Discuss the challenges involved in defining a microbial species and how HGT complicates matters, especially in the context of evolution ~~and phylogenetic distribution of microbial pathways~~. Comment on how HGT influences the maintenance of global biogeochemical cycles through time. Is it necessary to have a clear definition of microbial species? Why or why not?

A comparison of sequencing platforms and

bioinformatics pipelines for compositional

analysis of the gut microbiome

Imane Allali1,4, Jason W. Arnold1, Jeffrey Roach2, Maria Belen Cadenas1, Natasha Butz1, Hosni M. Hassan3,

Matthew Koci3, Anne Ballou3, Mary Mendoza3, Rizwana Ali3 and M. Andrea Azcarate-Peril1\*

* 16S sequencing has become the norm for identifying and classifying microbial species.
* Comparative studies of NGS tech emerging to make use of this technique allow for critique of existing data (Currently one big mess)
* Bioinformatics offers limited ability to reduce error
* There is bias introduced with primer design, library preparation, DNA isolation methods and PCR amplification.
* However, these do not impact the study results greatly.
* DNA isolation methods, and 16S rRNA primers are major sources of variation in sequencing results.
* Sequencing depth and sample storage has small but detectable effects.
* Roche GS Flex+, illumina Miseq and Ion torrent PGM generate high quality, comparable data
* Technical protocols and sequencing platforms have a variable impact on output: it depends.
* UPARSE bioinformatic pipeline leads to lower diversity than the QIIME pipeline
* Composition depends on the platform and bioinformatics pipeline
* Because NGS has rendered sequencing faster, more inexpensive as a tool, we are able to compare different samples and pipelines more rapidly than before in order to generate comparisons between them.
* So in conclusion, how you prepare samples, the pipelines used, etc all impact the diversity of your results at the end. These all need to be taken into account.

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan1, Paul J McMurdie2 and Susan P Holmes3

* We rely heavily on PCR amplified high throughout sequencing based on marker genes. To classify microbes.
* Analysis of marker genes begins with classifying OTUs: these are reads that differ by less than a certain %.
* The OTU is often treated as a certain species within the scope of a study
* Closed read reference methods use sequence similarity to classify OTUs. De novo methods are used to group OTUs by pairwise sequences. These are sequences that are functionally or structurally relevant in some way.
* ASVs are a new thing (amplicon sequence variants), where the similarity is not arbitrarily decided by %. They infer the biological sequences in the sample prior to introducing amplification and sequencing errors. They can detect differences in sequences by one nucleotide.
* ASVs are better at sensitivity and specificity that is equal to or better than OTUs at showing differences in ecological patterns.
* The benefits are that they combine the usefulness of closed reference and de novo OTUs for future reference. They are reusable over future studies and are not limited by incomplete reference datasets.
* De novo OTUs: emergent features of a data set with boundaries and membership depending on the data set in which they are defined. OTUs defined in two different data sets cannot be compared.
* Closed reference OTUs come from a reference database. Reference sequences in the database define and label associated closed-reference OTUs. OTUs from independent data sets can be compared. However, biological variation that is not represented is lost during assignment.
* ASVs: inferred via a de novo process, biological sequences are differentiated from errors based on likelihood of repeatedly occurring as an error-containing sequence. They cannot be performed independently on each read, and the smallest unit is a sample of reads. They are consistent labels that represent a biological reality outside of the analysis, which is the DNA of the organism.
* Consistent labels are important because of computational tractability. These allow OTUs to be split into independent assignment of subsets of data. For ASVs, as long as the sequencing effort devoted to each sample remains tractable, ASVs can be inferred from arbitrarily large data sets.
* Meta analysis is also now possible. Replication is also possible. Forward prediction is possible, where the composition of the microbial community is a predictive biomarker for phenomena. This is an exciting translational opportunity.
* Reference independence is also important. Measurement of diversity is improved because reference databases are incomplete. Unrepresentation of microbial species skews diversity measures.
* On that point, representation depending on different environments varies a great deal. This allows us to study various environments equally with different loci.
* This also guarantees that an observation of an OTU is real and not a false positive. ASVs are consistent labels derived from data alone, therefore remain consistent forever.

