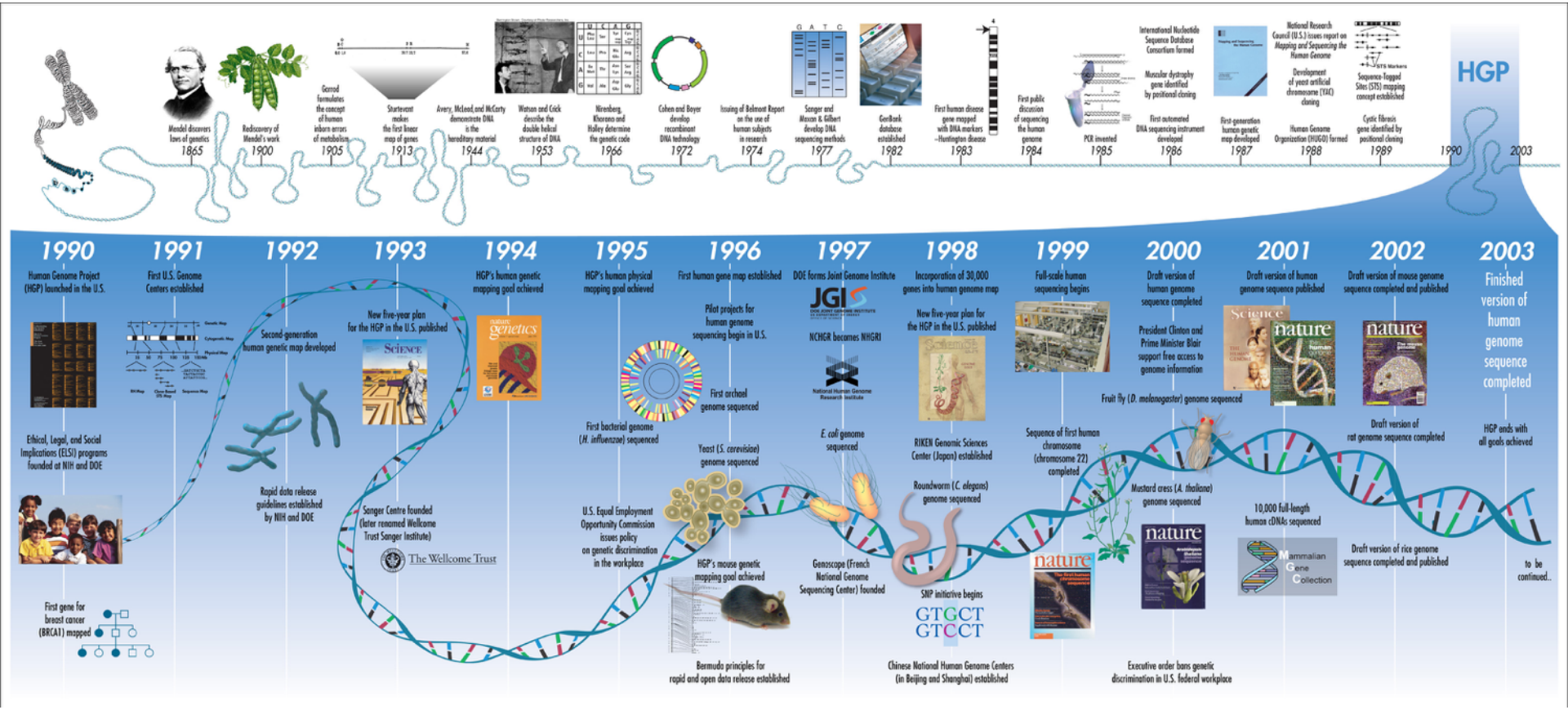
Bioinformatics

Lecturers:
Jonathan Warrell
Prashant Emani

Slides designed by: Prashant Emani, Jonathan Warrell, Declan Clarke

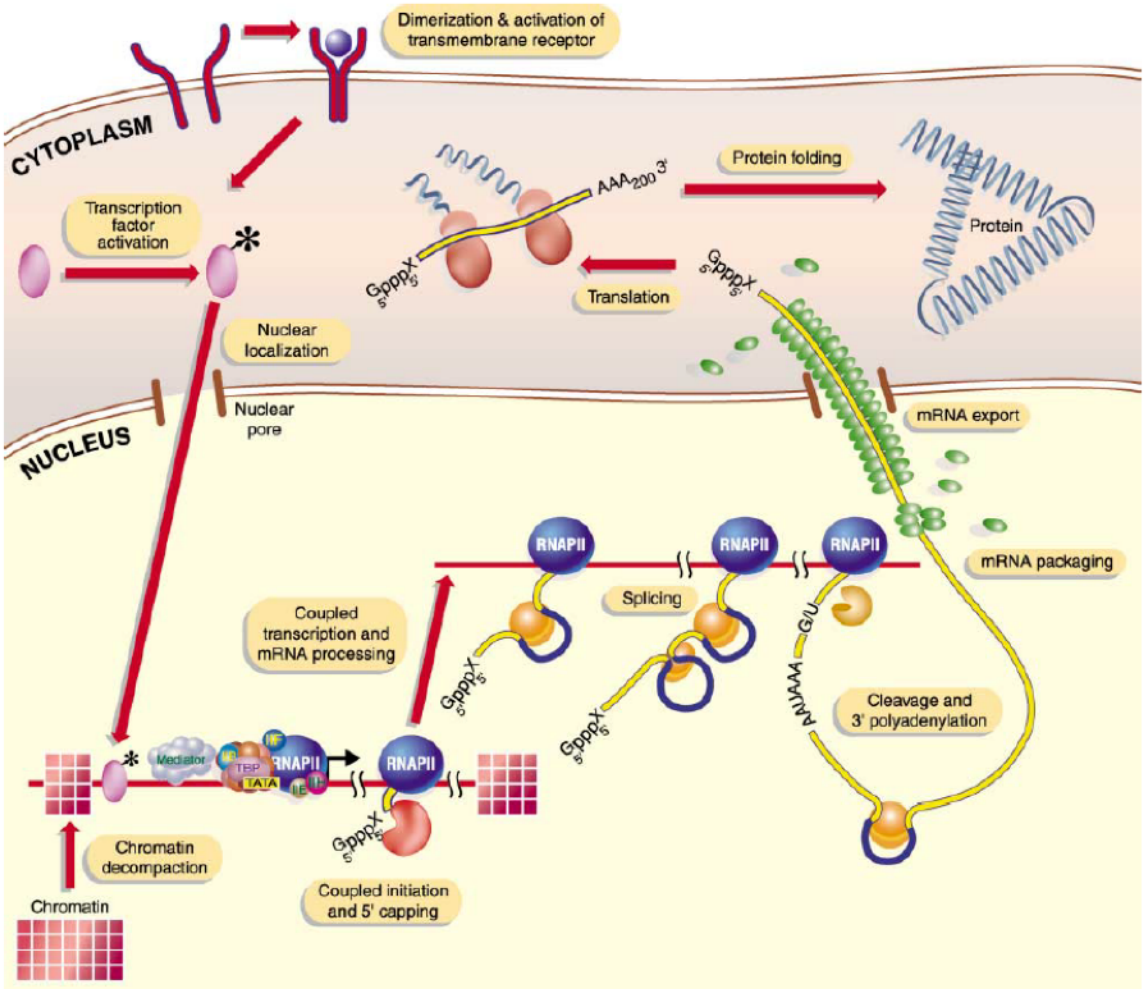
Date: June 11th, 2020

Human Genome Project



National Human Genome Research Institute (NHGRI) from Bethesda, MD, USA [CC BY 2.0 (https://creativecommons.org/licenses/by/2.0)]; Credit: Darryl Leja, NHGRI.

From mutations to function:
Multiple steps leading from genetic variation to cellular function



Orphanides and Reinberg **2002**, *Cell 108*, P. 439-451

Biological Assays:
Probing the various levels of cellular function

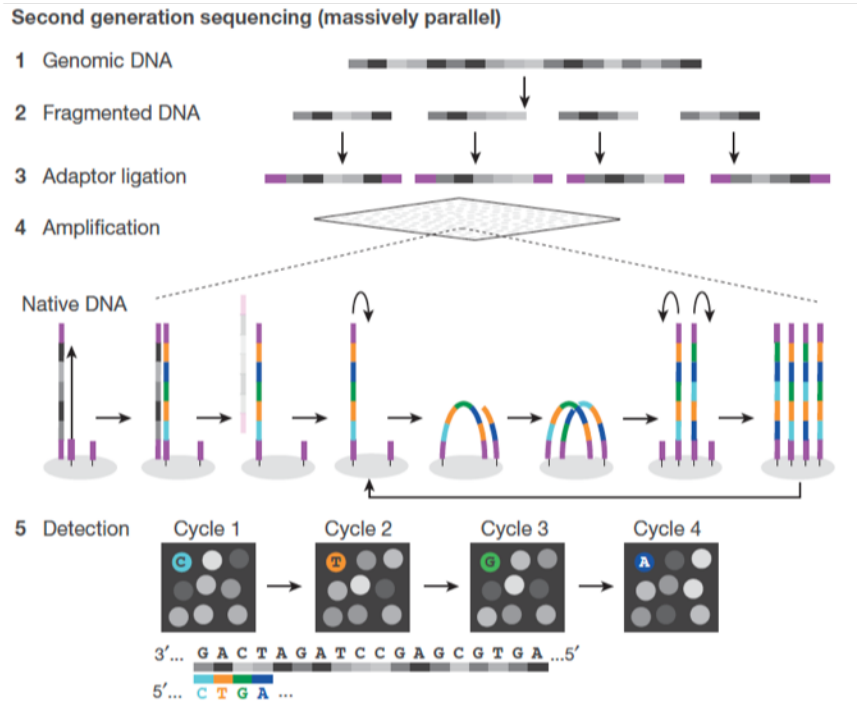
1. Genetic Assays:

a. <u>Whole Genome/Exome Sequencing:</u>

- Entire genome/exome sequenced
- Next-gen sequencing: DNA fragmented, adaptors added, amplified
- Detected using fluorescence-based methods
- Large sections of individual organism genomes can be sequenced

b. <u>DNA Microarray Sequencing:</u>

- Fragments hybridized to arrays of specific probes with mutations at sites polymorphic in a subset of human populations
- Results: mutation "hits"



