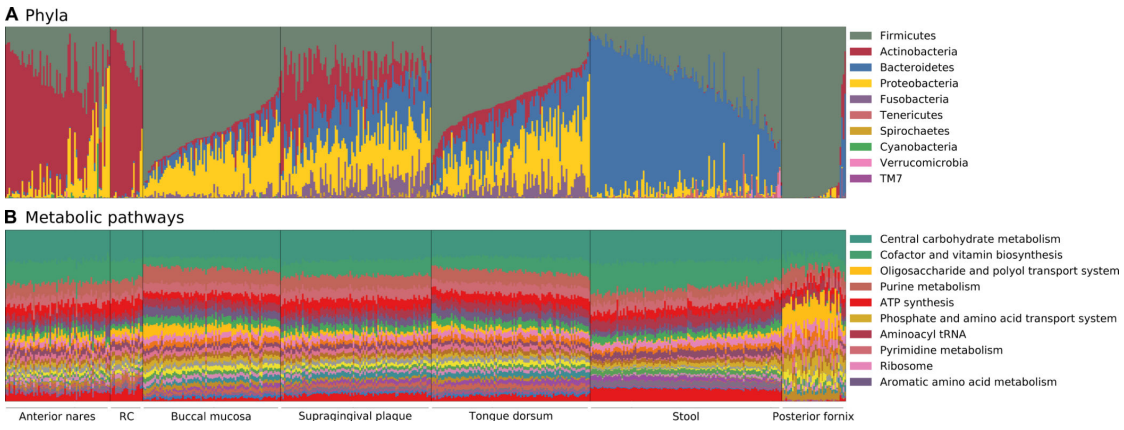
Lecture 8: Predicting and analyzing metagenomic composition from 16S survey data

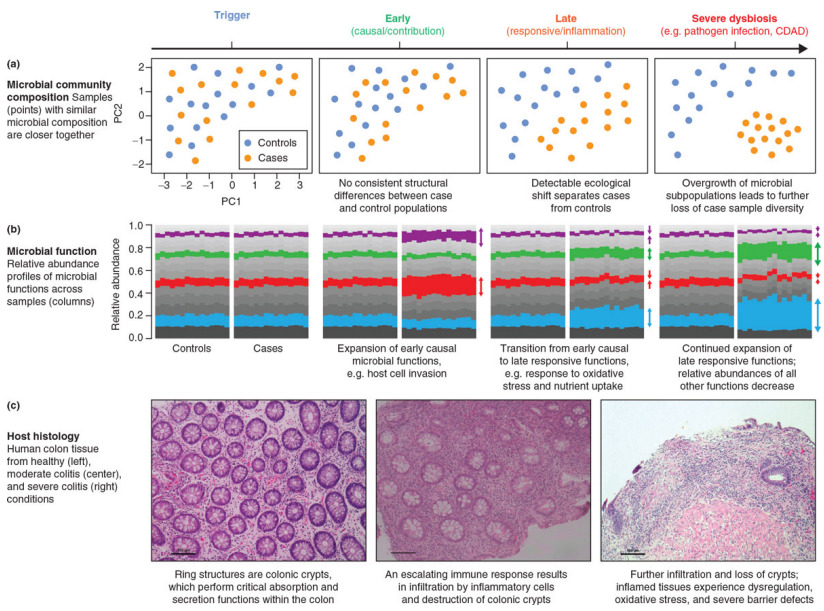
What can we tell about the taxonomic and functional stability of microbiota? Why?



