Sequencing on Cloud 9

Arya Kaul

May 2020

1 Background

1.1 Clouds

At any given point in time, roughly 70% of the Earth is covered by clouds.[1] Formed from an aggregate of cloud droplets and ice crystals, clouds play an indispensable role in the global hydrological cycle and in regulating the Earth's energy budget; as a result, accurately modeling their formation and geochemical properties constitutes a fundamental motivation in the field of atmospheric science.

1.2 Bioaerosols

In the 1970s it was shown that the bacteria *Pseudomonas Syringae* produces Ice Nucleation active (INA) proteins. These transmembrane proteins promote the ability of the bacterial cell to serve as an ice nucleator, and allow environmental water molecules to aggregate and freeze along the bacteria.[2] INA proteins were first thought to solely serve as vectors of attack for *Pseudomonas* to break plant cells open; however, atmospheric scientists have since demonstrated that this protein also allows the bacteria to serve as nucleators for clouds. Since this first discovery, a diverse assortment of microorganisms have also shown condensation and nucleation potential and are widely seen as playing a significant role in cloud formation.[2,3,4] Unfortunately, taxonomic characterization of these microorganisms and their biological properties have lagged behind. This is largely attributable to a focus on cultivable microorganisms and limited sampling.

1.3 Cumulonimbus

Clouds constitute a diverse genus of potential forms and properties dependent on the local atmospheric conditions that give rise to them. The World Meterological Organization currently recognize 10 canonical cloud types.[1] This proposal will focus on characterizing the microorganisms of the 9th cloud type, cumulonimbus, popularly known as thunderclouds. While further work should undoubtedly characterize the biomes of all 10 cloud types, properties of the cumulonimbus system better lend themselves to biological investigation. First, stability. The thunderstorm under investigation has a lifespan measured in months; significantly longer than other clouds with 'lives' typically measured in hours.[5] This consistency enables biological systems to fill more possible ecological niches as successive generations can adapt to the same environment. Second, cumulonimbus, when compared to other clouds, exhibits the largest vertical depth.[1] On average, thunderclouds are 12km tall and contain discrete layers representing distinct atmospheric properties. Prior work has indicated altitude is the largest determinant of aerosolized biological fauna, thus characterizing the complete thundercloud biome would produce generalizable knowledge about other cloud types whose altitudes are captured within the thundercloud.[6] Finally, the lightning that thunderclouds produce and how organisms respond to it could yield novel biological insight. Section 3.3 discusses this in detail.

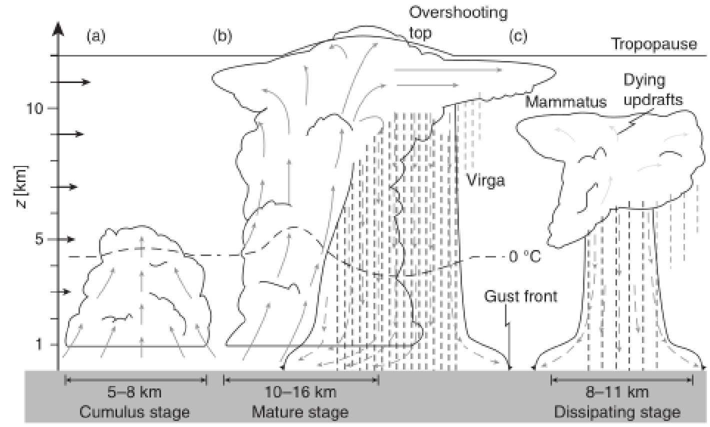

Figure 1: The three stages in the life cycle of an isolated single-cell thunderstorm. Gray arrows denote air motion and the black arrows along the y-axis represent strength of horizontal winds. Thin solid lines in (b) and (c) enclosing vertical broken lines represent the boundary of the precipitation-cooled air. Figure 10.1 in *An introduction to clouds*[1]

1.4 Lake Maracaibo & the Catatumbo Effect

In November 1997, NASA launched the Lightning Imaging Sensor on board the NASA Tropical Rainfall Measuring Mission. Analysis of 16 years worth of observations (1998-2013) reveal that Earth's top-ranked lightning hotspot is Lake Maracaibo in Venezuela, exhibiting an annual average of 233 flashes per square kilometer. Known as the Catatumbo Lightning phenomena, a number of unique geological and atmospheric conditions exist that produce reliable large thunderstorms in a seasonal manner [Figure 2].

