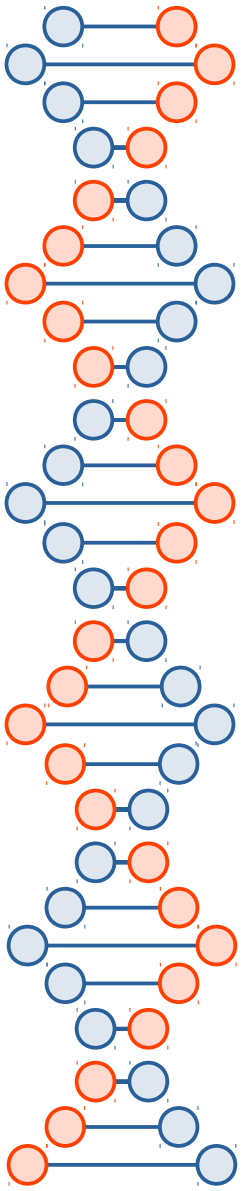


Omics Data Analysis

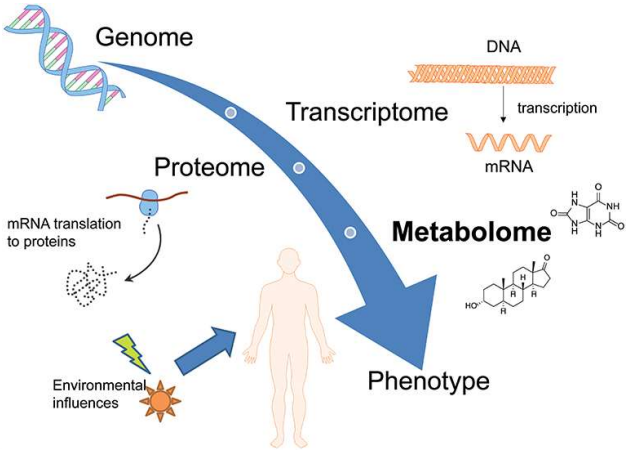
(Lunch & Learn)

Kith Pradhan

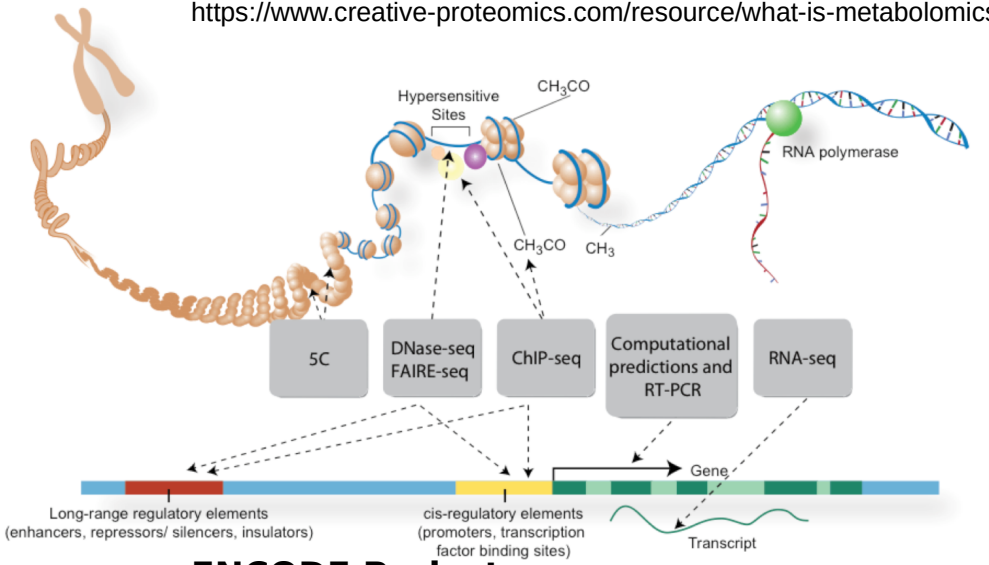
Nov 17, 2025

What is Omics

- The study of biological molecules at large scale
  - Genomics
  - Transcriptomics
  - Proteomics
  - Metabolomics
  - Epigenomics
  - Lipidomics
  - Microbiomics



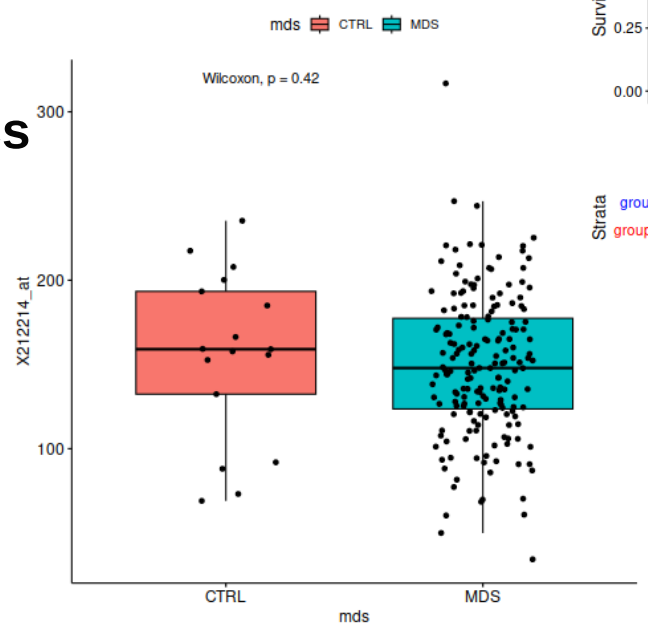
https://www.creative-proteomics.com/resource/what-is-metabolomics.html

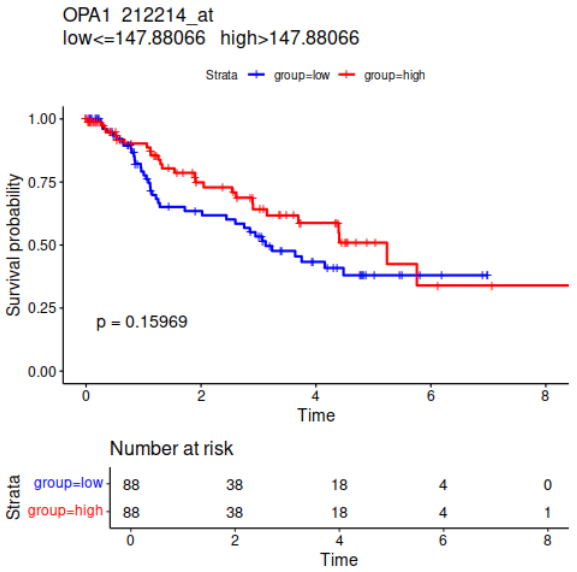


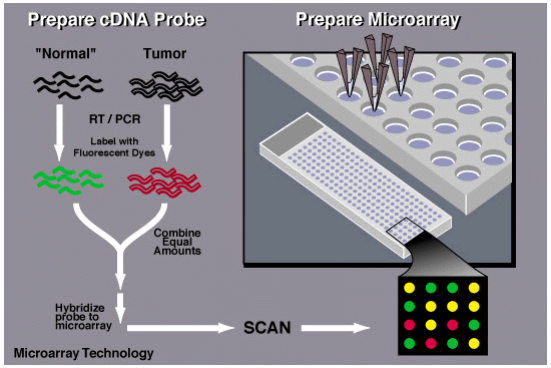
ENCODE Project

Example: Microarray

- The study of biological molecules at large scale
  - Genomics
  - **Transcriptomics**
  - Proteomics
  - Metabolomics
  - Epigenomics
  - Lipidomics
  - Microbiomics



