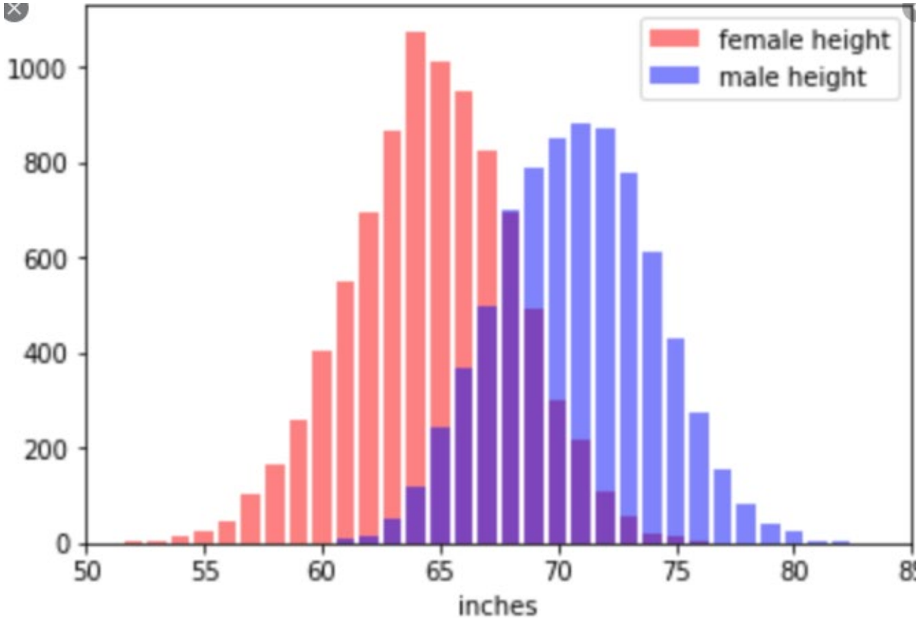
MAYO
CLINIC

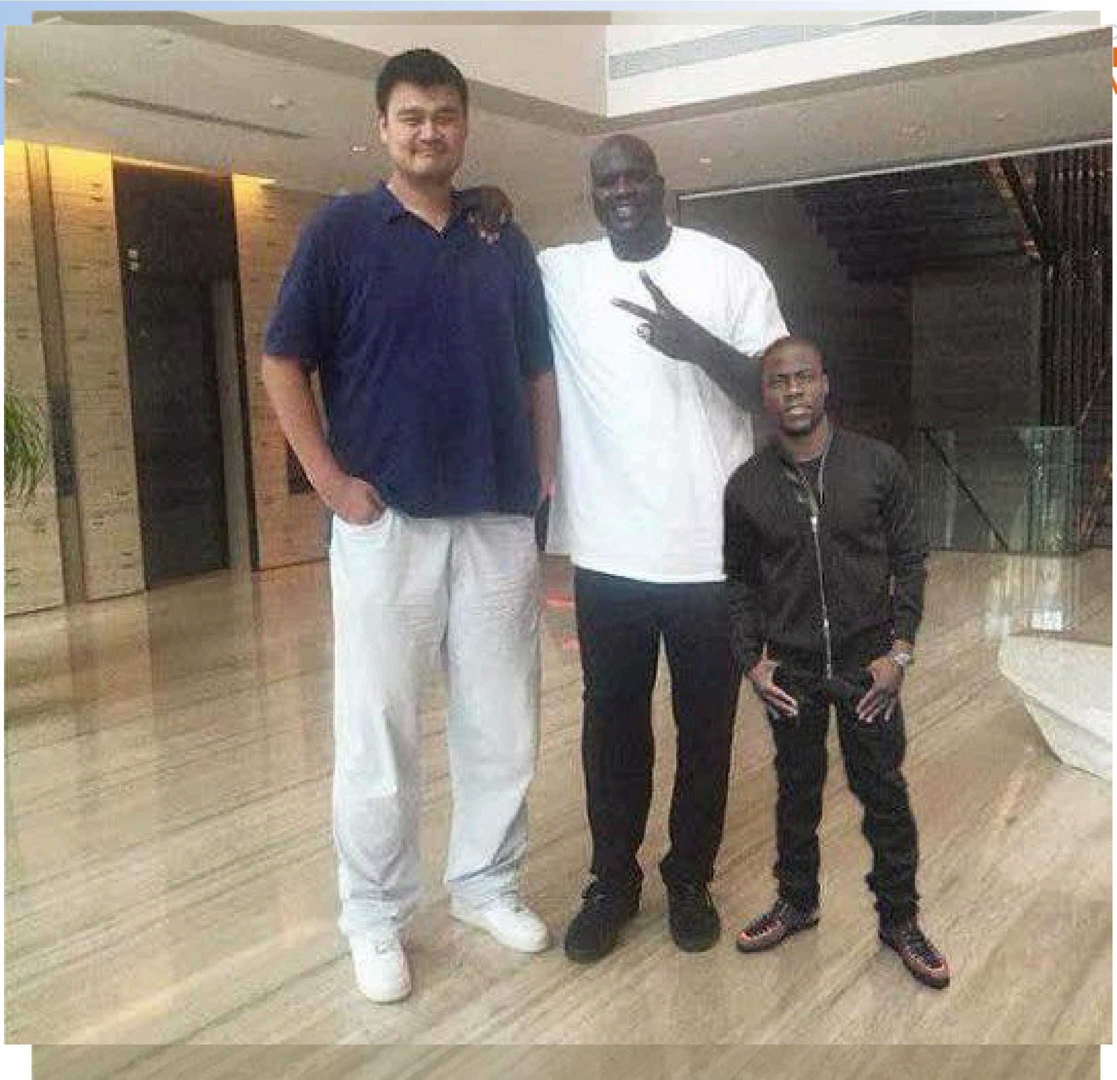
Confounding variable correction and outlier expression analysis

A. Collin Osborne
Senior Bioinformatician
Mayo Clinic Quantitative Health Sciences
Mayo Clinic Center for Individualized Medicine
Authored by: **Eric Klee, PhD and Garrett Jenkinson PhD**

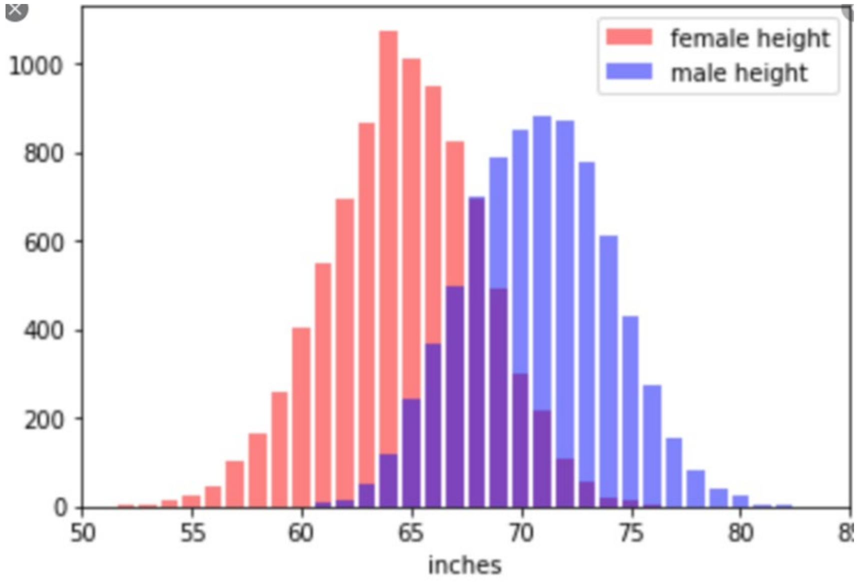
Confounding variables

- We sample a population many times and find the following data
- A 5’ tall man is unusually short, but if we did not factor in sex then we would not see 5’ tall adult as an outlier
- A 5’ male peds patient not an outlier so age also confounder



