A large-scale genome-based survey of acidophilic Bacteria suggests that genome streamlining is an adaption for life at low pH

Diego Cortez¹, Gonzalo Neira¹, Carolina González¹, Eva Vergara¹ and David S. Holmes^{1,2*}

¹Center for Bioinformatics and Genome Biology, Fundación Ciencia & Vida, Santiago, Chile

²Universidad San Sebastian, Santiago, Chile.

***Correspondence:**

David S. Holmes

dsholmes2000@yahoo.com

Center for Bioinformatics and Genome Biology

Fundación Ciencia & Vida

Zañartu 1482, Ñuñoa,

Santiago, Chile

Keywords: Genome Reduction; Genome Streamlining; Extremophile; Acidophile; Evolution of Acid Resistance; Chemolithoautotroph; Gene Gain and Loss; Protein Size Reduction and Expansion.

Abstract

Genome streamlining theory suggests that reduction of microbial genome size optimizes energy utilization in stressful environments. Although this hypothesis has been explored in several cases of low nutrient (oligotrophic) and high temperature environments, little work has been carried out on microorganisms from low pH environments and what has been reported is inconclusive. In this study, we performed a large-scale comparative genomics investigation of more than 260 bacterial high-quality genome sequences of acidophiles, together with genomes of their closest phylogenetic relatives that live at circum-neutral pH. A statistically supported correlation is reported between reduction of genome size and decreasing pH that we demonstrate is due to gene loss and reduced gene sizes. This trend is independent from other genome size constraints such as temperature and G+C content. Genome streamlining in the evolution of acidophilic Bacteria is thus supported by our results. Analyses of predicted COG categories and subcellular location predictions indicate that acidophiles have a lower representation of genes encoding extra-cellular proteins, signal transduction mechanisms and proteins with unknown function, but are enriched in inner membrane proteins, chaperones, basic metabolism, and core cellular functions. Contrary to other reports for genome streamlining, there was no significant change in paralog frequencies across pH. However, a detailed analysis of COG categories revealed a higher proportion of genes in acidophiles in the following categories: “Replication and repair”, “Amino acid transport” and “Intracellular trafficking”. This study brings increasing clarity regarding genomic adaptations of acidophiles to life at low pH while putting elements such as the reduction of average gene size under the spotlight of streamlining theory.

1. Introduction

Significant differences in genome sizes (number of base pairs per genome) have been detected between closely related lineages of prokaryotes isolated from a broad spectrum of environments and across multiple phylogenetic lineages, with genome sizes down to 1.2 Mb in free living Bacteria and differences of over 45% genome size between members from the same genus (Konstantinidis and Tiedje, 2004, Dufresne et. al., 2005, Lynch, 2006, Giovannoni et. al., 2014, Martinez-Cano et. al., 2015, Bentkowski et. al., 2015, Rodríguez-Gijón et. al., 2021). Small or reduced genomes, also termed streamlined genomes, have been widely observed in microorganisms adapted to live in low nutrient niches, such as cosmopolitan marine bacterioplankton (Giovannoni et. al., 2005, Schneiker et. al., 2006, Swan et. al., 2013, Luo et. al., 2014, Sun and Blanchard, 2014, Graham and Tully, 2021), rivers (Nakai et. al., 2016), slow growers in anoxic subsurfaces (Chivian et. al., 2008, McMurdie et. al., 2009), and in a wide range of extremophiles such as bacteria adapted to supersaturated silica (Saw et. al., 2008), halophiles (López-Pérez et. al. 2013, Min-Juan et. al., 2016), thermophiles (Sabath et. al., 2013, Saha et. al., 2015, Gu et. al., 2020), psychrophiles (Dsouza et. al., 2014, Goordial et. al., 2016), and alkaliphiles (Suzuki et. al., 2014). Differences in genome size have been reported for aerobes versus anaerobes (Nielsen et. al., 2021) and for microorganisms living in warmer versus cooler environments (Lear et. al., 2017, Sauer and Wang, 2019) and in bacterial pathogens (Murray et. al., 2021).

Streamlining theory proposes that genome reduction is a selective process these organisms undergo that promotes their evolutionary fitness (reviewed in Giovannoni et. al., 2014). The theory suggests that a smaller genome reduces the energy cost of replication and, by encoding fewer gene products, there is a concomitant reduction of cell size that could optimize transport and nutrient acquisition (Button, 1991, Sowell et. al., 2009). Some marine microorganisms with streamlined genomes have been found to have proportionately fewer genes encoding transcriptional regulators and an overall lower abundance of mRNA transcripts per cell, potentially reducing the cost of transcription and translation (Cottrell and Kirchman, 2016). These results are congruent with the observed correlation between regulatory network complexity and genome size (Konstantinidis and Tiedje, 2004). Genome size reduction is also observed in symbiotic microorganisms (Baker et. al., 2010, Gao et. al., 2014), but it has been theorized that this phenomenon differs to the streamlining of free-living bacteria as the former lose genes by genetic drift due to function redundancy between the host and the symbiont, while the latter would lose them by intense selective pressure (McCutcheon and Moran 2012, Giovannoni et. al., 2014), although recent evidence has argued otherwise (Gu et. al., 2020).

Any organism that grows optimally at low pH can technically be classified as an acidophile. However, because there are many neutrophiles (optimum growth ~pH 7) that successfully grow at around pH 6 or lower, it is useful from a practical point of view to define acidophiles as those microorganisms that grow optimally below pH 5 and make a distinction between moderate acidophiles that grow optimally between pH 5 and about pH 3.0 (Foster, 2004, Dopson, 2016, Benison et. al., 2021) and extreme acidophiles that grow below pH 3 (Johnson, 2007). The latter are particularly challenged for survival and growth as they face a proton concentration across their membranes of over 4 orders of magnitude (Baker-Austin and Dopson, 2007, Slonczewski et. al., 2009). Acidophilic microorganisms have been identified in all three domains of life (Johnson and Hallberg, 2003), but currently more genomic information is available for prokaryotic acidophiles (Archaea and Bacteria) (Cárdenas et. al., 2016, Neira et. al., 2020).

Our current understanding about genome streamlining in acidophiles comes from a limited number of observations. It has been reported that the genomes of several acidophilic microorganisms, such as

80 *Methylophilum*, *Ferroplasma* and *Leptospirillum* (domain Bacteria) and *Picrophilus* (domain
81 Archaea) are smaller (2.3, 1.9, 2.3 and 1.5 Mb, respectively) compared to their closest neutrophilic
82 phylogenetic relatives (Angelov and Liebl, 2006, Hou et. al., 2008, Ullrich et. al., 2016, Vergara et.
83 al., 2020). Genome reduction in acidophiles has been discussed as a mechanism to reduce energy
84 costs to survive in extremely low pH environments where organisms must deploy multiple energy-
85 intensive acid resistance mechanisms to maintain a circumneutral cytoplasmic pH (Hou et. al., 2008,
86 Ullrich et. al., 2016, Zhang et. al., 2017, Vergara et. al., 2020) while thriving in often nutrient scarce
87 and heavy metal polluted low pH environments (Johnson 1998, Dopson et. al., 2003, Johnson and
88 Hallberg, 2008). Despite this progress, there remains much to be discovered about genome reduction
89 in acidophiles. With the increased availability of genome sequences of acidophiles (Cárdenas et. al.,
90 2016, Neira et. al., 2020), we shed light on whether there is a statistically supported correlation of
91 genome reduction with low pH and, if so, what are its biological implications.

92 **2. Materials and Methods**

93 **2.1 Data procurement and management**

94 **2.1.1 Genome information**

95 Genomes of 345 bacterial acidophiles together with their associated growth and taxonomic data were
96 obtained from AcIDB (Neira et. al., 2020). This set of genomes was modified for the present study in
97 three ways: i) only free-living Bacteria were considered. For example, symbionts such as *Ca.*
98 *Micrarchaeum* were discarded; ii) organisms without an identified phylum affiliation were also
99 discarded and iii) seven new genomes and their associated metadata from acidophiles have been
100 added since the publication of AcIDB. This resulted in an initial dataset of 342 genomes of
101 acidophiles. In addition, 339 genomes were collected from non-acidophiles (growth optima, pH 5-8).
102 These included 222 genomes of neutrophiles (growth optima, pH 6-8) that were the closest
103 phylogenetic relatives to the acidophiles as identified using NCBI taxonomy (Schoch et. al., 2020),
104 GTDB (Chaumeil et. al., 2020) and AnnoTree (Mendler et. al., 2019), resulting in an equal
105 taxonomic representation of genomes of acidophiles and their neutrophilic phylogenetic relatives.
106 Genome sequences were downloaded from the National Center for Biotechnology Information
107 (NCBI) and the Joint Genome Institute (JGI). Genomes were filtered for quality using CheckM
108 v1.0.12 with cutoffs for completeness >80% and contamination <5% (Parks et. al., 2015). This
109 resulted in a final data set of 597 high quality bacterial genomes, comprising 264 genomes from
110 acidophiles (pH <5) and 333 genomes from non-acidophiles (pH 5-8). Genome information is
111 provided in Supplementary Table 1.

112 Genome average nucleotide identity (ANI) was determined using fastANI v1.3 with 4 threads (Jain
113 et. al., 2018). A cutoff of 95% average nucleotide identity was defined (Kim et. al., 2014) to group
114 identical or highly similar genomes into species clusters. Genomic characteristics, proteomic data and
115 associated metadata are reported as the means of each group for all plots. This reduced data bias due
116 to over-representation of some highly sequenced species.

117 **2.1.2 Growth pH and temperature**

118 Optimal growth pH and temperature of a species were downloaded from AcIDB (Neira et al., 2020).
119 For new species with sequenced genomes not yet deposited in AcIDB, information for optimal
120 growth pH and temperature was extracted from the literature. When no description of these optima
121 was available, they were defined as the midpoint of the growth range reported for the strain or closely

related strain as described by Neira et al., 2020. For metagenomes, the reported environmental data were used to determine optimum pH and temperature.

2.2 Proteome analyses

2.2.1 Protein annotations

Genome annotations were downloaded from NCBI (www.ncbi.nlm.nih.gov) or JGI (img.jgi.doe.gov). Genomes without an existing annotation were annotated with prokka v1.13.3 (Seemann, 2014). A proteome table was generated for each genome, that includes information for each predicted protein, including size, predicted subcellular localization, functional annotation with COGs and Pfams, COG category, presence of signal peptide and ortholog group. Unless stated, all software was run with default options.

2.2.2 Ortholog groups

To define ortholog groups, reciprocal BLASTP was performed within each genome by using all the proteins in its predicted proteome as queries against a database of the same proteins. A coverage of 50%, a sequence identity of 50% and an e-value of 10^{-5} were used as cutoffs (Tettellin et. al., 2005, Naz et. al., 2020). Protein pairs that follow these conditions were assigned to the same ortholog family if one or both were the best scored BLASTP hit of the other. Ortholog groups will also be referred as protein families.

2.2.3 Subcellular localization

Subcellular locations were assigned to each predicted protein using PSORTb v3.0 (Yu et. al., 2010), which predicts either cytoplasmatic, inner membrane, exported, outer membrane, periplasmic for gram negative Bacteria and cell wall for gram positive Bacteria. An “unknown” tag is assigned to proteins whose subcellular location could not be predicted. This was complemented with signal peptide identification, which was assigned using SignalP v5.0b that predicts the presence of signal peptides for translocation across the plasmatic membrane by either the Sec/SPI (standard system), Sec/SPII (lipoprotein signal peptide system) or the Tat/SPI (alternative system) translocation/signal peptidases (Almagro et. al., 2019). All three positive predictions were binned together and tagged as “Has Signal Peptide”. Proteins were sorted by both subcellular localization and signal peptide presence.

2.2.4 Pfam and COG functional annotations

Pfams were assigned to predicted proteins using Pfam_scan v1.6 (Finn et. al., 2016) under Pfam version 32.0 (El-Gebali et. al., 2019), which contains a total of 17929 different functional annotations including protein families and clans. An e-value of $<10^{-5}$ was applied as a cutoff for Pfam predictions of protein function. The pfam with the lowest e-value was assigned to each protein. COG annotations were assigned with the web tool eggNOG-mapper v5.0 (Huerta-Cepas et. al., 2019) under the December 2014 version of the COG database, which contains 4632 functional annotations (Galperin et. al., 2015). The percentage of ortholog groups that have a Pfam assignment (Mistry et. al., 2021) or a COG assignment (Galperin et. al., 2021) were calculated for each proteome. The percentage of ortholog groups belonging to each COG category was also calculated. In addition, Pfam assignments were used for the analysis of intra-protein family size variation and to determine the percentage of proteins with an annotation.