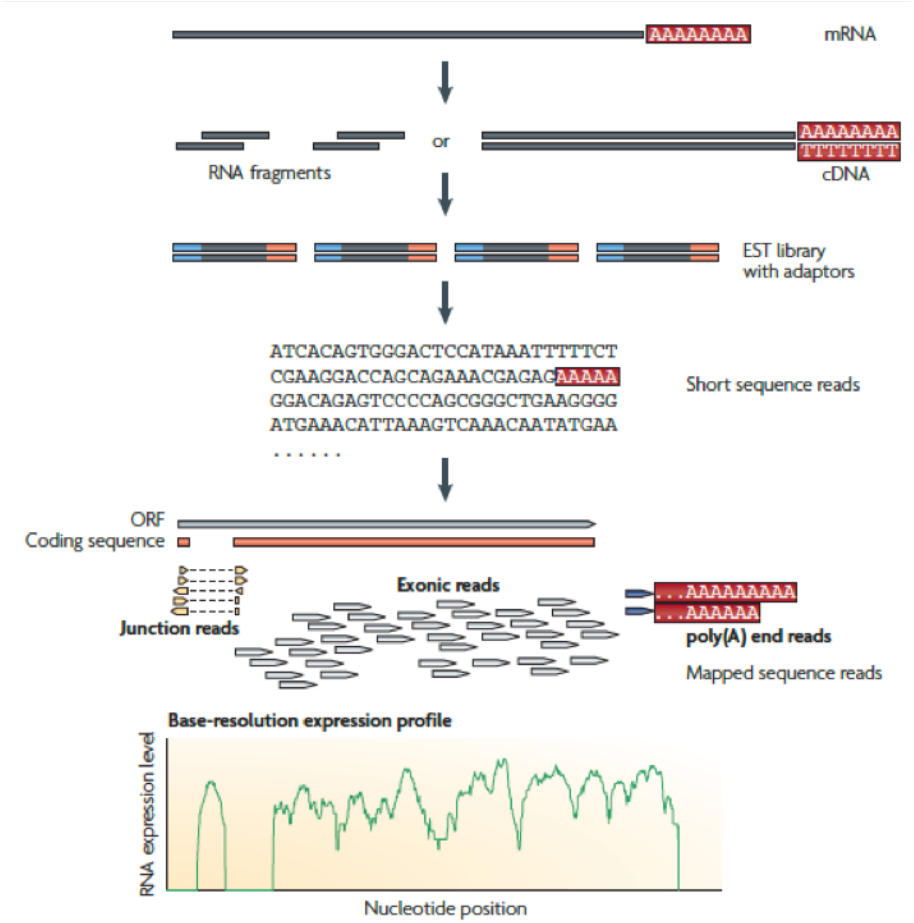
Shendure et al. **2017**, *Nat. Rev.* 550, Pgs. 345-353

Biological Assays (contd.)

2. Genomic Assays:

a. Transcriptomics/RNA-Seq (Tissues):

- Measures the amount of mRNA or total RNA
- For mRNA, purify with poly-A signatures
- Fragment, add adaptors, sequence, map to reference genome
- Stack reads and quantify genes and/or transcripts

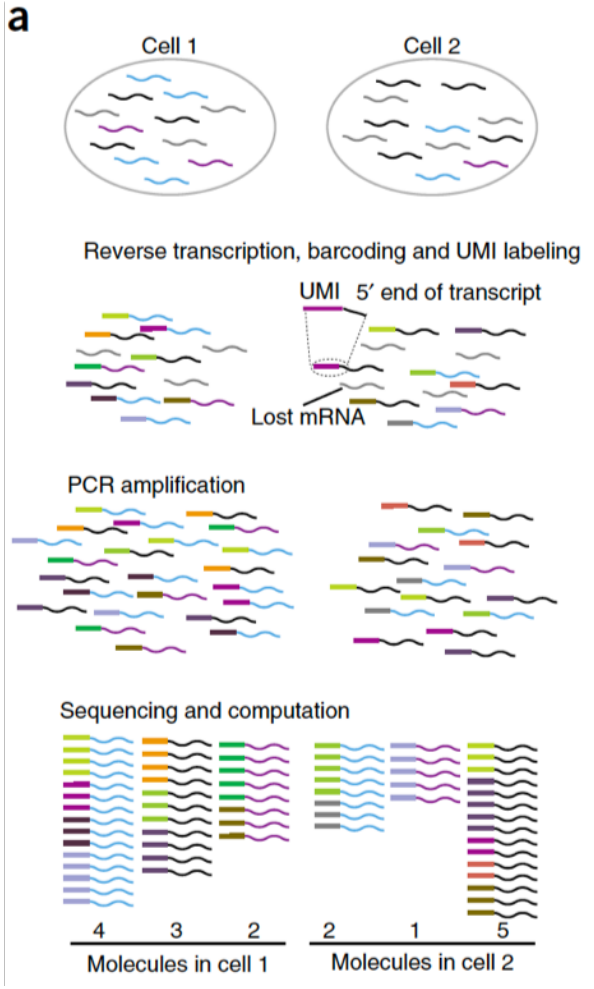


Nagalakshmi et al. **2009**, *Nat. Rev. Gen.* 10, Pgs. 57-63

Biological Assays (contd.)

b. RNA-Seq (Single cells):

- Measures the amount of mRNA in individual cells
- For mRNA, purify with poly-A signatures
- Fragment, add adaptors/barcodes, sequence, map to reference genome
- Whole genome amplification
- Stack reads and quantify genes

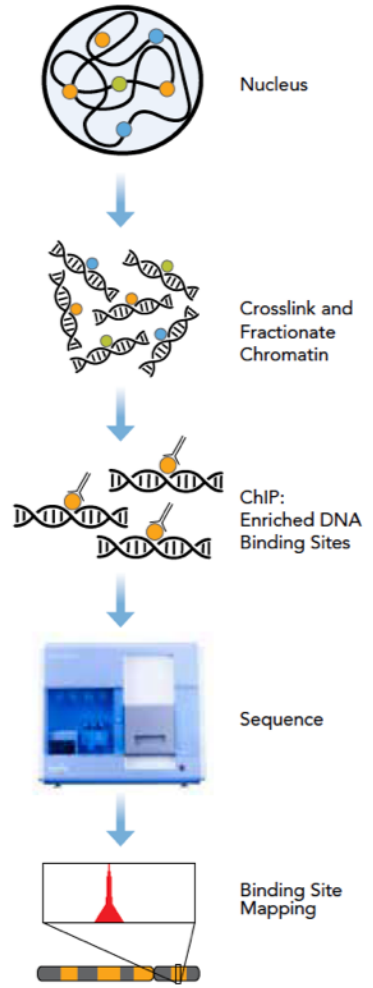


Islam et al. **2014**, *Nat. Methods* 11(2), Pgs. 163-166

Biological Assays (contd.)

c. Chromatin Immunoprecipitation with sequencing (ChIP-seq):

- Identifies regions of protein binding to DNA
- Helps identify transcription factor-binding sites, histone mark locations, etc.
- Use antibodies to pull down ("immunoprecipitated") proteins along with bound DNA
- DNA is processed in similar ways for sequencing and quantification



https://www.illumina.com/Documents/products/datasheets/datasheet_chip_sequence.pdf

Biological Assays (contd.)

d. Hi-C:

- Study of the 3D structure of DNA in the nucleus (*in situ*)
- Cross-link distal regions that are folded close to each other and sequence

e. DNase- and ATAC-seq:

- Finds regions of transcriptionally active, "open" DNA

f. Proteomics and Metabolomics

Other data-types

g. Trait / Phenotype data:

- Trait data can be continuous, categorical, univariate, multivariate …

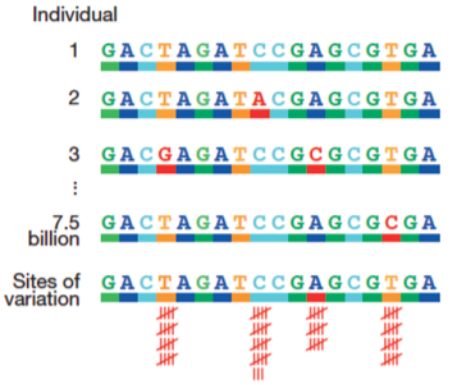
h. Imaging data:

