

LAB 6 – NEXT GENERATION SEQUENCING APPLICATIONS: RNA-SEQ AND METAGENOMICS

[SOFTWARE NEEDED: web access]

Part A: RNA-Seq (also called Whole Transcriptome Sequencing “WTS”)

RNA-seq is the study of an organism or community’s transcribed genes through next-generation sequencing of steady-state RNA from the sample of interest. Typically the RNA is reverse-transcribed to cDNA prior to sequencing. RNA-seq is a quantitative examination of the transcriptional profile of an organism or community: what genes are “on” as well as relative levels of transcription. This allows an examination of the functional activity of a sample: identifying allele specific gene expression and finding novel transcripts.

Comparison of RNA-seq samples taken from different tissues or taken from the same source under different conditions allows examination of differential gene expression in response to a developmental program or to environmental changes (what genes turn “on” or “off” under these conditions, how genes are regulated under these conditions).

Box 1. RNA-Seq Analysis

RNA-seq analysis can be thought of an extension of long-standing methods such as ESTs, SAGE, and MPSS (expressed sequence tags, serial analysis of gene expression, and massively-parallel signature sequencing, respectively) for gene expression analysis. The main difference is that the overall number of sequence “tags” that are generated for a given transcript population is far higher due to the efficiency of next generation sequencing machines at generating sequence cheaply, increasing accuracy and sensitivity.

extraction of poly-A RNAs

conversion into ds-cDNA and shearing

amplification and adapter ligation

sequencing

single end (SET)

paired-end (PET)

Figure 1: Protocol for RNA-seq analysis. RNA is extracted from a sample of interest, reverse transcribed to cDNA, sheared and converted into a sequencing library for either Roche 454 or Illumina sequencing, and sequenced. Paired end sequencing involves a slightly more complicated library preparation step, but allows better assembly of the final sequence data. (image courtesy of <http://cmb.molgen.mpg.de>)

With all of these methods, mRNA is converted into cDNA through a reverse transcription step. In the case RNA-seq, the double-stranded cDNA molecules are sheared into smaller fragments, to which adapters are then ligated. The fragments are then sequenced, either by single end reads or using the so-called paired-end methodology, in which short reads from both ends are generated. Because the spacing between the ends is approximately known in the latter case, a better mapping of ambiguous reads may be achieved.

In other kinds of expression profiling, “normalization” is an important procedure in order to make expression values between different experiments directly comparable (that is, we don’t want to be saying that genes are differentially expressed when in fact all we’re looking at is different amounts of input material). In the case of RNA-seq experiments, summarization and normalization are sometimes combined using “Fragments Per Kilobase of exon per Million fragments mapped (FPKM)”, or originally “Reads Per Kilobase of exon per Million fragments mapped (RPKM)” in the case of single end reads. A fragment would have two reads associated with it. Note that “normalization” in this context does not refer to the traditional statistical definition of scaling all numeric variables in the range [0,1]. Some methods for differential gene expression analysis require raw read counts, however (see Bioinformatic Methods II, Module 4 for more details!).

We will examine two different RNA-seq data sets. The first data set was generated from the rice species *Oryza glaberrima*. This is an African cultivated rice species whose genome has been sequenced, allowing all RNA-Seq reads generated using RNA samples from it to be mapped onto the genome and identified, and the second data set is from *Arabidopsis thaliana*. The rice experiment involved performing RNA-seq on different tissues in the rice plant, which allowed an examination of how gene expression levels vary in e.g. leaves, roots, etc.

1. Go to the Plant MPSS databases homepage, at <http://mpss.danforthcenter.org/>¹.
2. Scroll down to the Rice databases, and click on the [link](#) in the “RNA-seq DBs” (4th) column for “*Oryza glaberrima*” organism, which will take you to https://mpss.danforthcenter.org/dbs/index.php?SITE=rice_glab_RNAseq
 - a) How many genes have been identified on the *Oryza glaberrima* genome?
 - b) How many chromosomes does this species have?
3. Click on the “Library Information” link at the top of the page.
 - a) How many RNA-seq libraries are available for this rice strain?
 - b) What tissues are the libraries from? Why would you sequence different tissues from the same plant using RNA-seq?
4. Make a note of what the library names are, and which plant tissue they were generated from. Go back to the Rice Oglab home page (“Home/basic queries” link).
5. Click on one of the chromosomes (e.g. Chromosome 4) so that it expands in the viewer.
 - a) What do the red and blue bars mean? Why are they coloured differently? (hint: click on the “legend” link the text at the top to see a legend).
 - b) What would a pink region indicate?

¹ The MPSS database moved from the University of Delaware to the Danforth Center in mid-2016.



Figure 1: *Oryza glaberrima* chromosome 4 sequence view

6. Click somewhere on the magnified chromosome region to zoom in one step further.
 - a) *What do the grey bars represent?*
If you don't see any grey bars, move the view to the right or left 1 or 2 MB.
7. To look at a specific gene and its expression patterns, go back to the Home/basic queries tab and search for "Orgla03g0398400" in the "Protein or gene ID" search box. Use the defaults otherwise.
 - a) *What do the grey and white diamonds signify on the plot?*
 - b) *What is the predicted function for this gene?*
8. Click the "Library Abundances" tab. Change the Control Panel options to display the libraries separately (link in text at top of page) as "Sum of Abundances".
 - a) *What has changed? What does this new display show you?*
 - b) *Which sample has the highest sequence expression? Does this make sense given the predicted function of the gene?*