2.2.5 Paralog frequencies

Paralog families were defined as ortholog groups with two or more proteins from the same proteome. The percentage of proteins that belong in paralog families was calculated for each COG category in relation to the total number of proteins in the category. The same procedure was repeated for the full proteome.

2.3 Statistical analyses

A python script was developed to gather, filter, organize and analyze the data from the organisms' genomes and proteomes (van Rossum, 1995). Data distributions were statistically analyzed using the following methods. The scipy library (Virtanen et. al., 2020) was used for linear fittings (with the “*linregress*” module), binomial test (with the “*stats.binom_test*” module) and Pearson's linear correlation coefficient (with the “*stats.pearsonr*” module). A two-sided mode was used for all the tests. P-value thresholds used for statistical significance were 0.05, 0.01 and 0.001. For estimation of correlation in potentially heteroscedastic distributions, a Generalized Least Squares was applied using the module “*regression.linear_model.GLS*” within the statsmodels library (Seabold and Perktold, 2010). For multi-testing analyses, the false discovery rate (FDR) was used to determine statistical significance using the Benjamini/Hochberg procedure (Benjamini and Hochberg, 1995) with the “*stats.multitest.multipletests*” module also within the statsmodels library. A q-value of 0.05 was used for Pearson's correlation p-values. The q-value is the upper limit of the rate of the findings (null hypothesis rejections) that is expected to be a false positive. Principal component analysis (PCA) was performed with the “*decomposition.PCA*” module within the sklearn library (Pedregosa et. al., 2011). The number of components for dimensionality reduction was set to 2. Data was plotted using the matplotlib library (Hunter, 2007).

3. Results and Discussion

3.1 Phylogenetic distribution and associated metadata of genomes interrogated

From the 342 publicly available genomic sequences (264 high quality plus 78 low-quality genomes) of acidophilic Bacteria, 331 genomes with well-defined taxonomies (phylum and class) were mapped on to a rooted cladogram (Figure 1). The genome sequences come from 177 species distributed in 17 classes and 8 phyla out of a total of 37 recognized bacterial phyla (55 if candidate phyla are included) (Schoch et. al., 2020) (Figure 1 and Supplementary Table 1). The acidophiles are widely distributed in the cladogram supporting the idea that acidophile lineages have emerged independently multiple times during evolution (Cárdenas et. al., 2016, González et. al., 2016, Colman et. al., 2018, Khaleque et. al., 2019, Vergara et. al., 2020).

Figure 2 shows the distribution of acidophilic species with sequenced genomes by phylum across pH, where pH represents the optimum for growth for each species. The total number of species declines from about 60 species in the range pH 4-5 to about 10 at pH 0.5-1.5 (Figure 2A) consistent with the observation that species diversity declines in low pH environments (Bond et. al., 2000, Baker and Banfield, 2003, Johnson and Hallberg, 2003, Méndez-García et. al., 2014, Lukhele et. al., 2020, Hedrich and Schippers, 2021). These estimates are based on the distribution of acidophiles with publicly available sequenced genomes; the true richness of acidophile diversity is likely to be much higher and will probably increase as more acidic niches are sampled using metagenomics approaches.

Figure 2B shows the distribution of species by percentage across pH. The results have been divided into three sections (a-c) for discussion. Section (a) with a pH range of 1.0 to 2.0 is dominated by species in the phyla Proteobacteria, Firmicutes and Nitrospirae in approximately equal proportions around pH 2 and by Firmicutes at pH 1. Section (b) shows the species distribution in the range pH 2 to 4. Acidophilic species of the phylum Proteobacteria are the most prevalent in this range but exhibit a declining percentage with decreasing pH. Species of Actinobacteria and Verrucomicrobia are represented about equally but both phyla have few representatives below pH 2. Species of Aquificae are present in a low percentage (~ 3%) down to about pH 3, beyond which there are no representative genomes. Section (c) shows the species distribution in the range pH 4 to 5. All seven phyla (eight, if one includes the one species from Armatimonadetes) have species in this range but Acidobacteria show a declining percentage from pH 5 to pH 4 below which there are no representative genomes.

3.2 Genome size as a function of pH

A scatterplot of genome size across optimal growth pH shows declining genome sizes from about 4.5Mb for circum-neutrophiles to an average of about 3.4Mb for extreme acidophiles (Figure 3). There are no large genomes (>5Mb) for bacteria that grow below about pH 4, whereas large genomes including up to about 10Mb are present in acidophiles that grow between pH 4 and pH 5 and in neutrophilic relatives of the acidophiles that grow from pH 5 to pH 8. A linear regression model fitted to the data shows a tendency that is statistically significant with a positive Pearson's correlation coefficient of 0.19 and a p-value of 2.97×10^{-5} , implying genomes are smaller at lower pH. However, there is evidence of heteroscedasticity in the plot. We applied Generalized Least Squares Regression (GLS) to take into account heteroscedasticity, and a p-value of 1.8×10^{-3} was obtained supporting the proposed relationship between pH and genome size.

However, the presence of heteroscedasticity suggests the possibility that other variables, in addition to pH, may contribute to the determination of genome size. To address this issue, we investigated

potential contributions of growth temperature and genomic G+C content on the distribution of genome size across pH. Many acidophiles are also moderate or even extreme thermophiles (Johnson and Hallberg, 2003, Capece et. al., 2013, Colman et. al., 2018) and temperature has been suggested to be a driving force for genome reduction (Sabath et. al., 2013). Genome size has also been associated with G+C content, where organisms with relatively low genomic G+C content tend to have smaller genomes (Veloso et. al., 2005, Almpanis et. al., 2018).

We evaluated how these factors are correlated with genome size and pH. Temperature is negatively correlated with genome size (Pearson's correlation coefficient, -0.34; p-value, 2.9×10^{-13}) (Figure 4A) and G+C is positively correlated with genome size (Pearson's correlation coefficient, 0.48, p-value 1.9×10^{-25}) (Figure 4C). A negative correlation between genome size and temperature has recently been reported for extreme acidophiles of the *Acidithiobacillus* genus (Sriaporn et. al., 2021). However, no statistically supported correlation is observed between temperature and pH (Pearson's correlation coefficient, -0.01; p-value 0.84) (Figure 4B), nor between G+C content and pH (Pearson's correlation coefficient, -0.06; p-value 0.22) (Figure 4D). Therefore, while both temperature and G+C content have a strong influence on genome size, they appear to act independently of the relationship between pH and genome size.

To investigate further the interplay of pH, temperature and G+C content with genome size, we performed dimensionality reduction and visualization via principal component analysis (PCA) (Jolliffe, 2005). As seen in Figure 5, the directions of the loading vectors show temperature is negatively correlated with both G+C content and genome size, while genome size is positively correlated with both G+C content and pH. This is also depicted in how the smallest genomes are found in thermophiles (optimal temperature $>55^{\circ}\text{C}$, rightmost cluster) followed by extreme acidophiles (optimal pH <3 , upmost cluster), while the biggest genomes are found in a high G+C content group (leftmost cluster). Conversely, the orthogonality of the loading vectors suggests no correlation is observed between pH and temperature or between pH and G+C content. Therefore, when considering all variables at once, the same results are observed as when the variables were individually assessed (Figure 4), providing additional evidence that neither G+C content nor temperature affect the correlation between pH and genome size, rather multiple driving forces can independently exert their influence on genome size.

3.3 Genetic mechanisms involved in genome size changes

3.3.1 Hypothetical schema

Given the observation that genome size is negatively correlated with pH in acidophiles, we aimed to determine what genomic processes influence this relationship. Figure 6 shows a diagrammatic representation of genetic mechanisms that have been postulated to be involved in genome expansion or reduction in Bacteria and Archaea (Keeling and Slamovits, 2005, Sabath et. al., 2013, Giovannoni et. al., 2014, Gillings, 2017, Kirchberger et. al., 2020, Rodríguez-Gijón et. al., 2021, Westoby et. al., 2021). Genome size changes could result from having (i) changes in number of orthologous families (A, Figure 6) or paralogous genes (B, Figure 6); (ii) genome compaction/expansion resulting from changes in the number of intergenic nucleotides including alteration in the frequency of overlapping genes (C, Figure 6) (reviewed in Kirchberger et. al., 2020) and (iii) smaller or larger genes, including loss/gain of domains (D, Figure 6).

Based on the schema shown in Figure 6, we investigated the contribution of the different mechanisms in genome size changes in acidophiles across pH. Annotated open reading frames (ORFs) were used

as surrogates for “genes”. A caveat is that ORF prediction depends on the quality of the genome sequence, where poor quality genomes frequently have incorrectly annotated chimeric and truncated ORFs that confound subsequent identification of genes (Klassen and Currie, 2013). We minimized these potential errors by analyzing only genomes that had passed a high quality CheckM filter (Parks et. al., 2015). However, even high-quality genomes are prone to errors of ORF annotation especially in the identification of correct translation start sites (Korandla et. al., 2020) which will impact predictions of gene and intergenic spacer sizes. Currently, there are no computational program for ORF prediction that is flawless, including GenBank (Korandla et. al., 2020), and we expect that future work will improve the annotations of ORFs used in our study.

3.3.2 Reduction/expansion of gene (ORF) number

The number of protein coding genes (ORFs) of each genome under interrogation was plotted as a function of optimal growth pH of the species. The results indicate that there is a statistically significant reduction (Pearson’s coef.: 0.18; P-value: 1.25×10^{-4}) of the average number of ORFs per organism across pH from an average of about 4100 ORFs/organism at pH 7 to about 3200 ORFs/organism at pH 2 (Figure 7A). This has been regarded as possibly the most predominant mechanism for genome size changes (Konstantinidis and Tiedje, 2004) and this is likely also true for our dataset (Supplementary Figure 1).

3.3.3 Reduction of intergenic spacers as a possible contributor to genome compactness.

It is well established that bacteria have compact genomes with an average protein-coding density of 87 % with a typical range of 85–90 % (McCutcheon and Moran 2012). Genome size reduction could occur by decreasing the amount of DNA occupied by intergenic spacers, for example by promoting the frequency of overlapping genes (Veloso et. al., 2005, Saha et. al., 2015, Kreitmeier et. al., 2021). This strategy has been especially exploited in compacting viral genomes (Pavesi, 2021).

To evaluate whether a reduction in the fraction of the genome dedicated to non-protein coding DNA contributed to smaller genomes observed in acidophiles, we calculated the percentage of intergenic spaces (IG) dedicated to the total genome content across pH. IG was calculated as genome size (bp) - $\sum$ bps of all ORFs in a genome, expressed as a percentage of the total bps in the genome. A smaller % IG implies greater genome compaction. A tendency was observed for % IG to increase as pH growth optima declines (Figure 7B), however, this trend is not statistically significant (Pearson’s coef. = -0.11, p-value 0.06). A potential problem in the interpretation of this result stems from uncertainties in the identification of ORFs, most notably by errors in the identification of the correct site of initiation of protein coding regions (Korandla et. al., 2020). This influences the estimation of the percentage of intergenic genomic DNA.

3.3.4 Reduction/increase of gene (ORF) size

The average size of ORFs (as the number of amino acids of the predicted proteins) per genome was plotted as a function of pH (Figure 7C). There is a statistically supported positive correlation (p-value 4.03×10^{-8}) between average ORF size and pH, with an average size of 320 amino acids at pH 7 to 300 at pH 2. This indicates acidophiles have shorter proteins in average, which could be produced by a loss of larger proteins or by gene size reduction (Figure 6, mechanism D) or possibly both.

To quantify gene size reduction in acidophiles, we analyzed the protein sizes of several conserved Pfams in the dataset (Figure 8). We observed that the conserved Pfams with reduced protein sizes in acidophiles are over 5 times as many as the conserved Pfams with increased sizes (Figure 8 A,

binomial test p-value 2.1×10^{-13}). This result accounts mainly for changes in the predominant domain architectures, implying these proteins in acidophiles likely have fewer domains. For example, the biotin requiring enzyme was mainly found in single domain proteins below pH 5, while in neutrophiles it can often be found next to other domains such as the dihydrolipoamide acyltransferase (Supplementary Table 3). This inclination towards protein size reduction is also observed in a collection of conserved Pfams that are also in single copy and predominantly in single domain architectures (Figure 8 B, binomial test p-value 7.4×10^{-3}). This result accounts mainly for loop size reductions and domain size reductions. Such is the case of the ribosomal protein L19 that in acidophiles lacks long loops and is 4 amino acids shorter on average (Supplementary Table 4).