A Comparison of Methods for Clustering 16S rRNA

Sequences into OTUs

Wei Chen1,2, Clarence K. Zhang3, Yongmei Cheng1, Shaowu Zhang1, Hongyu Zhao2,3\*

* There are 5 X 10^30 microbial cells on earth. Their diversity is beyond any other living thing.
* Culture-dependent microbial studies were the first to be used, which captured ~1% of all microbes.
* Sequence-based classification allows the bypass of the cultivation limitation.
* Two major approaches exist: taxonomy-based and OTU-based methods. Taxonomic based are reference dependent. OTU-based methods classify based on cluster generation based on a distance matrix.
* Taxonomy dependent approaches can be applied to every new study, but it is difficult to characterize novel sequences, and the accuracy depends on the reference. Most references have good classification only at the genus or higher level. OTUs can assign OTUs without prior information on reference taxonomy.
* OTU based methods are useful for analyzing less characterized communities, but sequencing errors create an inflation of OTUs.
* Heterogenous evolution rates in 16s rRNA make it difficult to choose the threshold to define OTUs. Biologically meaningful information is also hard to decipher.
* One thing to note is that there is no consistent threshold that can be assigned to all methods. The minimum dissimilarity threshold sometimes exceeds the conventional 3% that can be assigned for low complexity sequences.
* The higher the data set complexity, the more stringent the dissimilarity cutoff has to be. If the wrong cutoff is used for a data set, it could lead to an under-estimation for the number of OTUs.
* There is a large need to develop methods for properly analysing 16S rRNA sequencing data.
* Majority of sequences are still unknown. This means that OTU-based methods are still going to be valid for future studies.
* Most OTU based algorithms either over or under estimate the number of OTUs if an improper dissimilarity threshold is chosen. This value should be chosen for each specific method, and in relation to the complexity of the datasets.
* Most OTU algorithms split samples from the same species into several sub clusters. Even if we tried to combine small clusters, some big clusters would come from different species.
* Sequence abundance affects clustering. Low abundance sequences should be filtered to increase the accuracy regarding number of OTUs.

Updating the 97% identity threshold for 16S ribosomal RNA OTUs

Robert C. Edgar1

* Basically that 97% identity threshold is an outdated clustering threshold, that was created when few 16S rRNA sequences were available.
* An optimal identity threshold is somewhere between 99-100% for the V4 hypervariable region.
* In fact, all optimal thresholds are higher than 97%. No algorithm is consistently better than the other. Therefore, they cannot be ranked.
* Optimal thresholds also differ by the data.
* There is enormous strain to strain variation within a species. The difference between a pathogenic and commensal strain can be large. OTUs that approximate species, even accurately, will lump distinct phenotypes into a single cluster.
* OTUs approximating strains rather than species are more valuable. This would be more biologically informative.
* Resolving strains can be done with the V4 variability, but require close to 100% threshold to be optimal. This is probably not possible in general, but some strains can be resolved for some species.
* Defining by strain raises new difficulty. Some strains have very similar phenotypes to each other. These should be in the same OTU. Others that are substantially different should be assigned a different OTU. Can we even annotate such distinctions?
* Complete genome assemblies are required for the greatest accuracy.
* Little information is available for subspecies occurrence and variation in vivo.
* Zero radius OTUs should replace 97% thresholds because there is an improvement in reusability, reproducibility and comprehensiveness.
* There is an increased tendency to split species and strains over multiple OTUs.
* Some lumping and splitting of strains is unavoidable at any threshold.
* Zero OTUs have a better balance of the two than 97% threshold, which tend to lump species together.
* ZOTUs can be comparable across datasets without needing to recluster.
* Longer sequences cause ZOTUs to produce more splitting than lumping. This can be address through downstream processing.