Lab Quiz
Question 1

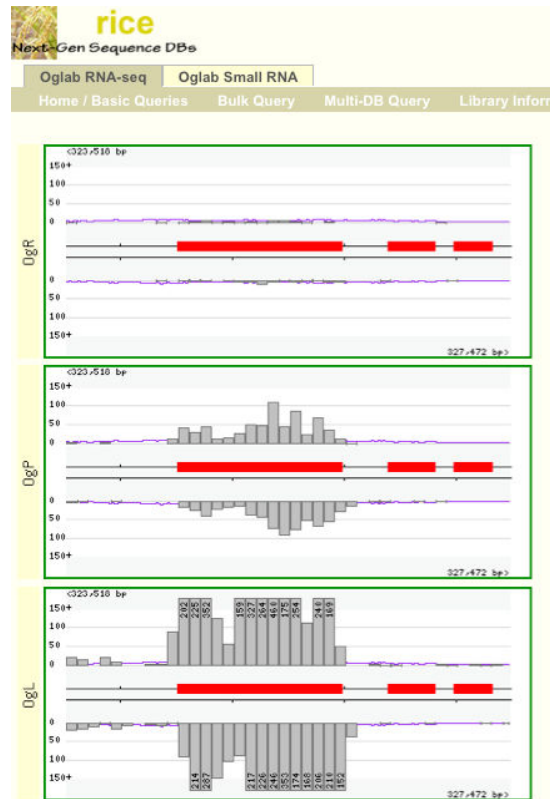


Figure 2: Graphical view of RNA-seq reads from three different tissue libraries mapped onto a gene of interest.

9. Go back to the “tabular view” for this gene. At the bottom of the page, look at the “RNA-seq Information”
 - a) What are these data? How do they relate to the graphs you looked at in step 8?

The second RNA-seq data set was compiled by Cheng et al. (2017) for *Arabidopsis thaliana*, which also has a complete genome sequence thereby allowing mapping of the RNA-seq reads to the genome. In this case, RNA-seq data came from 113 publicly-available RNA-seq data sets generated by different laboratories, collected from different parts of this model plant and, in some cases, sampled under different conditions. The goal of this “meta-analysis” was to update the *Arabidopsis* genome annotation (thereby creating the “Araport11” release) and to identify instances of alternative splicing. The process was also able to validate known transcript structures. Splicing refers to the removal of introns to generate the mature mRNA, which then serves as a template for translation. There are several different kinds of alternative splicing possible: alternative splicing at both intron acceptor and donor splice sites, alternative splice junctions, and alternative intronic sequences. All of these kinds of events would result in different transcripts being produced, which in turn might result in different proteins being produced.

Here, we will look at an example of alternative splicing, where a gene encodes different proteins depending on the introns and exons included or excluded during transcription. We’ll use TAIR’s Araport “JBrowse” instance (a tool for exploring genomic data and features) to examine the reads from the 113 RNA-seq samples mapped to the reference genome, grouped by tissue type.

10. Open a tab in your browser, and go to https://jbrowse.arabidopsis.org/index.html?data=Araport11&loc=Chr2%3A15203749..15208380&tracks=TAIR10_genome%2CAraport11_Loci%2CAraport11_gene_models%2Caerial_tophat%2Cleaf_tophat&highlight= (if clicking on this link does not work, copy-paste this link into your browser: ensure that all information is present and on one line – check in a text editor if necessary). This will open TAIR’s Araport JBrowse instance, centered on the gene At2g36270, which encodes the ABI5 protein, a transcription factor involved in abscisic acid (a plant hormone) response.
11. For TAIR’s JBrowse instance, Araport curated a collection of genomic information for Arabidopsis. If the above URL doesn’t preload the RNA-seq tracks we’re going to be looking at, you can add these using the **Available Tracks** panel on the left by using the checkboxes. The tracks may be reordered in the main panel on the right by dragging them up or down.