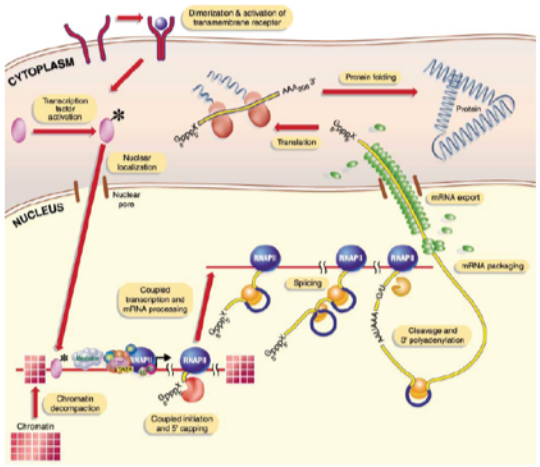
- Histological, MRI (sMRI / fMRI) data types

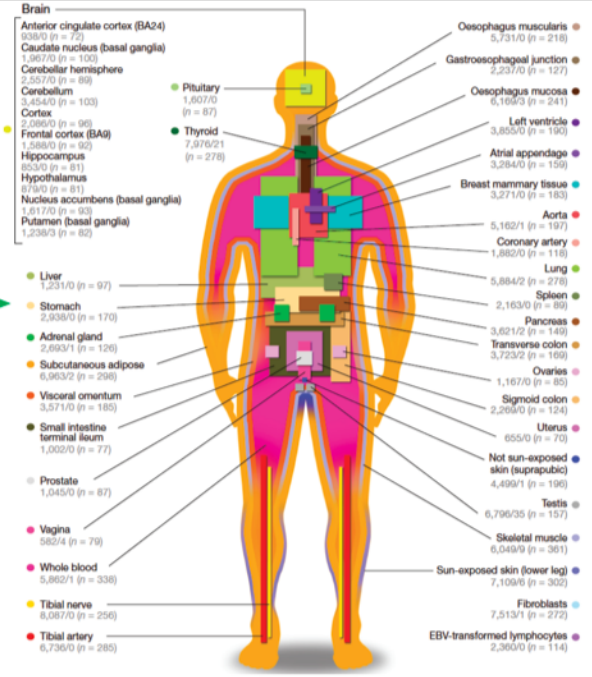
i. Medical records:

- Medical records can provide a rich source of trait data, and confounding factors (demographics, medical treatments…)

Genotype to Phenotype Hierarchy







Genetics Genomics Clinical

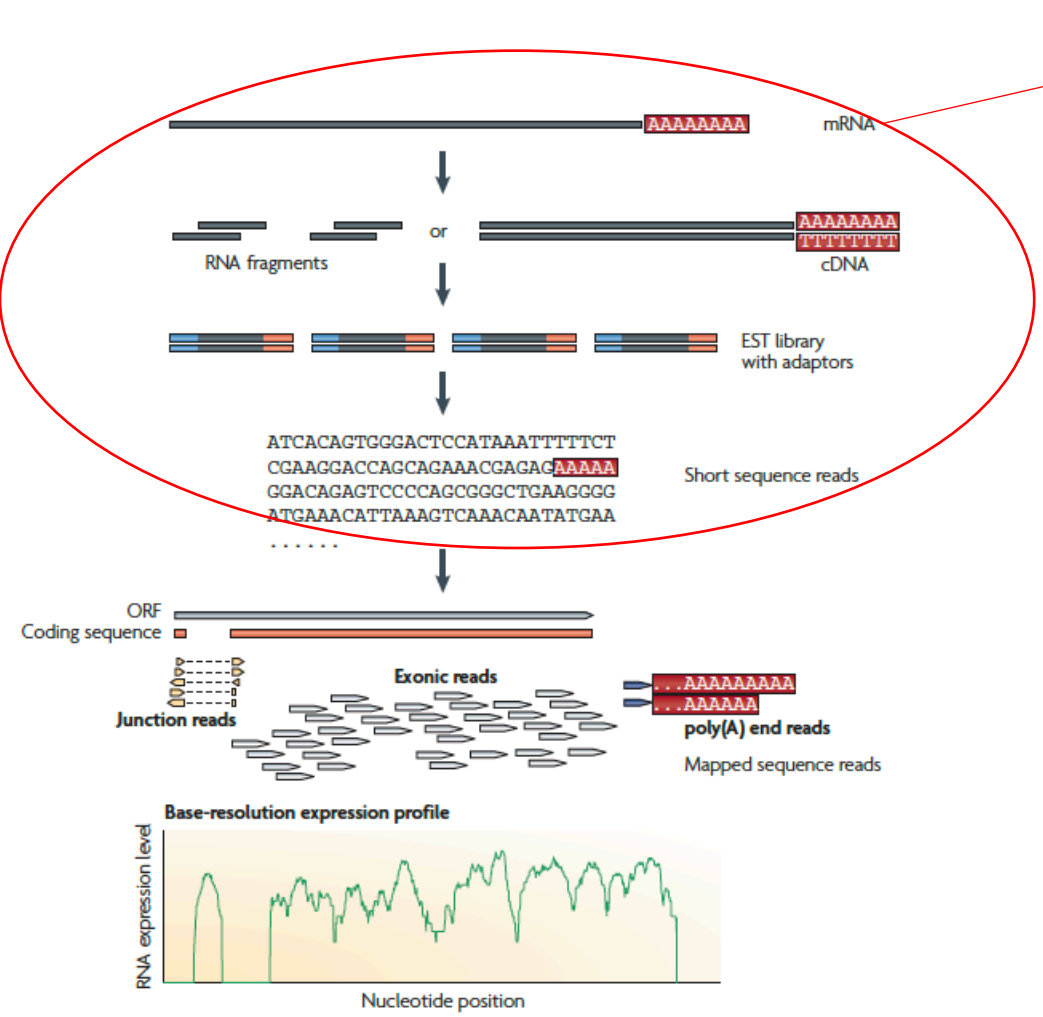
Image sources:

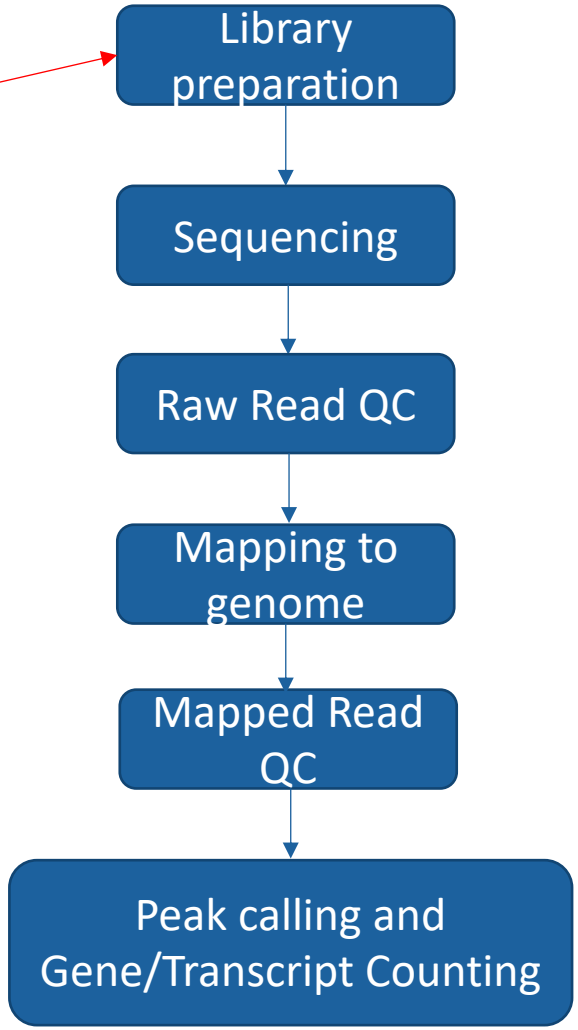
Shendure et al. **2017**, *Nat. Rev.* 550, P. 345-353

Orphanides and Reinberg **2002**, *Cell 108*, P. 439-451

GTEx Consortium **2017**, *Nature* 550, P. 204-213

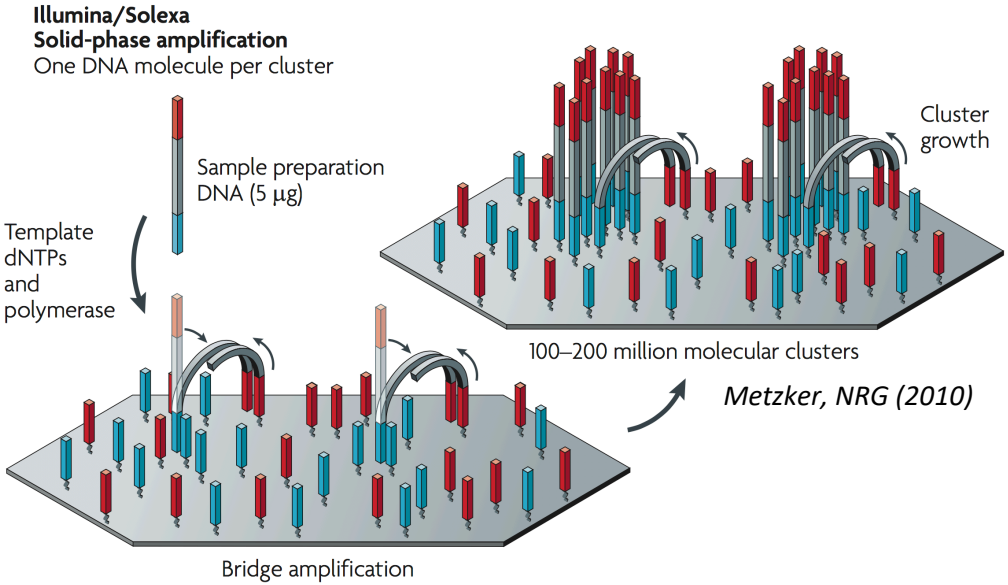
RNA-Seq for gene expression quantification