Recently, using atmospheric data collected from tethered balloons, a group has shown that they are able to reliably predict lightning strikes within Lake Maracaibo up to a few months in advance.[8] Using these balloons (or similar ones of my own fashioning), the following project is proposed.

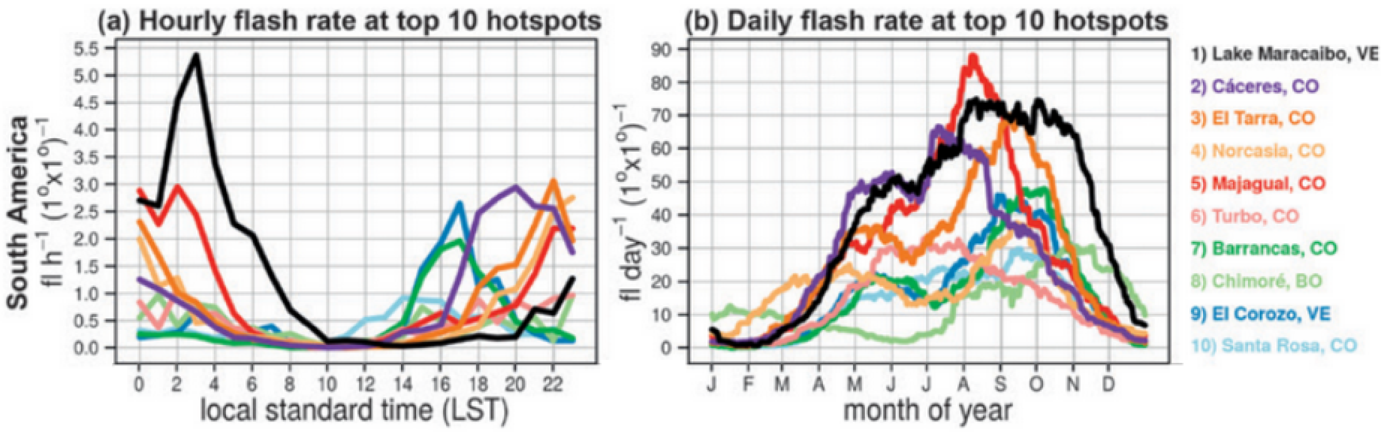

Figure 2: Hourly and daily flash rate densities of the top 10 lightning spots in South America. Values are calculated over a 1° box centered at the hotspot. Figure 3 in Albrecht et al.[7]

2 Project Proposal

Collect water/atmospheric air samples from the surface of Lake Maracaibo (9°, -71.65°) to 12km directly above the lake at intervals of 3km (total of 5 sampling zones). Perform this collection in mid October and early January at 0300 and 1600 local standard time.

2.1 Collection & Analysis Methodology

Water surface samples will be collected directly from the lake. Atmospheric samples will be harvested by the attachment of micro-weather stations attached to the lines of the tethered balloons.[8] Cooling condensation will be applied to atmospheric samples to harvest aerosolized water. All samples will be transferred to sterile cryovials, mixed with 10% glycerol for cryoprotection, flash-frozen in liquid nitrogen, and stored at -80°C.