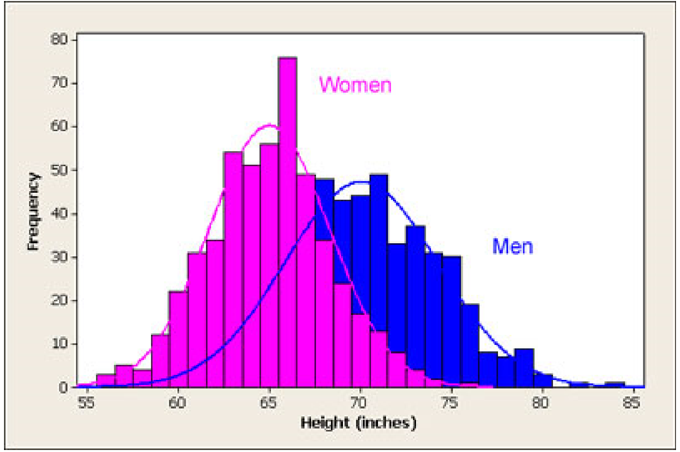


- Yao Ming is a male with 90” height
- Outlier detection has the goal of seeing this example and flagging it as an outlier or anomaly since *it is unlikely* within the population
- There are many methods for quantitatively deciding what is “unlikely” but we will discuss the broad class of methods based on statistical hypothesis tests/p-values



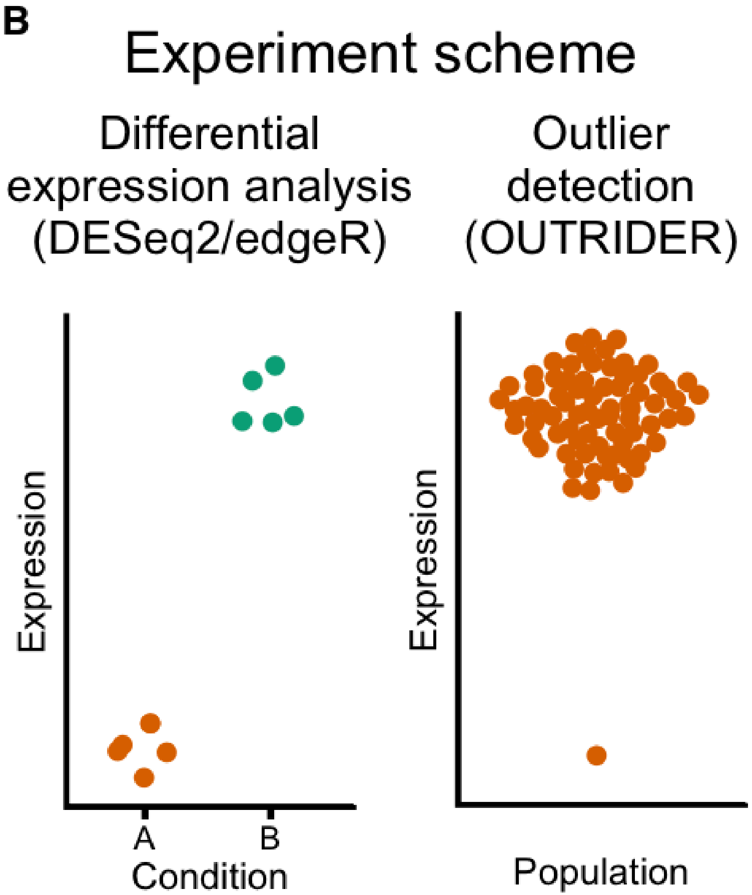
Statistical testing approach

- First define a model, for example a separate bell curve for men and women
- Collect "normal" data and fit models to the data
  - "normal" depends on scientific question; e.g., outlier in NBA versus general population
- Calculate p-values
- Conclude outlier when p-value sufficiently small



Traditional RNA-seq differential analysis

- Rather than outlier l test is a group-wise compari
  - Normal versus c
  - Lung versus brai
- Requires “replicates el/undiagnosed disease
- Not an individual-le
- This is more akin tc average taller than women and th ently (e.g., t-test or linear models)

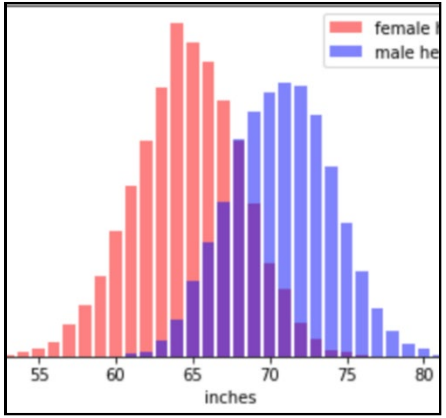


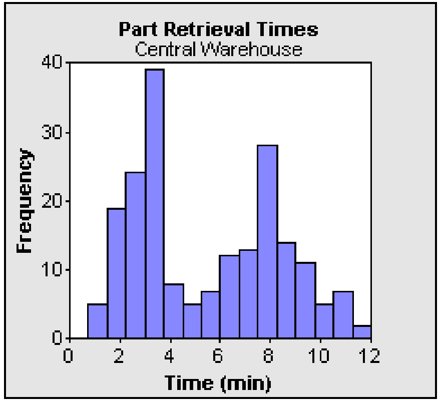
Closer look at the p-values

- Null: Yao Ming's 90 inch height was randomly selected from population of heights
- Suppose we estimate from data the average male is 70 inches and the standard deviation is 4 inches
- P-value for Yao Ming's height asks: "what is the probability that someone 90 inches *or taller* is randomly selected from the population"
  - Here we can calculate this as:
    zscore = (90-70)/4 = 5 → p-value = 0.00001
- Interpretation: it is highly unlikely that I randomly selected Yao Ming by chance from the population of heights
  - This is true. I specifically chose him due to his notoriously large height, and so outlier analysis has correctly identified an anomaly

Not all distributions are Gaussian

- The aforementioned procedure is general, but a good model should match the data generation process
- Normal distribution is good for heights but not the below graph
  - Note in the case of heights the situation was actually similar if we didn't account for sex
  - Height only looked Gaussian for each sex separately
- Z-scores sometimes used instead of p-value; incorrect unless Gaussian

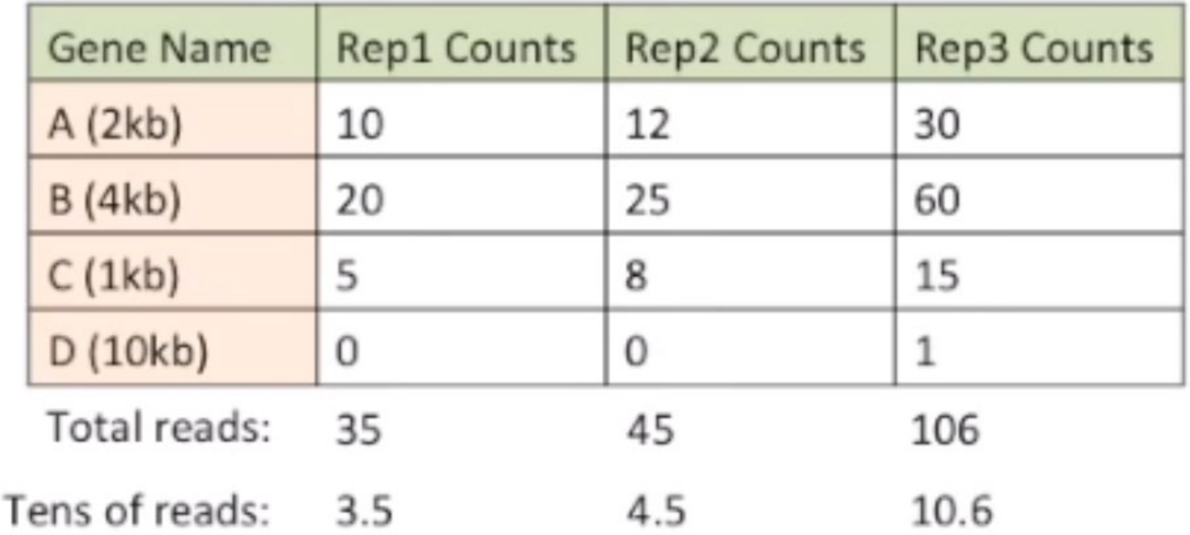




RNA-seq Expression

- RNA extracted from the cells and sequenced
- Each sequencing read can be mapped (not always uniquely) to a given transcript/gene
- We extract counts of reads coming from each gene
- Counts need context and are not useful in isolation
  - Long genes have more RNA-bases per transcript expressed
  - Samples sequenced to higher depths will have more total counts for technical and not biological reasons
- Models for RNA counts need to account for both