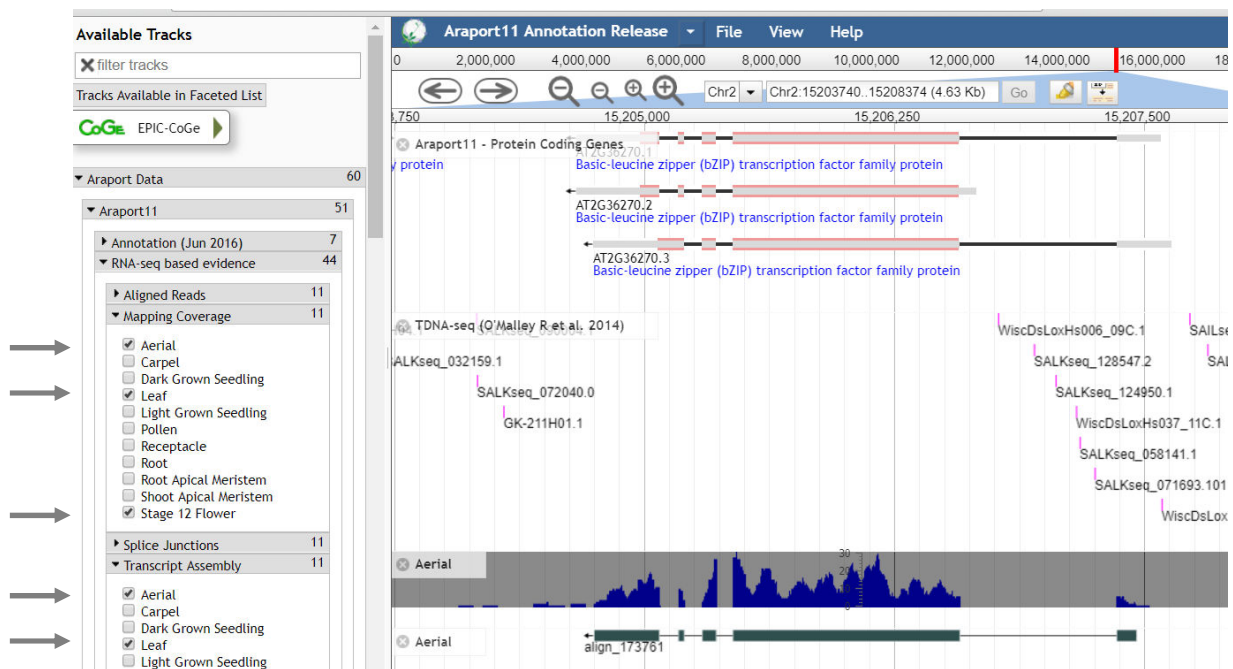


Figure 3: Track selection for *Arabidopsis thaliana* RNA-seq data sets. Choose the three RNA-seq “Mapping Coverage” data sets indicated with arrows: “Aerial”, “Leaf”, “Stage 12 Flower” along with the corresponding “Transcript Assembly” tracks. Make sure the “Protein Coding Genes” track is visible: this is found in the “Annotation” section of the **Available Tracks** panel.

12. The plot shown has the nucleotide positions of our gene of interest on the horizontal axis along the top. For each RNA-seq track chosen, RNA-seq reads that map to this gene are displayed as histograms showing the density of read coverage at that specific nucleotide. The taller the histogram, the more reads mapped there, and hence the higher the expression of that portion of the gene. Gaps in the histogram identify introns, or exons that were spliced out, in the mRNA sample analyzed.

- a) How many exons are expressed in the “Aerial” data set (blue in Figure 4)? Compare this with the gene models which shows the gene architecture, with exons and UTR highlighted in red/grey, respectively.
- b) How do the RNA-seq profiles from the “Leaf” and “Stage 12 Flower” samples differ for this gene? Based on the gene models, what do you think is happening?

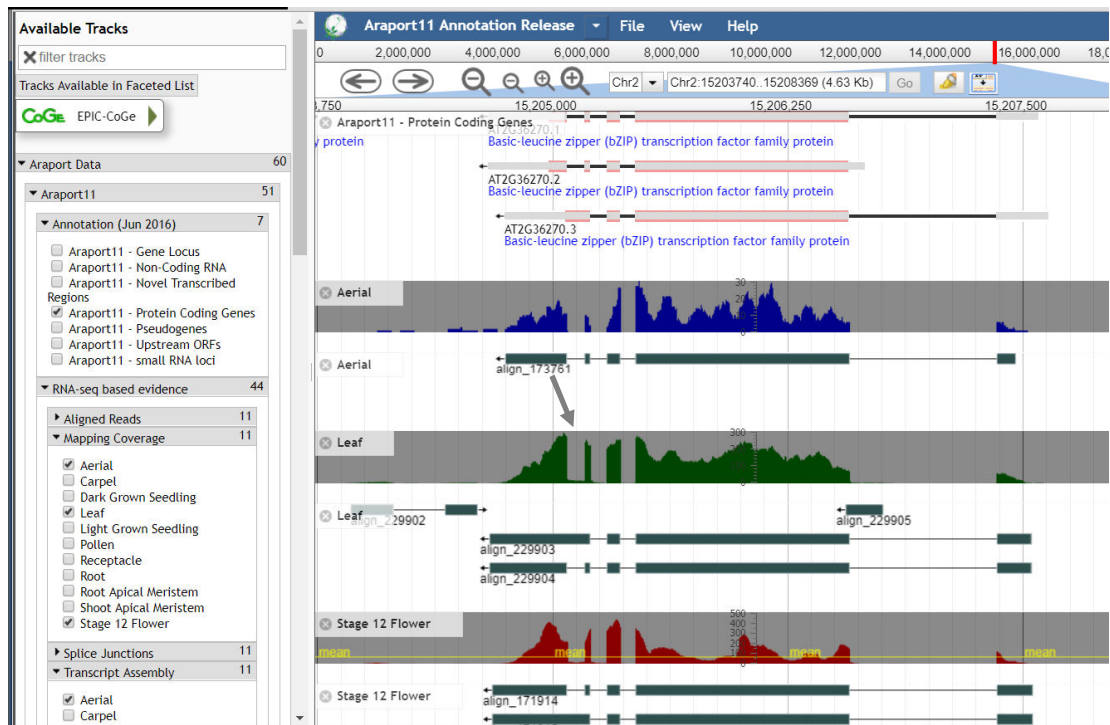


Figure 4: JBrowse map of three grouped RNA-seq data sets from TAIR’s Araport JBrowse instance for the gene *ABI5* (At2g36270), with CDS and gene model maps. The grey arrow above the Leaf track points to an intron retention event identified in, among others, leaf and stage 12 flower RNA-seq samples by the Araport11 release (Cheng et al. 2017). Note it is possible to edit the track colours; the display colour for the “Leaf” track has been set to green by editing the configuration file.

13. Add a few more RNA-seq coverage tracks, such as those for carpels or receptacles (these are flower parts), by using the checkboxes on the **Available Tracks** panel.
 - a) Do you see evidence for different alternative splicing possibilities in any other tissues? What kind of alternate splicing do you think is taking place? Try entering *Chr5:19864119..19866100*, which will call up another example of intron retention (or “exitron”) from the Cheng et al. 2017 paper.

Clearly identifying alternative splicing events genome-wide by hand would be a challenge, and thus researchers have developed algorithms to do this in an automated and statistically-significant way, such as with supersplat (Bryant et al., 2010, <http://www.ncbi.nlm.nih.gov/pubmed/20410051>), TAU, or others. Examining RNA-seq tracks in a genome browser can nonetheless be informative on a gene-by-gene basis.