Given the dearth of knowledge about the microbial fauna populating the thunderclouds above Lake Maracaibo, it would be prudent to utilize an unbiased sampling technique in the analysis. Luckily, a recent paper characterizing the microbial diversity present in the Earth's oceans made a compelling argument for the utilization of randomized single-cell sequencing to do exactly this.[9] First, current techniques for deconvoluting reads in metagenomic samples perform best when some organisms have been previously characterized; there is no evidence this will be the case for collected samples. Additionally, cleanly separating biosynthetic gene pathways becomes computationally intractable in such samples, such analysis will be necessary in Sections 3.2 and 3.3.1.[10]

Cryopreserved samples will be thawed and stained with SYTO-9 DNA stain for 10-60 min, after which a 40 μm filter and fluorescent-activated cell sorting (FACS) will filter microorganisms from other detritus. Upon individual cell sorting, individual nuclei will be lysed, amplified, and sequenced to generate Single Amplified Genomes (SAGs). All SAGs will be subject to paired-end sequencing on the Illumina platform. Sequencing will be conducted simultaneously at one facility to attempt to minimize batch effects. Upon sequencing, all SAGs will be assembled using SPAdes and assemblies will be analyzed for quality.[11] Only SAG assemblies with ≥50% predicted completion will be used for subsequent analysis.

Since sampling will take place during 2 months, at 2 times of day, and at 5 points, there will be a total of 20 (2*2*5) Populations of Single Amplified Genomes (PSAGs).

3 Specific Aims

Such data would represent the first characterization of the thundercloud biome, and also the very first interrogation of bioaerosols in a cultivation-independent manner. A number of hypotheses may be tested with the data; however, I believe the following questions represent the most exciting applications.

3.1 Meta-Population Analysis

To begin, population scale measures of the difference between each of the 20 PSAGs will be conducted. Specifically, within each PSAG, each SAG will be labeled with its triplicate sample collection label (January/October, Time of day, Altitude sampled from). Next, the average nucleotide identity (ANI) will be measured pairwise between each triplicate labelled SAG to each other SAG.[12] Distances between PSAGs will be computed as the average distance between the SAGs that make up each PSAG. The end result will be a 20x20 distance matrix comparing the distance between each PSAG to each other PSAG. To determine significance of each distance, label-shuffling and recomputation will be conducted 10,000 times to construct a null distribution. If the observed PSAG distance is greater than would be expected by random chance ($\leq$5% of simulations) then we conclude that the populations of single amplified genomes are significantly distinct. Additionally, hierarchical clustering will be performed to explore which factors explain the most difference between populations.

3.1.1 Thundercloud Uniqueness

The first question answered will be: *Does the unique environment within thunderclouds generate unique biological systems?* Mid-October samples are collected during peak lightning season, while January samples are taken in the middle of the Venezuelan dry season. [Figure 2] We hypothesize that the intensity and duration of thunderstorms will lead to a selection of microbes able to survive the conditions of a thunderstorm. If so, the most significant difference between PSAGs will occur between January populations and October populations. Further quantification may be performed by computing pairwise ANI solely using the January/October label and performing the same label-swapping experiment to determine significance. Such an experiment would be the first to test the effect of atmospheric change in bioaersols, and would encourage deeper analysis of the remaining 9 cloud biomes.

3.1.2 Atmospheric Selectors of Microbial Diversity

The next question answered will be: *What are the primary selectors of microbial diversity?* Prior longitudinal work analyzing the microbial composition of fog atop the puy de Dôme mountain in France concluded the primary factor determining atmospheric microbial fauna is altitude.[6] The authors attributed this distinction to a combination of selective pressures caused by increased UV exposure and cold temperatures; however, it was difficult to distinguish the relative contribution of each.