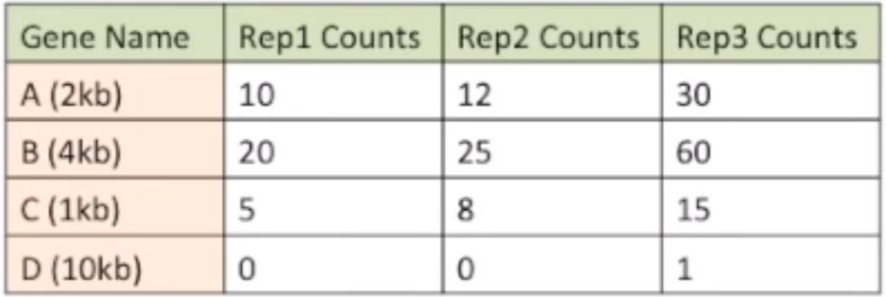
Over-simplified view: just use TPM



Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1
Total reads:	35	45	106
Tens of reads:	3.5	4.5	10.6

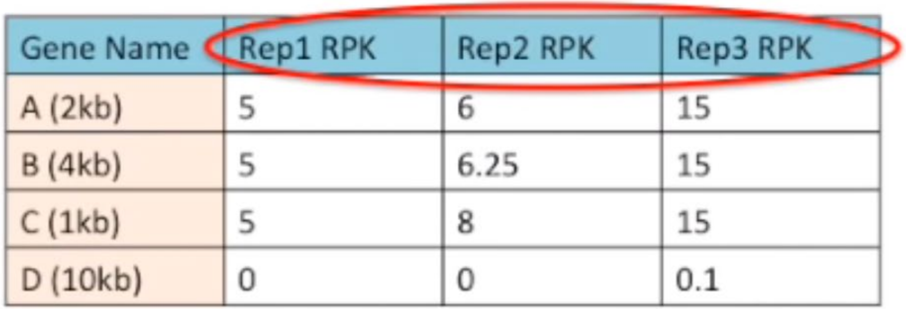
First normalize by gene length

Original data:



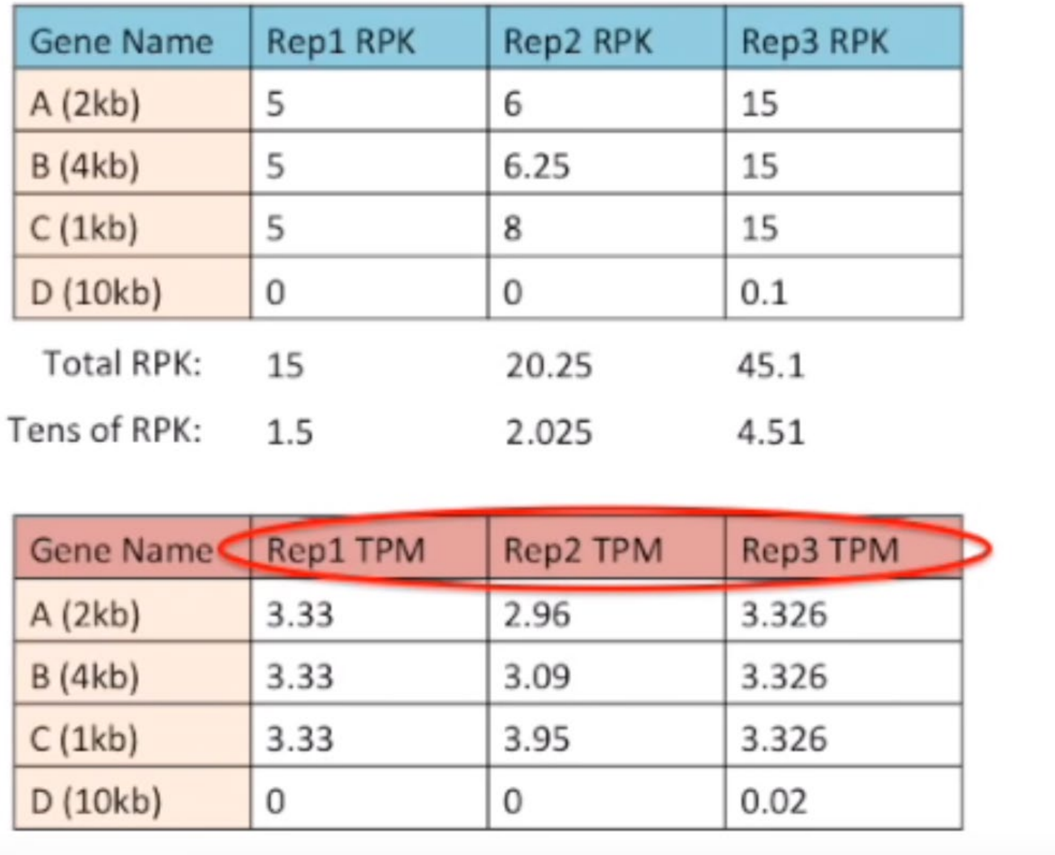
Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPK – scaled by gene length:



Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

Now normalize by sequencing depth



Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

Total RPK:	15	20.25	45.1
Tens of RPK:	1.5	2.025	4.51

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

Build model

- Use a Gaussian distribution as before, but now using TPMs that are comparable across genes and samples
- Most sophisticated RNA-seq analysis avoid this route and model the counts directly (often with negative binomial distribution) while accounting for gene length and sequencing depth but idea is similar
- Once you model your "normal" cohort, just do outlier detection via p-values as before
  - Trick in rare disease is to use patients as "normal" cohort

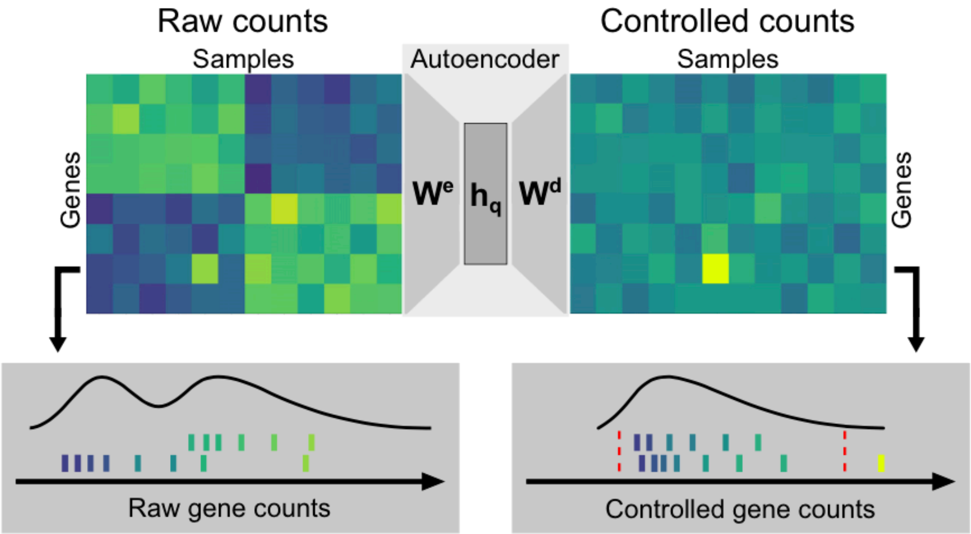
Confounding variables revisited

- Variables that are not of interest that affect your variable of interest
  - sex, age, smoking status, technical artifacts, cell admixture in whole blood, etc
- If your “normal” population was mostly males and you measured an average female’s height you might incorrectly label them an outlier
- One option is directly accounting for them as we did with sex in heights (e.g., PEER package)
  - Pros: intuitive model we control and interpret
  - Cons: requires much more data, and corresponding meta data (i.e., sex labels)