3.4 Gene representativity across pH

Having established that there is a statistically supported negative correlation between genome size and optimal pH for growth and that gene gain and loss events likely contributed to this correlation, we investigated in more detail what types of genes were involved these events.

3.4.1 Changes in ortholog groups representativity in acidophiles

To gain insight into the contribution of gains or losses of genes in the observed genome size changes of acidophiles (mechanism A, Figure 6), we first clustered the genes into ortholog families and systematically classified the predicted proteomes of each genome by (i) subcellular location and (ii) functional category as predicted by Pfam annotations (Mistry et. al., 2021) and COG categories (Galperin et. al., 2015). Subsequently, we mapped the frequencies of ortholog families of these categories in the genomes across pH.

3.4.1.1 Changes in ortholog frequencies by sub-cellular location

Figure 9 shows the frequency of occurrence of protein families with sub-cellular location and/or signal peptide predictions expressed as a percentage of the total protein families per genome. The frequency of proteins predicted to be in the cytoplasm does not change across pH (blue data points and line, Figure 9). However, there is a statistically significant decrease (Pearson's correlation coefficient 0.22, p-value 1.4×10^{-6}) in the frequency of proteins predicted to have a signal peptide with decreasing pH (red data points and line, Figure 9) and a statistically significant increase (Pearson's correlation coefficient -0.19, p-value 4.4×10^{-5}) in the frequency of proteins predicted to be in the inner membrane with decreasing pH (orange data points and line, Figure 9). There is a small, but nevertheless statistically significant decrease (Pearson's correlation coefficient 0.21, p-value 7.5×10^{-6}) in the frequency of proteins predicted to be in the category "periplasm, outer membrane, cell wall and exported" with decreasing pH (green data points and line, Figure 9).

The decrease in proportion of proteins with signal peptides at low pH (Figure 9) is consistent with the observation that there are correspondingly fewer proteins predicted in the category "periplasm, outer membrane, cell wall and exported" at low pH since most of these proteins require a signal peptide export mechanism to pass through the periplasmic membrane (Green and Mecsas 2016). We hypothesize that the decrease in relative frequency of proteins found outside the inner membrane in acidophiles could be due to physico-chemical challenges that such proteins would encounter as they are exposed to high concentrations of protons at low pH, potentially limiting the diversity of proteins that have evolved to survive such challenges (D'Abusco et. al., 2005, Chi et. al., 2007, Duarte et. al., 2009, 2011, Panja et. al., 2020, Chowhan et. al., 2021). We speculate that the observed enrichment of protein families predicted to be in the inner membrane in acidophiles (Figure 9) reflects the

importance of such proteins in acid stress management (Lund et. al., 2014, Zhang et. Al., 2016, Vergara et. al., 2020, Hu et. al., 2020).

3.4.1.2 Changes in ortholog frequencies by functional category

The contribution of gene gain or loss to genome size changes across pH were also analyzed using gene functional classification using COG and Pfam annotations. 25 functional categories are recognized in the 2014 COG database (Galperin et. al., 2015) and Pfam v32.0 contains a total of 17,929 families (El-Gebali et. al., 2019, <https://pfam.xfam.org>). The combination of COG and Pfam analyses provides deep and accurate coverage for searching for predicted protein function in our dataset. Figure 10 shows that the percentage of proteins per genome with a COG or Pfam annotation decreases at lower pH with statistical significance (Pearson's correlation coefficients 0.24 and 0.14, p-values 2×10^{-7} and 2.6×10^{-3}), which is not observed for small neutrophilic genomes (Supplementary Figure 3). This indicates that acidophiles have a higher proportion of putative protein coding genes that are not recognized by neither COG nor Pfam. These proteins can be classified as non-conserved, hypothetical proteins with no functional prediction, which do not have protein clusters with sufficient entries to have their own functional annotation in the COG or Pfam databases. It is possible that some of these represent poorly annotated sequences and pseudogenes. However, an intriguing possibility is that some could correspond to *bona fide* protein coding genes that are enriched in acidophiles. Their analysis could potentially yield clues about novel acid-tolerance mechanisms and other functions enriched in acidophiles. Examples of such proteins have recently been detected, although their function remain unknown (González et. al., 2016, Vergara et. al., 2020).

An analysis of the distribution of functional categories across pH using COGs shows that acidophiles are enriched in several functions that could possibly be attributed to their distinctive metabolisms and environmental challenges (Table 1). For example, enrichment in COG L (replication, recombination, and repair) and COG O (Chaperone, post-translational modification) might reflect their need for DNA repair and protein refolding when confronted by potentially damaging stresses such as low pH, high metal concentrations and oxidative stress (Crossman et. al., 2004, Baker-Austin and Dopson, 2007, Cárdenas et. al., 2012, Dopson and Holmes, 2014). The increase in frequency of COGs C, F and H (Energy production and transport; nucleotide metabolism and transport and coenzyme metabolism and transport, respectively) could reflect enzyme and pathway requirements associated with obligate autotrophic metabolism that has been found in many acidophiles (Johnson, 1998, Johnson and Hallberg 2008). As for COG J, it is possible that as ribosomal proteins are very conserved across prokaryotic life (Lecompte et. al., 2002), they are less likely to be discarded. Future research could investigate what are the functions in this category overrepresented in acidophiles.

On the contrary, genomes of acidophiles are depleted in COG T (Signal transduction mechanisms). A depletion of signal transduction mechanisms has been observed in some marine microbes especially those that are slow growing (Gifford et. al., 2013, Cottrell and Kirchman, 2016), in the streamlined genome of the extreme acidophile *Methylophilum inferorum* (Hou et. al., 2008) and in metagenomic profiling data of acidic environments (Chen et. al., 2015). The abundance of signal transduction mechanisms generally declines with decreasing genome size, as it has been found that the number of one and two component signal transduction systems is proportional to the square of the genome size (Konstantinidis and Tiedje, 2004, Ulrich et. al., 2005, Galperin, 2005). Extensive research has been conducted on the different signal pathways and regulatory networks of acidophiles (Rzhapishvskaya et. al., 2007, Shmaryahu et. al., 2009, Moinier et. al., 2017, Díaz et. al., 2018, Osorio et. al., 2019). However, additional research is needed to uncover what signal pathways are not present in these organisms. Acidophiles possess several features which may explain their

400 underrepresentation in proteins from this category, such as having small genomes, and having
401 relatively slow growth speeds (Fang et. al., 2006, Mykytczuk et. al., 2010).

402 The genomes of acidophiles also have a proportionately reduced number of COG S (unknown
403 function). These are proteins with unknown function that are conserved across multiple species and
404 so are distinct from the category described above (Figure 10) that are not conserved across multiple
405 species. As both are proteins with no known function, the representativity of unknown function
406 proteins remains relatively constant across pH, but a greater number of these proteins are in multiple
407 species in neutrophiles. It is possible that many functions assigned to COG S are found principally in
408 neutrophilic heterotrophs whose genome sequences are the most prevalent in databases (extrapolated
409 from the limited number of genomic sequences of acidophiles, Neira et. al., 2020) and therefore can
410 potentially dominate the COG database.

411 **3.4.2 Paralog frequency across pH**

412 We next examined whether the gain or loss of paralogs contributed to genome size changes
413 (mechanism B, Figure 6). In contrast to what has been described above concerning gain or loss of
414 specific COG and Pfam gene functions, here we explored how genome size could be influenced by
415 the expansion or contraction of the number of genes in such families. Gene duplication, followed by
416 functional diversification has been invoked as a major contributor to gene evolution (reviewed in
417 Innan and Kondrashov, 2010 and Copley, 2020) and gene paralogs can be present as a significant
418 proportion of a genome (Swan et. al., 2013). An increase in the number of paralogous protein copies
419 (including in- and out- paralogs and xenologs, Remm et. al., 2001, Darby et. al., 2017) has been
420 observed to be correlated with a better performance in a specific function, such as heavy metal
421 resistance or adaptation to other multiple stressors (Kondratyeva et. al., 1995, Dulmage et. al., 2018).
422 Relatively high paralog frequencies for proteins linked to acid resistance mechanisms have been
423 detected in acidophiles (Ullrich et. al., 2016, Vergara et. al., 2020).

424 We analyzed paralog frequency changes in genomes across pH by COG categories. The COG
425 annotation has been proved useful for gene enrichment analyses across several genomes (Galperin et.
426 al., 2021). As seen in Figure 11 and Supplementary Figure 5, acidophiles have relatively high paralog
427 frequencies in the COG categories “Replication, repair and recombination”, “Intracellular trafficking
428 and secretion” and “Energy production and conversion”, but low frequencies in the COG categories
429 “Signal transduction”, “Translation and ribosome” and “Amino acid metabolism”, as shown by
430 statistically significant correlations (p-value <0.01). Some of the results are in concordance with the
431 protein family representativity results (Table 1) which increases the importance of the putative
432 contribution of these functions on acidophilic survival and adaptation.

433 High paralog frequencies in the “Replication, repair and recombination” category in acidophiles
434 might be attributed to a large number of transposases and integrases. The high prevalence of mobile
435 elements in acidophilic genomes has been previously pointed out as a key factor for acidophilic
436 evolution (Aliaga et. al., 2009, Navarro et. al., 2013, Acuña et. al., 2013, Ullrich et. al., 2016, Zhang
437 et. al., 2017, Colman et. al., 2018, Vergara et. al., 2020). As discussed in the previous section (Table
438 1), DNA repair proteins might also be in several copies. These have been found to protect against
439 oxidative stress and heavy metal stress, which acidophiles are exposed to in higher levels (Crossman
440 et. al., 2004, Baker-Austin and Dopson, 2007, Cárdenas et. al., 2012).

441 The increased number of paralogous proteins from the “Intracellular trafficking and secretion”
442 category in the acidophile genomes could result from an abundance of type II secretory systems

involved in conjugation. It has been observed that these systems are frequently associated with mobile elements and are found to be particularly abundant in the flexible genomes of acidophiles (Acuña et. al., 2013, Beard et. al., 2021), suggesting that they are shared between organisms in a common econiche. In addition, vesicle related proteins might also be duplicated in acidophilic genomes, as studies show that vesicular transport (whose related functions belong in this category) is linked to biofilm formation (Jan, 2017), which in turn has been widely observed in acidophiles (Baker-Austin et. al., 2010, González et. al., 2013, Díaz et. al., 2018, Vargas-Straube et. al., 2020).

Similarly to the results of genome representativity (Table 1), the increased paralog frequencies of proteins from the “Energy production and conversion” category in acidophiles, might be related with their overrepresentation of chemolithotrophic metabolism. Some of the enzymes involved in iron or sulfur oxidation belong to this category, such as the cytochrome C, heterodisulfide reductase and quinone related proteins (Quatrini et. al., 2009, Zhan et. al., 2019). Additionally, several proteins in this category are involved in proton exporting functions, such as the H⁺-ATPase and the overall electron transfer chain proteins such as the ubiquinone oxidoreductase (Walker, 1992, Fütterer et. al., 2004, Feng et. al., 2015). This indicates that some genes in this category might be in high copy numbers to increase the acid resistance of acidophiles. Alternatively, it could be a consequence of the high energy requirements of maintaining a neutral internal pH (Baker-Austin and Dopson, 2007, Slonczewski et. al., 2009).

The reduced paralog frequencies in the “Signal transduction” category are concordant with their reduced genome representativity in acidophiles, and thus might be accounted by the same phenomena as previously exposed (Table 1).

The reduced number of paralogs in acidophiles in COG E “Amino acid transport and metabolism”, might be accounted for by a reduction in the number of amino acid importers that are not common in acidophiles. The predominancy of autotrophic metabolism in acidophiles could result in an inclination for these organisms towards biosynthesis of amino acids rather than uptake by active transporters. Additionally, uptake of amino acids could be harmful to acidophiles as organic acids carry protons into the cytoplasm of these organisms, short circuiting acid resistance mechanisms (Kishimoto et. al., 1990, Lehtovirta-Morley et. al., 2014, Carere et. al., 2021). The current hypothesis is that organic acids are protonated in the extremely acid medium where acidophiles grow (pH <3) becoming non-ionic and soluble in bacterial membranes, permitting diffusion into the cytoplasm (pH around 7) where they uncouple from the proton. A similar phenomenon could occur with amino acids but involving membrane transporters, as amino acids are unlikely to diffuse passively through the membrane.