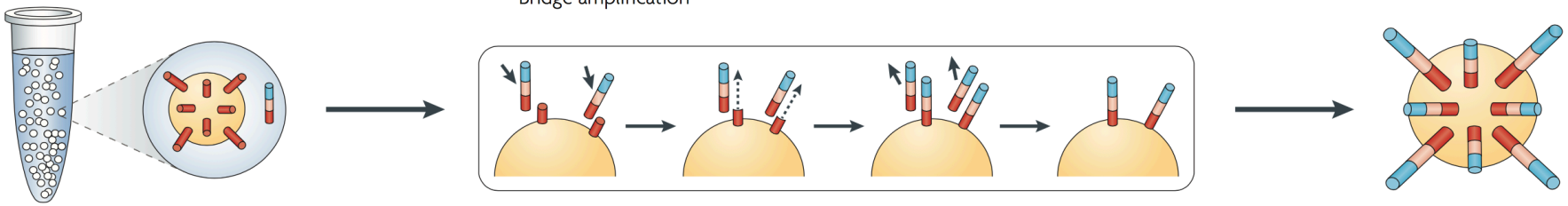


Next-generation sequencing

Examples of amplification strategies



Metzker, NRG (2010)



Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

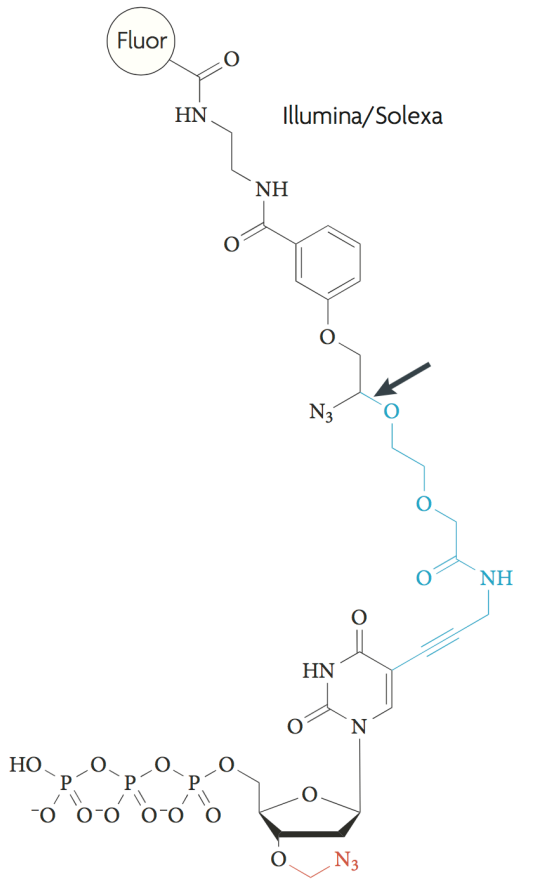
On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates

Final product
100–200 million beads with thousands of bound template

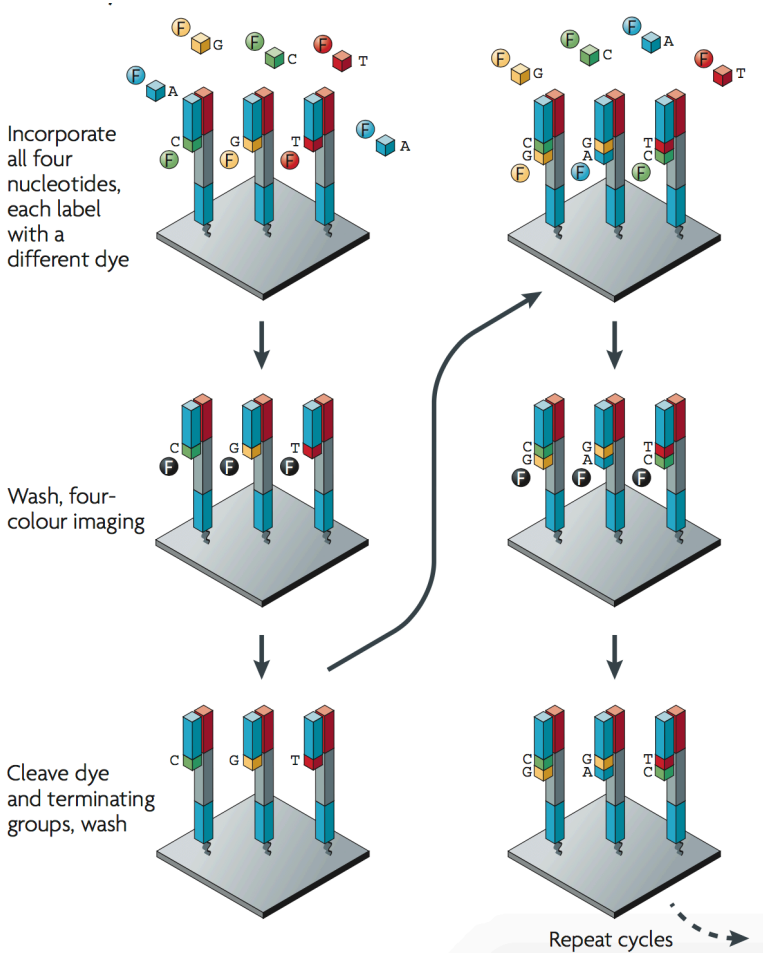
Goodwin, NRG (2016)

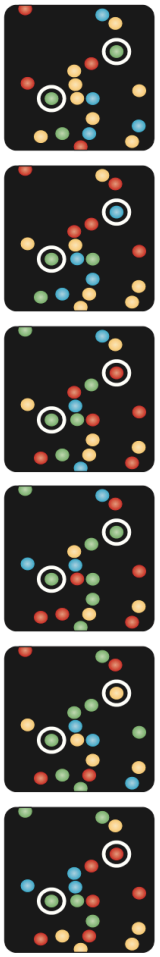
Next-generation sequencing

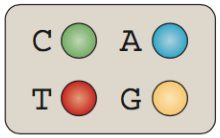
Reversible terminators and fluorescence detection



Metzker, NRG (2010)







Left:
CCCCCC

Right:
CATCGT

Raw Output from Sequencing: FASTQ Files

Base sequences and associated scores in ASCII format

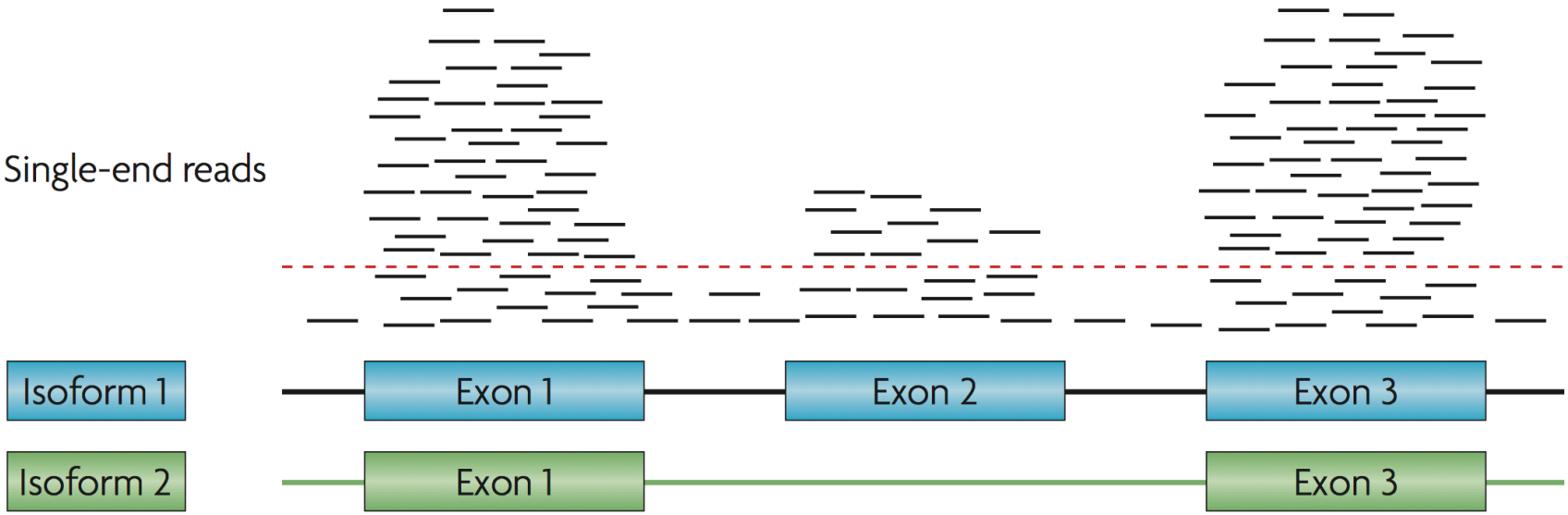
Robert Edgar, drive5.com

Common fastq software:

1. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
2. Cutadapt (https://doi.org/10.14806/ej.17.1.200)

Mapping/Alignment

Map reads (from FASTQ files) to human reference genome. Alignment can be performed using a number of available software suites. The output is ultimately provided as BAM file.