Part B: METAGENOMICS

Metagenomics is the study of mixed communities of organisms by sequencing DNA extracted from the community as a whole. The metagenome sequence provides information as to what kinds of species the community contains (“who is there”), as well as what metabolic functions these species are capable of (“what they might be able to do”). Metagenomics provides genomic rather than transcriptomic information; a gene with a predicted function present in a metagenome is not necessarily expressed, but it is present in the community so could be functionally important. Metagenomes are useful for comparing changes in community composition over time, for mapping RNA-seq reads or proteomic data, and for discovering novel genes.

Box 2. Metagenomics

The term metagenomics was coined in 1998 by Handelsman et al. (see [http://dx.doi.org/10.1016/S1074-5521\(98\)90108-9](http://dx.doi.org/10.1016/S1074-5521(98)90108-9)) and was widely popularized by Craig Venter’s ground-breaking metagenomics study of the ostensibly “dead” Sargasso Sea near Bermuda in 2004 (<http://dx.doi.org/10.1126/science.1093857>). This study showed a wide variety of prokaryotic – more than 1,800 – phylotypes present in its waters. Metagenomics has been defined as “the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species” (see <http://dx.doi.org/10.1371/journal.pcbi.0010024>).

Figure 1 shows the schematic of a typical metagenomics pipeline.

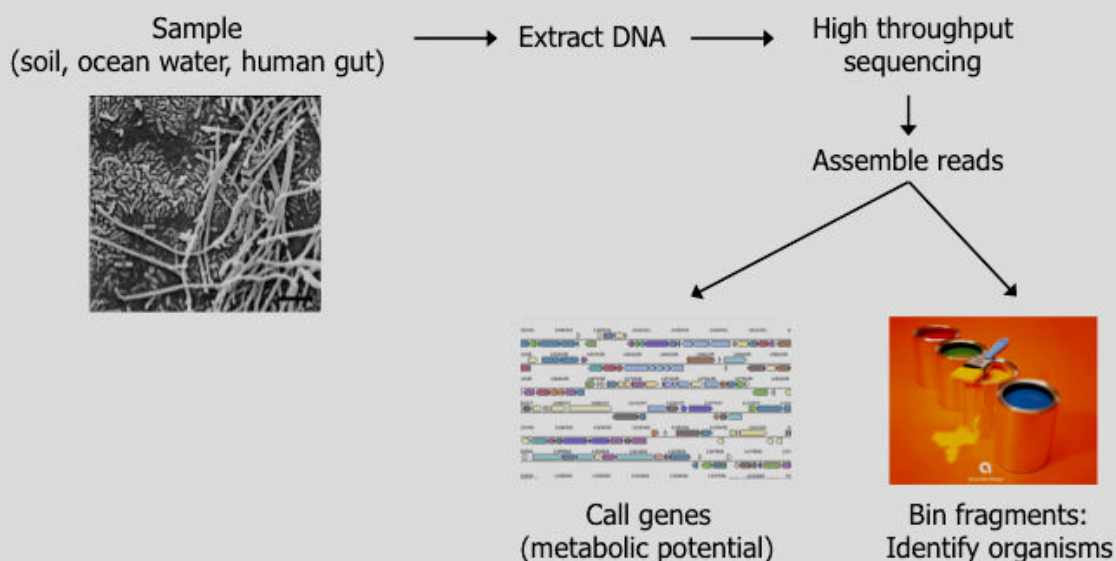


Figure 1: Metagenome sequencing protocol. DNA is extracted from a mixed community of interest for sequencing with next-generation sequencing, perhaps with a filtration step to size the species of interest. The assembled sequence gives a blueprint of the metabolic potential of the community, as well as a taxonomic profile (who is there, and what they are potentially capable of doing). Image courtesy of JGI.

Unlike earlier studies based on the sequencing of just 16S ribosomal rRNA sequences amplified from environmental samples, metagenomics also provides an indication of the metabolic potential of a community in terms of the metabolic intermediates from one species that could be used by another species that is not capable of synthesizing those on its own, as well as the reason why an individual species can survive in a particular niche (e.g., large numbers of iron scavenging systems in low pH environments where iron is not readily available).

The number of species in a community affects the number of whole genomes that can be assembled from the short “shot-gun” sequence reads from a metagenomics sequencing project. The fewer the number of species, the more whole genome sequence can be generated (see Figure 2).