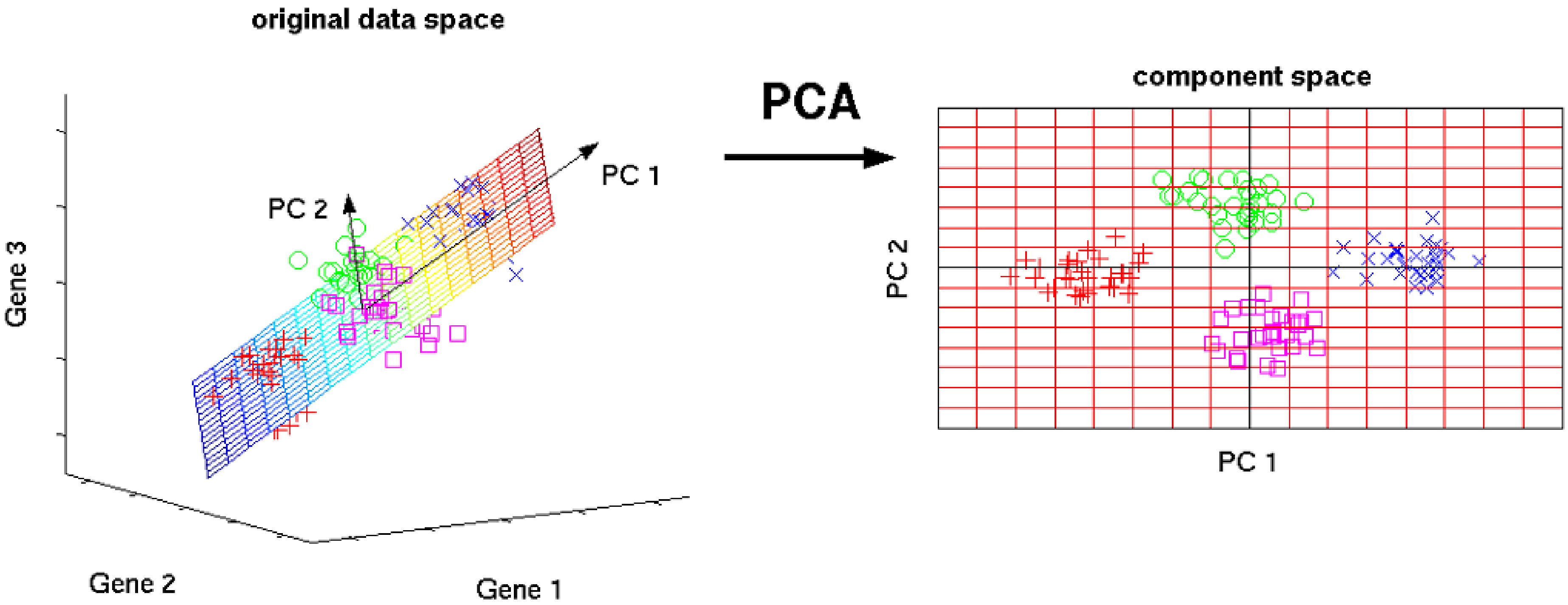
Autoencoder

"Black box" finds relationships between samples and corrects for them automatically

- Pros: no meta data needed, less samples needed, accounts for hidden confounders
- Cons: not easily interpretable output…need to "trust" the black-box



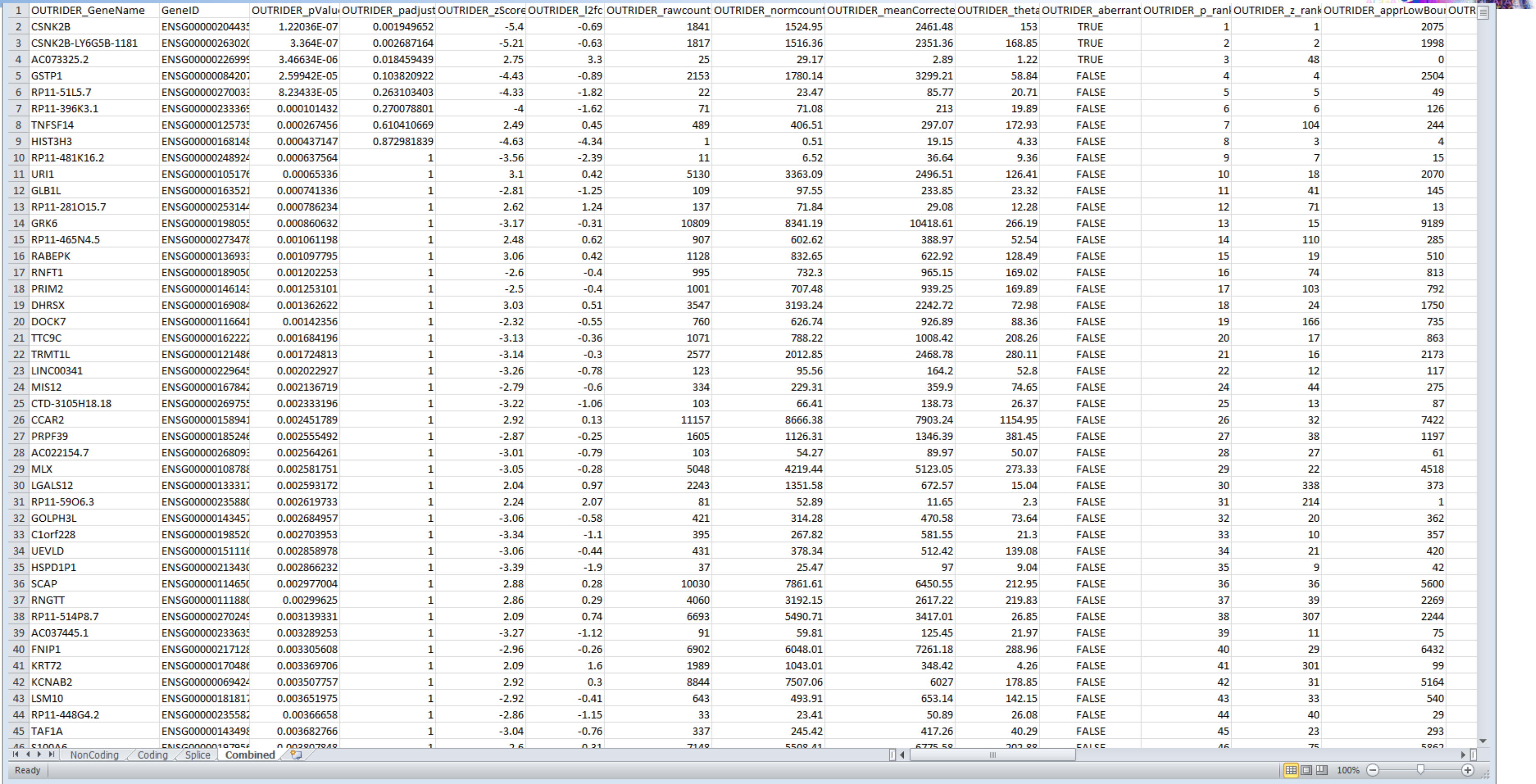
PCA (autoencoder initialization, $W^e=W^d$)



Outlier Case study

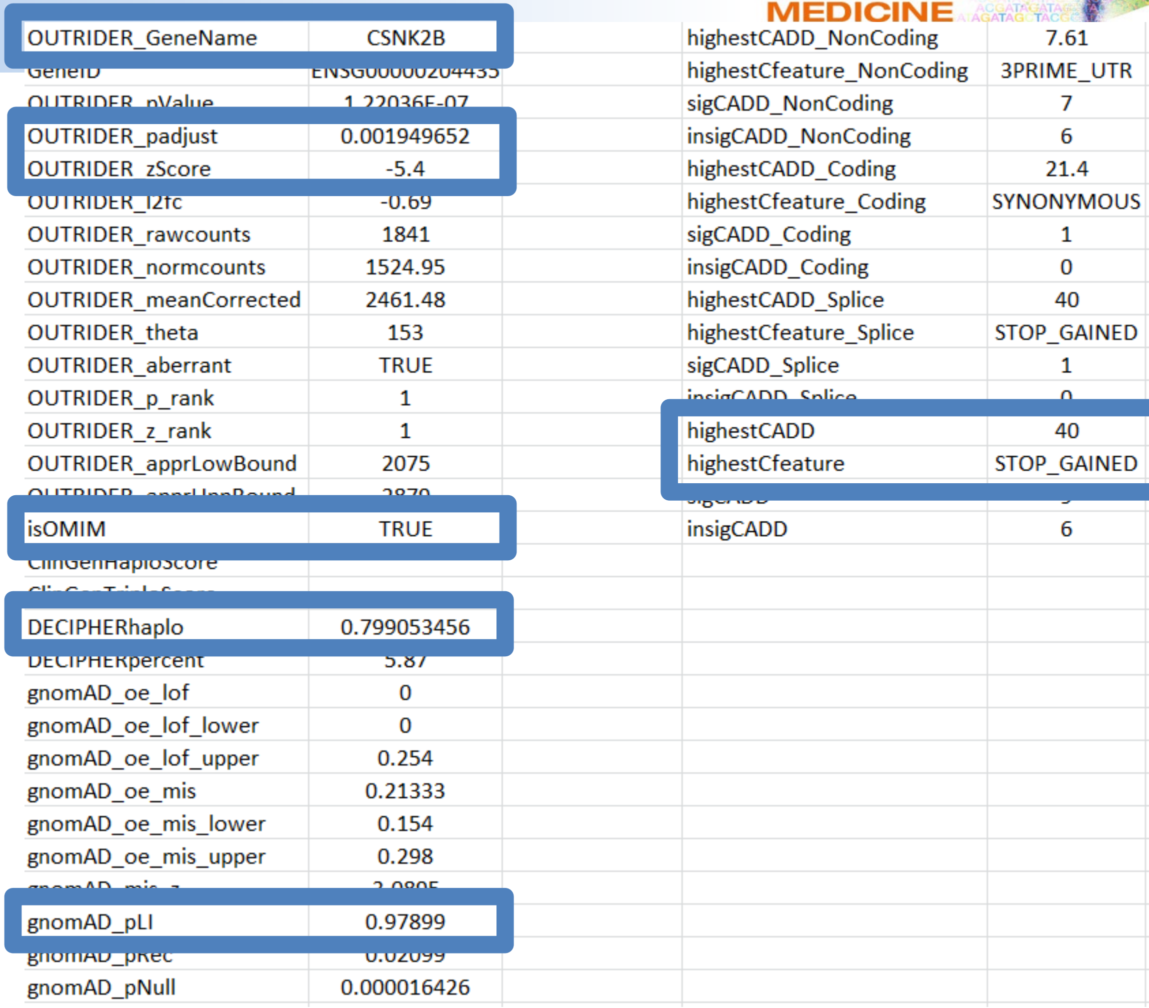
- 9 year old male, non-consanguineous family
- Mild global developmental delay
- General Convulsive Intractable Epilepsy
- Nevus sebaceous
- Constipation due to colonic dysmotility
- Outside diagnosis of mitochondrial disorder