As for COG J “Translation and ribosome”, their reduced paralog frequency is opposite to the increased representativity of protein families from this category in the genomes of acidophiles (Table 1). In other words, acidophiles tend to discard (or not evolve) duplicated genes from this category rather than losing core functions by relinquishing unique protein families. Further exploration is needed to determine what are the changes this category in acidophiles.

Concordantly, as there was an equilibrium between COG categories with increased and decreased paralog frequencies in acidophiles, the overall paralog frequency had no statistically significant correlation with optimal pH and remained at a relatively constant 8% average, ranging from 2% to 20% (Supplementary Figure 4). These relatively low percentages indicate that paralog frequencies are only a minor contributor to genome size changes in our dataset. Still, the constant paralog frequency across pH contradicts what has been found for other streamlined organisms, which have

relatively low number of paralogs (Giovannoni et. al., 2005, Swan et. al., 2013). This unusual finding could be partially a consequence of acid resistance genes in multiple copies that would compensate the evolutionary pressure of discarding paralogs.

4. Additional Discussion

We have shown acidophilic Bacteria possess several streamlining elements, such as having smaller genomes, fewer ORFs and an underrepresentation of signal transduction proteins (Gifford et. al., 2013, Giovannoni et. al., 2014, Cottrell and Kirchman, 2016). However, there are several streamlining elements that we could not identify in acidophiles, such as having lower intergenic space percentages, lower paralog frequencies and proportionately fewer pseudogenes (Giovannoni et. al., 2005, Swan et. al., 2013). This could be partially attributed to the high prevalence of HGT and recombination elements in acidophiles (Aliaga et. al., 2009, Navarro et. al., 2013, Acuña et. al., 2013, Ullrich et. al., 2016, Zhang et. al., 2017, Colman et. al., 2018, Vergara et. al., 2020). A high recombination activity is prone to increase the abundance of pseudogenes present in a genome (Holt et. al., 2009, Tutar, 2012) and could cause the observed high paralog frequencies in the Cog category L “Replication, recombination and repair”, which in turn increases the overall paralog frequencies of acidophiles. This is supported by the low paralog frequencies in COG category J “Translation and Ribosome”, which are amongst the most conserved proteins (Lecompte et. al., 2002) and thus could be an index of general paralog frequency tendencies. Additionally, streamlining as a phenomenon has been mainly described for extremely small genomes (<2Mb). While genomes as small as 1.7Mb exist in our dataset, most of the genomes are between 2-4 Mb, which could explain the absence of some streamlining elements in acidophiles.

What is observed for acidophiles then appears to differ from the classic examples of extremely streamlined organisms. However, as opposed to statistical analyses of multiple acidophilic clades, most of the studies that defined genome streamlining traits focus on a single clade and reflect on the underlying ecological variable to which attribute its genome reduction (Dufresne et. al., 2005, Giovannoni et. al., 2005, Chivian et. al., 2008, Sowell et. al., 2009, López-Pérez et. al., 2013, Luo et. al., 2014, Sun and Blanchard, 2014, Nakai et. al., 2016, Cottrell and Kirchman, 2016, Graham and Tully, 2021). The divergence in the observations from this study and others could be attributable to such difference, as single clade studies do not consider counter examples such as *Rhodococcus erythropolis*, an extreme oligotroph with a genome of over 7 Mb (Yano et. al., 2016, Retamal-Morales et. al., 2018). Nevertheless, streamlining in the evolution of acidophiles appears to be a less robust phenomenon than in thermophiles when comparing to other multi-clade statistical studies (Sabath et. al., 2013). This was also observed in our study, as shown by the stronger correlation between genome size and temperature (Figure 4A) than with pH (Figure 3) and the positioning of the lowest genome sizes in the PCA plot (Figure 5).

In terms of physiology, acidophiles possess several characteristics of streamlined Bacteria, such as relatively small cell sizes (Clark and Norris, 1996) and high generation times (Kishimoto and Tano, 1987, Fang et. al., 2006, Mykytczuk et. al., 2010). Chemolithoautotrophic metabolism is widespread amongst acidophiles (Johnson and Hallberg, 2008), which could be a bias in our study as the reduced genomes of acidophiles might be related to this overrepresentation of chemolithoautotrophs. However, some of the smallest genomes in free-living prokaryotes are heterotrophs (Giovannoni et. al., 2005, 2014) and are smaller than some of the smallest known genomes of chemolithoautotrophic prokaryotes besides methylotrophs (Raven et. al., 2013). Therefore, this is unlikely to be a major issue.

In agreement with what has been observed in Archaea (Colman et. al., 2018), the bacterial acidophiles are all nested within higher order neutrophilic lineages and no examples are observed of regression of acidophile lineages to neutrophiles, suggesting that the evolution of acidophilia is unidirectional. However, the current taxonomic distribution of acidophilic genomes is possibly affected by sampling bias, as acidic mine drainages are one of the most studied acidic environments (Johnson and Hallberg, 2003, Sharma et. al., 2016) which possibly produces an overrepresentation of organisms from these environments in the databases. Advances in metagenomics should attenuate this issue by increasing the genomic information from less studied acidophilic niches, such as deep-sea vents (Simmons and Norris, 2002, Reysenbach et. al., 2006) and to a lesser extent solfataric fields (Itoh et. al., 2011). Possibly, entirely novel acidophilic lineages from different phyla could be discovered.

Some of the genomic traits observed in acidophiles have not been described as general features of streamlined organisms, such as lower average protein sizes and higher representativity of inner membrane proteins. These features could be novel characteristics of streamlined organisms or perhaps are specific for acidophilic adaptation. The increased representativity of inner membrane proteins is likely to be specific for acidophiles, as no statistically supported correlation was found between the representativity of these proteins and genome size in neutrophiles (Supplementary Figure 2). This is also likely true for the lower representativity of proteins found outside the inner membrane of acidophiles. In contrast, average protein size has been analyzed in previous streamlining studies on adaptation to high temperatures (Sabath et. al., 2013). A decrease in average protein size was reported for thermophiles, and a conclusion regarding thermostability adaptations (Thompson and Eisenberg, 1999, Chakravarty and Varadarajan, 2000) was reached. However, protein size changes might be a major contributor to genome size changes besides gene gain or loss. Our discovery of a decrease in average protein size in acidophiles expands the possibility beyond thermophiles that protein size reduction might be a more general mechanism for genome streamlining in stressful environments. Further research on this feature is necessary to determine whether other streamlined organisms have smaller proteins than their counterparts. Nevertheless, smaller proteins in acidophiles could also be attributable to protein stability adaptations, such as the shorter loops observed for some proteins in the inner membrane of acidophiles (Duarte et. al., 2009, 2011). The investigation of which specific protein size changes or domain rearrangements might be attributable to a survival mechanism in acidic niches is a potential topic for future research.

Acidophiles pay the energetic toll of maintaining a proton gradient of several orders of magnitude across the inner membrane (Baker-Austin and Dopson, 2007, Slonczewski et. al., 2009). This, while proliferating in often nutrient scarce environments with multiple stressors (Johnson, 1998, Dopson et. al., 2003, Johnson and Hallberg, 2008). It is then congruent that these organisms would optimize transport and reduce replication costs to save energy by streamlining their genomes (Button, 1991, Sowell et. al., 2009). Several of our findings shed light on the ever-expanding knowledge about acidophiles ecology and the acid resistance systems that maintain this proton gradient. Mainly, the increased paralog frequencies in COG categories possibly related to energy production, DNA repair and biofilm formation. The investigation of which functions might be in greater copies in acidophiles is an interesting topic for future research, as it may uncover novel survival mechanisms for acidophiles. Similarly, acid related genes shared between acidophiles could be hidden amongst the proteins without functional annotation.

5. Conflict of Interest

575 The authors declare that the research was conducted in the absence of any commercial or financial
576 relationships that could be construed as a potential conflict of interest.

577 6. Author Contributions

578 DC, GN and DH designed the research. DC performed the research. DC, DH and GN analyzed the
579 data. DC and DH wrote the paper. GC and EV participated in the construction of the final
580 manuscript. All authors read and approved the final manuscript.

581 7. Acknowledgments

582 DH was supported by Fondecyt 1181717 and Programa de Apoyo a Centros con Financiamiento
583 Basal AFB170004 to Fundación Ciencia & Vida.

584 8. REFERENCES

- 585 Acuña, L. G., Cárdenas, J. P., Covarrubias, P. C., Haristoy, J. J., Flores, R., Nuñez, H., ... &
586 Rawlings, D. E. (2013). Architecture and gene repertoire of the flexible genome of the extreme
587 acidophile *Acidithiobacillus caldus*. *PLoS One*, 8(11). doi: 10.1371/journal.pone.0078237
- 588 Aliaga, D. S., Deneff, V. J., Singer, S. W., VerBerkmoes, N. C., Lefsrud, M., Mueller, R. S., ... &
589 Baker, B. J. (2009). Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing
590 “*Leptospirillum rubrum*” (Group II) and “*Leptospirillum ferrodiazotrophum*” (Group III) bacteria in
591 acid mine drainage biofilms. *Appl. Environ. Microbiol.*, 75(13), 4599-4615. doi:
592 10.1128/AEM.02943-08
- 593 Almagro, J. J., Tsigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... &
594 Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat.*
595 *Biotechnol.*, 37(4), 420-423. doi: 10.1038/s41587-019-0036-z
- 596 Alpanis, A., Swain, M., Gatherer, D., & McEwan, N. (2018). Correlation between bacterial G+ C
597 content, genome size and the G+ C content of associated plasmids and bacteriophages. *Microb.*
598 *Genom.*, 4(4). doi: 10.1099/mgen.0.000168
- 599 Angelov, A., & Liebl, W. (2006). Insights into extreme thermoacidophily based on genome analysis
600 of *Picrophilus torridus* and other thermoacidophilic archaea. *J. Biotechnol.*, 126(1), 3-10. doi:
601 10.1016/j.jbiotec.2006.02.017
- 602 Baker, B. J., & Banfield, J. F. (2003). Microbial communities in acid mine drainage. *FEMS*
603 *Microbiol. Ecol.*, 44(2), 139-152. doi: 10.1016/S0168-6496(03)00028-X
- 604 Baker, B. J., Comolli, L. R., Dick, G. J., Hauser, L. J., Hyatt, D., Dill, B. D., ... & Banfield, J. F.
605 (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl. Acad. Sci. USA*, 107(19), 8806-
606 8811. doi: 10.1073/pnas.0914470107
- 607 Baker-Austin, C., & Dopson, M. (2007). Life in acid: pH homeostasis in acidophiles. *Trends*
608 *Microbiol.*, 15(4), 165-171. doi: 10.1016/j.tim.2007.02.005