The unique conditions posed within thunderclouds allow a direct testing of both explanations. Amount of UV irradiation increases with altitude; however, contrary to popular belief, temperature does not always decrease similarly. The tropopause is the area between the the troposhere and stratosphere, and also happens to be a temperature inversion (a layer of air where temperature does not decrease with altitude).[1] The tropopause is the height limit of most thunderclouds, and occurs at roughly 12km above ground level [Figure 1]. If we observe a significant difference in PSAGs at the tropopause and the other altitudes, it would provide evidence that the primary selective factor of microbial fauna in the atmosphere is temperature. If no significant difference is observed, then that provides evidence that UV irradiation is the primary selective factor. (The effect of ultraviolet radiation will be discussed more in Section 3.2)

3.1.3 Bioaerosols within the Hydrological Cycle

Finally, *do bioaerosols participate in the hydrological cycle*? Though it is accepted that microorganisms play a role in cloud formation; it is unclear what the dominant mechanisms for bioaersol generation are and whether these bioaerosols participate in the hydrological cycle. Currently, it is hypothesized that bioaerosols are generated from a combination of aerosolized dry vegetation and film drops bubbling from nearby water sources and then precipitate to land to be aersolized again.[2] However, no data has been generated to support this hypothesis. By comparing the PSAGs taken from Maracaibo's lake water to PSAGs from the atmosphere one can quantify the similarity between atmospheric microbial fauna and terrestial microbial populations. If bioaerosols do indeed participate in the hydrological cycle, then we would expect to see increased similarity between atmospheric PSAGs and lake water PSAGs during the rainy season in October. If not, then aerosolized microbial populations should remain distinct from terrestial populations year-round.

3.2 DNA Damage

The ozone layer absorbs all UV-C radiation, and most UV-B radiation. This effect is substantially lessened as one approaches the stratosphere, and has been shown to influence microbial fauna at high altitudes.[6] UV radiation induces pyrimidine dimers and 6,4 photoproducts typically repaired by Nucleotide-Excision repair. In prokaryotic organisms, this repair pathway is carried out by the conserved UVr protein family.[13] Thus, the following question may be answered: *How do microorganisms adapt to continual UV exposure?*

To answer this, protein-coding gene prediction will be run on SGAs and predicted genes with significant homology to known UVr protein families will be computationally determined. Coding changes unique to the fauna at high altitudes would point to convergent evolution of UV tolerance. To increase the likelihood of identifying functional mutations, differences in each microbe's UVr proteins would be compared to the overall differences between the sampled genome and the closest characterized species match in genetic databases (as determined by whole genome ANI). If the mutations are neutral, then the changes to the UVr protein family should match the number of background mutations across the whole genome. Additionally, if we observe the same UVr mutations occurring between members of the same PSAG, then that points to convergenet evolution of UV tolerance mechanisms. If no coding changes are detected in the UVr protein family, it is possible that high altitude microbial life has evolved novel mechanisms that work in conjunction with the UVr pathway to repair or prevent DNA damage. In this case, the search for sequence similarity will be expanded to include orthologs known to confer protection, such as the *dsup* (damage suppressor) protein characterized in tardigrades.[14]

Understanding and characterizing the novel mechanisms evolved to tolerate high levels of UV radiation is an area of wide application. Given the rise of antibiotic-resistant bacteria, hospitals and wastewater treatment plants are increasingly relying on sterilization through applying ultraviolet germicidal irradiation. As a result, mapping the fitness landscape available to microorganisms to survive ultraviolet radiation and how we might circumvent their strategies is crucial. In addition, NASA, as part of their long term goal of exploration of the cosmos, is interested in fully characterizing the complete DNA damage prevention & repair pathways evolved by Earth-based life.[15] This analysis would help move the needle on both goals.

3.3 Electrobiology

Perhaps the most awe-inspiring feature of cumulonimbus is its lightning strikes. Though cloud-ground lightning is most readily visible to humans and subject to the bulk of analysis, cloud-cloud lightning actually represents the most common form of lightning.[1] Most thunderclouds exhibit a tripolar charge structure where the middle layer of the thundercloud retains a negative charge distinct from the positively charged top and bottom layers [Figure 3]. To equilibrate, a violent electrical discharge is generated. This discharge is visualized as lightning.