Nature. 2012; 486(7402): 207–214. doi:10.1038/nature11234

2

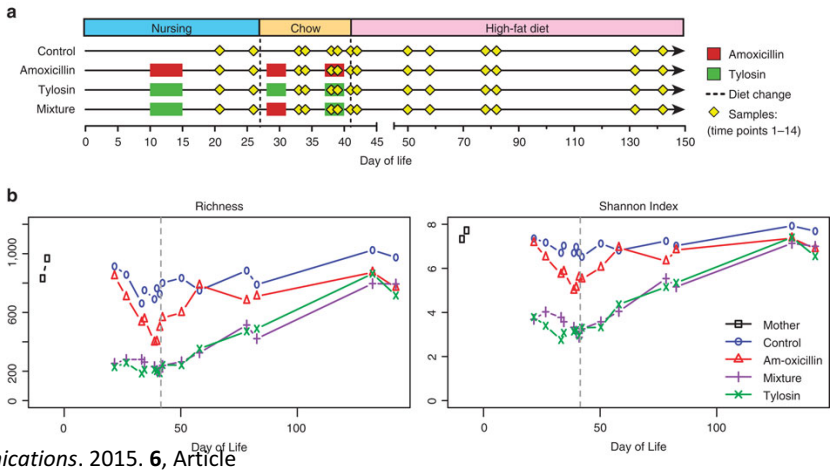
A model of functional dysbiosis in the human gut microbiome during initiation and progression of complex disease.



Genome Medicine, 2013; **5**:65
DOI: 10.1186/gm469

3

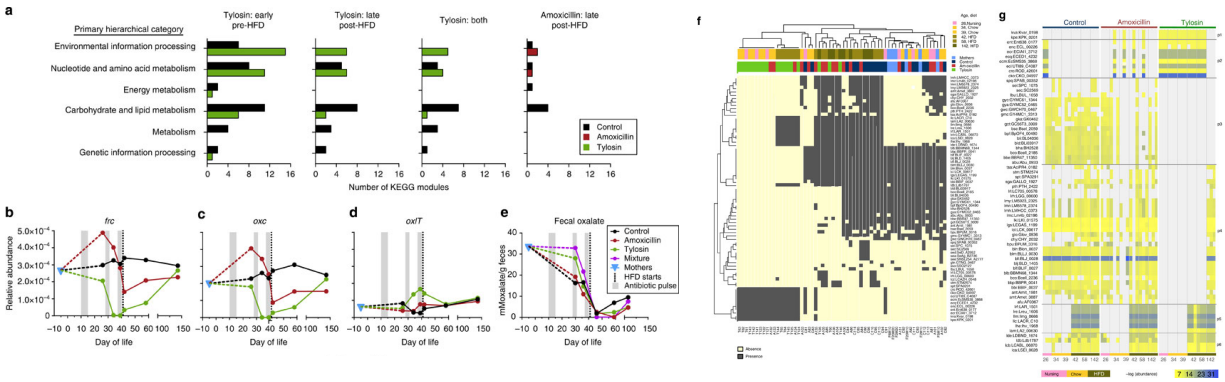
Functional differentiation in pulsed antibiotic treatment



Nature Communications. 2015. **6**, Article number: 7486. doi:10.1038/ncomms8486

4

Functional differentiation in pulsed antibiotic treatment



5

How to measure metagenomes?

- Unlike 16S rRNA gene sequencing, metagenomic sequencing is not targeted to a specific gene, but does an unbiased sample of the entire (bacterial) genomic DNA in a specimen.
- Typically shorter sequence reads are used to obtain >5Gb of data per sample.
- HiSeq instruments are typically more cost effective for metagenomic sequencing.
- This approach is also called shotgun whole metagenome sequencing or WmGS or WGS.