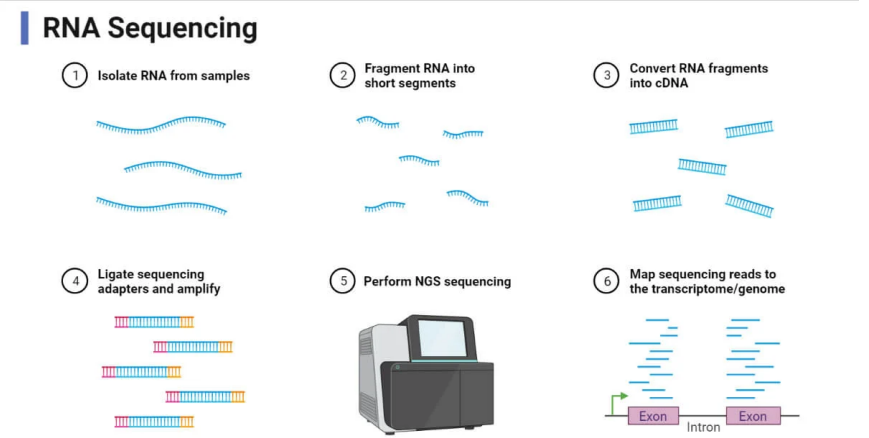


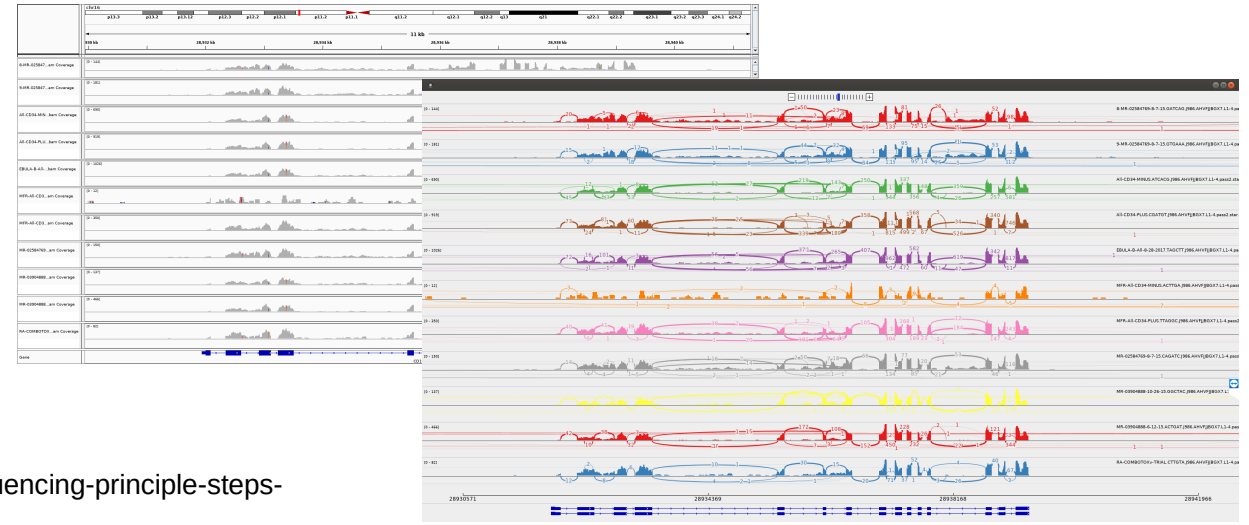


https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology

Example: NGS RNA-seq

- The study of biological molecules at large scale
  - Genomics
  - **Transcriptomics**
  - Proteomics
  - Metabolomics
  - Epigenomics
  - Lipidomics
  - Microbiomics

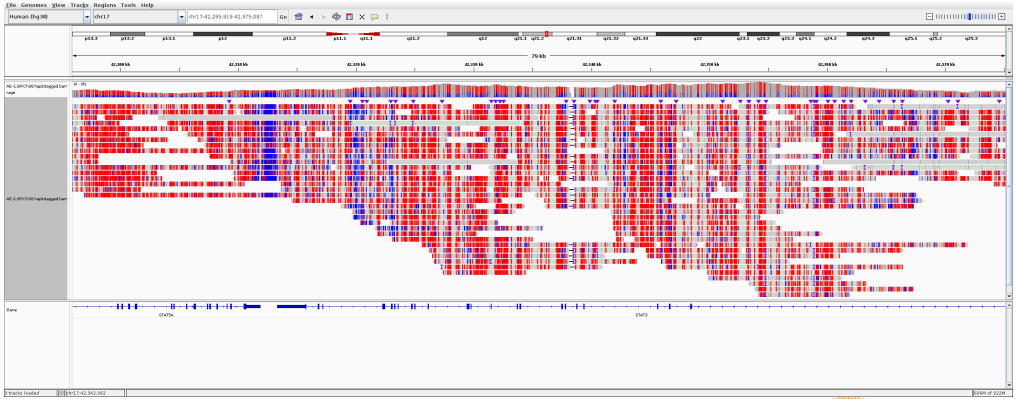


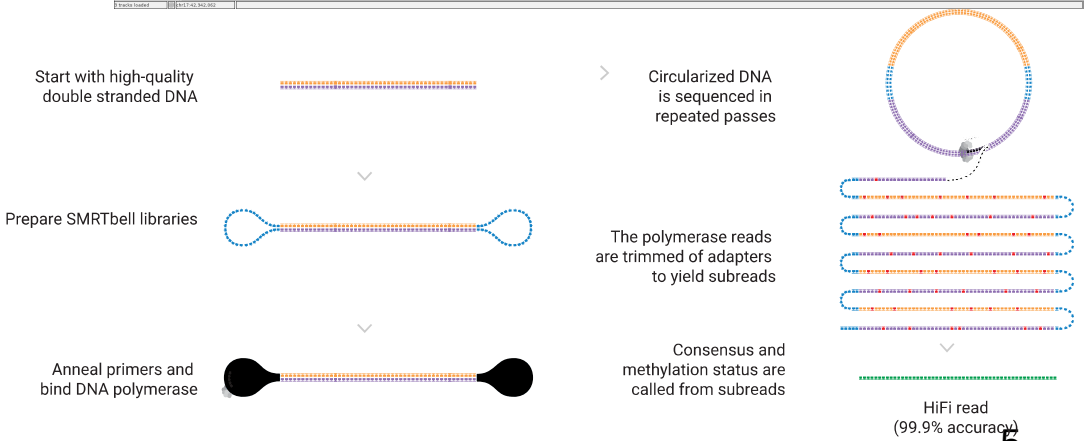


https://microbenotes.com/rna-sequencing-principle-steps-types-uses/

Example: PacBio HiFi-seq

- The study of biological molecules at large scale
  - Genomics
  - **Transcriptomics**
  - Proteomics
  - Metabolomics
  - **Epigenomics**
  - Lipidomics
  - Microbiomics





https://www.pacb.com/technology/hifi-sequencing/

The Biggest Issue in Omics Analysis?

- The differences between Omics technologies are vast and each requires specialized tools to process and analyze.
- But there is a common issue shared across all.
  - # of variables: N
  - # of observations: M
  - N >> M