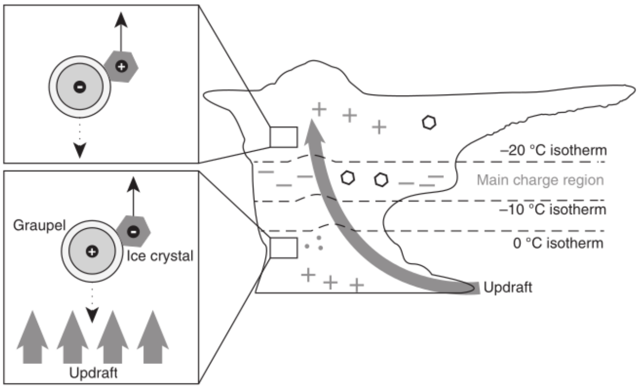

Figure 3: A possible generative mechanism of charge differential due to collision of graupel with an ice crystal. Graupel particles are more fragile versions of hail. Figure 10.5 in *An introduction to clouds*[1]

3.3.1 Membrane Permeabilization

Lightning strikes represent a rare but severe cause of injury and death worldwide. Non-fatal lightning injuries routinely present with long-term muscular and neurological conditions.[16] The current explanation for these conditions is electroporation affecting skeletal muscle and neuronal cells. Electroporation refers to the ability for cell membranes to permeabilize when exposed to a voltage differential. This fact has been exploited by biologists since the 80s to introduce foreign elements into cells; however, 40 years later the mechanism remains elusive.[17] Since the vast majority of lightning strikes occur within the cloud (cloud-cloud lightning), one can expect that the bioaerosols present within the cloud would be continually exposed to these same voltage differentials. Thus, the following question may be answered: *How do microorganisms adapt to constant lightning exposure?*

The obvious place to start would be the membrane itself. Towards this end, utilizing the protein-coding predictions described in Section 3.2, genes involved in fatty acid biosynthesis will be investigated. Specifically, the highly conserved set of proteins involved in the fatty acid synthase II (FASII) pathway. Prior work has already gone into purifying, characterizing, and understanding the mechanistic role of each protein.[18] Sequence similarity will be used to identify the homologs present in those SAGs collected during peak lightning exposure (October at 0300, see Figure 2). Specific coding changes to proteins within the FASII pathway conserved within lightning-exposed PSAGs will be collated and considered potential explanations for electrotolerance. To determine the significance of these mutations, similar mutations will be searched for within the PSAGs not exposed to lightning. If the mutations are neutral, then they should appear with similar frequency in other PSAGs. Further biological work will be required to causally link the identified mutations to electrotolerance; (for example inserting similar mutations into *E. Coli*) however, this work would provide needed preliminary evidence.

Understanding the mechanism behind resistance would help elucidate the mechanism behind electoporation, by identifying those factors required for electroporation. Such knowledge would prove useful for individuals interested in applying localized electroporation to selectively target disease cells with therapeutic intervention.[17]

3.3.2 Origin of Life

In 1953, Miller & Urey conducted their eponymous experiment attempting to recreate the conditions of prehistoric Earth.[19] One condition found to be necessary for the spontaneous generation of amino acids was electrical discharge. Further theoretical work indicated that lightning discharges were likely a dominant source of energy in prehisotric Earth.[20] One hypothesis for the origin of terrestrial life involves early protocells evolving in Earth's primordial oceans, with complex organic molecules synthesized in thunderclouds.[21,22] These protocells then aerosolized into the clouds, became exposed to these complex molecules, and then underwent rapid evolution to withstand the extreme conditions of the upper atmosphere. This rapid evolution is said to partly explain the abundance of speciation observed on planet Earth. If this hypothesis is to be believed, one would expect significant diversification in the fertile (yet harsh) environments provided by modern-thunderstorms. This brings us to the following question: *Do bioaerosol populations in thunderstorms exhibit unique diversification?*

