1 Slide 2

I want to start with the fundamental issue of what scientific questions are we asking of our sequence count data? What do we want to learn about our data?

2 Slide 3

The previous presentation focused on questions about differential expression/abundance These are questions of single genes/taxa for instance asking is the TNFα gene upregulated in tumor tissue or is the s pyogenes bacterium killed by an antibiotic Methods for DE/DA include ALDEx2 or DESeq2

3 Slide 4

[Read question] For instance, rather than ask if the TNFα gene is up/down regulated we want zoom out and know if an entire pathway represeninting a higher level biological process is changing

4 Slide 5

These higher level biological process questions are a part of what is termed differential set analysis This analysis asks the higher level questions like ... [read questions] Methods for DSA include ... [read]

5 Slide 6

We will follow previous presentations in focusing on the impact of scale on DSA methods Here I review the problem of scale in the analysis of sequence count data Consider an example experiment... [read] The problem of scale we want to analyze absolute abundances (DxN matrix W) But we can only know the absolute abundances if we know the composition (DxN matrix $W^{\|}$) and the scale (vector of length N, $W^{\perp}$) But we only measure the composition, meaning the scale and absolute abundances are known.

Mathematically the relationship can be represented as $W_{dn} = W^{\|}_{dn} \times W^{\perp}_{n}$ where...

6 Slide 7

CRITICAL: In the previous presentation we focused on LFCs, here we will again focus on LFCs as it is relevant for DSA The log fold change is represented as theta which is a diff in means ...

7 Slide 8

At the top here I rewritten the formula for LFC However we can also use the relationship that the absolute abundances W are a product of the composition and scale to decompose the lfc into its compositional and scale components. That is the LFC can be rewritten as ... TRANSITION: On the next slide I will restate this relationship again.

8 Slide 9

At the top we have the relationship between theta_{d} ...

We can represent the LFC and its composition/scale decomposition in vector form as ...

9 Slide 10

CRITICAL: I want to pause here to say that, to investigate DSA, we will need to distinguish between the True LFC and our estimate of the LFC.

Methods like DESeq2... [read]

CRITICAL: But we call the estimated LFC in the scale a scale assumption...

next part, read "The only way we cab estimate something that is unmeasured is by making an assumption, normalization make assumptions"

10 Slide 11

Here I show two scale assumption, [CRITICAL: mention ALDEx2 uses CLR and you've seen this in last presentation]

11 Slide 12

Our goal is to understand error in scale assumptions. Mathematically, this error can be expressed using the following formulation I want to ignore possible errors in the composition to focus only on errors in scale We can then say that the true LFC is equal to our estimate of the LFC plus error in our scale assumption Transition: I'll restate this formula again on the next slide

12 Slide 13

[Restate relationship]

I term a sensitivity analysis of error $\epsilon^{\perp}$ LFC Sensitivity analysis The idea of LFC sensitivity analysis is to understand how a method changes as a function of error in our estimate of the LFC in scale

13 Slide 14

14 Slide 15

[read]

15 Slide 16

Read, sets are biologically related genes/microbes, e.g., the apoptosis pathway

Read

16 Slide 17

Here I review the GSEA algorithm with gene label permutations

I will start by calculating the test statistic.

Step 1: Estimate LFCs e.g., with ALDEx2 and rank the LFCs in descending order

The bottom left plot shows 20 genes. The y-axis represents the LFC. The x's represent the 5 genes in the gene set of interest The Os represent the other genes.

The next step is to calculate a weighted running sum

This running sum is calculated by going through the genes in ranked order. The running sum increases by an amount proportional to the running sum if the gene is in the set CRITICAL: LFCs farther from zero increase the running sum more than genes close to 0.

The running sum decreases by a standard anout if the gene is not in the set.

The bottom right plot shows the running sum corresponding to the left plot. Notice there are 5 places where the running sum increases corresponding to the 5 genes in the gene set. The increases in the running sum differ because some genes' LFCs are farther away from zero than others

Finally the [test statistic] is the enrichment score, which is the max deviation from the running sum from zero this is indicated by the dotted line in the right plot.

17 Slide 18

After calculating the enrichment score we form a null distribution of enrichment scores through a permutation test

Critical: Here the permutation test is performed by permuting gene set labels. The first step of the test is to calculate the enrichment score as described.

Here I show a table representing 9 genes and a set of 4 genes. [this example is different from the previous example only because 20 genes are difficult to show in a table]

The gene set S is the 4 genes RB1, APP, AR and HTT and is shown in the second row of the table.

The permutation test involves randomly selecting gene sets of the same size, and recalculating the enrichment score This is shown in the final 4 rows.

The p-value is then the proportion of enrichment scores less than the observed, unpermuted enrichment score

18 Slide 19

[READ]

[READ]

[READ] - final part, this is an LFC sensitivity analysis, we look at how the results of GSEA change as a function of error in our scale assumption

19 Slide 20

here is an LFC sensitivity analysis using simualted data

The top figure shows the LFC for 60 genes in black, the x's again represent gebes in the set and the Os genes not in the set The blue and gold xs and os represent the same genes, but with scale error added.