Measuring the diversity of the human microbiota with

targeted next-generation sequencing

Francesca Finotello, Eleonora Mastrorilli and Barbara Di Camillo

* 16S rRNA seq is the status quo for identifying and quantifying human-resident bacteria.
* 16S rRNA sequence is highly conserved in species.
* PCR amplifies out the sequence from extracted DNA. The number of sequence reads is used as an approximate count for abundance.
* Alpha and beta diversity measure the diversity of a landscape. Alpha: average species diversity of sites belonging to a particular landscape. Compositional complexity in a sample. Beta: Differentiation between each site for diversity. Taxonomic differences between samples.
* Gamma: total diversity in a landscape.
* Diversity is not determined by a physical quantity that can be measured. Diversity is multi-faceted, and can be measured in different ways. Various mathematical formulas have been developed for this purpose. Each mathematical formula is better suited to answer a specific biological question than all of them.
* These mathematical models were developed for macroecology. The bias and error these formulas has not yet been studied.
* Alpha diversity is best used when all species are equally abundant. Richness does a direct count of all species. Underestimation of species richness occurs when sampling and sequencing is done. Prediction of true richness is done via mathematical correction.
* Alpha diversity indices take into account the abundance for each species. Evenness indices measure how evenly the abundances are distributed across different species. This measures not only diversity, but also the stability and reliance of an ecosystem.
* The estimation of true diversity is important because of a loss of species richness during sampling reflecting the losses of rare species.
* The limitation of sequencing depth leads to this. When a sample is sequenced, a max number of reads is generated from each sample, usually around 10^5. Because rare species are less counted per species, they are less likely to be represented in the final count than abundant ones.

A compendium of

multi-omic sequence information

from the Saanich Inlet water column

Alyse K. Hawley1,\*, Mónica Torres-Beltrán1,\*, Elena Zaikova2, David A. Walsh3,

Andreas Mueller1, Melanie Scofield1, Sam Kheirandish1, Chris Payne4, Larysa Pakhomova4,

Maya Bhatia1, Olena Shevchuk1, Esther A. Gies5, Diane Fairley1, Stephanie A. Malfatti6,

Angela D. Norbeck7, Heather M. Brewer7, Ljiljana Pasa-Tolic7, Tijana Glavina del Rio6,

Curtis A. Suttle1,4,8, Susannah Tringe6 & Steven J. Hallam1,9,10,11,12

* Oxygen minimal zones are expanding throughout the global ocean. This makes it difficult to define microbial metabolic networks that drive nutrient and energy cycles.
* Metagenomic sequencing allows for reconstruction of microbial metabolic potential at the gene and pathway level.
* These pathways have been around for very long and reflect storage of information in the environment.
* Just because the genes exist, do not mean however that there is active metabolism at a spot.
* Meta transcriptomic sequencing allows for gene expression to be followed.
* Meta proteomic sequencing allows for functional proteins that allow for catalysis to be tracked.
* The combination of omic information allows for tracking metabolism according to geochemical gradients.

**Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimatesemi\_**

**Victor Kunin,1 Anna Engelbrektson,1**

**Howard Ochman2 and Philip Hugenholtz1\***

* Pyrosequencing generates a large number of reads to provide high sample depth.
* The rare biosphere is much larger and more diverse than previously thought.
* However, pyrosequencing is prone to intrinsic errors that lead to overestimates in the number of rare phenotypes.
* This is because unlike genome sequencing, each pyroseq analysis is interpreted as a unique community member. This inflates diversity estimates.
* Quality filtering can help improve this problem by removing reads with one or more unresolved bases, errors in barcode/primer sequences, and abnormally long or short reads.
* Stringent read quality filtering and a low clustering threshold need to be applied to adjust estimates of diversity down.
* This is insufficient for ensuring accurate estimates of diversity. No greater than 97% identity should be used for avoiding overestimates of the rare biosphere.