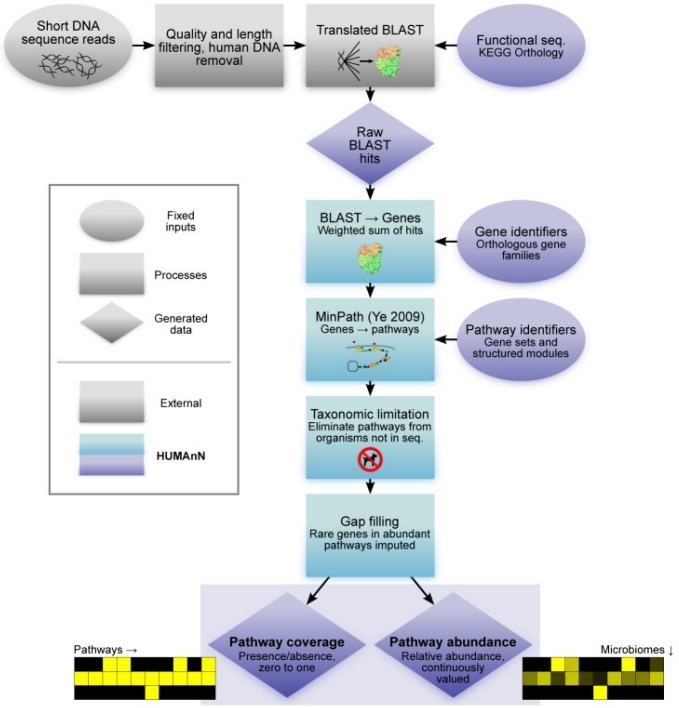
6

Metagenomic processing pipelines

- MG-RAST: *BMC Bioinformatics*, 2008; **9**:38. DOI**:** 10.1186/1471-2105-9-386
- SUPER-FOCUS: Bioinformatics, 2016; 32 (3): 354-361. DOI: 10.1093/bioinformatics/btv584
- HUMAnN: *PLoS Comput Biol*, 2012; 8(6):e1002358. DOI: 10.1371/journal.pcbi.1002358

7

Example pipeline: HUMAnN



8

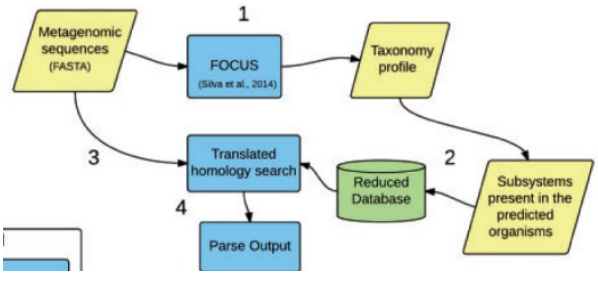
How is metagenome data representation different from 16S rRNA gene sequencing data?

- Taxonomic information is typically discarded, only functional data remain.
- The most fundamental unit of analysis is an individual gene, or orthologous group of genes.
- Genes may be grouped by pathways, systems, diseases, etc.
- Abundance or presence/absence of genes, pathways, etc. is captured in the data matrix.

9

Predicting metagenomes

- Metagenomic sequencing is considerably more expensive.
- The informatics processing is much more complicated.
- In the end, we rely on reference databases of known genes; no true de novo functional information is discovered.
- Some pipelines (e.g. SUPER FOCUS) require taxonomic information.



10

Idea: We can predict metagenomes from the 16S rRNA gene bacterial identification data

- 16S rRNA gene sequencing allows for indentification of microbiota.
- If we know the organism, we may have gene content of that organism or a related organism.
- We can use the information to infer the metagenomic content and use the abundances to reconstruct the metagenomic abundances.

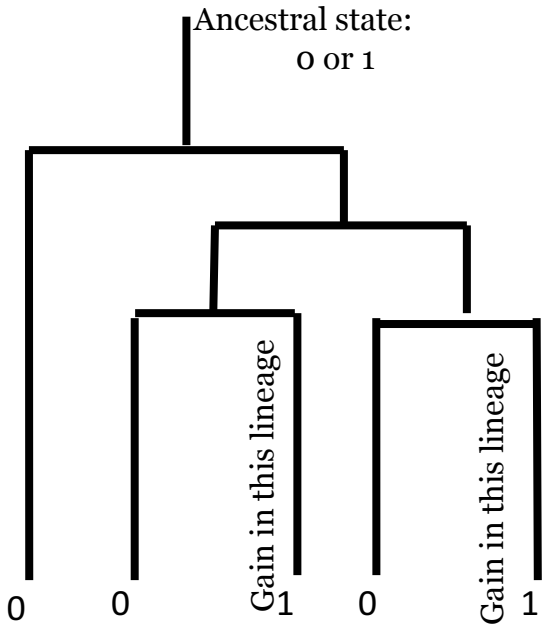
11

Key issues to address

- 16S rRNA gene may have multiple copies in some genomes
  - Solution: normalize the 16S data by multiplicity
- How do we infer metagenomic content of related organisms?
  - Solution: ancestral reconstruction