The bottom left figure shows the running sums, enrichment scores, p-values, and ϕ_S for each of the three values of ϵ Notice that only when we have scale error of 1, is the gene set enriched. Clearly scale error impacts the running sum and therefore the enrichment score

The bottom right plot shows the results of 1000s of simulations of different gene sets in this example. The PPV is the positive predictive value, the PPV represents the proportion of positives that are true positives. Notice when there is no scale error, meaning epsilonˆ⊥=0, that the ppv is 100\low as about 15\

20 Slide 21

I will present another GSEA LFC Sentivity Analysis, only this time with real data. I will use an experiment comparing gene pathways in healthy and normal adjacent to tumor tissue

21 Slide 22

Here is a visualization of the difference between healthy and normal adjacent to tumor tissue

Normal adjacent to tumor, or NAT tissue, is tissue that is visually healthy but is located near a tumor.

People use NAT tissue as a proxy for healthy tissue in cancer research. the goal of this experiment is to understand if NAT and healthy cells are really similar and if NAT tissue is a good proxy for healthy tissue in cancer research

22 Slide 23

Here are the results from the aran et al paper. They compared healthy and nat tissue in 7 tissue types representing the rows we will be focusing on the last row represeting thyroid tissue.

The x-axis represents a variety of gene sets representing core cellular functions.

THe red represents pathway enrochment in NAT, and the blue represents enrichment in healthy cells.

NAt and healthy tissue mainly seem to differ in that various immune-system related pathways are enriched in NAT.

23 Slide 24

Now I will reanalyze the results reported by aran et al for thyroid tissue.

Here I focus on two pathways. I will perform GSEA using an implementation of the GSEA algorithm called fgsea.

Here I show the results of fgsea for Myogenesis and inflamatory response. This function takes as input the LFCs and a list of pathways

[NES is what is called the normalized enrichment score] THe FDR adjusted p-value is also shown

Notice that both pathways are enriched, myogenesis is enriched in healthy tissue, while inflammatory repsonse is enriched in NAT tissue

However we need to consider error in our scale assumption.

I wrote a wrapper around the fgsea package which re-runs fgsea for different amounts of scale error.

[Here I chose -0.4, 0, and 0.4 ± 0.5 corresponds to the scale being 1.5 times lower or higher than originally assumed]

Notice that the myogenesis pathway is still significant for all errors. However, at $\epsilon = -0.4$ the infl response pathway is no longer enriched. LFC Sensitivity analysis suggests the results of this pathway are suspect, if scale error could be as large as a 1.5 increase/decrease

24 Slide 25

Here is a full LFC sensitivity analysis of thyroid NAT vs healthy tissue

The y axis shows different amounts of scale assumption error. The values in the y-axis, without going into detail, were chosed such that we are allowing up to a 0.9 log fold, meaing 2.5 fold increase or decrease in scale.

The X axis shows different gene sets.

The blue represents enrichment in healthy tissue, the red represents enrichment in NAT tissue, and the white represents no enrichment.

Gene sets show remarkably different degrees of senstivity of scale assumption error.

Myogensis is significant at all errors tested, while some immune response pathways change significance with only small amounts of scale error. Without LFC sensitivity analsyis, researchers currently only see the row represeting epilonˆ$\perp = 0$.

25 Slide 26

So far I have discussed GSEA with gene label permutations, but as we will see, gene label permutations do not account for gene expression correlation structure.

26 Slide 27

The question at hand is whether random gene sets make for an appropriate null distribution

Here I'm shoinf the random gene set permutation shema of GSEA

27 Slide 28

The core issue [READ]

28 Slide 29

Wu et al performed a simulation to demonstrate the issue of correlation.

[read]

up to 40\

29 Slide 30

Here I will present two methods that address correlation structure.

The first is gsea with sample-label permutations. Here in this slide I will compare gsea with gene vs sample label permutations. With gene label permutations first lfcs are estiamted, then the gene set S is permuted to get enrichment scores

With sample label permutations, first we estimate W, the matrix of absolute abundances. then we shuffle the sample labels , for instance, we shuffle whether each sample is labeled as coming from healthy or NAT tissue.

Finally, we re-estimate the LFCs over and over again for each permutation to get a null distribution of enrichment scores

CRITICAL: This accounts for correlation as it never changes the gene set and permutations are argued to maintain correlation structure.

30 Slide 31

The second method is different from GSEA. CAMERA is a method based on a two sample t-test

the two samples are the LFCs of the genes in and not in the gene sets represented by ...

The two-sample t-test statistic is adjusted by the average correaltion. here the average correlation is shown in red.

most ciritcally, the average correlation is a function of our estimate of the log absolute abundances

31 Slide 32

[read]

however gsea/camera take as input ... [read]

Critical: but unlike LFCs, here the scale term is a vector

32 Slide 33

Before GSEA with gene label permutations could be formulated as a function of the LFCs and error in scale could be represented by a single value epsilon$\perp$.

However for GSEA w/ slp and CAMERA we must represent these as a function of the log matrix of absolute abundances, However, this means we must consider error in our scale assumption made for each sample. We term this a scale sensitivity analysis.

for instance, in an experiment with 100 samples we make 100 scale assumptions and thus must consider a sensitivity analysis over a vector of 100 possible errors in these assumptions.

33 Slide 34

Here I will present two scale sensitivity anlyses, I will be using a real data set comparing healthy and tumor breast tissue.

I will consider two types of error. The first is constant error. For constant error, I consider that for healthy tissue there is no error, and for tumor samples there is constant error $\delta^{\perp}$.

For Sample-specific error, I will consider errors that vary sample by sample. However I will assume that the error lies within some bound -gamma to gamma.

34 Slide 35

Here are the results of the scale sensitivity analysis.

[review results]

35 Slide 36