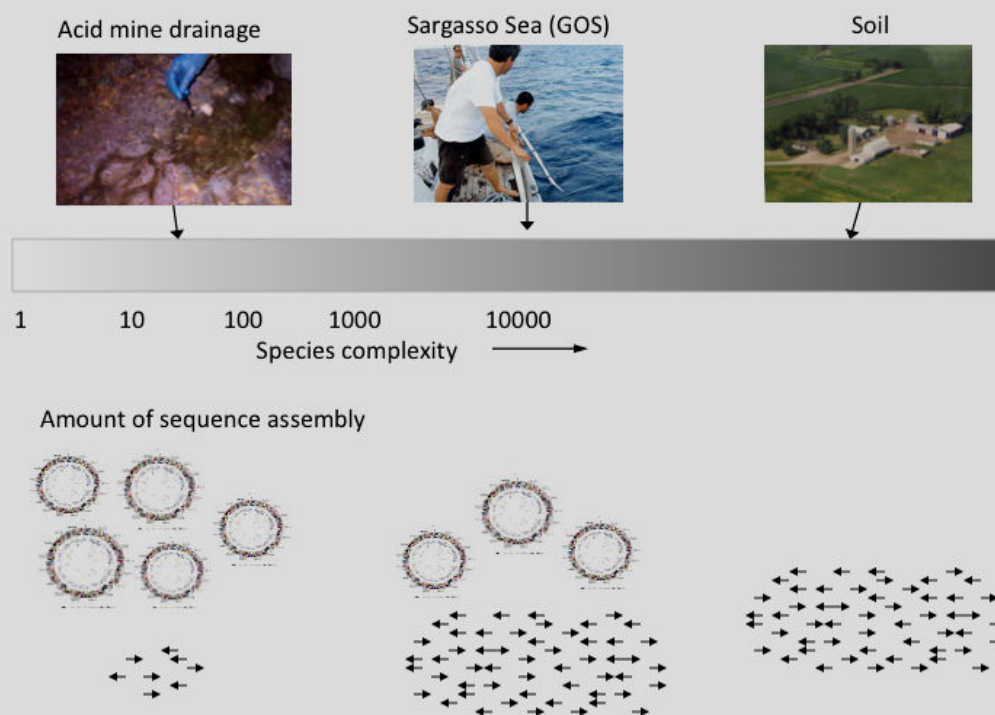


Figure 2: How community complexity affects the composition of a metagenome. Less complex communities (e.g., acid mine drainage biofilm) have only a few dominant organisms, whose genomes assemble relatively completely (represented by circles in the “Amount of sequence assembly” section). More complex communities (e.g., soil) assemble poorly, with many fragments of sequence that contain only a few genes. Image courtesy of JGI.

In this exercise, we will examine two metagenome sequences using the MG-RAST web server (<http://metagenomics.anl.gov/>).

The first metagenome comes from the drainage fluid from a metal ore mine in California. Acid mine drainage is a common environmental hazard formed from microbial activity on sulfide

mineral-containing rocks exposed to air and water. The community that can exist in these low pH conditions is very limited, and it is of interest to understand how these organisms create the acid mine drainage, as well as how a limited community can be used as a simplified system to understand more complex communities in general.

The second metagenome comes from Dr. J. Craig Venter's global ocean metagenomic survey, a pioneering metagenomics study that sought to "sequence the entire ocean". Oceanic environments contain a vast diversity of bacteria, archaea, and single celled eukaryotes, as well as larger marine life (fish, coral, etc.). This metagenome sequence is one of a series of 88 separate samples taken by Dr. Venter from his personal sailboat. All of the Global Ocean Survey (GOS) samples were filtered to ensure only single-cell organisms were sequenced.

Using the MG-RAST interface, we will examine the composition of these two communities based on their metagenome sequences, and compare the two samples both taxonomically (who is in the environments) and metabolically (what functions are taking place in these environments).

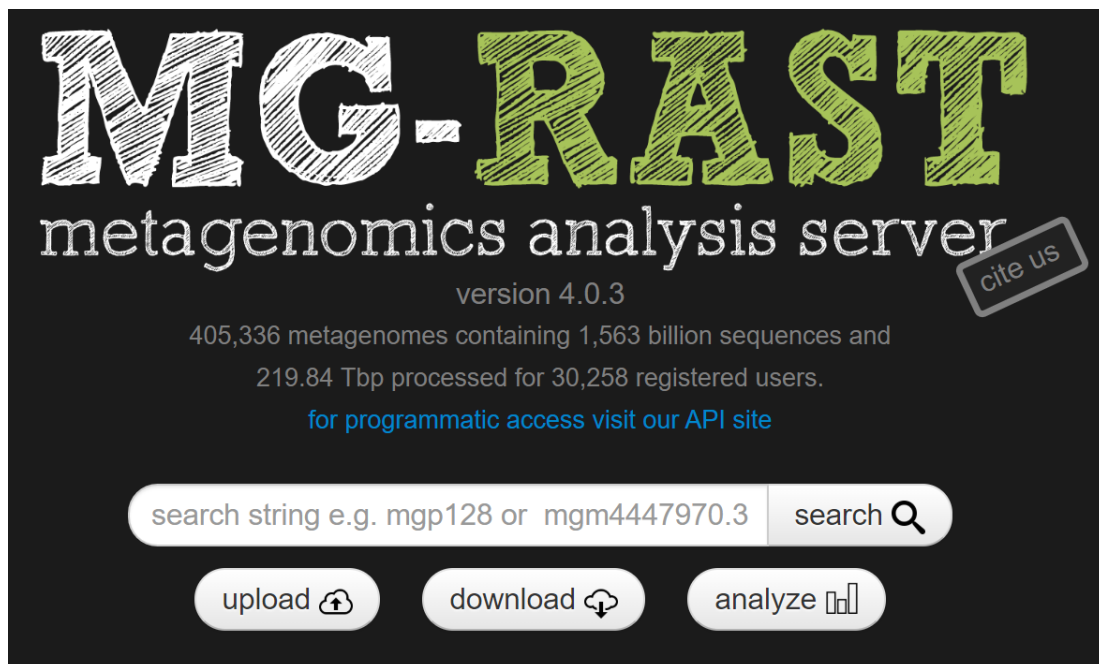


Figure 5: the MG-RAST server's homepage (<http://metagenomics.anl.gov/>)

1. Go to the MG-RAST website: <http://metagenomics.anl.gov/>.
 - a) *How many metagenomes are currently housed on the MG-RAST server?* ²
2. Search for **4441138.3** using the search function, this is the metagenome ID for UBA Acid Mine Drainage Biofilm. Leave this tab open for now.

² The number is constantly increasing – in Figure 5 you can see it's now around 405,000 (Dec. 2019)!