Alexander et al, NRG (2010)

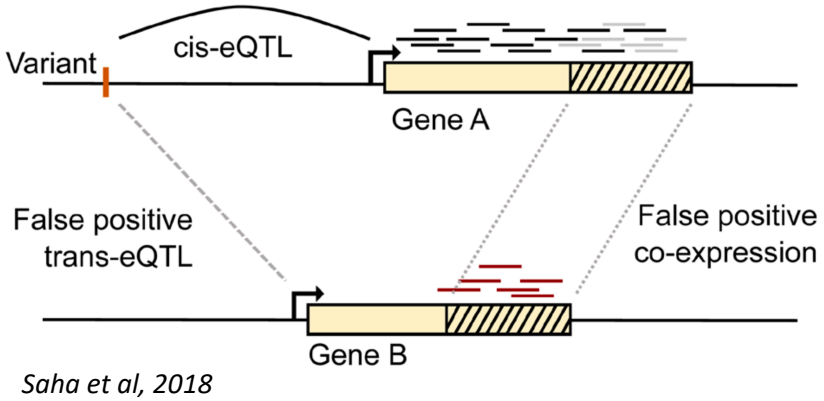
Commonly used aligners:

1. BWA-MEM (Li H. and Durbin R. (2009) *Bioinformatics*, 25:1754-60)
2. STAR aligner (Dobin et al (2013) *Bioinformatics,* 29(1):15–21)

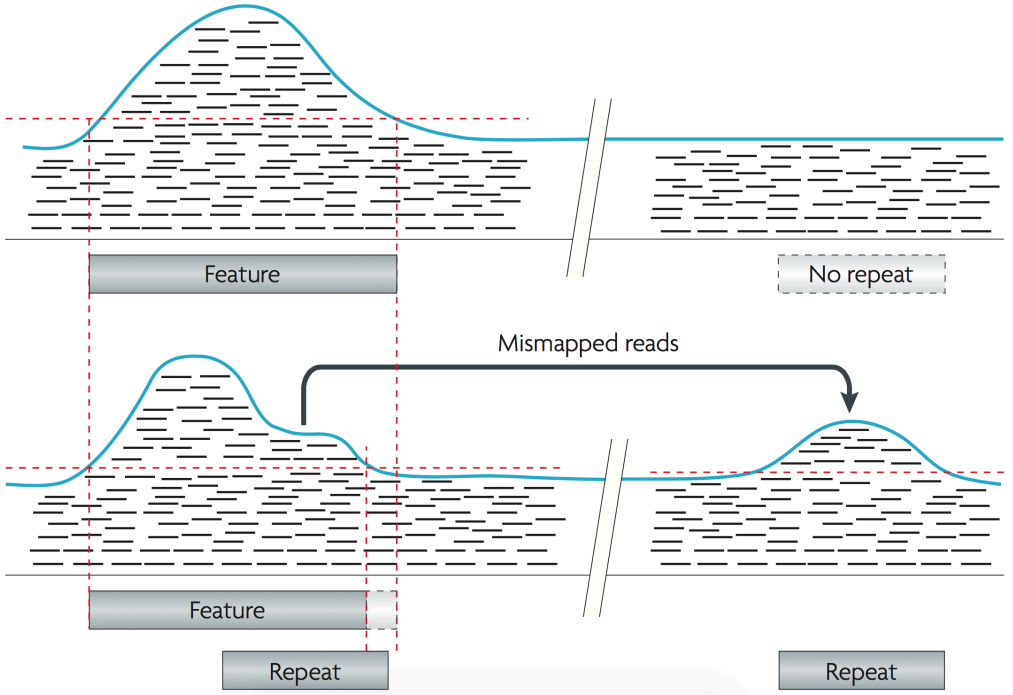
Mapped Read Quality Control

Experimental error may result in poor expression quantification. Many of the reads must be discarded entirely, and there may be a number of reasons for doing so:

+ failed to align well
+ failed to align at all (i.e., unmapped reads)
+ opposite problem: multi-mapping
+ duplicates



Saha et al, 2018



R Alexander, 2010

Peak Calling and Quantification

Once high-quality reads are aligned with confidence, how much is each gene expressed within a given tissue? As with quality control, a number of software tools may be used for peak calling and quantification (one example is MACS2).

A number of metrics are used in order to report expression levels:

- one popular choice is reads per kilobase of transcript, per Million mapped reads (RPKM -- needed in order to control for entire read count for a sample)
- Another approach is to use Transcripts Per Kilobase Million (TPM): first normalize for annotation length, and THEN normalize for sequencing depth

Output

A sample from a gene expression matrix

gene_id	Br1197	Br1413	Br1601	Br1414	...
ENSG00000000003	0.7618160961222128	0.36471935263577315	0.9208229763683788		...
ENSG00000000419	0.03893263241512441	-0.48841093328940016	0.7916386077433746		...
ENSG00000000457	0.11572878193547394	-0.25040077193169696	0.39399899592793247		...
ENSG00000000971	-0.7007855567241297	-0.6200928072637213	0.5372465676334013		...
ENSG00000001036	1.0561706239518178	1.19155226155541059	1.554426082403252		...
ENSG00000001084	-1.267521323193237	-0.46515292676660847	-0.8176429825378622		...
ENSG00000001167	-1.9716091285566475	-1.0884476945767128	-2.183683578885425		...
ENSG00000001460	1.7116753065097288	1.6207444157043982	1.4576844637815252		...
ENSG00000001461	-0.5751092812784382	-0.09876475664536664	-1.410523734759414		...
ENSG00000001497	0.7480146090187687	1.1415931121392715	1.5947594165725238		...
ENSG00000001561	-1.107411764001119	-1.00972011211089402	-0.9050773121085449		...
ENSG00000001629	-1.7116753065097285	-1.2880672130343624	-1.5205835031682469		...

Resources

Software tools:

Read mapping: STAR, BLAST, RSEM
Genetics / general: PLINK, HAIL, SAMtools
eQTL analysis: Fast-QTL, QTL-tools
GWAS/polygenic scores: GCTA, LDSR

Sources of data:

Genetics: 1000 Genomes, Hapmap
Genomics: ENCODE, ENCODEC, EnTEX, PsychENCODE,
GTEx, Epigenomics Roadmap
Large-scale collections: UKbiobank, TCGA (cancer), PGC (psychiatric)

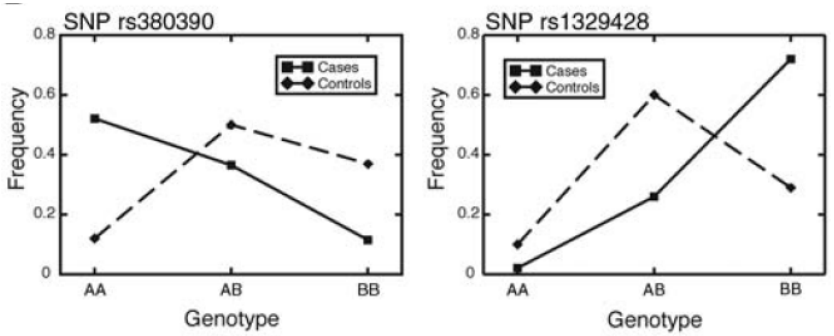
Machine Learning Applications:
Genome-Wide Association Studies (GWAS)

- Answers: Is a variant statistically more likely to occur in a case vs. a control?

- Use a linear model for a given phenotype:

$$Phenotype\ y_i = \sum_{j=1}^{M} \beta_j x_{ij} + \epsilon_i$$

  - For a categorical variable (disease or no disease), compare frequencies of occurrence.
  - For quantitative traits, use phenotype value.

