Below is an outline of review article for submission by invite to the “Genetics of Halophilic Microorganisms” special issue of the journal *Genes*:

**Putative title:** “Applying genome-resolved metagenomics to de-convolute the halophilic microbiome”

**Authors:** Gherman Uritskiy, James Taylor, Jocelyne DiRuggiero

**Abstract:**

High-throughput sequencing has rapidly advanced our understanding of microbiome community structure and function. In particular, whole metagenomic (WMG) sequencing shows great potential to thoroughly investigate the taxonomic composition and functional potential of microbial communities. Compared to the more established 16S rRNA amplicon sequencing, WMG sequencing potentially allows for reconstruction of the entire DNA content of any microbial community, including individual metagenome-assembled genomes (MAGs). However, many qualities of halophilic microbial communities are major deterrents for effective use of WMGs to elucidate their structure. Many key steps in WMG data analysis, including metagenomic assembly and extraction of MAGs, are much more difficult in communities with a high GC content and complex strain heterogeneity, as seen in sigh-salt environments. Additionally, many existing functional and taxonomic annotation pipelines rely heavily on robust databases of closely-related organisms, something that is not available for many novel halophiles. However, the WMG bioinformatics algorithms and pipelines have been rapidly improving, and now allow researchers to reliably de-convolute of even the most complex communities, opening avenues of halophilic community research that were previously unavailable. CRISPR array evolution, virus-host interactions, community functional potential, strain dispersal and selection, and novel genome discovery are only some of the possible aspects of halophilic microbiomes that WMGs enables. However, there still remain potential pitfalls and limitations of conventional WMG analysis being applied to these complex communities, which require more specialized experimental and analytical strategies to overcome. Finally, several second- and third-generation sequencing technologies that are conventionally not intended for microbiomes, are becoming available to study halophilic communities. In particular, metatranscriptomics coupled with WMG show promise to uncover the metabolic activity of halophilic communities as well as individual strains, long read technologies such as Nanopore and PacBio technologies can aid in the assembly of highly similar microbial strains, and Hi-C can assist in the binning closely-related strains, and possibly reveal the chromosomal arrangement of more community members.

**Knowledge gap:**

WMG sequencing is gradually taking over microbiome research, as it has many benefits over conventional amplicon sequencing. However, these advances have been stalling in the halophile research, as metagenomic sequence data from salt-adapted microbiomes is difficult to analyze. Recently developed methods and software have overcome this, however there is a lack of an outline detailing how WMG sequencing can benefit halophile research and how to apply apply it effectively.

**Sections overview:**

Introduction

Here we will introduce whole metagenome (WMG) sequencing and the strengths it has over the more popular 16S rRNA sequencing. Advantages (new things that we can learn) and drawbacks (cost and complexity considerations) of WMG will also be addressed (2 paragraphs). We also introduce genome-resolved metagenomics - a narrower aspect of WMG concerning binning and de-constructing a metagenome at the level of metagenome-assembled genomes (MAGs) (1 paragraph).

How did MAG analysis improve our understanding of halophilic communities?

This section will be a review of key publications that investigated halophilic microbiomes with WMG. In particular, emphasis will be placed on the success or failures of these studies to uncover various aspects of their respective communities. Functional potential reconstruction, relative success of MAG extraction, and the types of information gleaned from the analysis will be talked about. This section will have 4 paragraph outlining key papers that tacked these aspects of their microbiomes (MAGs, functional potential, taxonomic composition, viruses). *Figure 1*: a flowchart comparing WMG and amplicon sequencing, showing the overall conceptual workflow of both analyses and the various aspects of the community can be uncovered (taxonomy, functional potential, OTUs, differential taxa or pathway abundance, etc).

What are the current best practices for WMG experiment design and sequence analysis of halophilic communities?

Broad overview of WMG experimental design, cost vs benefit considerations, replication options. Possible applications: taxonomic and functional characterization, organism or functions detection, genome reconstruction, longitudinal community comparison, etc. In general, there are two major ways to approach WMG experimental design and analysis. The first is to sequence many samples with shallow coverage, and co-assemble the metagenome, and then bin the co-assembly. This has the advantage of being able to extract organisms that are too low in abundance in any one sample, however does not scale well to very large data sets and suffers from chimeric assembly and binning, particularly in halophilic communities due to low kmer diversity. The other option is to sequence a few samples with deep coverage, and process them individually. This will produce more robust assemblies and therefore MAGs (particularly the higher abundant taxa), however the lower abundance diversity will be lost due to low coverage. This section will also briefly go over the overall workflow of such analysis. There will be one paragraph introducing the issue (you can’t have everything), two paragraphs for each approach detailing their experimental and computation aspects, and a conclusion paragraph summarizing which approach works best in different situations. *Figure 2:* Two flowcharts showing the two common experimental designs and analysis workflows.

What are novel technologies and methods in other fields that can further our ability to de-convolute halophile WMG data?

This section will cover three next-generation sequencing types that are currently only starting to be applied to microbial systems, but show great potential, based on the existing limited research. The first is metatranscriptomics via the sequencing of total RNA, or ribosome-depleted RNA (in communities where this is possible). This method can reveal the transcriptional activity of the entire community, and when combined with MAGs, can be used to look at differential expression in individual strains. The second method is long read sequencing, particularly Nanopore, which rapidly improved in the last few years. The newest flowcells can produce reads up to 1-2 Mbp, which is getting close to the genome size of whole organisms, if we can nick their chromosome once. The error rates in the reads are also rapidly dropping, especially when combined with conventional Illumina sequence data. Considering the difficulty of assembling and binning halophilic metagenomes, this technology can help substantially. Finally, Hi-C has recently also been applied to metagenomic data, with some promising results. The chromosomal capture maps can be used to improve binning predictions by extracting clusters of contigs that have high connectivity. Additionally, by applying the Hi-C data to individual MAGs, we can glean the rough internal chromosomal conformation of the more abundant community members. There will be one paragraph for each technique. *Figure 3*: Flowchart illustrating the types of information that can be obtained from third-generation sequencing approaches (scaffold-based assembly, scaffold-based binning, connectivity-based binning, etc).

Conclusion: WMG is an incredibly valuable tool to study halophiles

The main purpose behind this review paper will be to introduce WMG to researchers working with high-salt environments, and provide them with a framework of what to expect if they decide to apply WMG to their system. We will also cover the general approaches to data analysis that may be useful to biologists with limited bioinformatics experience. Finally, we will speculate on the general direction in which halophile research is moving in the context of rapidly improving technologies.