3. Open a new tab and from the MG-RAST homepage search for **4441147.3** using the search function. This is one of the metagenome IDs for the “Global Ocean Sampling Expedition”. Click on the link in the Dataset column to see sample information.
 - a) *Where was this sample taken from?*

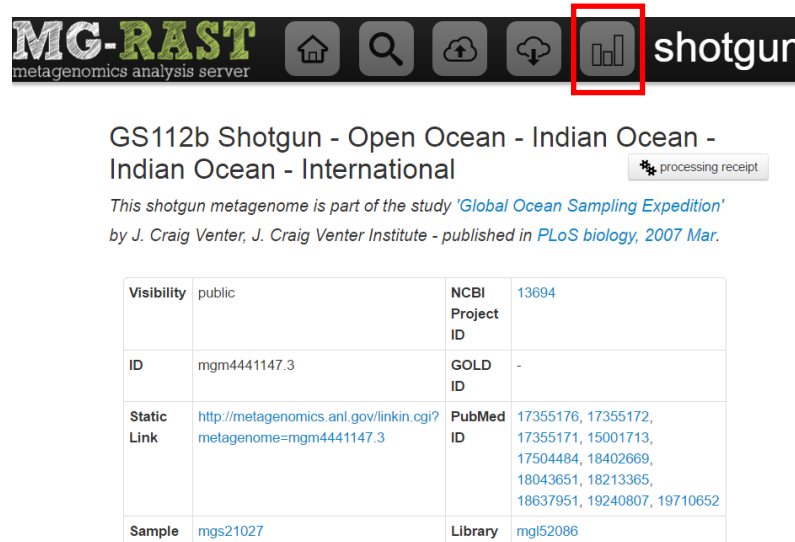


Figure 6: Initial data for a metagenome entry. Links on this page include those where you can download the sequence data, connections to the MG-RAST analysis page (boxed in red), and various links to this data set in other databases, including NCBI.

- b) For both metagenome records, scroll down to the “Sequence Breakdown” pie chart. *What does the GOS sample contain, in terms of its sequence? The Acid mine drainage sample? Which sample is better characterized based on these charts?*
 - c) There is a lot of information displayed on these entries: take some time to look through the various charts by clicking on the sections on the right side of the entry. *For the Taxonomic Distribution section, which domain is dominant in each metagenome? Which phylum? (hint – the charts are interactive, hover your mouse over the slices to see what they represent). Is one sample obviously more taxonomically diverse (many different groups) than the other?*
 - d) Scroll down to the rarefaction curve at the bottom of the entries. A rarefaction curve displays whether a sample has been sequenced to saturation: a steeper slope on the plot indicates that the sample has not yet been fully sequenced, while a slope that begins to flatten out indicates that the majority of the DNA in the sample has been sequenced. *Looking at the two plots, which metagenome sequence is closer to being a complete sequence for that community? Is that expected given the communities that were sequenced?*

Lab Quiz
Question 2

Comparative metagenomics

We have used the metagenomes' basic entries as a rough comparison tool: now let's look at the similarities and differences between these two samples in more detail.

4. Scroll to the top of either tab, and click on the bar-chart symbol (see red box, Figure 6). This will take you to the MG-RAST analysis page.
5. We'll retrieve the two metagenomes we were just looking at. Under the Metagenomes heading, type "acid" into the Enter filter box and press Enter (or just wait a second to see the records matching that keyword). Select the "UBA Acid Mine Drainage Biofilm" project from the list on the left hand side (see **Figure 7**), and click the →. Clear the **name: acid x** filter by clicking on the "x" beside the filter. Add the second metagenome we want to compare, the Indian Ocean sample. Type "ocean" in the Enter filter box and press Enter. Select the "GS112b Shotgun – Open Ocean – Indian Ocean – Indian Ocean – International" sample from the list on the left hand side, and click the →. On the right, our two metagenomes will be present.

Figure 7: Selecting metagenomes for comparison from the MG-RAST publicly available metagenome sequence data sets. Type "acid" in the filter box and press enter. Use > to add the "UBA Acid Mine Drainage Biofilm" metagenome to our analysis pipeline, as the first metagenome. Clear the filter and use "ocean" as the second filter to identify the ocean samples. Add the GS112b sample as our second metagenome for analysis with the > icon.

* Add "Subsystems" from the available databases in order to be able to do Step 8.

Provide an analysis name and click (you may need to scroll over to the right to see this checkmark beside the box where the two genomes' names appear). An **analysis panel** will appear with various options, see **Figure 8**.

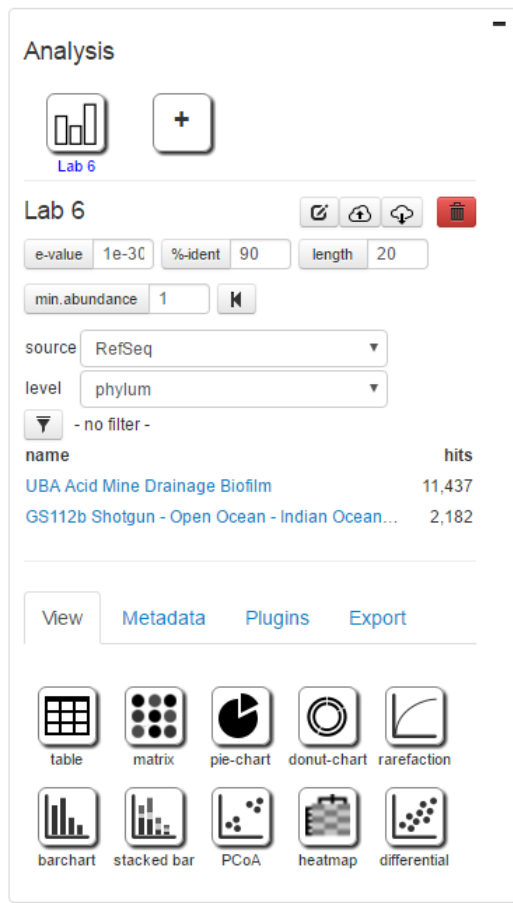



Figure 8: The MG-RAST analysis panel, where taxonomic and functional annotations for the sequence data can be determined, with user-applied requirements as per Step 6 of this lab (e-value, length of alignment, and % identity). This interface allows you to control the specificity of the results you then work with. It also allows comparison of metagenomes directly through the MG-RAST suite of tools via the icons at the bottom of the panel.