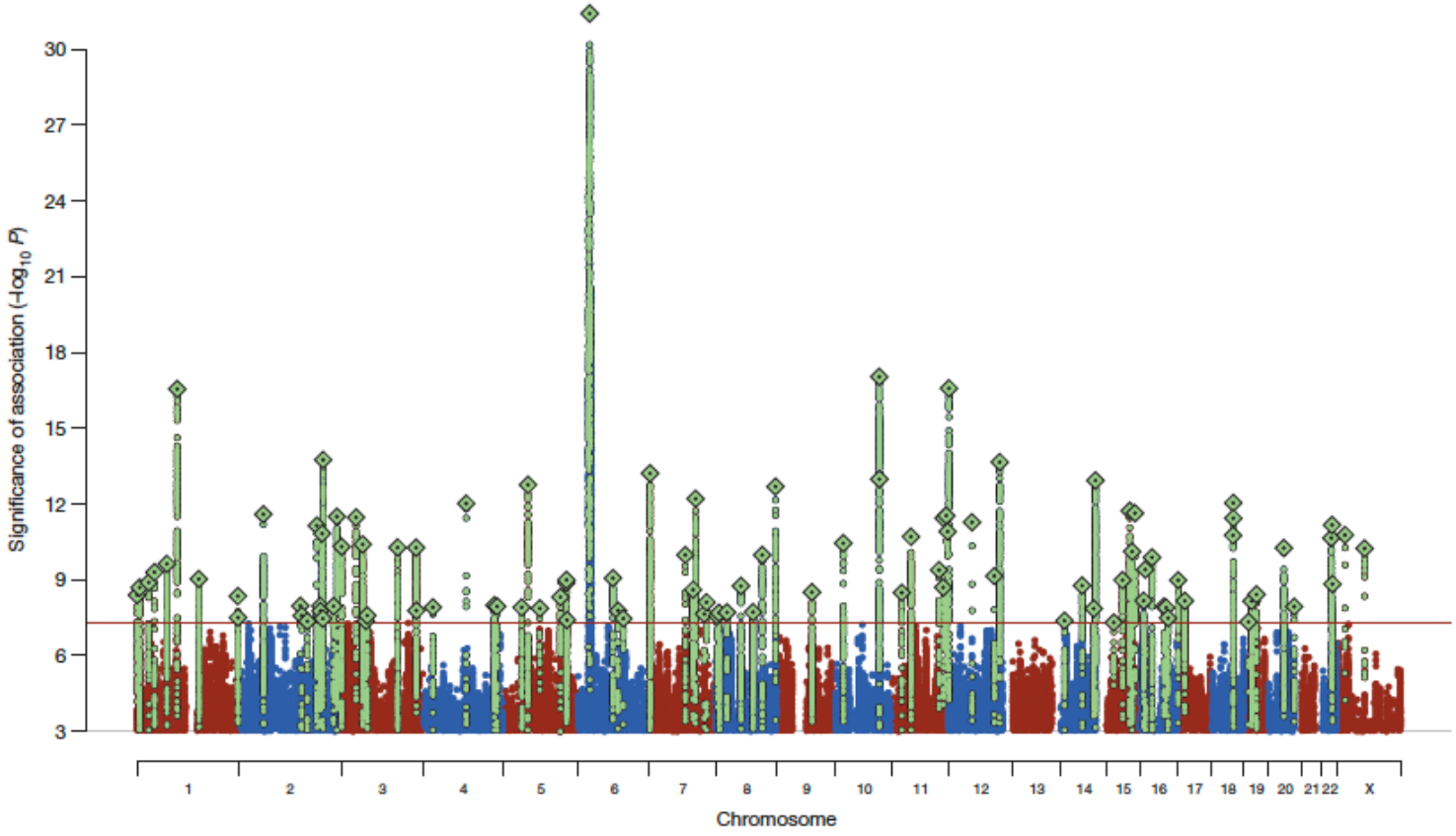


Klein et al **2002**, *Science*, 308, P. 385-389

- Carry out linear regression on each SNP and run a χ^2 test for significance

- Correct for multiple-testing burden!

Machine Learning Applications:
Genome-Wide Association Studies (GWAS) (contd.)



SCZ Working Group of the PGC **2014**, *Nature*, 511, P. 421-427

Machine Learning Applications:
Expression Quantitative Trait Loci (eQTL)

- Similar to a GWAS, except now uses gene expression as the target phenotype.

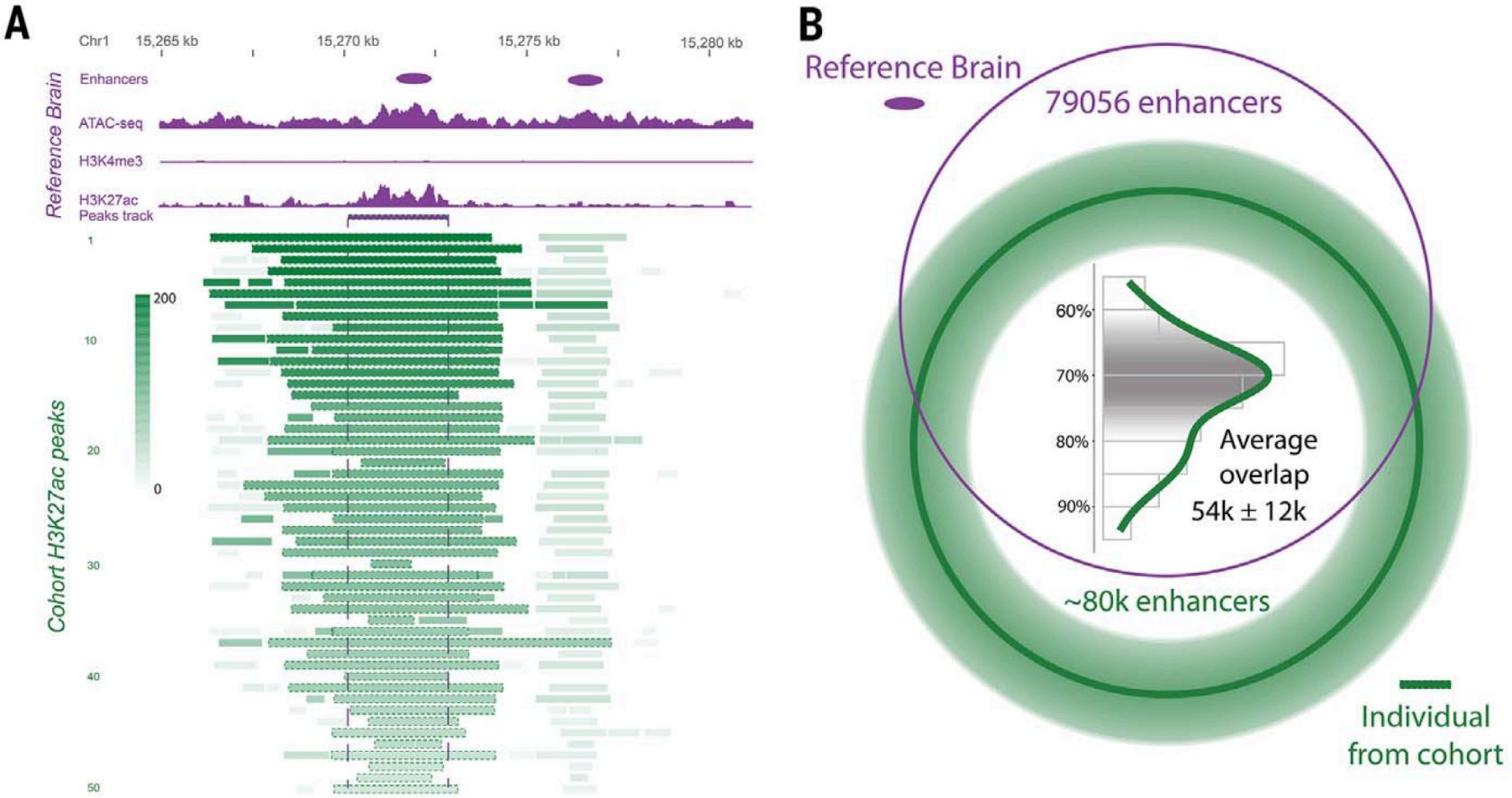
- Linear Regression to assess significance of the effect of the variant on gene expression:

$$Expression\ E_i = \sum_{j=1}^{M} \beta_j\, x_{ij} + \epsilon_i$$

- As in GWAS, need to correct for population structure:
  - If there is an accidental increase in frequency of a variant within high or low cohort due to relatedness, could mislead the analysis
  - Use PCs or PEER factors (confounding factors) to correct

- QTLs can be calculated for any intermediate phenotype: epigenetic modifications, changes in pathology image characteristics, changes in cell fractions, etc.

Machine Learning Applications:
Enhancer identification



Machine Learning Applications:
Biological Networks

- Structure in genomics data arises from various underlying networks; examples include:
  - Gene Regulatory Networks (GRNs):
    - Nodes: Genes
    - Edges: Transcription Factor -> Target gene linkages (directed)
  - Protein-Protein Interaction Networks (PPIs):
    - Nodes: Proteins
    - Edges: Physical interactions (undirected)
  - Gene co-expression networks:
    - Nodes: Genes
    - Edges: Correlation in expression (e.g. across subjects) (undirected, weighted)

Machine Learning Applications:
Weighted Gene Co-expression Network Analysis

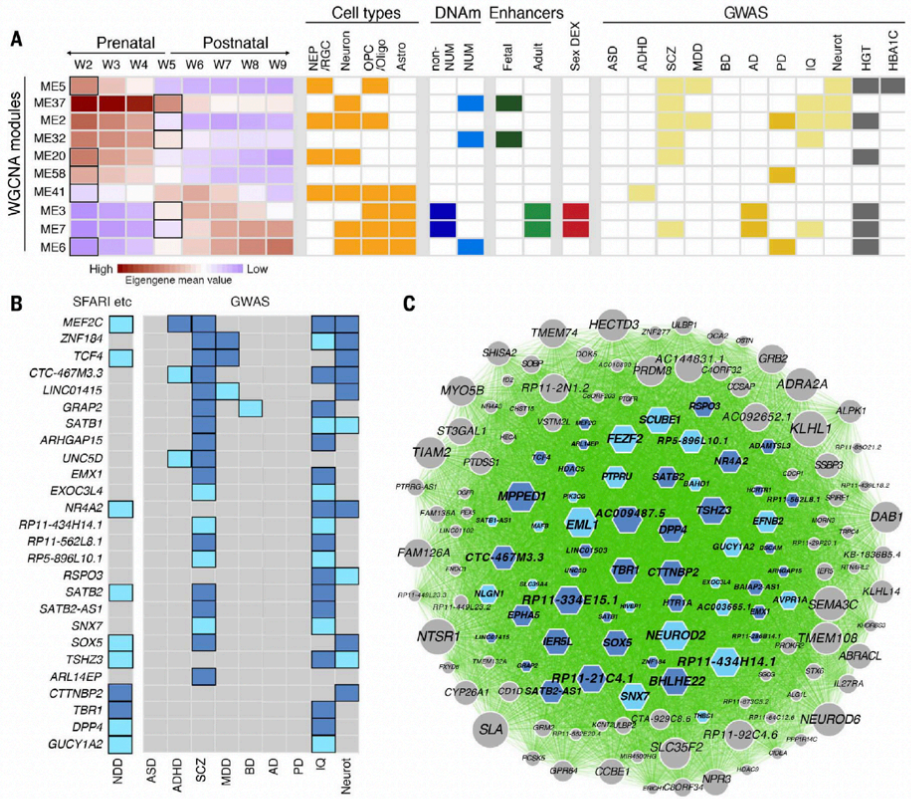
- Modules of genes may be defined via hierarchical clustering, as in Weighted Gene Co-expression Network Analysis (**WGCNA**, Zhang, 2005)
  - Start with a weighted network (correlation in expression values)
  - Compute the `Topological overlap' between all pairs of genes (proportional to # neighbors shared)
  - Build dendrogram using mean distance agglomerative clustering
  - Cut tree to produce final modules (Dynamic Tree Cut)

- Modules may represent functional gene groupings (e.g. cellular processes)
- WGCNA has been applied extensively in analysis of psychiatric data

Zhang, B. and Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).