Performance comparison of benchtop high-throughput sequencing platforms

Nicholas J Loman1, Raju V Misra2, Timothy J Dallman2, Chrystala Constantinidou1, Saheer Gharbia2,

John Wain2,3 & Mark J Pallen1

* Since arriving in the marketplace, these technologies have undergone sustained technical improvement, which, twinned with lively competition between alternative platforms
* three different benchtop high-throughput sequencing instruments are currently available, all capable of sequencing bacterial genomes in a matter of days
* 454 GS Junior from Roche was released in early 2010 and is a smaller, lower-throughput version of the 454 GS FLX machine lower set-up and running costs
* Ion Torrent Personal Genome Machine (PGM) was launched in early 2011 (ref. 5). Like the 454 GS Junior, this technology exploits emulsion PCR
* It also incorporates a sequencing-by-synthesis approach, but uses native dNTP chemistry and relies on a modified silicon chip to detect hydrogen ions released during base incorporation by DNA polymerase
* The MiSeq is based on the existing Solexa sequencing-by-synthesis chemistry6 but has dramatically reduced run times compared to the Illumina HiSeq
* The MiSeq had the highest throughput per run (1.6 Gb/ run, 60 Mb/h) and lowest error rates. The 454 GS Junior generated the longest reads (up to 600 bases) and most contiguous assemblies but had the lowest throughput (70 Mb/run, 9 Mb/h). Run in 100-bp mode, the Ion Torrent PGM had the highest throughput (80–100 Mb/h).

Introducing mothur: Open-Source, Platform-Independent,

Community-Supported Software for Describing and

Comparing Microbial Communities\_

Patrick D. Schloss,1,2\* Sarah L. Westcott,1,2 Thomas Ryabin,1 Justine R. Hall,3 Martin Hartmann,4

Emily B. Hollister,5 Ryan A. Lesniewski,6 Brian B. Oakley,7 Donovan H. Parks,8

Courtney J. Robinson,2 Jason W. Sahl,9 Blaz Stres,10 Gerhard G. Thallinger,11

David J. Van Horn,2 and Carolyn F. Weber12

* Mothur is a processing pipeline for analysis of sequencing data.
* It trims, screens and aligns sequences. It then calculates the distances and assigns operational taxonomic units to groups of sequences. It then describes the alpha and beta diversity of species.
* It is able to process pyrosequencing data, and in less than 2 hours for >200,000 sequences.

**Microbial diversity in the deep sea and the**

**underexplored ‘‘rare biosphere’’**

**Mitchell L. Sogin\*†, Hilary G. Morrison\*, Julie A. Huber\*, David Mark Welch\*, Susan M. Huse\*, Phillip R. Neal\*,**

**Jesus M. Arrieta‡§, and Gerhard J. Herndl‡**

* Microbial diversity in seawater are much more diverse than previously thought: it is one to two orders of magnitude higher than previously reported.
* Low abundance populations account for most of the observed phylogenetic diversity. This is known as the rare biosphere, and is the source of genomic innovation of the microbial world.
* Members in the rare biosphere are highly different from each other. At different points of earth’s history, they have had profound impacts on shaping the global metabolic processes.
* The figures are 100 times greater than cultivation surveys.
* At the genomic level, we compare three things: whole chromosome sequences, bacterial artificial chromosomes, and shotgun small inserts.
* There are unexpected levels of metabolic diversity and extensive horizontal gene transfer between species.
* There are these things called genetic reservoirs, and they shape the genome architecture of microbes in an environmental community through horizontal gene transfer.
* Microbes thus account for most of the genetic and metabolic variations in oceans
* There is much uncharted territory with regards to these.