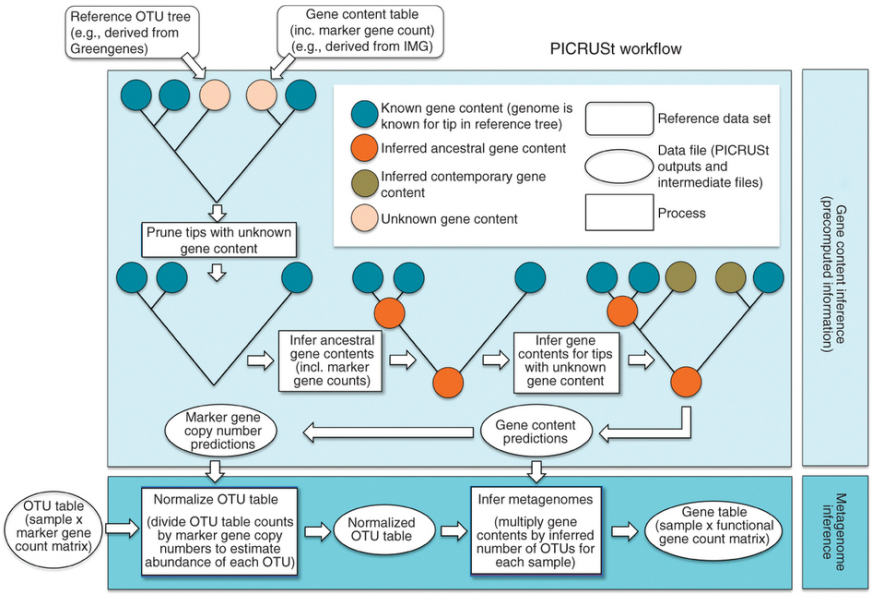
12

Ancestral reconstruction



13

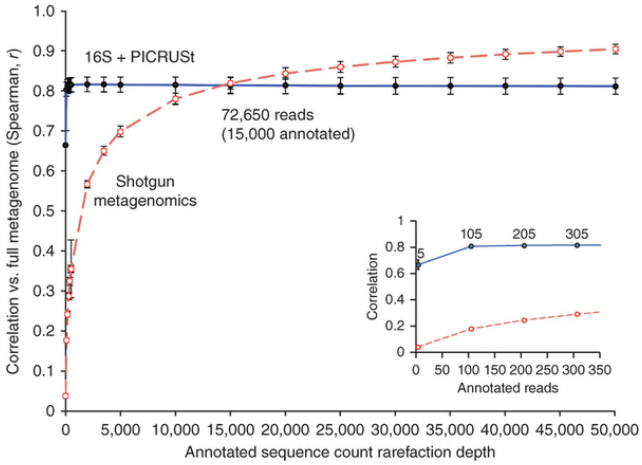
PICRUSt workflow



Nature Biotechnology. 2013; **31**, 814–821.

14

Performance of PICRUSt



Nature Biotechnology. 2013; **31**, 814–821.

How do we analyze the metagenome data?

- Analyses we discussed before are still applicable
  - Multivariate analysis
  - Hypothesis testing
  - Machine learning approaches
- Gene set enrichment analysis:
  - Determine if the number of significant genes within a category is greater than expected by chance.
  - Can be accomplished with Fisher-exact test, hypergeometric test

Gene set enrichment analysis (GSEA)

- Derives from univariate ranking of individual genes.
- Assesses the significance of the over-representation of top-ranked genes in lists of genes *a priori* grouped (pathways).
- Developed to overcome analytic challenges:
  - No significant genes;
  - Too many significant genes;
  - Inability of univariate methods to assign further meaning to significant genes.

Two versions

- Based on the entire list of effect sizes;
- Based on pre-determined significance threshold.