- 609 Baker-Austin, C., Potrykus, J., Wexler, M., Bond, P. L., & Dopson, M. (2010). Biofilm development
610 in the extremely acidophilic archaeon 'Ferroplasma acidarmanus' Fer1. *Extremophiles*, 14(6), 485-
611 491. doi: 10.1007/s00792-010-0328-1
- 612 Beard, S., Ossandon, F. J., Rawlings, D. E., & Quatrini, R. (2021). The Flexible Genome of
613 Acidophilic Prokaryotes. *Curr. Issues in Mol. Biol.*, 40(1), 231-266. doi: 10.21775/cimb.040.231
- 614 Benison, K. C., O'Neill, W. K., Blain, D., & Hallsworth, J. E. (2021). Water activities of acid brine
615 lakes approach the limit for life. *Astrobiology*, 21(6), 729-740. doi: 10.1089/ast.2020.2334
- 616 Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful
617 approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, 57(1), 289-300. doi: 10.1111/j.2517-
618 6161.1995.tb02031.x
- 619 Bentkowski, P., Van Oosterhout, C., & Mock, T. (2015). A model of genome size evolution for
620 prokaryotes in stable and fluctuating environments. *Genome Biol. Evol.*, 7(8), 2344-2351. doi:
621 10.1093/gbe/evv148
- 622 Bond, P. L., Druschel, G. K., & Banfield, J. F. (2000). Comparison of acid mine drainage microbial
623 communities in physically and geochemically distinct ecosystems. *Appl. Environ. Microbiol.*, 66(11),
624 4962-4971. doi: 10.1128/AEM.66.11.4962-4971.2000
- 625 Button, D. K. (1991). Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic
626 capacity, and the meaning of the Michaelis constant. *Appl. Environ. Microbiol.*, 57(7), 2033-2038.
627 doi: 10.1128/aem.57.7.2033-2038.1991
- 628 Capece, M. C., Clark, E., Saleh, J. K., Halford, D., Heinl, N., Hoskins, S., & Rothschild, L. J. (2013).
629 Polyextremophiles and the constraints for terrestrial habitability. In Polyextremophiles (pp. 3-59).
630 Springer, Dordrecht. doi: 10.1007/978-94-007-6488-0_1
- 631 Cárdenas, J. P., Moya, F., Covarrubias, P., Shmaryahu, A., Levicán, G., Holmes, D. S., & Quatrini,
632 R. (2012). Comparative genomics of the oxidative stress response in bioleaching microorganisms.
633 *Hydrometallurgy*, 127, 162-167. doi: 10.1016/j.hydromet.2012.07.014
- 634 Cárdenas, J. P., Quatrini, R., & Holmes, D. S. (2016). Progress in acidophile genomics. *Acidophiles:*
635 *life in extremely acidic environments*, 179-197. doi: 10.21775/9781910190333
- 636 Carere, C. R., Hards, K., Wigley, K., Carman, L., Houghton, K. M., Cook, G. M., & Stott, M. B.
637 (2021). Growth on formic acid is dependent on intracellular pH homeostasis for the
638 thermoacidophilic methanotroph *Methylophilum* sp. RTK17.1. *Front. Microbiol.*, 12. doi:
639 10.3389/fmicb.2021.651744
- 640 Chakravarty, S., & Varadarajan, R. (2000). Elucidation of determinants of protein stability through
641 genome sequence analysis. *Febs Lett.*, 470(1), 65-69. doi: 10.1016/S0014-5793(00)01267-9
- 642 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to
643 classify genomes with the Genome Taxonomy Database. *Bioinform.*, 36(6), 1925-1927. doi:
644 10.1093/bioinformatics/btz848

- 645 Chen, L. X., Hu, M., Huang, L. N., Hua, Z. S., Kuang, J. L., Li, S. J., & Shu, W. S. (2015).
646 Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine
647 drainage. *ISME J*, 9(7), 1579-1592. doi: 10.1038/ismej.2014.245
- 648 Chi, A., Valenzuela, L., Beard, S., Mackey, A. J., Shabanowitz, J., Hunt, D. F., & Jerez, C. A.
649 (2007). Periplasmic proteins of the extremophile *Acidithiobacillus ferrooxidans*: a high throughput
650 proteomics analysis. *Mol. Cell. Proteom.*, 6(12), 2239-2251. doi: 10.1074/mcp.M700042-MCP200
- 651 Chivian, D., Brodie, E. L., Alm, E. J., Culley, D. E., Dehal, P. S., DeSantis, T. Z., ... & Moser, D. P.
652 (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science*,
653 322(5899), 275-278. doi: 10.1126/science.1155495
- 654 Chowhan, R. K., Hotumalani, S., Rahaman, H., & Singh, L. R. (2021). pH induced conformational
655 alteration in human peroxiredoxin 6 might be responsible for its resistance against lysosomal pH or
656 high temperature. *Sci. Rep.*, 11(1), 1-10. doi: 10.1038/s41598-021-89093-8
- 657 Clark, D. A., & Norris, P. R. (1996). *Acidimicrobium ferrooxidans* gen. nov., sp. nov.: mixed-culture
658 ferrous iron oxidation with *Sulfobacillus* species. *Microbiology*, 142(4), 785-790. doi:
659 10.1099/00221287-142-4-785
- 660 Colman, D. R., Poudel, S., Hamilton, T. L., Havig, J. R., Selensky, M. J., Shock, E. L., & Boyd, E. S.
661 (2018). Geobiological feedbacks and the evolution of thermoacidophiles. *ISME J*, 12(1), 225-236.
662 doi: 10.1038/ismej.2017.162
- 663 Copley, S. D. (2020). Evolution of new enzymes by gene duplication and divergence. *FEBS*
664 *J.*, 287(7), 1262-1283. doi: 10.1111/febs.15299
- 665 Cottrell, M. T., & Kirchman, D. L. (2016). Transcriptional control in marine copiotrophic and
666 oligotrophic bacteria with streamlined genomes. *Appl. Environ. Microbiol.*, 82(19), 6010-6018. doi:
667 10.1128/AEM.01299-16
- 668 Crossman, L., Holden, M., Pain, A., & Parkhill, J. (2004). Genomes beyond compare. *Nat. Rev.*
669 *Microbiol.*, 2(8), 616-617. doi: 10.1038/nrmicro961
- 670 D'Abusco, A. S., Casadio, R., Tasco, G., Giangiacomo, L., Giartosio, A., Calamia, V., ... & Politi, L.
671 (2005). Oligomerization of *Sulfolobus solfataricus* signature amidase is promoted by acidic pH and
672 high temperature. *Archaea*, 1(6), 411-423. doi: 10.1155/2005/543789
- 673 Darby, C. A., Stolzer, M., Ropp, P. J., Barker, D., & Durand, D. (2017). Xenolog
674 classification. *Bioinform.*, 33(5), 640-649. doi: 10.1093/bioinformatics/btw686
- 675 Díaz, M., Castro, M., Copaja, S., & Guiliani, N. (2018). Biofilm formation by the acidophile
676 bacterium *Acidithiobacillus thiooxidans* involves c-di-GMP pathway and Pel
677 exopolysaccharide. *Genes*, 9(2), 113. doi: 10.3390/genes9020113
- 678 Dopson, M., Baker-Austin, C., Koppineedi, P. R., & Bond, P. L. (2003). Growth in sulfidic mineral
679 environments: metal resistance mechanisms in acidophilic micro-organisms. *Microbiology*, 149(8),
680 1959-1970. doi: 10.1099/mic.0.26296-0

- 681 Dopson, M., & Holmes, D. S. (2014). Metal resistance in acidophilic microorganisms and its
682 significance for biotechnologies. *Appl. Microbiol. Biotechnol.*, 98(19), 8133-8144. doi:
683 10.1007/s00253-014-5982-2
- 684 Dopson, M. (2016). Physiological and phylogenetic diversity of acidophilic bacteria. R. Quatrini,
685 D.B. Johnson (Eds.), *Acidophiles: Life in Extremely Acidic Environments*, Caister Academic Press,
686 Norfolk UK (2016), pp. 79-92. doi: 10.21775/9781910190333
- 687 Dsouza, M., Taylor, M. W., Turner, S. J., & Aislabie, J. (2014). Genome-based comparative analyses
688 of Antarctic and temperate species of *Paenibacillus*. *PloS One*, 9(10). doi:
689 10.1371/journal.pone.0108009
- 690 Duarte, F., Araya-Secchi, R., González, W., Perez-Acle, T., González-Nilo, D., & Holmes, D. S.
691 (2009). Protein function in extremely acidic conditions: Molecular simulations of a predicted
692 aquaporin and a potassium channel in *Acidithiobacillus ferrooxidans*. In *Advanced Materials*
693 *Research* (Vol. 71, pp. 211-214). Trans Tech Publications Ltd. doi:
694 10.4028/www.scientific.net/AMR.71-73.211
- 695 Duarte, F., Sepulveda, R., Araya, R., Flores, S., Perez-Acle, T., Gonzales, W., ... & Holmes, D. S.
696 (2011). Mechanisms of protein stabilization at very low pH. In *Proc. 19th International*
697 *Biohydrometallurgy Symposium, Changsha, China* (pp. 349-353).
- 698 Dufresne, A., Garczarek, L., & Partensky, F. (2005). Accelerated evolution associated with genome
699 reduction in a free-living prokaryote. *Genome Biology*, 6(2), R14. doi: 10.1186/gb-2005-6-2-r14
- 700 Dulmage, K. A., Darnell, C. L., Vreugdenhil, A., & Schmid, A. K. (2018). Copy number variation is
701 associated with gene expression change in archaea. *Microb. Genom.*, 4(9). doi:
702 10.1099/mgen.0.000210
- 703 El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... & Finn, R. D. (2019).
704 The Pfam protein families database in 2019. *Nucleic Acids Res.*, 47(D1), D427-D432. doi:
705 10.1093/nar/gky995
- 706 Fang, H. H., Zhang, T., & Li, C. (2006). Characterization of Fe-hydrogenase genes diversity and
707 hydrogen-producing population in an acidophilic sludge. *J. Biotechnol.*, 126(3), 357-364. doi:
708 10.1016/j.jbiotec.2006.04.023
- 709 Feng, S., Yang, H., & Wang, W. (2015). System-level understanding of the potential acid-tolerance
710 components of *Acidithiobacillus thiooxidans* ZJJN-3 under extreme acid stress. *Extremophiles*, 19(5),
711 1029-1039. doi: 10.1007/s00792-015-0780-z
- 712 Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., ... & Salazar, G. A.
713 (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*,
714 44(D1), D279-D285. doi: 10.1093/nar/gkv1344
- 715 Foster, J. W. (2004). *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nat. Rev.*
716 *Microbiol.*, 2(11), 898-907. doi: 10.1038/nrmicro1021

- 717 Fütterer, O., Angelov, A., Liesegang, H., Gottschalk, G., Schleper, C., Schepers, B., ... & Liebl, W.
718 (2004). Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. *Proc.*
719 *Natl. Acad. Sci. USA*, 101(24), 9091-9096. doi: 10.1073/pnas.0401356101
- 720 Galperin, M. Y. (2005). A census of membrane-bound and intracellular signal transduction proteins
721 in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol.*, 5(1), 1-19. doi: 10.1186/1471-
722 2180-5-35
- 723 Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded microbial genome
724 coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, 43(D1),
725 D261-D269. doi: 10.1093/nar/gku1223
- 726 Galperin, M. Y., Wolf, Y. I., Garushyants, S. K., Vera Alvarez, R., & Koonin, E. V. (2021).
727 Nonessential Ribosomal Proteins in Bacteria and Archaea Identified Using Clusters of Orthologous
728 Genes. *J. Bacteriol.*, 203(11), e00058-21. doi: 10.1128/JB.00058-21
- 729 Gao, Z. M., Wang, Y., Tian, R. M., Wong, Y. H., Batang, Z. B., Al-Suwailem, A. M., ... & Qian, P.
730 Y. (2014). Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont
731 “*Candidatus Synechococcus spongiarum*”. *MBio*, 5(2), e00079-14. doi: 10.1128/mBio.00079-14
- 732 Gifford, S. M., Sharma, S., Booth, M., & Moran, M. A. (2013). Expression patterns reveal niche
733 diversification in a marine microbial assemblage. *ISME J*, 7(2), 281-298. doi: 10.1038/ismej.2012.96
- 734 Gillings, M. R. (2017). Lateral gene transfer, bacterial genome evolution, and the
735 Anthropocene. *Ann. N. Y. Acad. Sci.*, 1389(1), 20-36. doi: 10.1111/nyas.13213
- 736 Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., ... & Rappé, M. S.
737 (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738), 1242-1245.
738 doi: 10.1126/science.1114057
- 739 Giovannoni, S. J., Thrash, J. C., & Temperton, B. (2014). Implications of streamlining theory for
740 microbial ecology. *ISME J*, 8(8), 1553-1565. doi: 10.1038/ismej.2014.60
- 741 González, A., Bellenberg, S., Mamani, S., Ruiz, L., Echeverría, A., Soulère, L., ... & Vera, M.
742 (2013). AHL signaling molecules with a large acyl chain enhance biofilm formation on sulfur and
743 metal sulfides by the bioleaching bacterium *Acidithiobacillus ferrooxidans*. *Appl. Microbiol.*
744 *Biotechnol.*, 97(8), 3729-3737. doi: 10.1007/s00253-012-4229-3
- 745 González, C., Lazcano, M., Valdés, J., & Holmes, D. S. (2016). Bioinformatic analyses of unique
746 (orphan) core genes of the genus *Acidithiobacillus*: functional inferences and use as molecular probes
747 for genomic and metagenomic/transcriptomic interrogation. *Front. Microbiol.*, 7, 2035. doi:
748 10.3389/fmicb.2016.02035
- 749 Goordial, J., Raymond-Bouchard, I., Zolotarov, Y., de Bethencourt, L., Ronholm, J., Shapiro, N., ...
750 & Whyte, L. (2016). Cold adaptive traits revealed by comparative genomic analysis of the
751 eurypsychrophile *Rhodococcus* sp. JG3 isolated from high elevation McMurdo Dry Valley
752 permafrost, Antarctica. *FEMS Microbiol. Ecol.*, 92(2). doi: 10.1093/femsec/fiv154
- 753 Graham, E. D., & Tully, B. J. (2021). Marine Dadabacteria exhibit genome streamlining and
754 phototrophy-driven niche partitioning. *ISME J*, 15(4), 1248-1256. doi: 10.1038/s41396-020-00834-5