ENSEMBL	ENTREZGENE	SYMBOL	Mock1.1	Mock2.1	ZIKV1.1	ZIKV2.1	Mock1.2	Mock2.2	ZIKV1.2
ENSG00000010030	51513	ETV7	0.0708594	0	0.3114447	0.2252182	0.0493682	0.0367962	0.2537942
ENSG00000010072	83932	SPRTN	3.1900906	2.6378208	3.0984427	3.2892438	3.6348061	3.2145277	4.6315328
ENSG00000010165	51603	METTL13	4.9670487	4.2128095	3.6224677	3.5393062	4.4813932	4.3178785	3.6262744
ENSG00000010219	8798	DYRK4	2.0031906	2.0030852	3.4851214	2.9813266	2.0685315	2.0866457	3.0435702
ENSG00000010244	7756	ZNF207	12.419449	12.024094	12.350847	13.169165	15.314333	14.552613	16.34819
ENSG00000010256	7384	UQCRC1	40.287622	40.417455	40.098703	38.056824	38.66833	40.203852	43.152384
ENSG00000010270	83930	STARD3N	19.151117	17.581794	20.533327	19.142134	22.498407	21.205717	25.820423
ENSG00000010278	928	CD9	1.3767652	1.0982165	0.3922089	0.7563255	0.9651221	1.1430112	0.4565833
ENSG00000010282	57467	HHATL	0	0.0626603	0	0	0.0202698	0.020144	0.0208408
ENSG00000010292	9918	NCAPD2	32.54455	29.572793	14.275639	14.808391	30.03627	28.103656	14.74612
ENSG00000010295	25900	IFFO1	4.9790368	6.1973561	6.5368079	6.3027053	5.0757457	5.4397063	6.9248392
ENSG00000010310	2696	GIPR	0.4262339	0.3824975	0.3122339	0.3386833	0.4536887	0.3842655	0.368934
ENSG00000010318	51533	PHF7	2.5745529	2.5632081	1.3344099	1.6468745	2.3408447	2.3207032	1.1192984
ENSG00000010319	56920	SEMA3G	0.0671404	0.1446025	0	0	0.0779619	0.0958788	0.0721422
ENSG00000010322	11188	NISCH	29.013	32.390026	41.020053	39.998457	25.416185	27.365164	38.895627
ENSG00000010327	23166	STAB1	0	0	0.0174577	0.0168325	0.0147589	0.0068753	0.032246
ENSG00000010361	80199	FUZ	5.6453597	6.5091477	3.4462152	3.7456973	4.8336793	5.6212736	3.635468
ENSG00000010379	6540	SLC6A13	0.0792931	0.042694	0.2613846	0.4200393	0.082866	0.0892141	0.3408008
ENSG00000010404	3423	IDS	16.637373	17.058433	19.773289	21.232514	18.890098	17.876043	25.156345
ENSG00000010438	5646	PRSS3	0.2021879	0.1088647	0	0.321315	0.0939104	0.0874943	0.1689724
ENSG00000010539	7752	ZNF200	1.8396994	1.8361512	1.9475265	2.020396	2.1310069	2.0951206	2.5794192
ENSG00000010610	920	CD4	0.2936423	0.1976334	0.201661	0.1166633	0.2727766	0.3017913	0.184052
ENSG00000010626	10233	LRRC23	3.8557374	3.8555346	3.08676	2.2759283	3.1532244	3.4991297	2.8737798
ENSG00000010671	695	BTK	0	0	0.0404027	0	0.0256175	0.0318231	0.0219493
ENSG00000010704	3077	HFE	0.0592847	0	0.0651427	0.0628098	0.013768	0.0128274	0.0389286

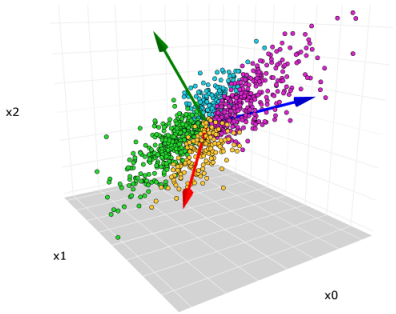
N >> M

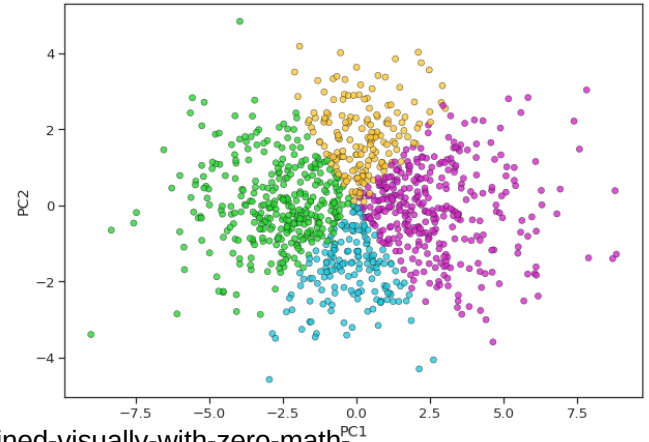
- How do we address this?
  1. Dimensionality reduction
  2. Choosing the right test
  3. Dealing with the multiple comparisons in N variables

ENSEMBL	ENTREZGENE	SYMBOL	Mock1.1	Mock2.1	ZIKV1.1	ZIKV2.1	Mock1.2	Mock2.2	ZIKV1.2
ENSG00000010030	51513	ETV7	0.0708594	0	0.3114447	0.2252182	0.0493682	0.0367962	0.2537942
ENSG00000010072	83932	SPRTN	3.1900906	2.6378208	3.0984427	3.2892438	3.6348061	3.2145277	4.6315328
ENSG00000010165	51603	METTL13	4.9670487	4.2128095	3.6224677	3.5393062	4.4813932	4.3178785	3.6262744
ENSG00000010219	8798	DYRK4	2.0031906	2.0030852	3.4851214	2.9813266	2.0685315	2.0866457	3.0435702
ENSG00000010244	7756	ZNF207	12.419449	12.024094	12.350847	13.169165	15.314333	14.552613	16.34819
ENSG00000010256	7384	UQCRC1	40.287622	40.417455	40.098703	38.056824	38.66833	40.203852	43.152384
ENSG00000010270	83930	STARD3N	19.151117	17.581794	20.533327	19.142134	22.498407	21.205717	25.820423
ENSG00000010278	928	CD9	1.3767652	1.0982165	0.3922089	0.7563255	0.9651221	1.1430112	0.4565833
ENSG00000010282	57467	HHATL	0	0.0626603	0	0	0.0202698	0.020144	0.0208408
ENSG00000010292	9918	NCAPD2	32.54455	29.572793	14.275639	14.808391	30.03627	28.103656	14.74612
ENSG00000010295	25900	IFFO1	4.9790368	6.1973561	6.5368079	6.3027053	5.0757457	5.4397063	6.9248392
ENSG00000010310	2696	GIPR	0.4262339	0.3824975	0.3122339	0.3386833	0.4536887	0.3842655	0.368934
ENSG00000010318	51533	PHF7	2.5745529	2.5632081	1.3344099	1.6468745	2.3408447	2.3207032	1.1192984
ENSG00000010319	56920	SEMA3G	0.0671404	0.1446025	0	0	0.0779619	0.0958788	0.0721422
ENSG00000010322	11188	NISCH	29.013	32.390026	41.020053	39.998457	25.416185	27.365164	38.895627
ENSG00000010327	23166	STAB1	0	0	0.0174577	0.0168325	0.0147589	0.0068753	0.032246
ENSG00000010361	80199	FUZ	5.6453597	6.5091477	3.4462152	3.7456973	4.8336793	5.6212736	3.635468
ENSG00000010379	6540	SLC6A13	0.0792931	0.042694	0.2613846	0.4200393	0.082866	0.0892141	0.3408008
ENSG00000010404	3423	IDS	16.637373	17.058433	19.773289	21.232514	18.890098	17.876043	25.156345
ENSG00000010438	5646	PRSS3	0.2021879	0.1088647	0	0.321315	0.0939104	0.0874943	0.1689724
ENSG00000010539	7752	ZNF200	1.8396994	1.8361512	1.9475265	2.020396	2.1310069	2.0951206	2.5794192
ENSG00000010610	920	CD4	0.2936423	0.1976334	0.201661	0.1166633	0.2727766	0.3017913	0.184052
ENSG00000010626	10233	LRRC23	3.8557374	3.8555346	3.08676	2.2759283	3.1532244	3.4991297	2.8737798
ENSG00000010671	695	BTK	0	0	0.0404027	0	0.0256175	0.0318231	0.0219493
ENSG00000010704	3077	HFE	0.0592847	0	0.0651427	0.0628098	0.013768	0.0128274	0.0389286

1. Dimensionality Reduction

- Huge # of variables to look over
- Hard to show more than 3 vars with traditional plots
- Not all the variables are equally important
- Can we get away with a smaller representation that only contains relevant informatio