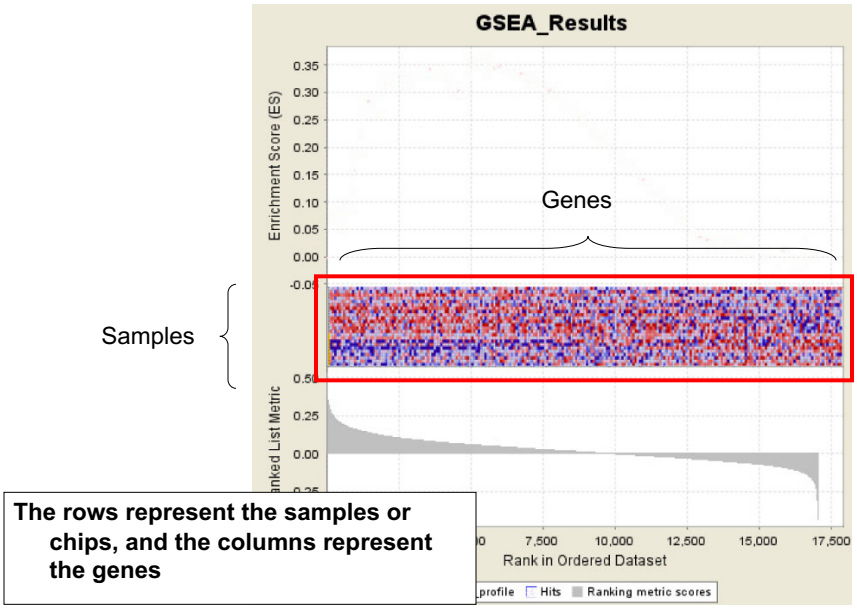
Based on pre-determined significance threshold

- Determine a significance threshold;
- Create a contingency table of the number of genes in or out of the pathway vs the number of significant vs non-significant genes;
- Use a contingency table association test, like chi-square or Fisher exact test.

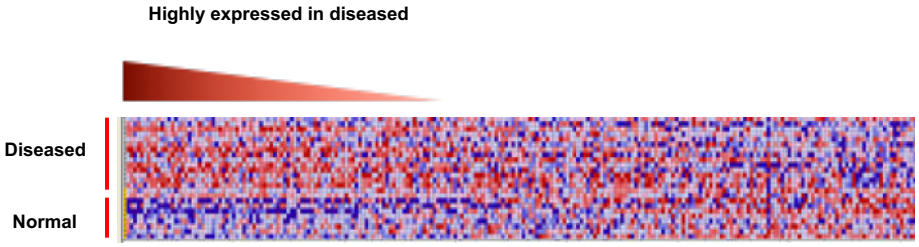
	In pathway	Not in pathway	
Significant	10	1	11
Not significant	10	999	1009
	20	1000	1020

GSEA

- Formally, GSEA considers a pre-defined list of genes and determines whether the members of this list are over-represented (enriched) at the top of the ranked list of genes.
- The ranking is based on association to a phenotype of interest.
- Can use effect size (absolute value) as the ranking.