- 755 Green, E. R., & Meccas, J. (2016). Bacterial secretion systems: an overview. *Microbiol. Spectr.*, 4(1),
756 4-1. doi: 10.1128/microbiolspec.VMBF-0012-2015
- 757 Gu, J., Wang, X., Ma, X., Sun, Y., Xiao, X., & Luo, H. (2021). Unexpectedly high mutation rate of a
758 deep-sea hyperthermophilic anaerobic archaeon. *ISME J*, 15(6), 1862-1869. doi: 10.1038/s41396-
759 020-00888-5
- 760 Hedrich, S., & Schippers, A. (2021). Distribution of acidophilic microorganisms in natural and man-
761 made acidic environments. *Curr. Issues in Mol. Biol.*, 40(1), 25-48. doi: 10.21775/cimb.040.025
- 762 Holt, K. E., Thomson, N. R., Wain, J., Langridge, G. C., Hasan, R., Bhutta, Z. A., ... & Parkhill, J.
763 (2009). Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars
764 Paratyphi A and Typhi. *BMC Genom.*, 10(1), 1-12. doi: 10.1186/1471-2164-10-36
- 765 Hou, S., Makarova, K. S., Saw, J. H., Senin, P., Ly, B. V., Zhou, Z., ... & Wolf, Y. I. (2008).
766 Complete genome sequence of the extremely acidophilic methanotroph isolate V4,
767 *Methylophilum infernorum*, a representative of the bacterial phylum Verrucomicrobia. *Biol.*
768 *Direct*, 3(1), 26. doi: 10.1186/1745-6150-3-26
- 769 Hu, W., Feng, S., Tong, Y., Zhang, H., & Yang, H. (2020). Adaptive defensive mechanism of
770 bioleaching microorganisms under extremely environmental acid stress: Advances and
771 perspectives. *Biotechnol. Adv.*, 42, 107580. doi: 10.1016/j.biotechadv.2020.107580
- 772 Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... &
773 von Mering, C. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated
774 orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, 47(D1), D309-
775 D314. doi: 10.1093/nar/gky1085
- 776 Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, 9(3), 90-95. doi:
777 10.1109/MCSE.2007.55
- 778 Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and
779 distinguishing between models. *Nat. Rev. Genet.*, 11(2), 97-108. doi: 10.1038/nrg2689
- 780 Itoh, T., Yamanoi, K., Kudo, T., Ohkuma, M., & Takashina, T. (2011). *Aciditerrimonas ferrireducens*
781 gen. nov., sp. nov., an iron-reducing thermoacidophilic actinobacterium isolated from a solfataric
782 field. *Int. J. Syst. Evol. Microbiol.*, 61(6), 1281-1285. doi: 10.1099/ijs.0.023044-0
- 783 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High
784 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat.*
785 *Commun.*, 9(1), 1-8. doi: 10.1038/s41467-018-07641-9
- 786 Jan, A. T. (2017). Outer membrane vesicles (OMVs) of gram-negative bacteria: a perspective
787 update. *Front. Microbiol.*, 8, 1053. doi: 10.3389/fmicb.2017.01053
- 788 Johnson, D. B. (1998). Biodiversity and ecology of acidophilic microorganisms. *FEMS Microbiol.*
789 *Ecol.*, 27(4), 307-317. doi: 10.1111/j.1574-6941.1998.tb00547.x
- 790 Johnson, D. B., & Hallberg, K. B. (2003). The microbiology of acidic mine waters. *Res. Microbiol.*,
791 154(7), 466-473. doi: 10.1016/S0923-2508(03)00114-1

- 792 Johnson, D. B. (2007). Physiology and ecology of acidophilic microorganisms. *Physiology and*
793 *Biochemistry of Extremophiles*, 255-270. doi: 10.1128/9781555815813.ch20
- 794 Johnson, D. B., & Hallberg, K. B. (2008). Carbon, iron and sulfur metabolism in acidophilic micro-
795 organisms. *Adv. Microb. Physiol.*, 54, 201-255. doi: 10.1016/S0065-2911(08)00003-9
- 796 Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*.
797 doi: 10.1002/0470013192.bsa501
- 798 Keeling, P. J., & Slamovits, C. H. (2005). Causes and effects of nuclear genome reduction. *Curr.*
799 *Opin. Genet. Dev.*, 15(6), 601-608. doi: 10.1016/j.gde.2005.09.003
- 800 Khaleque, H. N., González, C., Kaksonen, A. H., Boxall, N. J., Holmes, D. S., & Watkin, E. L.
801 (2019). Genome-based classification of two halotolerant extreme acidophiles, *Acidihalobacter*
802 *prosperus* V6 (= DSM 14174= JCM 32253) and *Acidihalobacter ferrooxidans* V8 (= DSM 14175=
803 JCM 32254) as two new species, *Acidihalobacter aeolianus* sp. nov. and *Acidihalobacter*
804 *ferrooxydans* sp. nov., respectively. *Int. J. Syst. Evol. Microbiol.*, 69(6), 1557-1565. doi:
805 10.1099/ijsem.0.003313
- 806 Kim, M., Oh, H. S., Park, S. C., & Chun, J. (2014). Towards a taxonomic coherence between average
807 nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of
808 prokaryotes. *Int. J. Syst. Evol. Microbiol.*, 64(Pt_2), 346-351. doi: 10.1099/ijse.0.059774-0
- 809 Kirchberger, P. C., Schmidt, M. L., & Ochman, H. (2020). The ingenuity of bacterial genomes. *Annu.*
810 *Rev. Microbiol.*, 74, 815-834. doi: 10.1146/annurev-micro-020518-115822
- 811 Kishimoto, N., & Tano, T. (1987). Acidophilic heterotrophic bacteria isolated from acidic mine
812 drainage, sewage, and soils. *J. Appl. Microbiol.*, 33(1), 11-25. doi: 10.2323/jgam.33.11
- 813 Kishimoto, N., Inagaki, K., Sugio, T., & Tano, T. (1990). Growth inhibition of *Acidiphilium* species
814 by organic acids contained in yeast extract. *J. Biosci. Bioeng.*, 70(1), 7-10. doi: 10.1016/0922-
815 338X(90)90021-N
- 816 Klassen, J. L., & Currie, C. R. (2013). ORFcor: identifying and accommodating ORF prediction
817 inconsistencies for phylogenetic analysis. *PLoS One*, 8(3), e58387. doi:
818 10.1371/journal.pone.0058387
- 819 Kondratyeva, T. F., Muntyan, L. N., & Karavaiko, G. I. (1995). Zinc-and arsenic-resistant strains of
820 *Thiobacillus ferrooxidans* have increased copy numbers of chromosomal resistance
821 genes. *Microbiology*, 141(5), 1157-1162. doi: 10.1099/13500872-141-5-1157
- 822 Konstantinidis, K. T., & Tiedje, J. M. (2004). Trends between gene content and genome size in
823 prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. USA*, 101(9), 3160-3165. doi:
824 10.1073/pnas.0308653100
- 825 Korandla, D. R., Wozniak, J. M., Campeau, A., Gonzalez, D. J., & Wright, E. S. (2020). AssessORF:
826 combining evolutionary conservation and proteomics to assess prokaryotic gene
827 predictions. *Bioinform.*, 36(4), 1022-1029. doi: 10.1093/bioinformatics/btz714

- 828 Kreitmeier, M., Ardern, Z., Abele, M., Ludwig, C., Scherer, S., & Neuhaus, K. (2021). Shadow
829 ORFs illuminated: long overlapping genes in *Pseudomonas aeruginosa* are translated and under
830 purifying selection. *bioRxiv* [Preprint]. Available at:
831 <https://www.biorxiv.org/content/10.1101/2021.02.09.430400> doi: 10.2139/ssrn.3866842
- 832 Lear, G., Lau, K., Perchec, A. M., Buckley, H. L., Case, B. S., Neale, M., ... & Lewis, G. (2017).
833 Following Rapoport's Rule: the geographic range and genome size of bacterial taxa decline at warmer
834 latitudes. *Environ. Microbiol.*, 19(8), 3152-3162. doi: 10.1111/1462-2920.13797
- 835 Lecompte, O., Ripp, R., Thierry, J. C., Moras, D., & Poch, O. (2002). Comparative analysis of
836 ribosomal proteins in complete genomes: an example of reductive evolution at the domain
837 scale. *Nucleic Acids Res.*, 30(24), 5382-5390. doi: 10.1093/nar/gkf693
- 838 Lehtovirta-Morley, L. E., Ge, C., Ross, J., Yao, H., Nicol, G. W., & Prosser, J. I. (2014).
839 Characterisation of terrestrial acidophilic archaeal ammonia oxidisers and their inhibition and
840 stimulation by organic compounds. *FEMS Microbiol. Ecol.*, 89(3), 542-552. doi: 10.1111/1574-
841 6941.12353
- 842 López-Pérez, M., Ghai, R., Leon, M. J., Rodríguez-Olmos, Á., Copa-Patiño, J. L., Soliveri, J., ... &
843 Rodríguez-Valera, F. (2013). Genomes of “Spiribacter”, a streamlined, successful halophilic
844 bacterium. *BMC Genom.*, 14(1), 787. doi: 10.1186/1471-2164-14-787
- 845 Lukhele, T., Selvarajan, R., Nyoni, H., Mamba, B. B., & Msagati, T. A. (2020). Acid mine drainage
846 as habitats for distinct microbiomes: current knowledge in the era of molecular and omic
847 technologies. *Curr. Microbiol.*, 77(4), 657-674. doi: 10.1007/s00284-019-01771-z
- 848 Lund, P., Tramonti, A., & De Biase, D. (2014). Coping with low pH: molecular strategies in
849 neutrophilic bacteria. *FEMS Microbiol. Rev.*, 38(6), 1091-1125. doi: 10.1111/1574-6976.12076
- 850 Luo, H., Swan, B. K., Stepanauskas, R., Hughes, A. L., & Moran, M. A. (2014). Evolutionary
851 analysis of a streamlined lineage of surface ocean Roseobacters. *ISME J*, 8(7), 1428-1439. doi:
852 10.1038/ismej.2013.248
- 853 Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu. Rev.*
854 *Microbiol.*, 60, 327-349. doi: 10.1146/annurev.micro.60.080805.142300
- 855 Martínez-Cano, D. J., Reyes-Prieto, M., Martínez-Romero, E., Partida-Martínez, L. P., Latorre, A.,
856 Moya, A., & Delage, L. (2015). Evolution of small prokaryotic genomes. *Front. Microbiol.*, 5, 742.
857 doi: 10.3389/fmicb.2014.00742
- 858 McCutcheon, J. P., & Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nat.*
859 *Rev. Microbiol.*, 10(1), 13-26. doi: 10.1038/nrmicro2670
- 860 McMurdie, P. J., Behrens, S. F., Müller, J. A., Göke, J., Ritalahti, K. M., Wagner, R., ... &
861 Spormann, A. M. (2009). Localized plasticity in the streamlined genomes of vinyl chloride respiring
862 Dehalococcoides. *PLoS Genet.*, 5(11). doi: 10.1371/journal.pgen.1000714
- 863 Méndez-García, C., Mesa, V., Sprenger, R. R., Richter, M., Diez, M. S., Solano, J., ... & Ferrer, M.
864 (2014). Microbial stratification in low pH oxic and suboxic macroscopic growths along an acid mine
865 drainage. *ISME J*, 8(6), 1259-1274. doi: 10.1038/ismej.2013.242

866 Mendler, K., Chen, H., Parks, D. H., Lobb, B., Hug, L. A., & Doxey, A. C. (2019). AnnoTree:
867 visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids*
868 *Res.*, 47(9), 4442-4448. doi: 10.1093/nar/gkz246