A compendium of

geochemical information from the

Saanich Inlet water column

Mónica Torres-Beltrán1,\*, Alyse K. Hawley1,\*, David Capelle2, Elena Zaikova3,

David A. Walsh4, Andreas Mueller1, Melanie Scofield1, Chris Payne2, Larysa Pakhomova2,

Sam Kheirandish1, Jan Finke2, Maya Bhatia1, Olena Shevchuk1, Esther A. Gies5,

Diane Fairley1, Céline Michiels1, Curtis A. Suttle1,2,6, Frank Whitney7, Sean A. Crowe1,2,8,

Philippe D. Tortell2,6,9 & Steven J. Hallam1,8,9,10,11

this geochemical compendium is paired with a cognate compendium of multi-omic sequence

information (DNA, RNA, protein) focused on microbial diversity, abundance and function.38 Combined,

these compendiums provide a community-driven framework for observing and predicting microbial

community responses to changing levels of oxygen deficiency extensible to open ocean OMZs.

**Extensive mosaic structure revealed by the complete**

**genome sequence of uropathogenic Escherichia coli**

**R. A. Welch\*, V. Burland†‡, G. Plunkett III†, P. Redford\*, P. Roesch\*, D. Rasko§, E. L. Buckles¶, S.-R. Liou†**\_**, A. Boutin†\*\*,**

**J. Hackett†,††, D. Stroud†, G. F. Mayhew†, D. J. Rose†, S. Zhou†‡‡, D. C. Schwartz†‡‡, N. T. Perna§§, H. L. T. Mobley§,**

**M. S. Donnenberg¶, and F. R. Blattner†**

**A three-way genome comparison of**

**the CFT073, enterohemorrhagic *E. coli* EDL933, and laboratory**

**strain MG1655 reveals that, amazingly, only 39.2% of their combined**

**(nonredundant) set of proteins actually are common to all**

**three strains.**

**The pathogen genomes are as different from each**

**other as each pathogen is from the benign strain.**

**Striking differences exist**

**between the large pathogenicity islands of CFT073 and two other**

**well-studied uropathogenic *E. coli* strains, J96 and 536.**

**Comparisons**

**indicate that extraintestinal pathogenic *E. coli* arose independently**

**from multiple clonal lineages.**

**The different *E. coli* pathotypes**

**have maintained a remarkable synteny of common, vertically**

**evolved genes, whereas many islands interrupting this common**

**backbone have been acquired by different horizontal transfer**

**events in each strain.**

The Microbial Engines That Drive

Earth’s Biogeochemical Cycles

Paul G. Falkowski,1\* Tom Fenchel,2\* Edward F. Delong3\*

* Metabolic processes evolved early on in the earth’s evolutionary history exclusively through microbes.
* These evolutions came to alter the chemical makeup of virtually all of earth’s surface.
* Our current environment is the product of the historically integrated outcomes of microbial experimentation and evolution throughout history.
* This is the wellspring of life on the planet, and it is resilient to the most extreme of changes: microbes are likely to outlast humanity.
* They are essentially earth’s core biological machinery.
* We now analyze these machines through genomic sequencing.
* The grand challenge exists in studying how these machines evolved, how they interact with each other, and how they maintain earth’s elemental cycling.
* Contemporary pathways responsible for redox chemistry in microbes are encoded by key genes that encode the proteins that are the microbial machines.
* These are highly conserved at the secondary and primary structure. However, lateral gene transfer and extensive selection obscure the origins of these, including the order in which they appear.
* As such, it is very challenging to reconstruct electron transfer reactions.
* Identical or nearly identical pathways can be used in both forward and reverse reactions to maintain cycles.
* Molecular nutrient tension must exist in the environment for chemical reactions to progress favorably. Thus, there is often a synergistic special association of multiple species within an environment.
* Metabolic pathways driving biogeochemical cycles are not necessarily directly related. There can be multispecies interactions that drive these cycles through catalysis.
* Metabolic machines co-evolved. This means that cycles influence the outcomes of each other. This happens because microbes evolved to make use of multiple substrates at once for energetic needs. Sometimes intermediates generated from one pathway can be used for another.
* The outcome of all this is that the earth in time is able to create self-sustaining elemental cycling.
* The total life on earth thus derive two primary energy inputs: geological processes and solar energy.
* ATP complex probably evolved early on, with one of the first ancestors. However, oxygenic photosynthesis is probably one of the last.
* We currently roughly have an idea for how important each metabolic pathway is, but we don’t really have an answer for the order by which they evolved.
* Molecular evidence suggests that early cellular evolution was communal with lots of horizontal gene transfer.
* Entire metabolic pathways are frequently transferred over horizontally. Nutritional or bioenergetic serve as major drivers for retaining such genes.
* The full genome sequences can be mirrored by shotgun sequencing.
* The absolute number of genes and proteins encoding these machines is unknown. But, the number of protein families increases with genome size. This is related to evolutionary rate but not core metabolic processes.
* Discovery of new genes and protein families occurs linearly with new sampling efforts.
* There is a hypothesis that there is essentially a limitless diversity in nature. However, a more realistic illustration of the number of protein families in existence is that they eventually follow a trend of depreciating increases.
* So on the fact that you can tell that for different environments different microbes are going to exist. This is generally true, but if we look at this more closely, we notice that this is not always the case. The habitat’s selective pressure plays a greater role in the selection of protein in an area.
* Viable bacteria of any functional type can be recovered from almost any environment, even when the environment doesn’t seem to support their growth.