To answer this question, within-PSAG distances will be computed. Specifically, the ANI between every pairwise combination of genomes within a given population will be computed. The result will be a distribution for each PSAG serving as a proxy for ecological diversity in a given sampling point. More diverse systems would exhibit less sequence similarity between members, and thus have a left-skewed within-PSAG distribution. To determine if thundercloud populations are significantly enriched for diversity, PSAG membership will be scrambled and within-PSAG distances recomputed. If the true distribution of within-PSAG distances remains significantly more left-skewed than the simulations then we may conclude that bioaerosols residing in thunderclouds exhibit significantly more diversification.

Testing this hypothesis has the added benefit of informing current exobiologists of the best places to search for extraterrestial life. There is evidence for lightning on Jupiter, Saturn, Uranus and Neptune, and it is possible on Venus and Titan.[23] Though theoretical exobiologists have postulated that the clouds of these solar bodies might contain organic compounds and life; this work would help future astronomical missions calibrate and search for life within these extraterrestial clouds.[21]

4 Broader Impact

Environmental sequencing has been applied to great effect in a variety of ecosystems; however, one of the most ubiquitous ecosystems on Earth has been tragically overlooked. The proposed work would motivate broader interest in investigating the interplay of biology in our atmosphere, and how this biology reacts to the remarkable conditions within cumulonimbus.

Though the data collected may be immediately used to resolve a number of questions facing biology, the proposed study also has the ability to reap benefits beyond the scientific.

Too often, scientists are content with enclosing themselves within their respective field, generating and consuming only that knowledge which immediately pertains to their area of interest. The Universe is a system; and all scientists, regardless of research subfield or departmental affiliation, are after fundamental truths regarding this same shared home. Actively engaging with one another and consciously crossing knowledge domains is not only critical to develop an accurate picture of our Universe, but also to construct a cohesive pan-scientific identity.

At the time of this writing, humanity is in the midst of combating a global pandemic. If there is one truth to derive from this moment of peril it is that science *matters*. Informed decisions, driven not by emotion, but by hard data and critical thinking *matter*. This is not some novel realization, nor a divine epiphany. For hundreds of years, scientists have wielded the scientific method as a torch. Systematically dispelling shadowy premonitions with the dispassionate light of scientific rationality and understanding.

If humanity is to overcome any of the multitude of existential threats threatening our survival (climate change, famine, disease, etc.) it will be on the back of what the scientific method represents: informed, deliberate, and rational decisionmaking. It is time for scientists to descend from their ivory towers and encourage society to share in enlightenment with them.

The impotent response to this pandemic from governments the world over only highlights how desparate the need is and how far we have yet to go. It is my sincere hope that the proposed work would encourage scientists from disparate disciplines to communicate with one another; and in doing so, understand that there exist more similarities that bind us than differences that divide us.

> Clouds are the patron goddesses of idle fellows
>
> *Aristophanes*

References

[1] Ulrike Lohmann, Felix Luond, and Fabian Mahrt. An introduction to clouds, 2016.

[2] Janine Fröhlich-Nowoisky, Christopher J. Kampf, Bettina Weber, J. Alex Huffman, Christopher Pöhlker, Meinrat O. Andreae, Naama Lang-Yona, Susannah M. Burrows, Sachin S. Gunthe, Wolfgang Elbert, and et al. Bioaerosols in the earth system: Climate, health, and ecosystem interactions. *Atmospheric Research*, 182:346–376, Dec 2016.

[3] S. S. Hirano and C. D. Upper. Bacteria in the leaf ecosystem with emphasis on pseudomonas syringae—a pathogen, ice nucleus, and epiphyte. *Microbiology and Molecular Biology Reviews*, 64(3):624–653, Sep 2000.