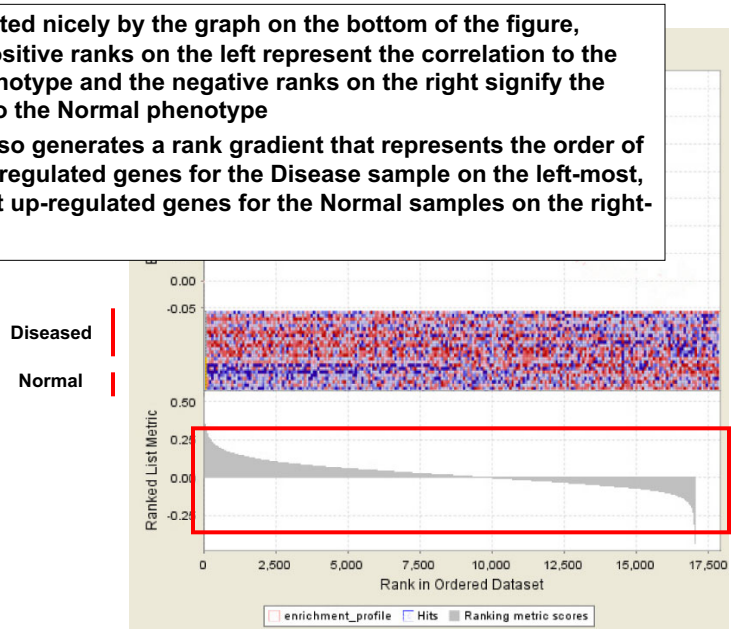


The rows represent the samples or chips, and the columns represent the genes

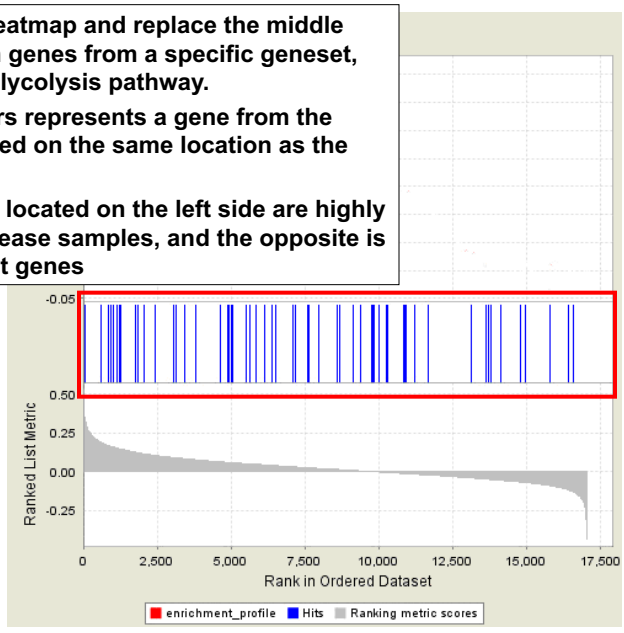


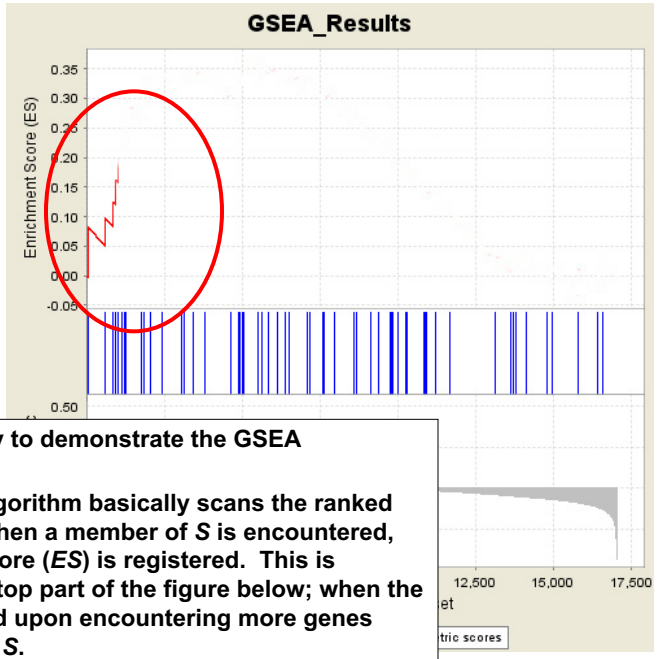
- **Genes on the left side are highly expressed on the top half (indicated by red color) and lowly expressed on the bottom half (indicated by blue color). The reverse is shown on the right-most genes**
- **Created a gradient or ranked list corresponding to the degree of correlation with the two phenotypes**

- **This is depicted nicely by the graph on the bottom of the figure, where the positive ranks on the left represent the correlation to the Disease phenotype and the negative ranks on the right signify the correlation to the Normal phenotype**
- **The graph also generates a rank gradient that represents the order of the most up-regulated genes for the Disease sample on the left-most, and the most up-regulated genes for the Normal samples on the right-most**