Machine Learning Applications:
Weighted Gene Co-expression Network Analysis

- Example of WGCNA from PsychENCODE developmental analysis (Li *et al.* Science 2018)
  - Module ME37 is enriched for psychiatric and cognitive GWAS hits, and prenatal expression



Li, M., Santpere, G., Kawasawa, Y.I., Evgrafov, O.V., Gulden, F.O., Pochareddy, S., Sunkin, S.M., Li, Z., Shin, Y., Zhu, Y. and Sousa, A.M., 2018. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, *362*(6420), p.eaat7615.

Machine Learning Applications:
Dimensionality reduction: t-SNE

- Dimensionality reduction techniques can also be used to find structure in genomics data
- t-Distributed Stochastic Nearest Neighbor Embedding (t-SNE, van der Maaten and Hinton, 2008) finds an embedding which preserves pairwise probabilities of choosing neighboring points under Gaussian (original) and Student-t (target) distributions resp.

Conditional neighbor choice:
(in original space)

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}, \qquad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

Joint neighbor choice:
(in low-dimensional space)

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), pp.2579-2605.

Machine Learning Applications: Dimensionality reduction: t-SNE

- t-SNE optimizes KL-divergence between source/target distributions

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for t=1 **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end

- Use of Student-t distribution helps constrain outliers
- Method only explicitly preserves local structure; global structure may reflect initialization (e.g. using PCA)

Machine Learning Applications:
Dimensionality reduction: UMAP

- UMAP similarly embeds high-dimensional data in a low-dimensional space (McInnes *et al.* 2018)
- Minimizes KL-divergence between weighted graphs in high and low-dimensional spaces, in which a weight is placed between each data point and its k-nearest neighbors using:

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

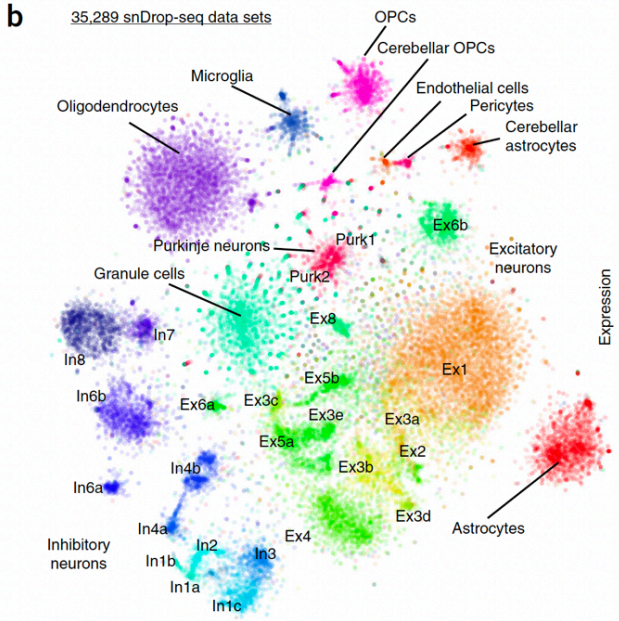
  - where σ_i=diameter of k-neighborhood, ρ_i=distance to nearest neighbor
- Can be motivated by topological data analysis (preserves topological structure, defined using 'fuzzy' sets)
- Global distances and coordinates not meaningful

McInnes, L.,*et al.*, 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

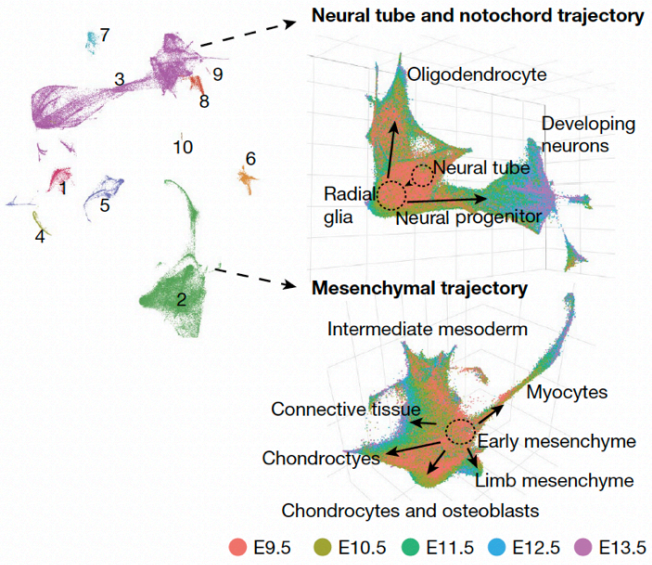
Machine Learning Applications: Dimensionality reduction: t-SNE/UMAP

- t-SNE and UMAP have been used extensively in recent genomics studies; for example to cluster cell types in brain and embryological single-cell data

tSNE, Lake *et al.*, 2016 (Human brain)



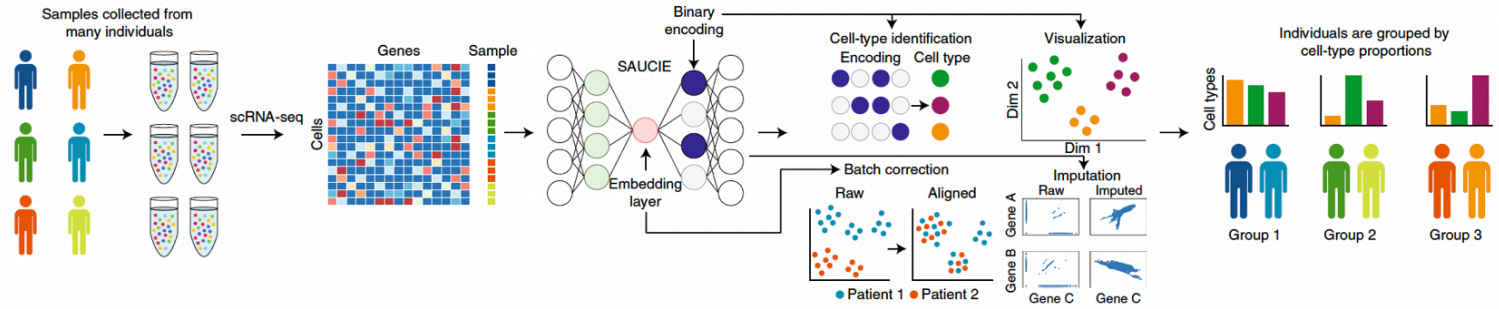
UMAP, Cao *et al.*, 2018 (mouse embryos)



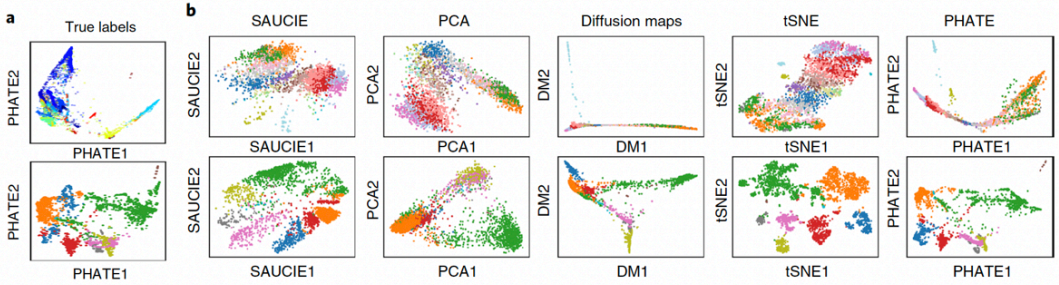
Lake, B.B., et el., 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotech.*, *36*(1), pp.70-80.
Cao, J., et al., 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), pp.496-502.