869 Min-Juan, X. U., Jia-Hua, W. A. N. G., Xu-Liang, B. U., He-Lin, Y. U., Peng, L. I., Hong-Yu, O. U.,
870 ... & Ping, A. O. (2016). Deciphering the streamlined genome of *Streptomyces xiamenensis* 318 as
871 the producer of the anti-fibrotic drug candidate xiamenmycin. *Sci. Rep.*, 6, 18977. doi:
872 10.1038/srep18977

873 Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., ... &
874 Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.*, 49(D1), D412-
875 D419. doi: 10.1093/nar/gkaa913

876 Moinier, D., Byrne, D., Amouric, A., & Bonnefoy, V. (2017). The global redox responding
877 RegB/RegA signal transduction system regulates the genes involved in ferrous iron and inorganic
878 sulfur compound oxidation of the acidophilic *Acidithiobacillus ferrooxidans*. *Front. Microbiol.*, 8,
879 1277. doi: 10.3389/fmicb.2017.01277

880 Murray, G. G., Charlesworth, J., Miller, E. L., Casey, M. J., Lloyd, C. T., Gottschalk, M., ... &
881 Weinert, L. A. (2021). Genome reduction is associated with bacterial pathogenicity across different
882 scales of temporal and ecological divergence. *Mol. Biol. Evol.*, 38(4), 1570-1579. doi:
883 10.1093/molbev/msaa323

884 Mykytczuk, N. C. S., Trevors, J. T., Ferroni, G. D., & Leduc, L. G. (2010). Cytoplasmic membrane
885 fluidity and fatty acid composition of *Acidithiobacillus ferrooxidans* in response to pH
886 stress. *Extremophiles*, 14(5), 427-441. doi: 10.1007/s00792-010-0319-2

887 Nakai, R., Fujisawa, T., Nakamura, Y., Nishide, H., Uchiyama, I., Baba, T., ... & Niki, H. (2016).
888 Complete genome sequence of *Aurantimicrobium minutum* type Strain KNCT, a planktonic
889 ultramicrobacterium isolated from river water. *Genome Announc.*, 4(3), e00616-16. doi:
890 10.1128/genomeA.00616-16

891 Navarro, C. A., von Bernath, D., & Jerez, C. A. (2013). Heavy metal resistance strategies of
892 acidophilic bacteria and their acquisition: importance for biomining and bioremediation. *Biol. Res.*,
893 46(4), 363-371. doi: 10.4067/S0716-97602013000400008

894 Naz, K., Ullah, N., Zaheer, T., Shehroz, M., Naz, A., & Ali, A. (2020). Pan-genomics of model
895 bacteria and their outcomes. In *Pan-genomics: Applications, Challenges, and Future Prospects* (pp.
896 189-201). Academic Press. doi: 10.1016/B978-0-12-817076-2.00009-3

897 Neira, G., Cortez, D., Jil, J., & Holmes, D. S. (2020). AciDB 1.0: a database of acidophilic
898 organisms, their genomic information and associated metadata. *Bioinform.*, 36(19), 4970-4971. doi:
899 10.1093/bioinformatics/btaa638

900 Nielsen, D. A., Fierer, N., Geoghegan, J. L., Gillings, M. R., Gumerov, V., Madin, J. S., ... &
901 Westoby, M. (2021). Aerobic bacteria and archaea tend to have larger and more versatile
902 genomes. *Oikos*, 130(4), 501-511. doi: 10.1111/oik.07912

- 903 Osorio, H., Mettert, E., Kiley, P., Dopson, M., Jedlicki, E., & Holmes, D. S. (2019). Identification
904 and unusual properties of the master regulator FNR in the extreme acidophile *Acidithiobacillus*
905 *ferrooxidans*. *Front. Microbiol.*, *10*, 1642. doi: 10.3389/fmicb.2019.01642
- 906 Panja, A. S., Maiti, S., & Bandyopadhyay, B. (2020). Protein stability governed by its structural
907 plasticity is inferred by physicochemical factors and salt bridges. *Sci. Rep.*, *10*(1), 1-9. doi:
908 10.1038/s41598-020-58825-7
- 909 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM:
910 assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
911 *Genome Res.*, *25*(7), 1043-1055. doi: 10.1101/gr.186072.114
- 912 Pavese, A. (2021). Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic
913 Review. *Genes*, *12*(6), 809. doi: 10.3390/genes12060809
- 914 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J.
915 (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, *12*,
916 2825-2830.
- 917 Quatrini, R., Appia-Ayme, C., Denis, Y., Jedlicki, E., Holmes, D. S., & Bonnefoy, V. (2009).
918 Extending the models for iron and sulfur oxidation in the extreme acidophile *Acidithiobacillus*
919 *ferrooxidans*. *BMC Genom.*, *10*(1), 1-19. doi: 10.1186/1471-2164-10-394
- 920 Raven, J. A., Beardall, J., Larkum, A. W., & Sánchez-Baracaldo, P. (2013). Interactions of
921 photosynthesis with genome size and function. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, *368*(1622),
922 20120264. doi: 10.1098/rstb.2012.0264
- 923 Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-
924 paralogs from pairwise species comparisons. *J. Mol. Biol.*, *314*(5), 1041-1052. doi:
925 10.1006/jmbi.2000.5197
- 926 Retamal-Morales, G., Heine, T., Tischler, J. S., Erler, B., Gröning, J. A., Kaschabek, S. R., ... &
927 Tischler, D. (2018). Draft genome sequence of *Rhodococcus erythropolis* B7g, a biosurfactant
928 producing actinobacterium. *J. Biotechnol.*, *280*, 38-41. doi: 10.1016/j.jbiotec.2018.06.001
- 929 Reysenbach, A. L., Liu, Y., Banta, A. B., Beveridge, T. J., Kirshtein, J. D., Schouten, S., ... &
930 Voytek, M. A. (2006). A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal
931 vents. *Nature*, *442*(7101), 444-447. doi: 10.1038/nature04921
- 932 Rodríguez-Gijón, A., Nuy, J. K., Mehrshad, M., Buck, M., Schulz, F., Woyke, T., & Garcia, S. L.
933 (2021). A genomic perspective on genome size distribution across Earth's microbiomes reveals a
934 tendency to gene loss. *bioRxiv* [Preprint]. Available at:
935 <https://www.biorxiv.org/content/10.1101/2021.01.18.427069> doi: 10.1101/2021.01.18.427069
- 936 Rzhapishvskaya, O. I., Valdés, J., Marcinkeviciene, L., Gallardo, C. A., Meskys, R., Bonnefoy, V., ...
937 & Dopson, M. (2007). Regulation of a novel *Acidithiobacillus caldus* gene cluster involved in
938 metabolism of reduced inorganic sulfur compounds. *Appl. Environ. Microbiol.*, *73*(22), 7367-7372.
939 doi: 10.1128/AEM.01497-07

- 940 Sabath, N., Ferrada, E., Barve, A., & Wagner, A. (2013). Growth temperature and genome size in
941 bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation.
942 *Genome Biol. Evol.*, 5(5), 966-977. doi: 10.1093/gbe/evt050
- 943 Saha, D., Panda, A., Podder, S., & Ghosh, T. C. (2015). Overlapping genes: a new strategy of
944 thermophilic stress tolerance in prokaryotes. *Extremophiles*, 19(2), 345-353. doi: 10.1007/s00792-
945 014-0720-3
- 946 Sauer, D. B., & Wang, D. N. (2019). Predicting the optimal growth temperatures of prokaryotes
947 using only genome derived features. *Bioinform.*, 35(18), 3224-3231. doi:
948 10.1093/bioinformatics/btz059
- 949 Saw, J. H., Mountain, B. W., Feng, L., Omelchenko, M. V., Hou, S., Saito, J. A., ... & Galperin, M.
950 Y. (2008). Encapsulated in silica: genome, proteome and physiology of the thermophilic bacterium
951 *Anoxybacillus flavithermus* WK1. *Genome Biol.*, 9(11), R161. doi: 10.1186/gb-2008-9-11-r161
- 952 Schneiker, S., dos Santos, V. A. M., Bartels, D., Bekel, T., Brecht, M., Buhrmester, J., ... &
953 Goesmann, A. (2006). Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium
954 *Alcanivorax borkumensis*. *Nat. Biotechnol.*, 24(8), 997-1004. doi: 10.1038/nbt1232
- 955 Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., ... & Karsch-
956 Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and
957 tools. *Database*, 2020, 1-21. doi: 10.1093/database/baaa062
- 958 Seabold, S., & Perktold, J. (2010, June). Statsmodels: Econometric and statistical modeling with
959 python. In Proceedings of the 9th Python in Science Conference (Vol. 57, p. 61). Scipy.
- 960 Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinform.*, 30(14), 2068-2069.
961 doi: 10.1093/bioinformatics/btu153
- 962 Sharma, A., Parashar, D., & Satyanarayana, T. (2016). Acidophilic Microbes: Biology and
963 Applications. Grand Challenges in Biology and Biotechnology, 215–241. doi: 10.1007/978-3-319-
964 13521-2_7
- 965 Shmaryahu, A., Lefimil, C., Jedlicki, E., & Holmes, D. S. (2009). Small regulatory RNAs in
966 *Acidithiobacillus ferrooxidans*: case studies of 6S RNA and Frr. In Advanced Materials
967 Research (Vol. 71, pp. 191-194). Trans Tech Publications Ltd. doi:
968 10.4028/www.scientific.net/AMR.71-73.191
- 969 Simmons, S., & Norris, P. (2002). Acidophiles of saline water at thermal vents of Vulcano,
970 Italy. *Extremophiles*, 6(3), 201-207. doi: 10.1007/s007920100242
- 971 Slonczewski, J. L., Fujisawa, M., Dopson, M., & Krulwich, T. A. (2009). Cytoplasmic pH
972 measurement and homeostasis in bacteria and archaea. *Adv. Microb. Physiol.*, 55, 1-317. doi:
973 10.1016/S0065-2911(09)05501-5
- 974 Sowell, S. M., Wilhelm, L. J., Norbeck, A. D., Lipton, M. S., Nicora, C. D., Barofsky, D. F., ... &
975 Giovanonni, S. J. (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient
976 extremes in the Sargasso Sea. *ISME J*, 3(1), 93-105. doi: 10.1038/ismej.2008.83

977 Sriaporn, C., Campbell, K. A., Van Kranendonk, M. J., & Handley, K. M. (2021). Genomic
978 adaptations enabling Acidithiobacillus distribution across wide-ranging hot spring temperatures and
979 pHs. *Microbiome*, 9(1), 1-17. doi: 10.1186/s40168-021-01090-1

980 Sun, Z., & Blanchard, J. L. (2014). Strong genome-wide selection early in the evolution of
981 Prochlorococcus resulted in a reduced genome through the loss of a large number of small effect
982 genes. *PLoS One*, 9(3). doi: 10.1371/journal.pone.0088837

983 Suzuki, S., Kuenen, J. G., Schipper, K., Van Der Velde, S., Ishii, S. I., Wu, A., ... & Kamagata, Y.
984 (2014). Physiological and genomic features of highly alkaliphilic hydrogen-utilizing
985 Betaproteobacteria from a continental serpentinizing site. *Nat. Commun.*, 5(1), 1-12. doi:
986 10.1038/ncomms4900

987 Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., ... &
988 Thompson, B. P. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic
989 bacteria in the surface ocean. *Proc. Natl. Acad. Sci. USA*, 110(28), 11463-11468. doi:
990 10.1073/pnas.1304246110

991 Tettelin, H., Maignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & Fraser, C.
992 M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications
993 for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. USA*, 102(39), 13950-13955. doi:
994 10.1073/pnas.0506758102

995 Thompson, M. J., & Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism
996 for enhancing protein thermostability. *J. Mol. Biol.*, 290(2), 595-604. doi: 10.1006/jmbi.1999.2889

997 Tutar, Y. (2012). Pseudogenes. *Comp. Funct. Genomics*, 2012. doi: 10.1155/2012/424526

998 Ullrich, S. R., González, C., Poehlein, A., Tischler, J. S., Daniel, R., Schlömann, M., ... & Mühling,
999 M. (2016). Gene loss and horizontal gene transfer contributed to the genome evolution of the extreme
1000 acidophile “Ferroplasma”. *Front. Microbiol.*, 7, 797. doi: 10.3389/fmicb.2016.00797

1001 Ulrich, L. E., Koonin, E. V., & Zhulin, I. B. (2005). One-component systems dominate signal
1002 transduction in prokaryotes. *Trends Microbiol.*, 13(2), 52-56. doi: 10.1016/j.tim.2004.12.006

1003 van Rossum, G. (1995). Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en
1004 Informatica (CWI), Amsterdam."