Outlier-centric View



	OUTRIDER_GeneName	GeneID	OUTRIDER_pValu	OUTRIDER_padjust	OUTRIDER_zScore	OUTRIDER_l2fc	OUTRIDER_rawcount	OUTRIDER_normcount	OUTRIDER_meanCorrecte	OUTRIDER_theta	OUTRIDER_aberrant	OUTRIDER_p_rank	OUTRIDER_z_rank	OUTRIDER_apprLowBou	OUTR
2	CSNK2B	ENSG00000204435	1.22036E-07	0.001949652	-5.4	-0.69	1841	1524.95	2461.48	153	TRUE	1	1	2075	
3	CSNK2B-LY6G5B-1181	ENSG00000263020	3.364E-07	0.002687164	-5.21	-0.63	1817	1516.36	2351.36	168.85	TRUE	2	2	1998	
4	AC073325.2	ENSG00000226999	3.46634E-06	0.018459439	2.75	3.3	25	29.17	2.89	1.22	TRUE	3	48	0	
5	GSTP1	ENSG00000084207	2.59942E-05	0.103820922	-4.43	-0.89	2153	1780.14	3299.21	58.84	FALSE	4	4	2504	
6	RP11-51L5.7	ENSG00000270033	8.23433E-05	0.263103403	-4.33	-1.82	22	23.47	85.77	20.71	FALSE	5	5	49	
7	RP11-396K3.1	ENSG00000233369	0.000101432	0.270078801	-4	-1.62	71	71.08	213	19.89	FALSE	6	6	126	
8	TNFSF14	ENSG00000125735	0.000267456	0.610410669	2.49	0.45	489	406.51	297.07	172.93	FALSE	7	104	244	
9	HIST3H3	ENSG00000168148	0.000437147	0.872981839	-4.63	-4.34	1	0.51	19.15	4.33	FALSE	8	3	4	
10	RP11-481K16.2	ENSG00000248924	0.000637564	1	-3.56	-2.39	11	6.52	36.64	9.36	FALSE	9	7	15	
11	URI1	ENSG00000105176	0.00065336	1	3.1	0.42	5130	3363.09	2496.51	126.41	FALSE	10	18	2070	
12	GLB1L	ENSG00000163521	0.000741336	1	-2.81	-1.25	109	97.55	233.85	23.32	FALSE	11	41	145	
13	RP11-281O15.7	ENSG00000253144	0.000786234	1	2.62	1.24	137	71.84	29.08	12.28	FALSE	12	71	13	
14	GRK6	ENSG00000198055	0.000860632	1	-3.17	-0.31	10809	8341.19	10418.61	266.19	FALSE	13	15	9189	
15	RP11-465N4.5	ENSG00000273478	0.001061198	1	2.48	0.62	907	602.62	388.97	52.54	FALSE	14	110	285	
16	RABEPK	ENSG00000136933	0.001097795	1	3.06	0.42	1128	832.65	622.92	128.49	FALSE	15	19	510	
17	RNFT1	ENSG00000189050	0.001202253	1	-2.6	-0.4	995	732.3	965.15	169.02	FALSE	16	74	813	
18	PRIM2	ENSG00000146143	0.001253101	1	-2.5	-0.4	1001	707.48	939.25	169.89	FALSE	17	103	792	
19	DHRSX	ENSG00000169084	0.001362622	1	3.03	0.51	3547	3193.24	2242.72	72.98	FALSE	18	24	1750	
20	DOCK7	ENSG00000116641	0.00142356	1	-2.32	-0.55	760	626.74	926.89	88.36	FALSE	19	166	735	
21	TTC9C	ENSG00000162222	0.001684196	1	-3.13	-0.36	1071	788.22	1008.42	208.26	FALSE	20	17	863	
22	TRMT1L	ENSG00000121486	0.001724813	1	-3.14	-0.3	2577	2012.85	2468.78	280.11	FALSE	21	16	2173	
23	LINC00341	ENSG00000229645	0.002022927	1	-3.26	-0.78	123	95.56	164.2	52.8	FALSE	22	12	117	
24	MIS12	ENSG00000167842	0.002136719	1	-2.79	-0.6	334	229.31	359.9	74.65	FALSE	24	44	275	
25	CTD-3105H18.18	ENSG00000269755	0.002333196	1	-3.22	-1.06	103	66.41	138.73	26.37	FALSE	25	13	87	
26	CCAR2	ENSG00000158941	0.002451789	1	2.92	0.13	11157	8666.38	7903.24	1154.95	FALSE	26	32	7422	
27	PRPF39	ENSG00000185246	0.002555492	1	-2.87	-0.25	1605	1126.31	1346.39	381.45	FALSE	27	38	1197	
28	AC022154.7	ENSG00000268093	0.002564261	1	-3.01	-0.79	103	54.27	89.97	50.07	FALSE	28	27	61	
29	MLX	ENSG00000108788	0.002581751	1	-3.05	-0.28	5048	4219.44	5123.05	273.33	FALSE	29	22	4518	
30	LGALS12	ENSG00000133317	0.002593172	1	2.04	0.97	2243	1351.58	672.57	15.04	FALSE	30	338	373	
31	RP11-59O6.3	ENSG00000235880	0.002619733	1	2.24	2.07	81	52.89	11.65	2.3	FALSE	31	214	1	
32	GOLPH3L	ENSG00000143457	0.002684957	1	-3.06	-0.58	421	314.28	470.58	73.64	FALSE	32	20	362	
33	C1orf228	ENSG00000198520	0.002703953	1	-3.34	-1.1	395	267.82	581.55	21.3	FALSE	33	10	357	
34	UEVLD	ENSG00000151116	0.002858978	1	-3.06	-0.44	431	378.34	512.42	139.08	FALSE	34	21	420	
35	HSPD1P1	ENSG00000213430	0.002866232	1	-3.39	-1.9	37	25.47	97	9.04	FALSE	35	9	42	
36	SCAP	ENSG00000114650	0.002977004	1	2.88	0.28	10030	7861.61	6450.55	212.95	FALSE	36	36	5600	
37	RNGTT	ENSG00000111880	0.00299625	1	2.86	0.29	4060	3192.15	2617.22	219.83	FALSE	37	39	2269	
38	RP11-514P8.7	ENSG00000270249	0.003139331	1	2.09	0.74	6693	5490.71	3417.01	26.85	FALSE	38	307	2244	
39	AC037445.1	ENSG00000233635	0.003289253	1	-3.27	-1.12	91	59.81	125.45	21.97	FALSE	39	11	75	
40	FNIP1	ENSG00000217128	0.003305608	1	-2.96	-0.26	6902	6048.01	7261.18	288.96	FALSE	40	29	6432	
41	KRT72	ENSG00000170486	0.003369706	1	2.09	1.6	1989	1043.01	348.42	4.26	FALSE	41	301	99	
42	KCNAB2	ENSG00000069424	0.003507757	1	2.92	0.3	8844	7507.06	6027	178.85	FALSE	42	31	5164	
43	LSM10	ENSG00000181817	0.003651975	1	-2.92	-0.41	643	493.91	653.14	142.15	FALSE	43	33	540	
44	RP11-448G4.2	ENSG00000235582	0.00366658	1	-2.86	-1.15	33	23.41	50.89	26.08	FALSE	44	40	29	
45	TAF1A	ENSG00000143498	0.003682766	1	-3.04	-0.76	337	245.42	417.26	40.29	FALSE	45	23	293	
46	S100A6	ENSG00000197956	0.003807848	1	2.6	0.31	7148	5508.41	6775.58	202.88	FALSE	46	75	5862	