Machine Learning Applications: Dimensionality reduction: Deep learning

- Related deep-learning methods have been proposed which use autoencoders to perform dimensionality reduction (Amodio *et al.* '19)



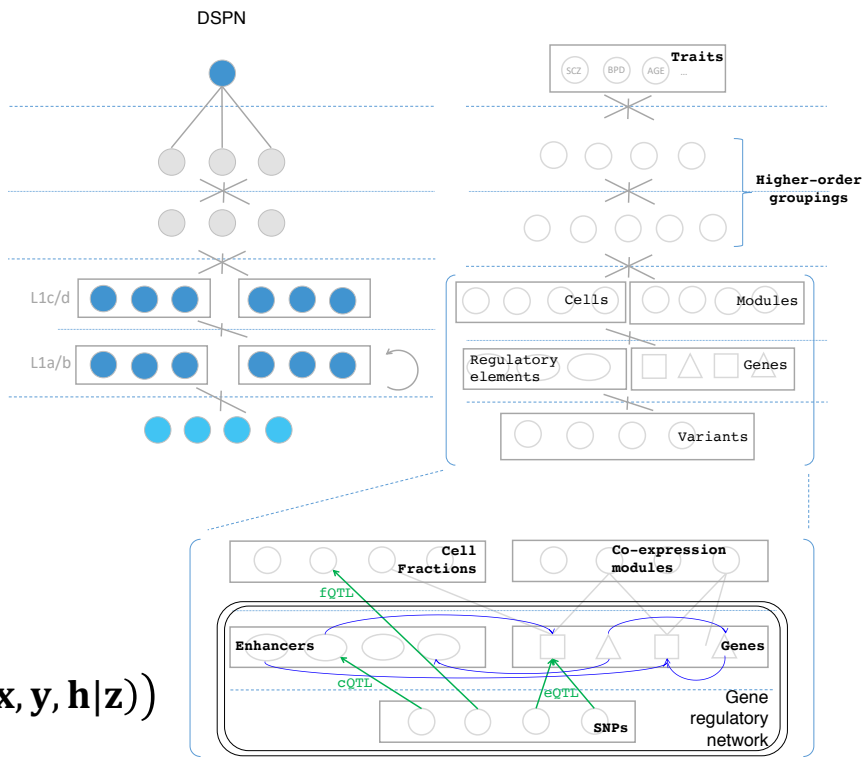
- Comparison of autoencoders (SAUCIE) with other methods, including diffusion maps on single cell data



Amodio, *et al.*, 2019. Exploring single-cell data with deep multitasking neural networks. *Nature methods*, pp.1-7.

Machine Learning Applications:
Deep Structured Phenotype Network (DSPN)

Gene regulatory network builds skeleton



Boltzmann machine

y: phenotypes

h: hidden units (e.g., circuits)

x: intermediate phenotypes (e.g., genes, enhancers)

z: genotypes (e.g., SNPs)

W: weights (e.g., regulatory network)

Energy model:

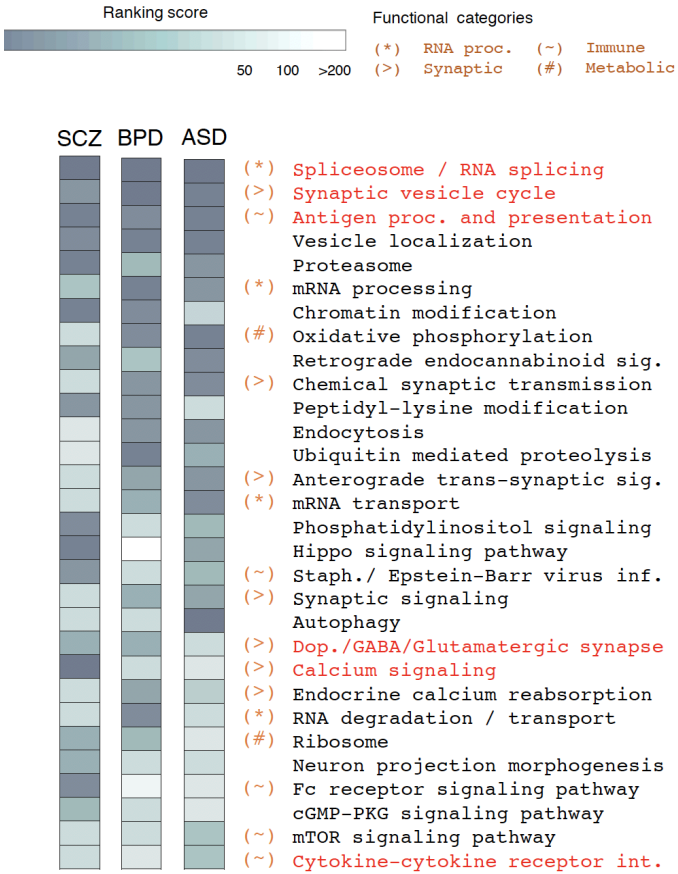
$$p(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) \propto \exp\left(-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z})\right)$$

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{W}_1\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_2\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_3\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_4\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_5\mathbf{y} - \boldsymbol{Bias}$$

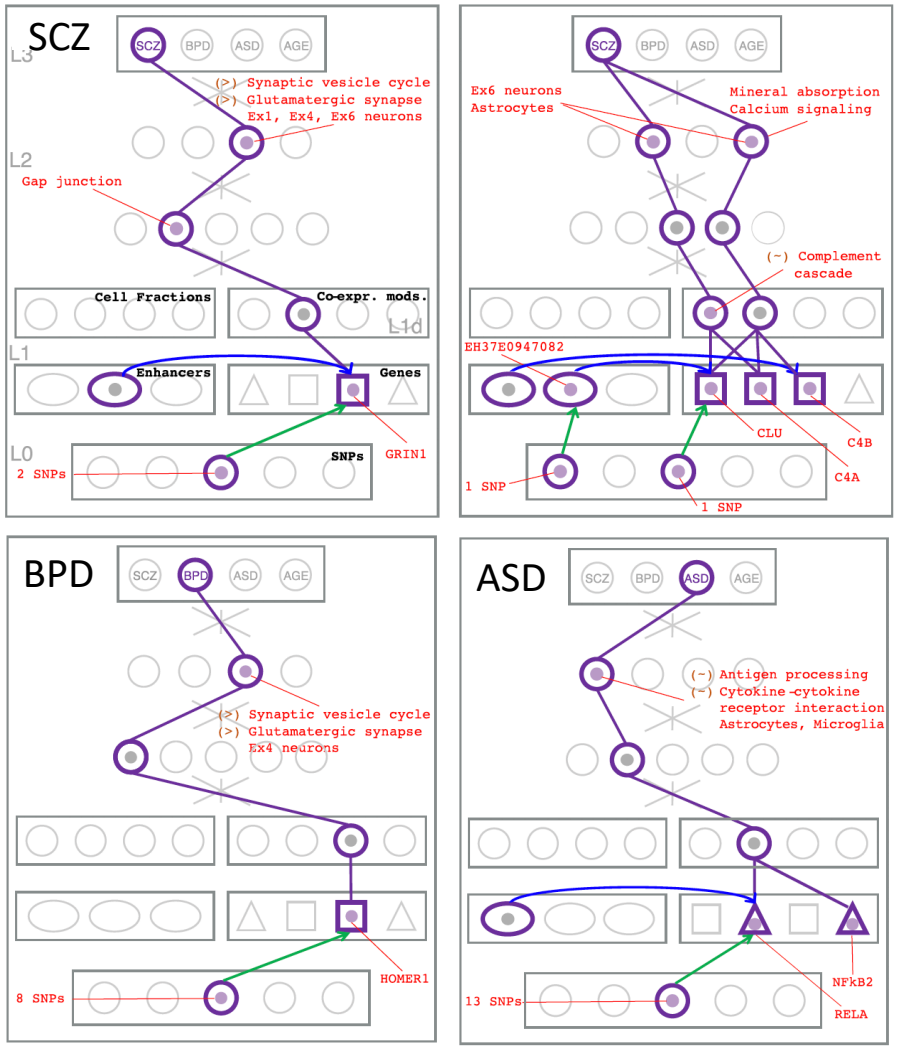
[Wang et al. ('18) Science]

Machine Learning Applications:
DSPN discovers enriched pathways and linkages to genetic variation

Cross-disorder MOD/HOG enrichment ranking



[Wang et al. ('18) Science]



References

- Bioinformatics references:

  - GTEx Consortium. "Genetic effects on gene expression across human tissues." *Nature* 550, no. 7675 (2017): 204.
  - Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C., Clarke, D., Gu, M., Emani, P., Yang, Y.T. and Xu, M., 2018. Comprehensive functional genomic resource and integrative model for the human brain. *Science*, *362*(6420).
  - Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer et al. "The UK Biobank resource with deep phenotyping and genomic data." *Nature* 562, no. 7726 (2018): 203.
  - Alon, Uri. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.

References (contd.)

- Nagalakshmi et al. **2009**, "RNA-Seq: a revolutionary tool for transcriptomics" *Nat. Rev. Gen.* 10, Pgs. 57-63
- Islam et al. **2014**, "Quantitative single-cell RNA-seq with unique molecular identifiers" *Nat. Methods* 11(2), Pgs. 163-166
- Shendure et al. **2017**, "DNA sequencing at 40: past, present and future" *Nat. Rev.* 550, Pgs. 345-353
- Alexander, Roger P., et al. "Annotating non-coding regions of the genome." Nature Reviews Genetics 11.8 (2010): 559.
- Goodwin, Sara, John D. McPherson, and W. Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics 17.6 (2016): 333.
- Metzker, Michael L. "Sequencing technologies—the next generation." Nature reviews genetics 11.1 (2010): 31.
- Robert Edgar, drive5.com (credit attribution information not provided: image taken from: https://www.drive5.com/usearch/manual/fastq_files.html)
- Saha, Ashis, and Alexis Battle. "False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors." F1000Research 7 (2018).