1005 Vargas-Straube, M. J., Beard, S., Norambuena, R., Paradela, A., Vera, M., & Jerez, C. A. (2020).
1006 High copper concentration reduces biofilm formation in Acidithiobacillus ferrooxidans by decreasing
1007 production of extracellular polymeric substances and its adherence to elemental sulfur. *J.*
1008 *Proteomics*, 225, 103874. doi: 10.1016/j.jprot.2020.103874

1009 Veloso, F., Riadi, G., Aliaga, D., Lieph, R., & Holmes, D. S. (2005). Large-scale, multi-genome
1010 analysis of alternate open reading frames in bacteria and archaea. *OMICS*, 9(1), 91-105. doi:
1011 10.1089/omi.2005.9.91

1012 Vergara, E., Neira, G., González, C., Cortez, D., Dopson, M., & Holmes, D. S. (2020). Evolution of
1013 Predicted Acid Resistance Mechanisms in the Extremely Acidophilic Leptospirillum Genus. *Genes*,
1014 11(4), 389. doi: 10.3390/genes11040389

1015 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van
1016 Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*
1017 *Methods*, 17(3), 261-272. doi: 10.1038/s41592-019-0686-2

1018 Walker, J. E. (1992). The NADH: ubiquinone oxidoreductase (complex I) of respiratory chains. *Q.*
1019 *Rev. Biophys.*, 25(3), 253-324. doi: 10.1017/S003358350000425X

1020 Westoby, M., Nielsen, D. A., Gillings, M. R., Litchman, E., Madin, J. S., Paulsen, I. T., & Tetu, S. G.
1021 (2021). Cell size, genome size, and maximum growth rate are near-independent dimensions of
1022 ecological variation across bacteria and archaea. *Ecol. Evol.*, 11(9), 3956-3976. doi:
1023 10.1002/ece3.7290

1024 Yano, T., Funamizu, Y., & Yoshida, N. (2016). Intracellular accumulation of trehalose and glycogen
1025 in an extreme oligotroph, *Rhodococcus erythropolis* N9T-4. *Biosci. Biotechnol. Biochem.*, 80(3),
1026 610-613. doi: 10.1080/09168451.2015.1107467

1027 Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., ... & Brinkman, F. S. (2010).
1028 PSORTb 3.0: improved protein subcellular localization prediction with refined localization
1029 subcategories and predictive capabilities for all prokaryotes. *Bioinform.*, 26(13), 1608-1615. doi:
1030 10.1093/bioinformatics/btq249

1031 Zhan, Y., Yang, M., Zhang, S., Zhao, D., Duan, J., Wang, W., & Yan, L. (2019). Iron and sulfur
1032 oxidation pathways of *Acidithiobacillus ferrooxidans*. *World J. Microbiol. Biotechnol.*, 35(4), 1-12.
1033 doi: 10.1007/s11274-019-2632-y

1034 Zhang, X., Liu, X., Liang, Y., Fan, F., Zhang, X., & Yin, H. (2016). Metabolic diversity and adaptive
1035 mechanisms of iron- and/or sulfur-oxidizing autotrophic acidophiles in extremely acidic
1036 environments. *Environ. Microbiol. Reports*, 8(5), 738-751. doi: 10.1111/1758-2229.12435

1037 Zhang, X., Liu, X., Liang, Y., Guo, X., Xiao, Y., Ma, L., ... & Zhang, Y. (2017). Adaptive evolution
1038 of extreme acidophile *Sulfobacillus thermosulfidooxidans* potentially driven by horizontal gene
1039 transfer and gene loss. *Appl. Environ. Microbiol.*, 83(7), e03098-16. doi: 10.1128/AEM.03098-16
1040

Figure Captions

Figure 1. Taxonomic distribution of acidophilic genomes interrogated. A rooted cladogram displaying phyla, classes, and metadata of acidophiles with genomic data. The acidophiles are classified into those that grow optimally at pH <3 or at pH 3-5. The cladogram was constructed using AnnoTree (Mendler et. al., 2019) as a guide for phylogenetic positioning and rooted as described by Parks et. al., 2018. Phyla with acidophiles were broken down into classes. Lineages with known acidophiles are highlighted and their branches are shown with thick red lines. Dashed lines connect the acidophilic lineages with the taxon's information when necessary. Growth pH pie charts represent the percentage of species that grow optimally at pH <3 (red) and at pH 3-5 (yellow). For both pH ranges, the percentage of acidophilic species by phyla are shown in the blue box. Genome source pie charts represent the percentage of acidophilic genomes sequenced from laboratory pure strains (dark green) versus metagenome assemblies (grey). The totals of both pie charts for all the phyla combined are shown in the yellow box. Ph. = Phylum; Sph. = Superphylum. *Mean values for the acidophiles in the taxon. A more detailed table with the classes' information can be found in Supplementary Table 2.

Figure 2. Distribution of acidophilic species with sequenced genomes by phylum across pH. Phylum *Armatimonadetes* has only one acidophilic species and is not shown. (A) Histogram of species number grouped by phyla across pH in overlapping increments of one pH unit. Phyla are color coded. (B) Cumulative plot of relative abundance (%) of acidophiles across pH. Percentages indicate species that can live at or below a given pH. Color coding of phyla is the same as A. (a), (b) and (c) indicate pH ranges 1-2, 2-4 and 4-5 respectively.

Figure 3. Scatterplot of genome size (Mb) of bacterial acidophiles and their most closely related extant, circum-neutral relatives versus optimal growth pH. Each point corresponds to a different species. A linear regression curve has been fitted to the data with a Pearson's correlation coefficient of 0.19 and a p-value of 2.97×10^{-5} . Generalized Least Squares (GLS) p-value was 1.8×10^{-3} .

Figure 4. Scatterplots showing correlation of genome size and pH versus optimal growth temperature and G+C content of the species in the dataset. (A) Genome size vs optimal growth temperature. Pearson's correlation coefficient is -0.34 with p-value 2.9×10^{-13} . (B) Optimal growth pH versus optimal growth temperature. Pearson's correlation coefficient is -0.01 with p-value 0.84. (C) Genome size versus G+C content. Here, data were separated by pH ranges. Pearson's correlation coefficients were 0.34 and 0.50, with p-values 4.7×10^{-3} and 1.5×10^{-22} respectively for pH 0-4 and pH 4-8. The overall Pearson's correlation coefficient and p-value were 0.48 and 1.91×10^{-25} , respectively. (D) Optimal growth pH versus G+C content. Pearson's correlation coefficient is -0.06 with p-value 0.22.

Figure 5. Principal component analysis of multiple variables potentially influencing genome size. Dimensionality reduction was performed by PCA, inputting the optimal growth pH, optimal growth temperature, G+C content and genome size of each species in the dataset. A biplot was constructed showing the loadings of each variable as arrows at the center of the plot and the distribution of the principal components. The average genome size of each species is shown as a color scale. Three clusters within the dotted circles are highlighted for their distinctive features.

Figure 6. Diagrammatic representation of genetic mechanisms involved in genome size changes. **Top row**, five genes of a hypothetical genome. Orange boxes indicate paralogous genes. **Middle row**, processes involved in genome size changes where A and B represent gene loss/gain of single

copy genes or paralogous genes respectively, C shows intergenic space reduction or expansion, which we refer to as genome compaction, and D shows gene size reduction or increase. **Bottom row** reduced or streamlined genome relative to the starting genome shown in top row; alternatively, the starting genome before expansion to genome shown in top row. Large blue arrows indicate time or direction of evolutionary events. Small dotted bidirectional arrows show hypothetical insertion or deletion events.

Figure 7. Factors influencing genome size of acidophiles across optimal growth pH. Every point corresponds to the average for a different species. **(A)** Number of genes (ORFs, open reading frames) across pH. Pearson's correlation coefficient is 0.18 with p-value 1.25×10^{-4} . **(B)** Intergenic space vs pH. Intergenic space is defined as genome size minus the sum of the nucleotide length of all protein coding genes as defined by ORFs of a genome divided by genome size, in percentage. A stricter genome quality filter of 97% completeness and 2% contamination was used in this analysis to minimize missannotation errors due to fragmented genomes. Pearson's correlation coefficient is -0.11 with p-value 0.06. **(C)** Average ORF length per genome across pH. Pearson's correlation coefficient is 0.25 with p-value 4.03×10^{-8} .

Figure 8. Protein size versus pH correlations for conserved Pfams. **(A)** Pfams present in over 90% of species and in a pH span of at least 6 pH units were selected for analysis. For each Pfam, the Pearson's correlation coefficient for protein size vs organism optimal growth pH was calculated, using the species averages as data. Each point corresponds to a different Pfam. Positive correlations (91 red points to the right) indicate Pfams whose proteins are shorter at low pH while negative correlations (17 purple points to the left) are Pfams whose proteins are larger at low pH. The 25 Pfams with the lowest p-values are listed in Supplementary Table 3. **(B)** Analog to **(A)**, but for a list of Pfams that in addition to being present in over 90% of the species and in a span of at least 6 pH units were also in a unique copy in the genomes (proteins with the Pfam per genome < 1.1) and only one domain architecture was dominant in the proteins. These Pfams are listed in Supplementary table 4. For both plots, an FDR q-value of 0.05 was used for statistical significance. Significant correlations are shown as big points which are red for positive correlations and purple for negative correlations. Non-significant correlations are shown as small grey points.

Figure 9. Subcellular localization and signal peptide presence of protein families across pH. PSORTb and SignalP were used to predict subcellular location of proteins and signal peptide, respectively. Each point corresponds to a species, and either subcellular localization or signal peptide presence are expressed in terms of percentage of the protein families (ortholog groups). Linear regression curves have been plotted for each category. Pearson's correlation coefficient and p-value respectively are -0.01 and 0.77 for cytoplasmic, -0.19 and 4.4×10^{-5} for inner membrane, 0.21 and 7.5×10^{-6} for Periplasmic, Outer membrane, Cell wall and Exported, and 0.22 with 1.4×10^{-6} for proteins with a signal peptide.

Figure 10. Percentage of protein families with functional classification across pH. Each point corresponds to a species. Blue data points and the blue line correspond to proteins with a COG annotation and orange data points and the orange line correspond to proteins with a Pfam annotation. Pearson's correlation coefficients and p-values are respectively 0.24 and 2×10^{-7} for proteins with a COG annotation, and 0.14 with 2.6×10^{-3} for proteins with a Pfam annotation.

Figure 11. Paralog frequency vs pH by COG category. The percentage of genes (relative to the proteome size) belonging to paralog families (paralog frequency) were calculated for each COG category. Categories where the paralog frequency had a statistically significant correlation with pH

1128 (p-value <0.01) are shown. The mean duplication frequencies at pH 1 and 7 are displayed, calculated
1129 with linear regression (Supplementary Figure 5). ** p-value<0.01, *** p-value<0.001.
1130

Tables

Table 1 | Genomic representativity of protein families by function as defined by COG categories in acidophile genomes

COG Category	Pearson's correlation coefficient	p-value
Increased representativity in acidophiles (p-value<0.01)		
(L) Replication, recombination, and repair	-0.25	3.6*10 ⁻⁸
(F) Nucleotide metabolism and transport	-0.21	5.4*10 ⁻⁶
(C) Energy production and conversion	-0.21	8.0*10 ⁻⁶
(H) Coenzyme metabolism and transport	-0.19	3.0*10 ⁻⁵
(D) Cell cycle control and cell division	-0.16	5.2*10 ⁻⁴
(J) Translation and ribosome	-0.15	1.1*10 ⁻³
(O) Chaperones, post-translational mod.	-0.13	6.3*10 ⁻³
Decreased representativity in acidophiles (p-value<0.01)		
(S) Function unknown	0.30	1.3*10 ⁻¹⁰
(T) Signal transduction mechanisms	0.26	3.4*10 ⁻⁸

Phylum and class

Phyla with acidophiles

Distribution of optimal pH by phylum

Totals

- Species with optimal growth pH <3
- Species with optimal growth pH 3-5

- Laboratory grown strains
- Metagenome assemblies

A **B**

Genome
reduction
or
streamlining

Genome
expansion

A **B**

A