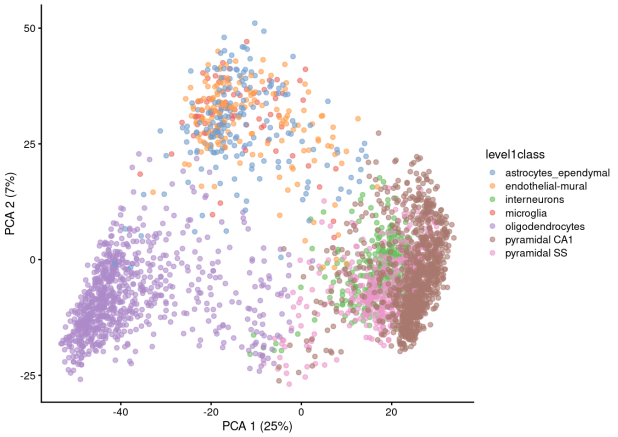


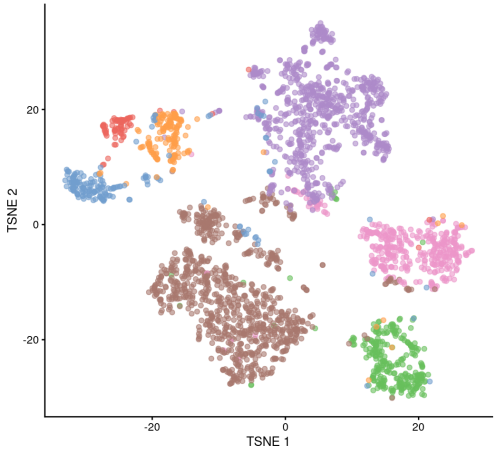


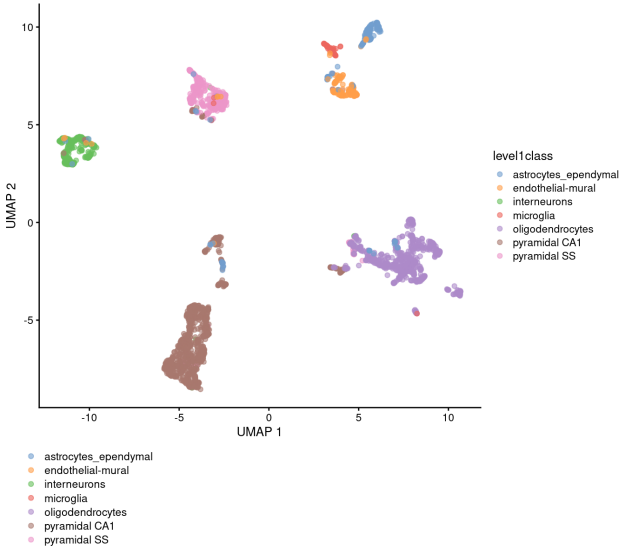
https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d/

Dimensionality Reduction

- Common dimensionality reduction techniques:
  - PCA
  - t-SNE
  - UMAP





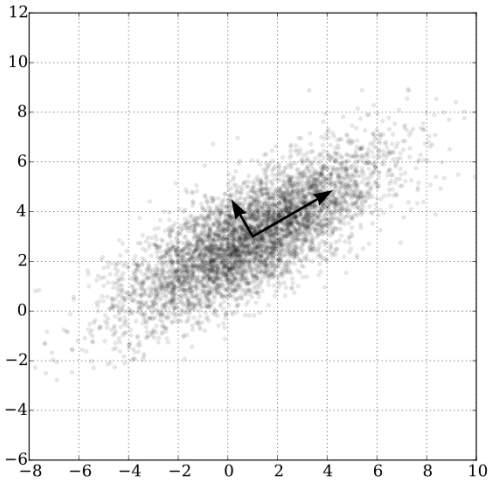


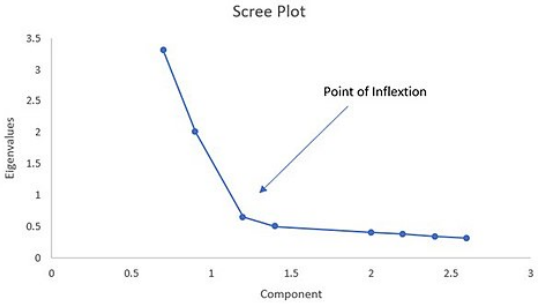
https://bioconductor.org/books/3.15/OSCA.basic/dimensionality-reduction.html

Principle Component Analysis(PCA)

(Pearson 1901)

- Linear projections along lines of greatest variance
- Principal components(the new orthogonal axes) are ordered by the amount of variance explained.
- The idea is that we can use a smaller set of these PCs to capture most of the information of the original data.



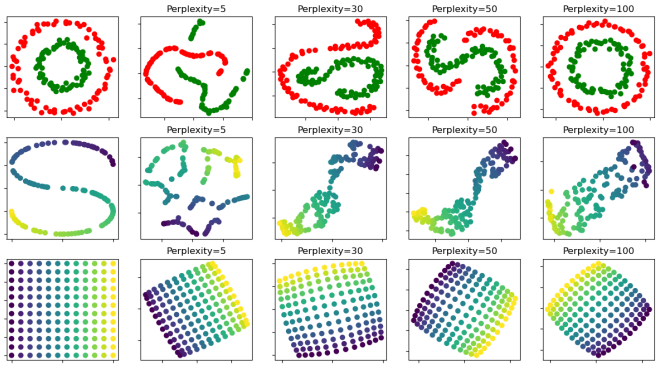


https://en.wikipedia.org/wiki/Principal_component_analysis

t-distributed Stochastic Neighbor Embedding (t-SNE)

(Maaten & Hinton 2008)

- A Non-linear approach to dimension reduction
- Idea: Distances between the points in the high dim space should be similar to distance between points in the low dim space
  - High dim: Distances represented by Gaussian
  - Low dim: Distance represented by t-distribution
  - Loss function: Measures differences between the distance representations in high and low dimensions
    - Kullback-Leibler divergence
- Optimize: find a low dim representation that minimizes the loss function.
  - Start with a random low dim representation
  - Iteratively shift points in the low dim space so that loss function goes down
    - move along the gradient of the Kullback-Leibler divergence
- Key hyperparameter
  - Perplexity: The width of the gaussian or t-distribution.
    - A wider bell shaped curve means we're looking at more points at each step.

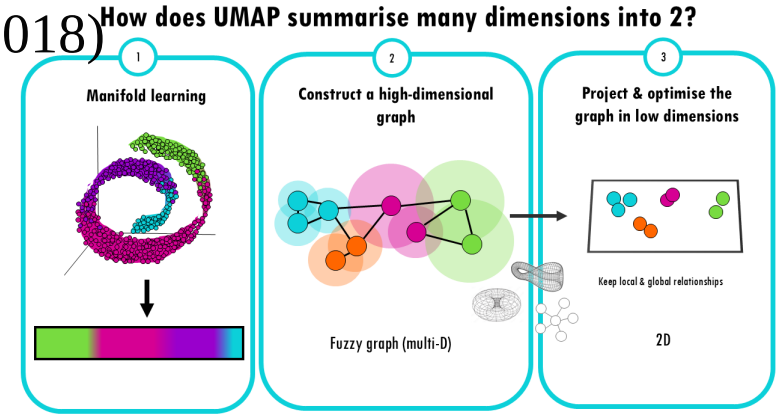


https://scikit-learn.org/stable/auto_examples/manifold/plot_t_sne_perplexity.html

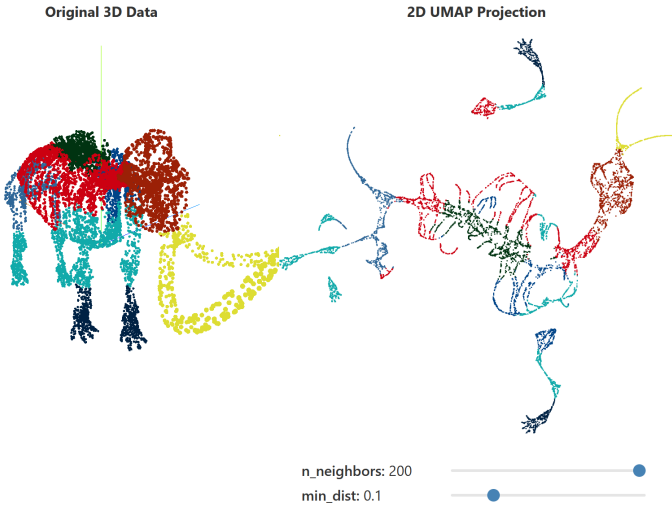
Uniform Manifold Approximation and Projection (UMAP)

(Innes & Healy 2018)