6. To compare the two samples on a taxonomic level, first adjust the analysis parameters as follows: change **e-value** to **1e-30** (click the e-value button to enter that value once you've type it into the box, or press enter), **% Ident** to **90%**, and **Length** to **20**. These parameters determine which annotations will be accepted and displayed in the analysis. The default parameters are quite lax (recall the second lecture, slide 29: 1e-20 is probably the minimum we should be using to have longer matches to NCBI's RefSeq sequences that are being used to categorize the metagenomes by MG-RAST!).
 - a) *What would changing the % Identity cutoff to 80% mean? Why would you not want to do this for your metagenome? Why might it be useful?*
7. Click the donut-chart icon  and choose "phylum" as the level.
 - a) *What level of taxonomy is being displayed?*
 - b) *What are the groups coloured by?*
 - c) *Which sample is on the inside of the donut, and which sample is on the outside?*
 - d) *From this visualization, do the two samples' communities overlap a lot? At all? Name one taxonomic category where both metagenomes are represented.*

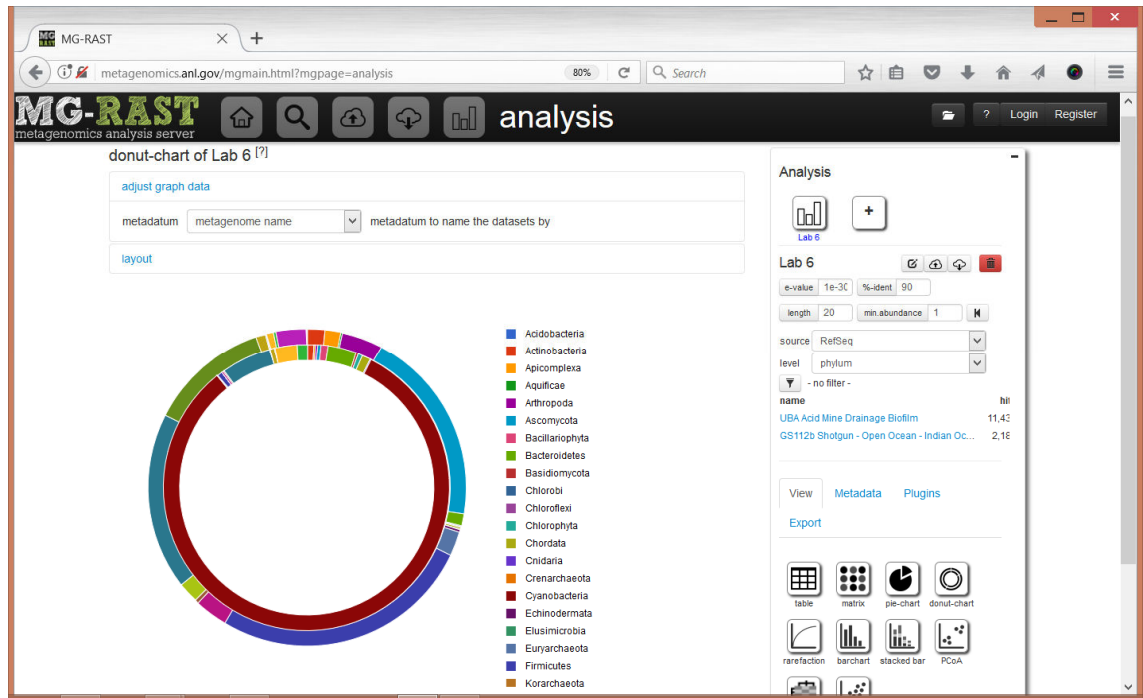



Figure 9: Phylogenetic analysis of taxonomic groups identified from two metagenomes.

8. Now let's compare the two communities from a functional viewpoint. Switch the "source" to "Subsystems": this is MG-RAST's in-house method of classification. Keep the other parameters (e-value, %-ident, length, min. abundance) the same as in Step 6 (level should be Level1). If you don't see "Subsystems" in the "source" dropdown, you may need to add it by choosing "Subsystems" in the available databases part of the Analysis page and clicking "add".

Select the heatmap icon . (If you don't see the heatmap or get an error message, try starting the analysis again by clicking on the red garbage can icon...you'll be taken back to the step shown in Figure 7, and you'll have to adjust the e-value parameters etc. again). You may want to adjust the parameters of the [layout](#) in order to be able to see all of the Subsystem categories in the graphical output. In order to be able to see something like **Figure 10**, use a height of 2000, a width of 800, and a row title width of 300. If you're red-green colour blind, change the heatmap colour to yellow and blue. It might help to [adjust graph data](#) so that the metagenome IDs are shown in the graphical output, instead of their names.