[4] Anne-Marie Delort, Mickael Vaïtilingom, Pierre Amato, Martine Sancelme, Marius Parazols, Gilles Mailhot, Paolo Laj, and Laurent Deguillaume. A short overview of the microbial population in clouds: Potential roles in atmospheric chemistry and nucleation processes. *Atmospheric Research*, 98(2–4):249–260, Nov 2010.

[5] Luiz Machado, W. Rossow, Roberto Guedes, and AW Walker. Life cycle variations of mesoscale convective systems over the americas. *Monthly Weather Review - MON WEATHER REV*, 126, Jun 1998.

[6] Pierre Amato, Marius Parazols, Martine Sancelme, Paolo Laj, Gilles Mailhot, and Anne-Marie Delort. Microorganisms isolated from the water phase of tropospheric clouds at the puy de dome: major groups and growth abilities at low temperatures. *Fems Microbiology Ecology*, 59(2):242–254, Feb 2007.

[7] Rachel I. Albrecht, Steven J. Goodman, Dennis E. Buechler, Richard J. Blakeslee, and Hugh J. Christian. Where are the lightning hotspots on earth? *Bulletin of the American Meteorological Society*, 97(11):2051–2068, Nov 2016.

[8] Á. G. Muñoz, J. Díaz-Lobatón, X. Chourio, and M. J. Stock. Seasonal prediction of lightning activity in north western venezuela: Large-scale versus local drivers. *Atmospheric Research*, 172–173:147–162, May 2016.

[9] Maria G. Pachiadaki, Julia M. Brown, Joseph Brown, Oliver Bezuidt, Paul M. Berube, Steven J. Biller, Nicole J. Poulton, Michael D. Burkart, James J. La Clair, Sallie W. Chisholm, and et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell*, 179(7):1623–1635.e11, Dec 2019.

[10] Ramunas Stepanauskas. Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology*, 15(5):613–620, Oct 2012.

[11] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, and et al. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, May 2012.

[12] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1), Nov 2018.

[13] A. Kuzminov. Recombinational repair of dna damage in escherichia coli and bacteriophage lambda. *Microbiology and Molecular Biology Reviews*, 63(4):751–+, Dec 1999.

[14] Takuma Hashimoto, Daiki D. Horikawa, Yuki Saito, Hirokazu Kuwahara, Hiroko Kozuka-Hata, Tadasu Shin-I, Yohei Minakuchi, Kazuko Ohishi, Ayuko Motoyama, Tomoyuki Aizu, and et al. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nature Communications*, 7(1), Sep 2016.

[15] Michael Johnson. Studying dna breaks to protect future space travelers, May 2019.

[16] Amber E. Ritenour, Melinda J. Morton, John G. McManus, David J. Barillo, and Leopoldo C. Cancio. Lightning injury: A review. *Burns*, 34(5):585–594, Aug 2008.

[17] Martin P. Stewart, Robert Langer, and Klavs F. Jensen. Intracellular delivery by membrane disruption: Mechanisms, strategies, and concepts. *Chemical Reviews*, 118(16):7409–7531, Jul 2018.

[18] Yong-Mei Zhang and Charles O. Rock. Membrane lipid homeostasis in bacteria. *Nature Reviews Microbiology*, 6(33):222–233, Mar 2008.

[19] S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529, May 1953.

[20] Christopher Chyba and Carl Sagan. Electrical energy sources for organic synthesis on the early earth. *Origins of Life and Evolution of the Biosphere*, 21(1):3–17, Jan 1991.

[21] Verne R. Oberbeck, John Marshall, and Thomas Shen. Prebiotic chemistry in clouds. *Journal of Molecular Evolution*, 32(4):296–303, Apr 1991.

[22] David J. Smith. Microbes in the upper atmosphere and unique opportunities for astrobiology research. *Astrobiology*, 13(10):981–990, Oct 2013.

[23] Karen L. Aplin. Atmospheric electrification in the solar system. *Surveys in Geophysics*, 27(1):63–108, Jan 2006.