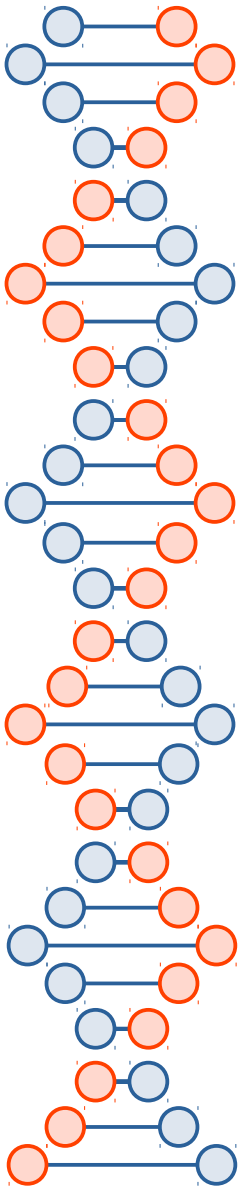
- A Non-linear approach to dimension reduction
- Idea: Distances between the points in the high dim space should be similar to distance between points in the low dim space
  - High dim: weighted sum of k nearest neighbor graphs
  - Low dim: weighted sum of k nearest neighbor graphs
  - Loss function: Measures differences between the distance representations in high and low dimensions
    - Cross entropy of adjacency matrices
- Optimize: find a low dim representation that minimizes the loss function.
  - Start with a random low dim representation
  - Iteratively shift points in the low dim space so that loss function goes down
    - Minimize cross entropy by stochastic gradient descent
- Key hyperparameter
  - n_neighbors: how many neighbors to look at



https://biostatsquid.com/umap-simply-explained/



https://pair-code.github.io/understanding-umap/

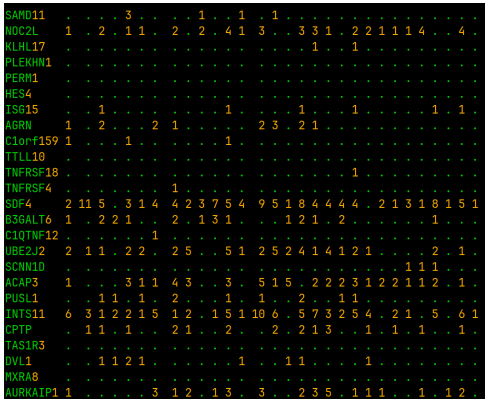


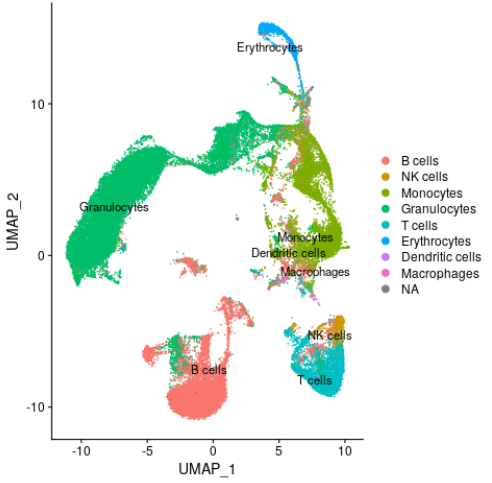
Dimensionality Reduction

- Pros
  - PCA
    - Easy to interpret
    - Fast
  - T-SNE
    - Captures non-linear relationships between variables
    - Preserves local structure
  - UMAP
    - Captures non-linear relationships between variables
    - Preserves local and global structure
    - Scalable
    - Stable
- Cons
  - PCA
    - Only linear relationships
  - t-SNE
    - Very slow
    - Changing the perplexity hyperparameter gives very different results across runs
  - UMAP
    - Not as good as t-SNE if you're ONLY interested in local structure.

Dimensionality Reduction: example1

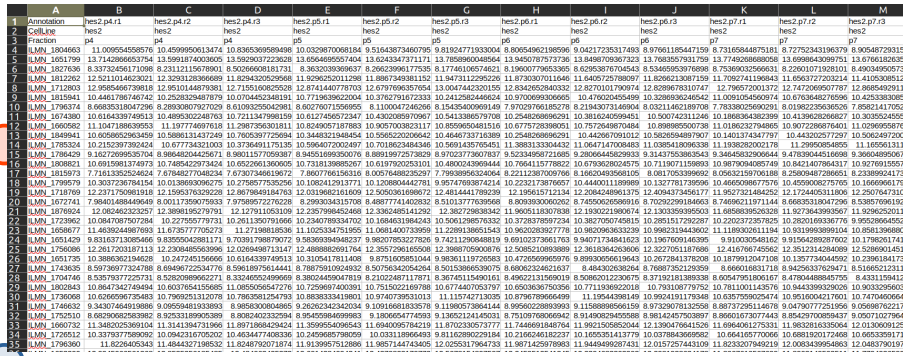
- Single-cell sequencing
  - We are able to capture gene expression for each individual cell
  - Too many genes to be able to map out everything
  - A 2D UMAP (or t-SNE) of all cells allows us to:
    - Visually examine the individual cells
    - Pick out clusters of similar cells
    - Identify celltypes of the clusters and their relationships to one another
      - Manual celltype type ID
      - Automatic ID by Machine learning algorithms



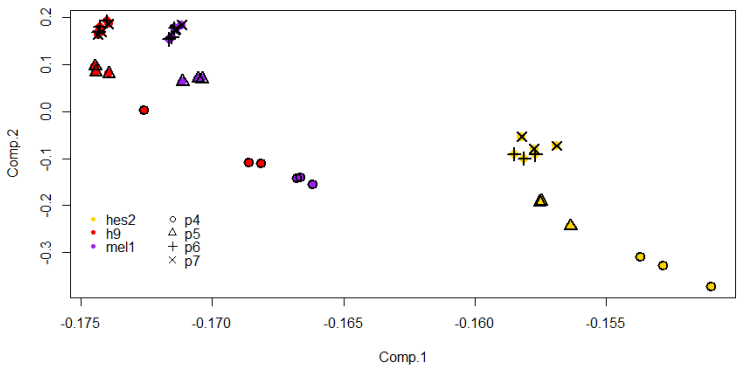


Dimensionality Reduction: example 2

- RNA-seq experiment with a batch effect
  - Examining pluripotency of 3 different cell lines
- A PCA plot of the samples gives us a global picture of gene expression
  - This allows us to identify batch effects visually
    - Once you know they are there, they can be corrected.