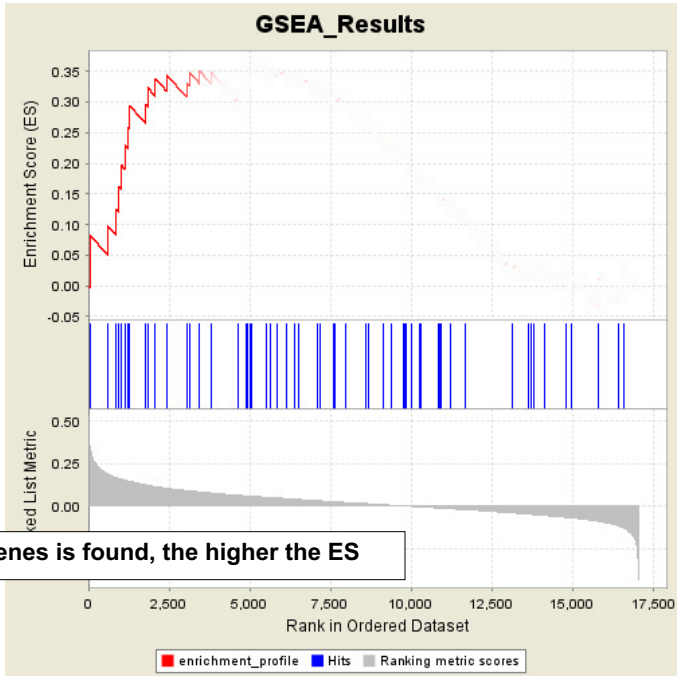


- **Now, let's hide the heatmap and replace the middle part of the figure with genes from a specific geneset, say genes from the Glycolysis pathway.**
- **Each vertical blue bars represents a gene from the pathway, being mapped on the same location as the whole dataset**
- **Again, genes that are located on the left side are highly expressed on the Disease samples, and the opposite is true for the right-most genes**

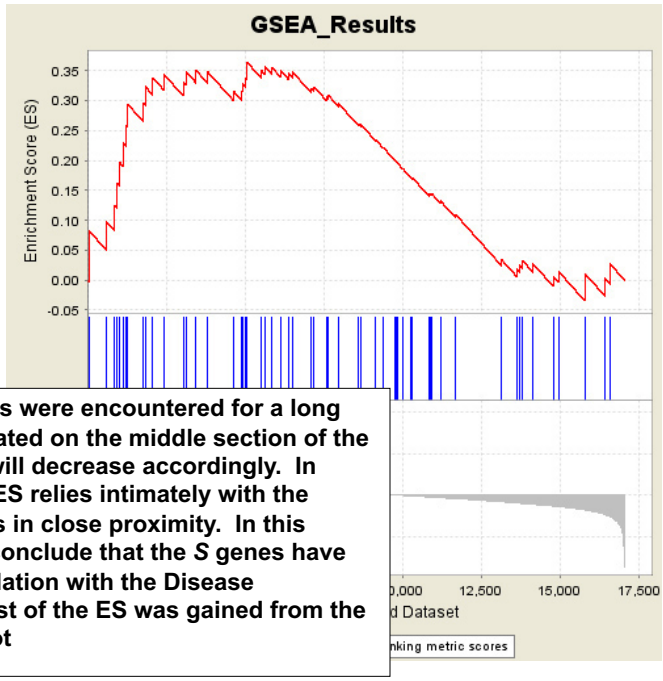




- **Now, we are ready to demonstrate the GSEA algorithm.**
- **The walk down algorithm basically scans the ranked gene list *L*, and when a member of *S* is encountered, an Enrichment Score (*ES*) is registered. This is illustrated on the top part of the figure below; when the ES started to build upon encountering more genes from the GeneSet *S*.**



- **The more *S* genes is found, the higher the ES**



- **But, when no *S* genes were encountered for a long walk down, as indicated on the middle section of the middle plot, the ES will decrease accordingly. In other words, a high ES relies intimately with the clustering of *S* genes in close proximity. In this example, we would conclude that the *S* genes have high degree of correlation with the Disease phenotype since most of the ES was gained from the left portion of the plot**

Advantages of GSEA

- Agnostic to the type of gene set and the source of annotation
- Operates on any ordered gene list
- Does not require the choice of a gene selection threshold or the explicit definition of a statistically significant marker set
- Uses distribution-free, non-parametric, permutation-based test procedures with increased statistical power
- Incorporates the permutation of phenotype labels thereby preserving the "biological" correlation structure of the markers
- Takes into account multiple hypotheses testing of multiple gene sets