Remapping the body of the world

Lecture 02, Hallam, 2018, MICB 425.

* We can see the regulation of genes that encode metabolic pathways as a microbial food web, and is co-ordinated through feedback loops at every level.
* Also touches on a lot of what was said in the prev paper.
* In ecological design, there is redundancy. Metabolic pathways are not always needed but are held.
* Horizontal gene transfer is the mechanism by which distribution of metabolic diversity across networks occur.
* Organisms thus encode part but not all of a single process.

Foundational Questions

• *What is the taxonomic and functional structure of the*

*ecosystem?*

• *How does this structure change in response to*

*environmental perturbation?*

• *What are the ecological and biogeochemical*

*consequences of this change?*

• *What are relevant units of selection, conservation or*

*utilization within microbial communities?*

With respect to the microbial community

The use of multi-omics: DNA, RNA, protein help us

* answer these qns
* It is evident that we need to use more than a single method for evaluating what is in the community.
* FISH-SIMS: fluorescence in situ hybridization with secondary ion mass spec allows for analysis of colony composition through space.
* Stable isotope probing

Challenges

• Analysis is a Big Data problem

• Volume - billions of sequences

• Variety - environmental conditions, functional and taxonomic information,

distributed metabolic pathways

• Computation is heterogeneous

• Software - many different software have to work together

• Compute - local resources or High-performance compute (HPC)

• Parallelism - local (CPU) and cloud (distributed)

* *Need to maintain a data model throughout.*
* Assembly Problem

• Complex such as soil

* metagenomes don’t
* assemble very well.
* • Need a lot more raw
* sequence data and
* computational power.
* • Use unassembled short
* read data for gene
* finding and annotation.

The habitat and nature of early life

**E. G. Nisbet**\* **& N. H. Sleep**†

* The question of what is life can be answered from the view point that growth is life.
* The point where autocatalysis becomes recognized as life is hard to determine.
* Life on earth is thought to have begun by the last common ancestor, which is really a community of microbes that shared genomic information through horizontal gene transfer.
* This common ancestor split into archaea and bacteria.
* One of the things that horizontal gene transfer does is it links parts of the phylogenetic tree that would otherwise be very far apart from each other.
* A better model would be a mangrove instead of a tree or a delta, where clearly defined flows exist, but much cross-over also occurs.
* Molecular phylogeny can be calibrated by reference to the geological record.
* The first life we can image probably existed in a special zone neither too hot or cold, and also had many of the essential items needed for life to replicate itself.
* Evolution meant that once the first biofilms began to form and die off, these created new niches within which new kinds of life could evolve into.