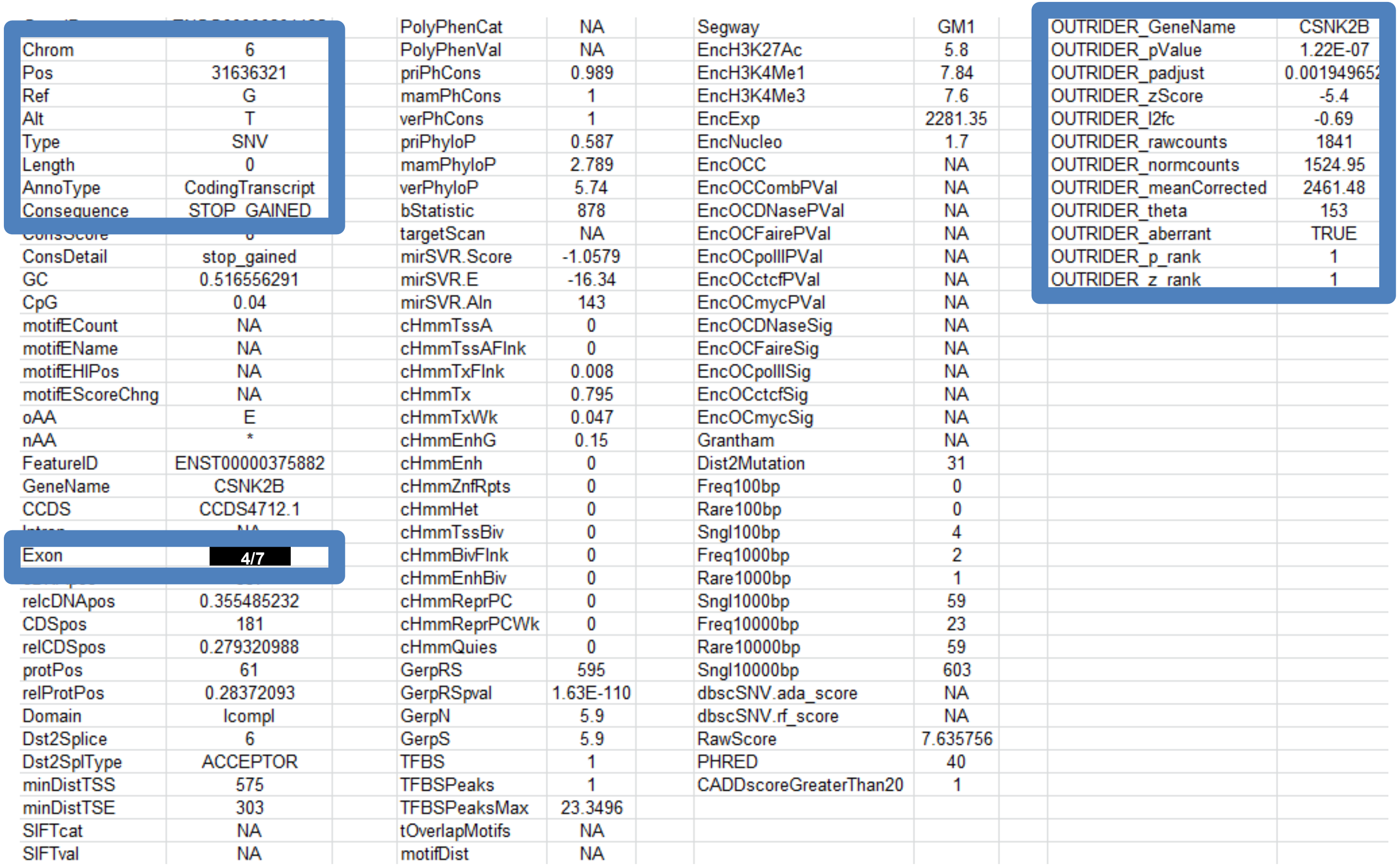
Outlier-centric View (top hit)



OUTRIDER_GeneName	CSNK2B
GeneID	ENSG00000204435
OUTRIDER_pValue	1.22036E-07
OUTRIDER_padjust	0.001949652
OUTRIDER_zScore	-5.4
OUTRIDER_l2fc	-0.69
OUTRIDER_rawcounts	1841
OUTRIDER_normcounts	1524.95
OUTRIDER_meanCorrected	2461.48
OUTRIDER_theta	153
OUTRIDER_aberrant	TRUE
OUTRIDER_p_rank	1
OUTRIDER_z_rank	1
OUTRIDER_apprLowBound	2075
OUTRIDER_apprUppBound	2870
isOMIM	TRUE
ClinGenHaploScore	
ClinGenTriploScore	
DECIPHERhaplo	0.799053456
DECIPHERpercent	5.87
gnomAD_oe_lof	0
gnomAD_oe_lof_lower	0
gnomAD_oe_lof_upper	0.254
gnomAD_oe_mis	0.21333
gnomAD_oe_mis_lower	0.154
gnomAD_oe_mis_upper	0.298
gnomAD_mis_z	3.0805
gnomAD_pLI	0.97899
gnomAD_pRec	0.02099
gnomAD_pNull	0.000016426

highestCADD_NonCoding	7.61
highestCfeature_NonCoding	3PRIME_UTR
sigCADD_NonCoding	7
insigCADD_NonCoding	6
highestCADD_Coding	21.4
highestCfeature_Coding	SYNONYMOUS
sigCADD_Coding	1
insigCADD_Coding	0
highestCADD_Splice	40
highestCfeature_Splice	STOP_GAINED
sigCADD_Splice	1
insigCADD_Splice	0
highestCADD	40
highestCfeature	STOP_GAINED
sigCADD	9
insigCADD	6