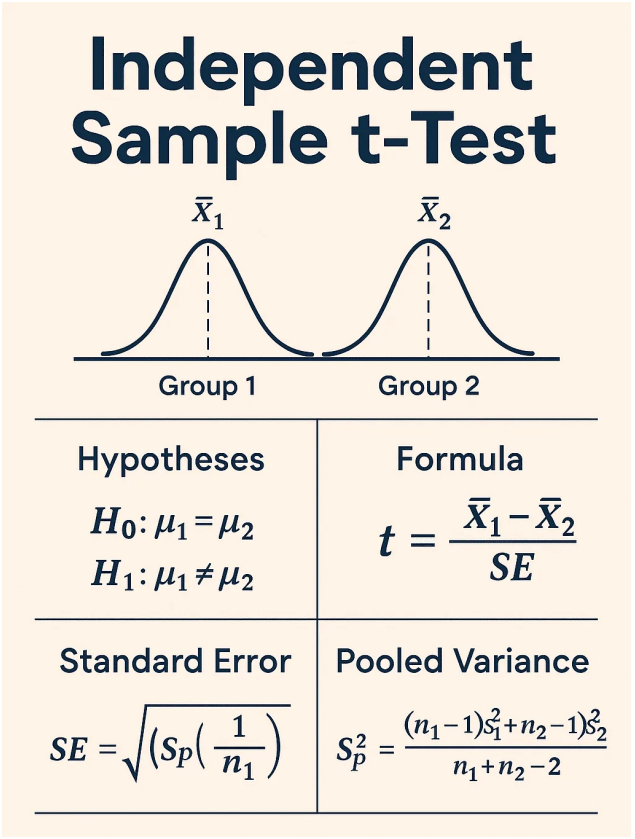


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Annotation	hes2.p4.r1	hes2.p4.r2	hes2.p4.r3	hes2.p5.r1	hes2.p5.r2	hes2.p5.r3	hes2.p6.r1	hes2.p6.r2	hes2.p6.r3	hes2.p7.r1	hes2.p7.r2	hes2.p7.r3
2	CellLine	hes2	hes2	hes2	hes2	hes2	hes2	hes2	hes2	hes2	hes2	hes2	hes2
3	Fraction	p4	p4	p4	p5	p5	p5	p6	p6	p6	p7	p7	p7
4	ILMN_1804663	11.0095545588576	10.4599950613474	10.8365369589498	10.0329870068184	9.51643873460795	9.81924771933004	8.80654962198696	9.04217235317493	8.97661185447159	8.73165844875181	8.72752343196379	8.90548729315
5	ILMN_1651799	13.7142866653754	13.5991874003605	13.5929037223628	13.6564695557404	13.6243347371171	13.785896004856	13.9450787573736	13.8498709367323	13.7683557931759	13.7749268688058	13.6998643099751	13.6766182635
6	ILMN_1827636	8.33732456171098	8.23112115678901	8.50266608181731	8.363203936963	8.26623996177535	8.17746106574621	8.19600779653365	8.62953876704543	8.53465953976698	8.75369032566631	8.22601071928101	8.49034950573
7	ILMN_1812262	12.5211014623021	12.3293128366689	11.8294320529568	12.9296252011298	11.8867349381152	11.9473112295226	11.8730307011646	11.6405725788097	11.8206213087159	11.709274119684	11.6563727203214	11.4105308551
8	ILMN_1712803	12.9585466739818	12.9510144879381	12.7155160825528	12.8741440778703	12.6797696357654	13.0047442320155	12.8342652840332	12.8270101790974	12.8289678310747	12.796572001372	12.7472069507787	12.8685492911
9	ILMN_1815941	10.4461786746742	10.2528329487879	10.0704452348191	10.7719639622004	10.3762791672333	10.2412582446624	10.9700699306665	10.476020455499	10.3286936246542	11.0091054560974	10.6763648276596	10.4253383085
10	ILMN_1796374	8.66835318042796	8.28930807927029	8.61093255042981	8.60276071556955	8.11000472462666	8.15435400969149	7.97029766185278	8.21943073146904	8.03211462189708	7.78339025690291	8.01982235636526	7.8582141705
11	ILMN_1674380	10.6164339749513	10.4895302248763	10.721134799815	10.6127456572347	10.4302085970967	10.5413386579708	10.2548268696291	10.3816240599451	10.500742311246	10.186836438239	10.4139628266827	10.3035524555
12	ILMN_1660582	11.1047188639553	11.197774697618	11.2987356301811	10.8249057187883	10.9057003823117	10.8559950481516	10.677572839805	10.7572649870484	10.898985500738	11.0186232794865	10.9072286876401	11.0296955876
13	ILMN_1949941	10.6058652963459	10.5896131437249	10.7605397725694	10.3448321948454	10.5065222006642	10.464673716389	10.2548268696291	10.446067091012	10.5826594897907	10.140137434777	10.4432025777297	10.5062497200
14	ILMN_1785324	10.215239739242	10.6777343210003	10.3736491175135	10.5964072002497	10.7018623484346	10.5691435765451	11.3883133304432	11.0647147008483	11.0385418096338	11.1938282002178	11.29950854855	11.1655613
15	ILMN_1786429	9.16272699535704	8.98648204425671	8.98011577059387	8.94551699350076	8.88919972573829	8.97023773607837	9.52334958721685	9.28066445829933	9.34145532863543	9.34645832906644	9.47839044516698	9.3660489506
16	ILMN_1808821	10.6915981374973	10.7485422973424	10.6522661360605	10.7318139885267	10.6197920253101	10.480024396944	10.7664115778822	10.6778362824575	10.7119071159893	10.9879094085749	10.8421407864317	10.9276915557
17	ILMN_1815973	7.71613352524624	7.67848277048234	7.67307346619672	7.8607766156316	8.00574688235297	7.79938956324064	8.22112387009766	8.16620493568105	8.0817053399692	8.05632159706188	8.25809487286651	8.2338992417
18	ILMN_1799579	10.3037236784154	10.0138693096275	10.275857753526	10.1082412913771	10.1208804442781	9.95747693874214	10.2232173876657	10.444001118989	10.1327781739596	10.4665098677576	10.4559008275765	10.1666966175
19	ILMN_1718769	12.2371750981918	12.1595376329228	12.8678849184763	12.0319682161609	12.5056391698672	12.481444178923	12.1956157121134	12.2084248961375	12.4094373456177	11.9527321484252	12.1724405311806	12.2507647315
20	ILMN_1672741	7.98401488449649	8.00117359075933	7.97589572276228	8.2993034315708	8.48877741402832	8.51013777639568	8.8093930060262	8.7455062586916	8.70292299184663	8.74696211971144	8.66835318047296	8.5385769619
21	ILMN_1876924	12.082462323257	12.3898195279791	12.127911053109	12.2357998452468	12.3362485141292	12.382729838342	11.9605118307838	12.1930221980674	12.1303359395503	11.6858839526328	11.9273643993567	11.9296252011
22	ILMN_1723982	10.0847087507284	10.2275557779731	10.2611350791666	10.2340788334702	10.1694631984243	10.5061298576332	10.3728378597234	10.3827050745815	10.2851517292287	10.2202372357825	10.2820169336776	9.9552866455
23	ILMN_1658677	11.4639244907693	11.673577705273	11.271988181536	11.1025334751955	11.0681400733959	11.228913865143	10.9620283927778	10.9820963633239	10.9982319443402	11.1189326111194	10.9319953899104	10.8581396880
24	ILMN_1651429	9.83163713085466	9.83555042881171	9.70391798879072	9.58369394948237	9.98207853227826	9.742112908488 19	9.69102373661763	9.94071734841623	10.196760946395	9.910030548162	9.91564289287602	10.179826174
25	ILMN_1756086	12.2617203187113	12.2308485563996	12.0269498713147	12.4888862691764	12.3557296165508	12.3988705900876	12.5085210893889	12.3618364263606	12.3227051187686	12.416766745962	12.3512314284089	12.5286901451
26	ILMN_1651735	10.3886302194628	10.247245156666	10.6164339749513	10.3105417811408	9.87510605851044	9.98361119726583	10.4726599695976	9.89930656619643	10.2672841378208	10.1879912047108	10.1357734094592	10.2396184171
27	ILMN_1743635	8.59736977324788	8.69496722534776	8.59618975614441	8.78875910924932	8.50756342054264	8.50153866539075	8.68062324621637	8.484302683264	8.76887352129359	8.666016831718	8.94256337629471	8.51665212311
28	ILMN_1704746	8.53579377225731	8.52820989662271	8.33246552499669	8.38024459047819	8.21022487117871	8.36745115490161	8.49622131569019	8.50862012230675	8.37192181389338	8.60547951800167	8.47904488845755	8.4331159412
29	ILMN_1802843	10.8647342749494	10.6037654155685	11.0855096547276	10.7296977400391	10.7515022169788	10.677440705379 7	10.6503636750356	10.7711936922018	10.793108779752	10.7811001143576	10.9443399329026	10.9033295603
30	ILMN_1736068	10.6266596735883	10.7969251312078	10.7863581254793	10.8833333419801	10.9740739531013	11.11574213035	10.8796789666499	11.19543398149	10.9924191179348	10.6357559025474	10.9516004217601	10.7470406664
31	ILMN_1746632	9.34307464919886	9.09559481933893	8.96583008804865	9.26262342342034	9.10916681833578	9.11980573864144	8.99560228893993	9.11588898566159	8.97329078132558	8.88737295114678	9.04790777251956	9.0569876221
32	ILMN_1752510	8.68290682583982	8.92533189905389	8.90824023329 94	8.95405984699983	9.18066547745993	9.13652124145031	8.75109768066942	8.91490829455088	8.98142457503897	8.89601673077443	8.85429700859437	9.0907102796
33	ILMN_1660732	11.3482025369104	11.3141394731966	11.8971868429424	11.3599554096543	11.6940095784219	11.8702330573777	11.7446691848764	11.9921505852044	12.1390476641526	11.6964061275331	11.9832816335064	12.0130091257
34	ILMN_1726512	10.3379377589092	10.0942316705202	10.4634477408336	10.2459665798059	10.033118966493	9.81162890229184	10.2166246182237	10.1655351413779	10.0378843669582	10.664165770066	10.6881920172468	10.6653359171
35	ILMN_1796360	11.822640543	11.4844327198532	11.8248792071874	11.9139957512886	11.9857144743405	12.0255317964733	11.9871425978963	11.9449498287431	12.0157257443109	11.8233207949219	12.0083439654863	12.0483790197