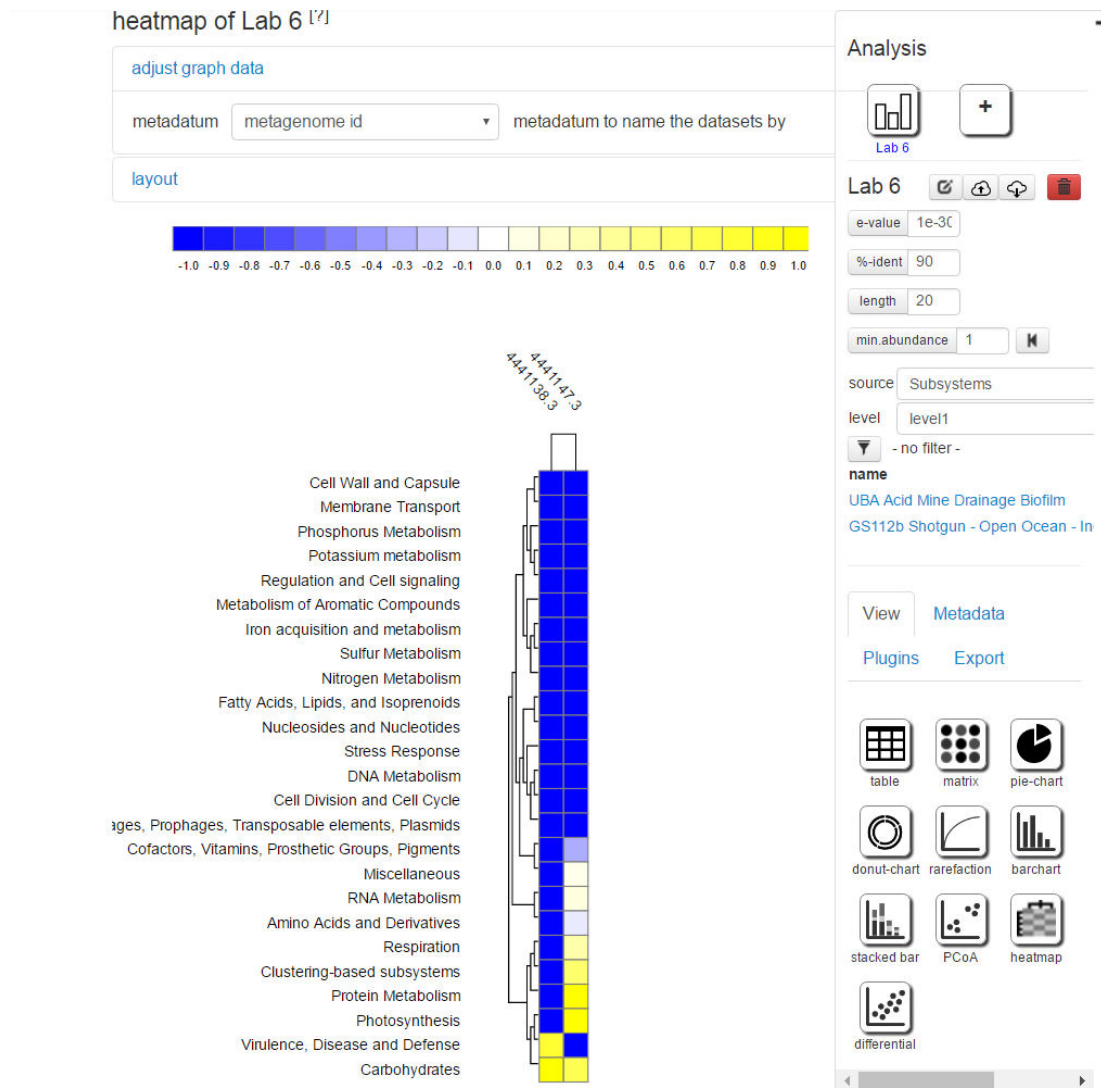


Figure 10: Heat map comparison of two metagenome's function profiles at subsystem level 1.

9. The heat map displays functional categories and the relative enrichment of each metagenomes' proteins in those categories.
 - a) From the scale, what does the colour red/blue mean? Green/yellow?
 - b) Name the functional category or categories that the GOS metagenome (4441147.3) is most enriched in compared to the Acid Mine Drainage community (4441138.3). Why might this be? What about for the Acid Mine drainage?
10. In the parameters section of the Analysis panel, change the level the heat map is grouped by to "level 2". Level 2 is a more specific category scheme of the MG-RAST subgroups, while the Level 1 categories are very large, general headings. (e.g., a Level 1 category is "Carbohydrates", while a Level 2 category within the Level 1 category of "Carbohydrates" is "Polysaccharides"). Change the minimum abundance from 1 to 5 by entering 5 in the **min. abundance** box. By doing so, we're removing low frequency

occurrences of metagenomics sequences to focus on the more prevalent sequences in the communities we're examining.

- a) With the min. abundance set to 5, and looking at Level 2 what are the highest enriched subcategories for each metagenome? Can you connect these to the location the sample was taken from? Try using the filter to see e.g. just the Carbohydrate subcategories in the heatmap.

Lab Quiz
Question 3

End of Lab!

Where to get it:

MG-RAST <http://metagenomics.anl.gov/>

Lab 6 Objectives

By the end of Lab 6 (comprising the lab including its boxes, and the lecture), you should:

- understand some of the current technologies for performing next generation sequencing (NGS);
- be familiar with how NGS assemblies are generated, and be aware of some of the pitfalls associated with assembling short reads and know how paired-end sequencing can solve some of these problems;
- know how to explore RNA-seq data on a Genome Browser and understand what an RNA-seq track can be used for in terms of ascertaining the expression level of a gene and what might be happening at the level of alternative splicing;
- be acquainted with the concept of metagenomes and understand how these can be used to assess species diversity of a community and metabolic capacity of an organism or community.

Do not hesitate to check with the Discussion Forums on Coursera if any of the above material is unclear after reading the relevant material.

Further Reading

El-Metwally S, Hamza T, Zakiria M, Helmy M. 2013. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PLoS Computational Biology
<http://dx.doi.org/10.1371/journal.pcbi.1003345>.

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, & Town CD. 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. The Plant Journal, 89(4), 789-804.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.

Nakano M, Nobuta K, Vemaraju K, *et al.* 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Research 34:D731-5.