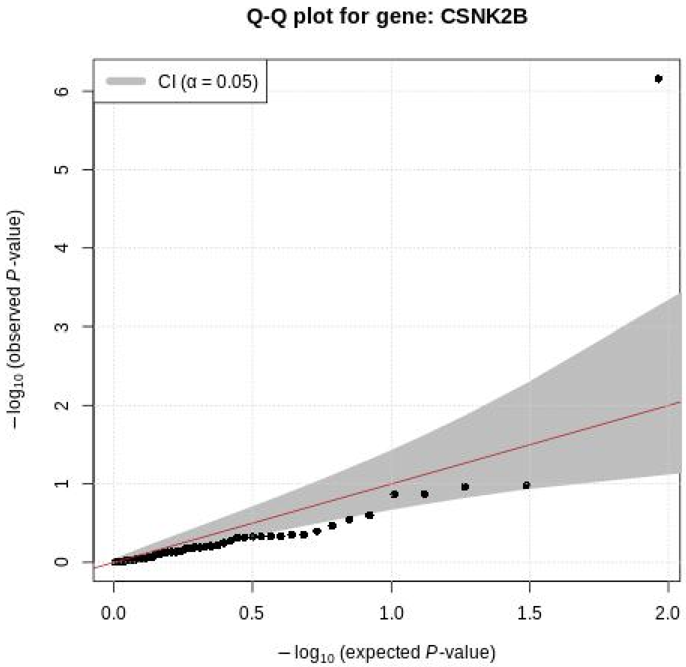
Variant-centric View



[illegible]	[illegible]	PolyPhenCat	NA	Segway	GM1	OUTRIDER_GeneName	CSNK2B
Chrom	6	PolyPhenVal	NA	EncH3K27Ac	5.8	OUTRIDER_pValue	1.22E-07
Pos	31636321	priPhCons	0.989	EncH3K4Me1	7.84	OUTRIDER_padjust	0.001949652
Ref	G	mamPhCons	1	EncH3K4Me3	7.6	OUTRIDER_zScore	-5.4
Alt	T	verPhCons	1	EncExp	2281.35	OUTRIDER_l2fc	-0.69
Type	SNV	priPhyloP	0.587	EncNucleo	1.7	OUTRIDER_rawcounts	1841
Length	0	mamPhyloP	2.789	EncOCC	NA	OUTRIDER_normcounts	1524.95
AnnoType	CodingTranscript	verPhyloP	5.74	EncOCCombPVal	NA	OUTRIDER_meanCorrected	2461.48
Consequence	STOP_GAINED	bStatistic	878	EncOCDNasePVal	NA	OUTRIDER_theta	153
ConsScore	[illegible]	targetScan	NA	EncOCFairePVal	NA	OUTRIDER_aberrant	TRUE
ConsDetail	stop_gained	mirSVR.Score	-1.0579	EncOCpolIIPVal	NA	OUTRIDER_p_rank	1
GC	0.516556291	mirSVR.E	-16.34	EncOCctcfPVal	NA	OUTRIDER_z_rank	1
CpG	0.04	mirSVR.Aln	143	EncOCmycPVal	NA		
motifECount	NA	cHmmTssA	0	EncOCDNaseSig	NA		
motifEName	NA	cHmmTssAFlnk	0	EncOCFaireSig	NA		
motifEHIPos	NA	cHmmTxFlnk	0.008	EncOCpolIISig	NA		
motifEScoreChng	NA	cHmmTx	0.795	EncOCctcfSig	NA		
oAA	E	cHmmTxWk	0.047	EncOCmycSig	NA		
nAA	*	cHmmEnhG	0.15	Grantham	NA		
FeatureID	ENST00000375882	cHmmEnh	0	Dist2Mutation	31		
GeneName	CSNK2B	cHmmZnfRpts	0	Freq100bp	0		
CCDS	CCDS4712.1	cHmmHet	0	Rare100bp	0		
Intron	NA	cHmmTssBiv	0	Sngl100bp	4		
Exon	4/7	cHmmBivFlnk	0	Freq1000bp	2		
cDNApos	[illegible]	cHmmEnhBiv	0	Rare1000bp	1		
relcDNApos	0.355485232	cHmmReprPC	0	Sngl1000bp	59		
CDSpos	181	cHmmReprPCWk	0	Freq10000bp	23		
relCDSpos	0.279320988	cHmmQuies	0	Rare10000bp	59		
protPos	61	GerpRS	595	Sngl10000bp	603		
relProtPos	0.28372093	GerpRSpval	1.63E-110	dbscSNV.ada_score	NA		
Domain	lcompl	GerpN	5.9	dbscSNV.rf_score	NA		
Dst2Splice	6	GerpS	5.9	RawScore	7.635756		
Dst2SplType	ACCEPTOR	TFBS	1	PHRED	40		
minDistTSS	575	TFBSPeaks	1	CADDscoreGreaterThan20	1		
minDistTSE	303	TFBSPeaksMax	23.3496				
SIFTcat	NA	tOverlapMotifs	NA				
SIFTval	NA	motifDist	NA				

Case summary

- OUTRIDER:
  - p-value = 1E-7
  - zScore = -5.4
  - l2fc = -0.69
- Gene's top CADD variant is case solving variant
- Nonsense variant p.Glu61*, consistent with the near complete allelic loss of expression



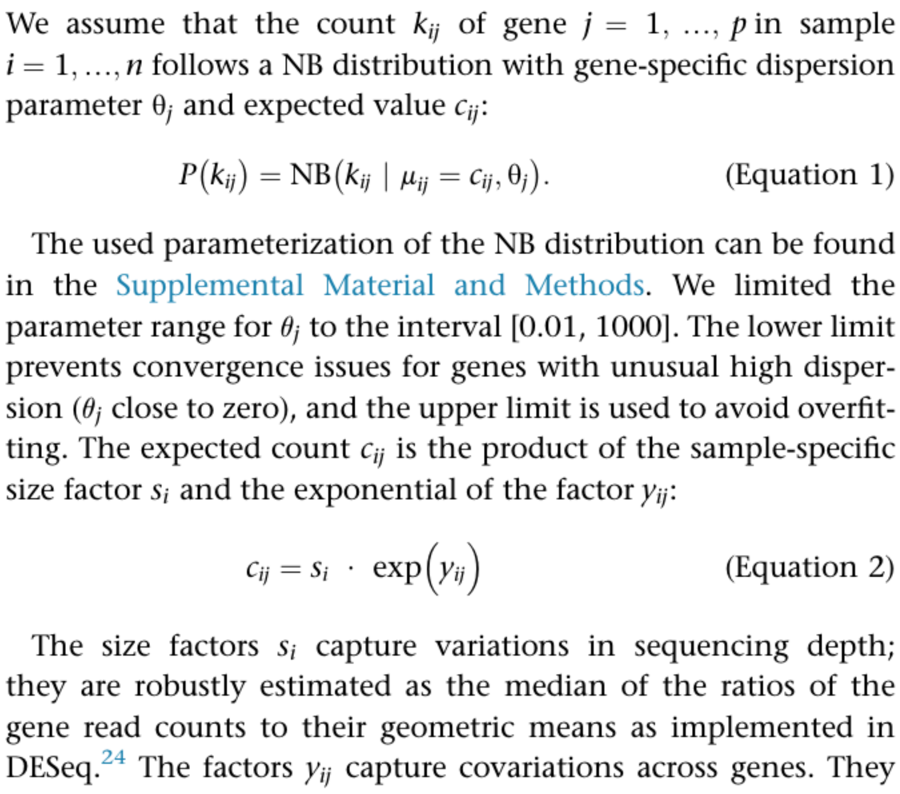
Questions?

Slides curtesy of Garrett Jenkinson

MAYO CLINIC

OUTRIDER

CENTER FOR INDIVIDUALIZED MEDICINE



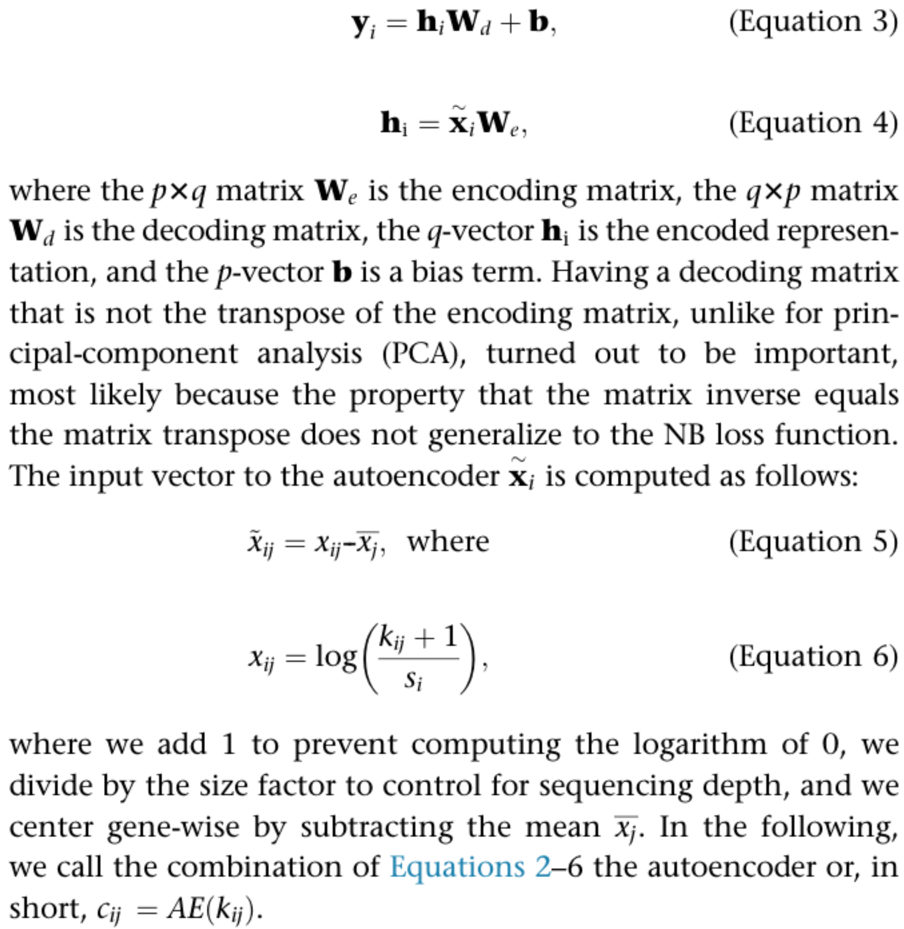
We assume that the count k_{ij} of gene $j = 1, \ldots, p$ in sample $i = 1, \ldots, n$ follows a NB distribution with gene-specific dispersion parameter θ_j and expected value c_{ij}:

$$P(k_{ij}) = \mathrm{NB}(k_{ij} \mid \mu_{ij} = c_{ij}, \theta_j). \qquad \text{(Equation 1)}$$