2. Choosing the right test

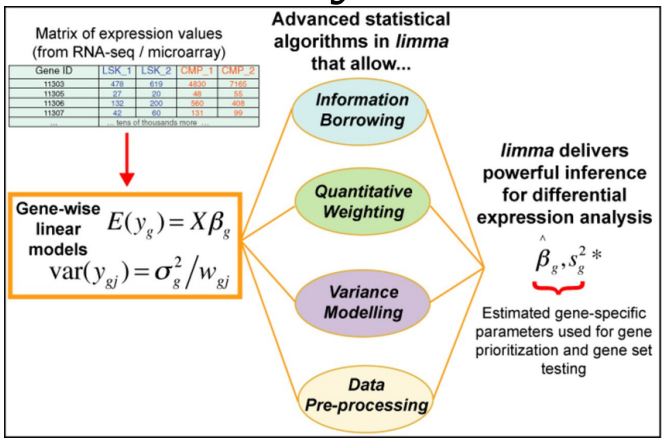
- A t-test is good for large sample sizes or if data is normally distributed
  - This is usually not the case in Omics datasets
- Each Omics technology needs its own particular methodology
  - 1. A stat that fits the data
  - 2. An "information borrowing" technique to deal with the N >> M problem



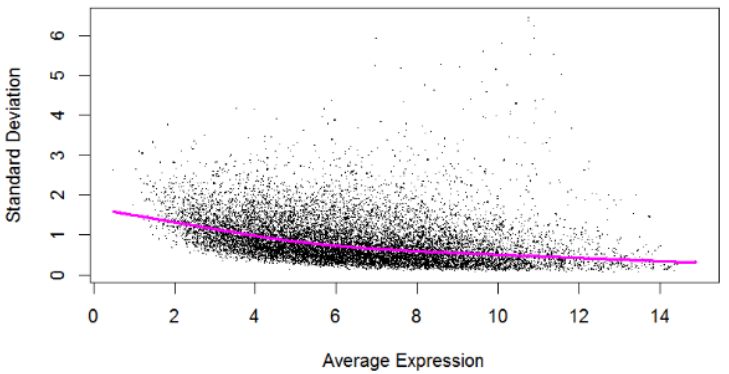
https://www.statisticalaid.com/independent-sample-t-test/

Example 1: LInear Models for MicroArray data (LIMMA)

- The underlying statistical framework uses empirical Bayes linear modeling.
- Conceptually, for simple two group comparisons, it can be thought of as a moderated t-test
  - Looks for differences in gene expression across experimental groups.
  - But uses an empirical Bayesian approach to "shrink" variance



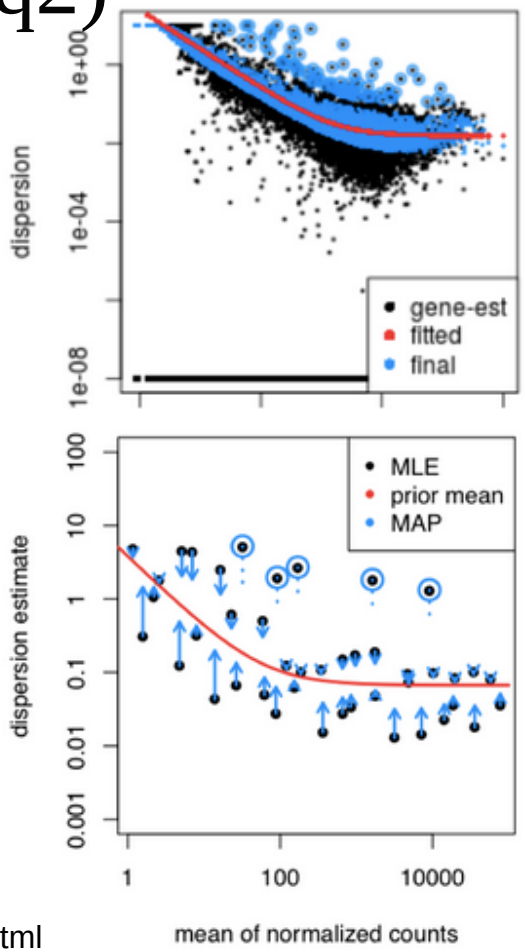
$$T_{(gene)} = \frac{\acute{D} - \acute{C}}{f\left(Var\left(D, C\right) + \alpha\right)}$$



https://pmc.ncbi.nlm.nih.gov/articles/PMC4402510/

Example 2: Differential Expression analysis for Sequencing data (DESeq2)

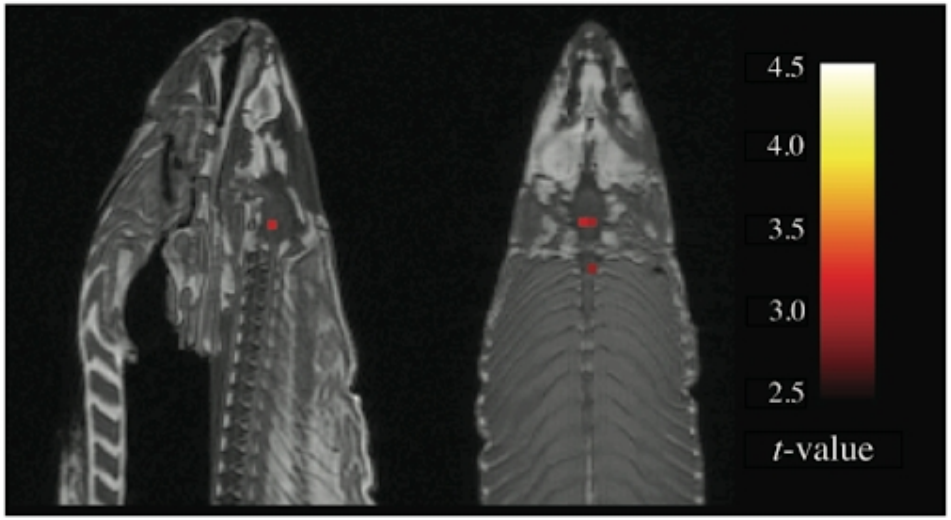
- RNA-seq gives us integer counts of the number of read fragments that map within a gene's boundary
- Negative Binomial Regression
  - Like Poisson, but allows for over/under dispersion through parameter (θ)
  - Uses similar genes to estimate the dispersion parameter (θ)



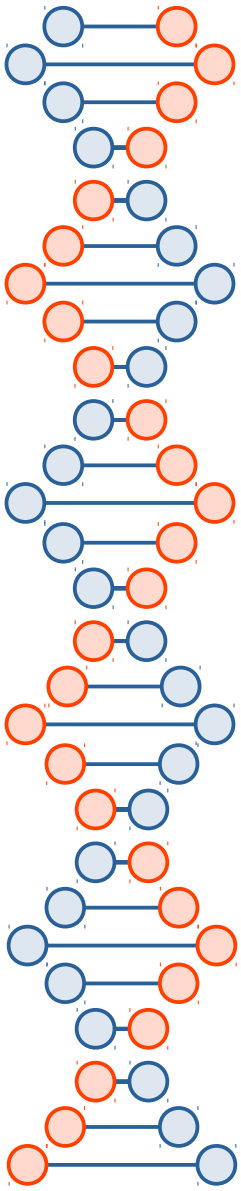
https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