The used parameterization of the NB distribution can be found in the Supplemental Material and Methods. We limited the parameter range for θ_j to the interval [0.01, 1000]. The lower limit prevents convergence issues for genes with unusual high dispersion (θ_j close to zero), and the upper limit is used to avoid overfitting. The expected count c_{ij} is the product of the sample-specific size factor s_i and the exponential of the factor y_{ij}:

$$c_{ij} = s_i \cdot \exp(y_{ij}) \qquad \text{(Equation 2)}$$

The size factors s_i capture variations in sequencing depth; they are robustly estimated as the median of the ratios of the gene read counts to their geometric means as implemented in DESeq.[24] The factors y_{ij} capture covariations across genes. They



$$\mathbf{y}_i = \mathbf{h}_i \mathbf{W}_d + \mathbf{b}, \qquad \text{(Equation 3)}$$

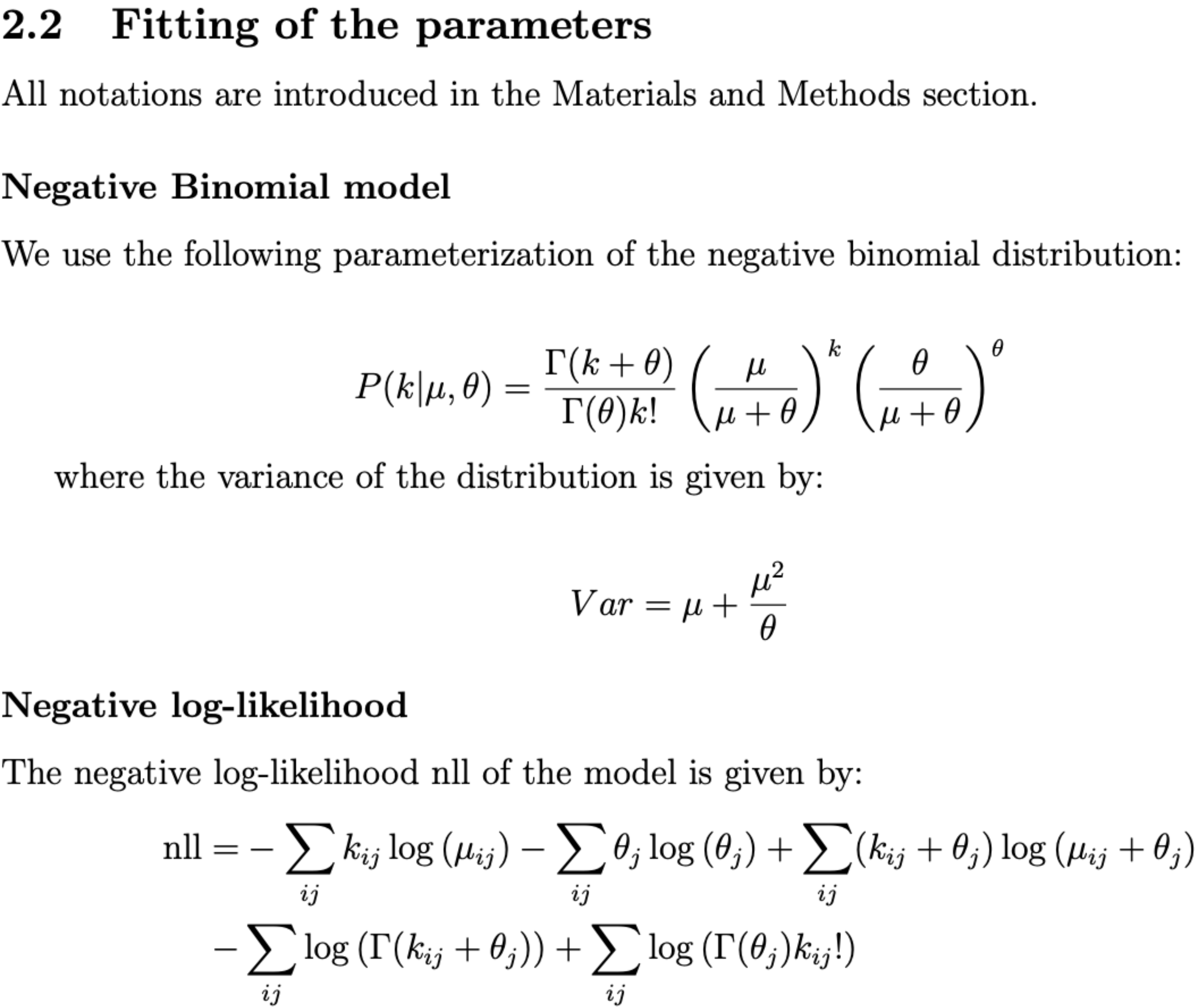
$$\mathbf{h}_{\mathrm{i}} = \tilde{\mathbf{x}}_i \mathbf{W}_e, \qquad \text{(Equation 4)}$$

where the $p \times q$ matrix $\mathbf{W}_e$ is the encoding matrix, the $q \times p$ matrix $\mathbf{W}_d$ is the decoding matrix, the q-vector $\mathbf{h}_{\mathrm{i}}$ is the encoded representation, and the p-vector $\mathbf{b}$ is a bias term. Having a decoding matrix that is not the transpose of the encoding matrix, unlike for principal-component analysis (PCA), turned out to be important, most likely because the property that the matrix inverse equals the matrix transpose does not generalize to the NB loss function. The input vector to the autoencoder $\tilde{\mathbf{x}}_i$ is computed as follows:

$$\tilde{x}_{ij} = x_{ij} - \overline{x}_j, \quad \text{where} \qquad \text{(Equation 5)}$$

$$x_{ij} = \log\left(\frac{k_{ij} + 1}{s_i}\right), \qquad \text{(Equation 6)}$$

where we add 1 to prevent computing the logarithm of 0, we divide by the size factor to control for sequencing depth, and we center gene-wise by subtracting the mean $\overline{x}_j$. In the following, we call the combination of Equations 2–6 the autoencoder or, in short, $c_{ij} = AE(k_{ij})$.



2.2 Fitting of the parameters

All notations are introduced in the Materials and Methods section.

Negative Binomial model

We use the following parameterization of the negative binomial distribution:

$$P(k|\mu,\theta) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)k!}\left(\frac{\mu}{\mu+\theta}\right)^{k}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}$$

where the variance of the distribution is given by:

$$Var = \mu + \frac{\mu^2}{\theta}$$

Negative log-likelihood

The negative log-likelihood nll of the model is given by:

$$\begin{aligned}\text{nll} = &-\sum_{ij} k_{ij}\log(\mu_{ij}) - \sum_{ij}\theta_j \log(\theta_j) + \sum_{ij}(k_{ij}+\theta_j)\log(\mu_{ij}+\theta_j)\\ &-\sum_{ij}\log(\Gamma(k_{ij}+\theta_j)) + \sum_{ij}\log(\Gamma(\theta_j)k_{ij}!)\end{aligned}$$

For the optimization of the model only the first and third term of the nll need to be considered, as all other terms are independent of $\mathbf{W}_e$ and $\mathbf{W}_d$, yielding the following truncated form of the negative log likelihood:

$$\mathrm{nll}_{\mathbf{W}} = -\sum_{ij} \left[k_{ij} \log\left(\mu_{ij}\right) - \left(k_{ij} + \theta_j\right) \log\left(\mu_{ij} + \theta_j\right)\right] \tag{1}$$

We use L-BFGS to fit the autoencoder model as described in Methods. We implemented the following gradients.

The expectations μ_{ij} are modeled by:

$$\mu_{ij} = s_i e^{y_{ij}}$$

Hence, $\mathrm{nll}_{\mathbf{W}}$ can be rewritten as:

$$\mathrm{nll}_{\mathbf{W}} = -\sum_{ij} \left[k_{ij} \log(s_i) + y_{ij} - (k_{ij} + \theta_j) \cdot \left(\log(s_i) + y_{ij} + \log\left(1 + \frac{\theta_j}{s_i \cdot e^{y_{ij}}}\right)\right)\right]$$

In the following the y_{ij} are the elements of the $\mathbf{Y}$ defined as:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}_e\mathbf{W}_d^T + \mathbf{b}, \tag{2}$$

where the element (i, j) of the matrix $\mathbf{X}$ is given by: $\log\left(\frac{k_{ij}+1}{s_i}\right) - \bar{x}_j$.