3. Dealing with multiple comparisons

- Type 1 error: Incorrectly rejecting the null hypothesis when the null is true.
  - Finding signal when there really is none there
- If you set significance level (α) at 0.05, and run 50,000 tests across all genes
  - How many false positives do you expect?
    - 50000*0.05 = 2500
- How do you address it?
  - Do nothing
  - Control for FWER
  - Control for FDR



http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf



Dealing with multiple comparisons

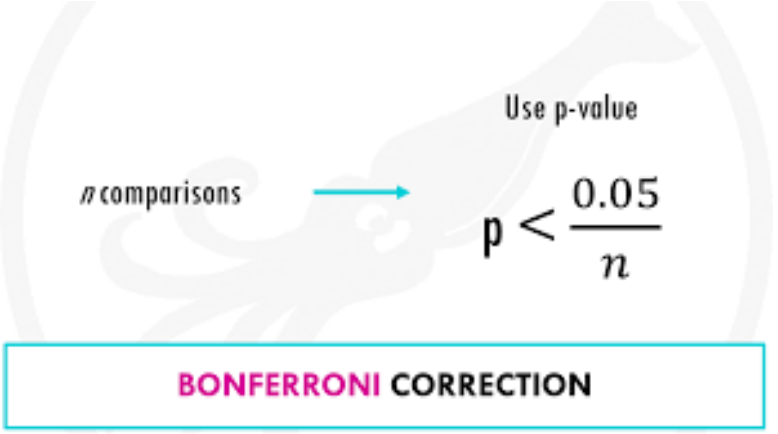
- Do nothing:
  - Ok if you have a specific location of interest known beforehand
    - Do NOT look at all the results then select the location of interest
  - Ok if there are a small set of interesting locations known beforehand
    - Show them all
  - Do NOT only show the significant results

Dealing with multiple comparisons

Control for Family-Wise Error Rate(FWER):

- Make at most X false positives across all tests
- How many expected false positives in 100 tests at alpha=0.05?
- How many expected false positives in 100 tests at FWER=0.05?

- Bonferroni
  - Instead of selecting 0.05 as your type I error rate, choose 0.05 / the number of tests
  - Makes no assumptions about correlation structure in the multiple tests
  - This is the most conservative way to deal with FWER, and is very hard to meet the significance threshold.



https://biostatsquid.com/multiple-testing-correction-fdr/

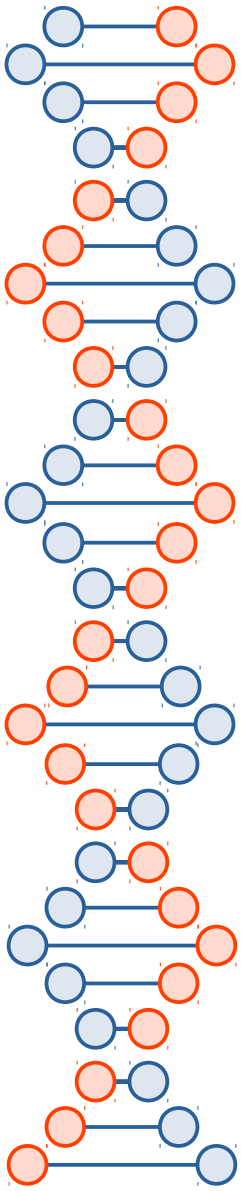
Dealing with multiple comparisons

Control for False Discovery Rate(FDR)

- Type I error rate:
  - In all the tests I've run, how many false positives were there?
- FDR:
  - Among the tests that where deemed significant, how many were incorrect.

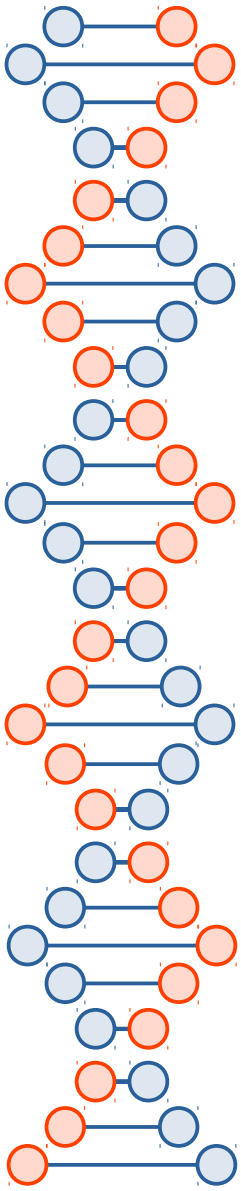
- Benjamini-Hochberg Method
  - The least conservative technique and the most widely used in omics studies
  - Define q = V/R where
    V = # false positives,
    R = # discoveries (rejected H_0s).
    - Order the P-values in ascending order, smallest first: P(1) < … < P(N).
    - For α, find the largest k such that P(k) ≤ (k/N)α.
    - Then reject all H0(i) for i = 1, …, k.

https://www.youtube.com/watch?v=-oIkIdhSNeU



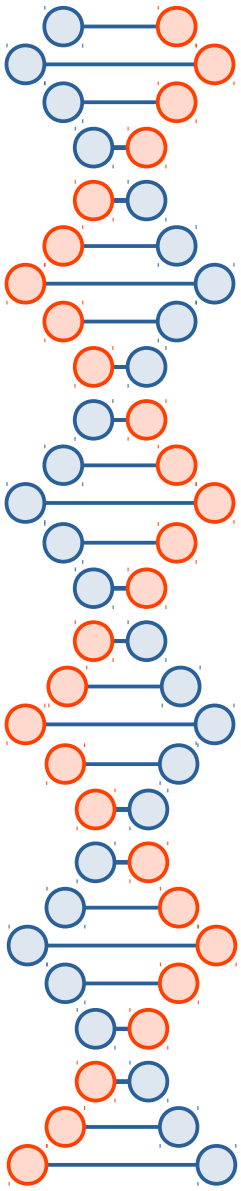
Typical workflow in an Omics study

1. Plot global signal of all samples
2. Look for differential signals or associations across conditions
3. Make sure your method deals with multiple comparisons



Typical workflow in an Omics study

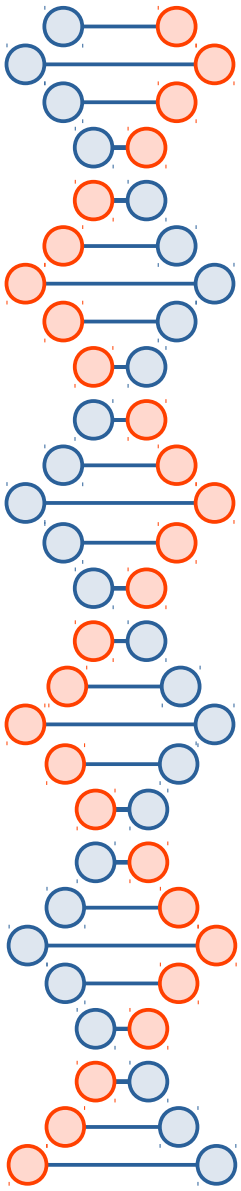
1. Plot global signal of all samples
   - Find outliers or potential problems with the way the data was prepared/processed
   - See if there are batch effects that needs to be addressed
   - See if there are differences between experimental conditions



Typical workflow in an Omics study

2. Look for differential signals or associations across conditions

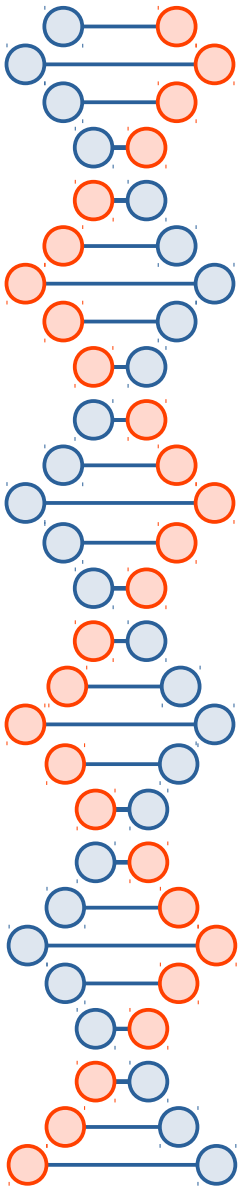
- Use an R package tailored to your protocol
  - deseq2
  - LIMMA
  - methylkit
  - more
- Use specialized software developed for your protocol
  - Homer
  - RMats
  - Bismark
  - more
- Write your own



Typical workflow in an Omics study

3. Make sure your method deals with multiple comparisons

- Most of the well known bioinformatics software packages already take care of this
- At the very least, adjust for FDR



Common filetypes for Omics

- What is a fastq file?
- What is a fasta file?
- What do you get when you map a fastq to a fasta?
- What is a bam file?
- What is a gtf file?
- What is a gene count file?