Prediction of Optimal Growth Temperature using only Genome Derived Features

David B. Sauer[1]* and Da-Neng Wang[1]*

[1]Department of Cell Biology, and The Helen L. and Martin S. Kimmel Center for Biology and Medicine, Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, New York, United States of America

*Corresponding authors

E-mail: david.sauer@med.nyu.edu (D.B.S), da-neng.wang@med.nyu.edu (DN.W.)

Abstract

Optimal growth temperature is a fundamental characteristic of all living organisms. Knowledge of this temperature is central to the study the organism, the thermal stability and temperature dependent activity of its genes, and the bioprospecting of its genome for thermally adapted proteins. While high throughput sequencing methods have dramatically increased the availability of genomic information, the growth temperatures of the source organisms are often unknown. This limits the study and technological application of these species and their genomes. Here, we present a novel method for the prediction of growth temperatures of prokaryotes using only genomic sequences. By applying the reverse ecology principle that an organism's genome includes identifiable adaptations to its native environment, we can predict a species' optimal growth temperature with an accuracy of 4.69 °C root-mean-square error and a correlation coefficient of 0.908. The accuracy can be further improved for specific taxonomic clades or by excluding psychrophiles. This method provides a valuable tool for the rapid calculation of organism growth temperature when only the genome sequence is known.

Author Summary

The optimal growth temperature is a fundamental characteristic of all living organisms. It is the temperature at which the organism grows at the greatest rate, and is a consequence of adaptations of that organism to its native environment. These adaptations are contained within the genome of the organism, and therefore species from varying environments have distinct genomic characteristics. Here we use those genomic characteristics to predict a species' optimal growth temperature. This provides a novel tool for describing a key parameter of the species' native environment when it is otherwise unknown. This is particularly valuable as the rate of genome sequencing has increased, while the determination of growth temperature remains laborious.

Introduction

Growth conditions of an organism are essential to its characterization. However, these values may be unknown in organisms which are difficult to culture, “unculturable”, or otherwise poorly characterized. Reverse ecology posits that the evolutionary effects of an organism’s native environment is reflected by adaptations in its genome [1]. Therefore, an organism’s native environment can be identified by comparing its genome to the genomes of other organisms from a range of environments. Notably, this is done without experimental manipulation or interrogation of the organism beyond genome sequencing. Such reverse ecology strategies have been successful in studying adaptation to soil conditions [2], salinity [3], and temperature [4].

Of these environmental pressures, temperature, being a description of the internal energy of the environment, is a particularly strong driving force for adaptation. Prokaryotes are often viable over a range of temperatures, which varies by species. For a particular organism, increasing temperature beyond it’s growth range, corresponding to increased internal energy, can lead to loss of protein and nucleic acid structure. Conversely, a sub-optimal temperature leads to reduced enzyme kinetics and stiffening lipid membranes. Each of these biological consequences may be deleterious to un-adapted organisms. Therefore, it is perhaps not surprising that an organism’s optimal growth temperature (OGT) correlates to quantifiable properties (features) in the organism’s nucleotide and protein sequences. Features correlated with OGT can be identified in the genomic [5], tRNA [6,7], rRNA [6–8], open reading frame [9,10], and in the proteomic sequences [10–13]. Correlations between OGT and tRNA G+C content [6,7] or the charged versus polar amino acid ratios [14] are well known.

Clearly, OGT is a necessary parameter for analyzing physiological processes of an organism or activities of its genes and proteins. [15,16]. However, the experimental determination of OGT is laborious [17,18], and sometimes unattainable [19]. Also, recorded OGT or environmental temperature may be inconsistently measured, particularly in genetic samples not obtained from pure culture [20]. Further, for metagenomic samples the conditions during collection may significantly differ from the originating species' growth environment. This can be due to the organism or its genetic material being found distant from its originating environment [21], or the collected genomic material may be from organisms which are inviable [22]. Even in pure culture in the laboratory, the experimental growth conditions can vary greatly [23] and may not be at the source organisms' OGT [24].

While many previous studies have aimed to identify genes and proteins [25], mutations [16], and mechanisms [15] that drive thermal adaptation, there is also great value in using these adaptive differences to provide data of an organism's native environment when it may not be otherwise known or well-described. A number of parameters have been identified which correlate with OGT [14]. However, those correlations are often weak and therefore of limited predictive value alone. Here, we aim to predict a prokaryotic species' OGT only from its genomic sequence. We set out to develop a novel tool for the ecological characterization of a species based solely on its genome, the study of thermoadaptation, and bioprospecting for thermoadapted genes.

Results

Prokaryote genome redundancy is highly skewed

Of the initial 8270 prokaryotic species with a reported OGT, genome sequences were available for 2708 species. These sequenced species were composed of 2538 Bacteria and 170 Archaea, with OGTs ranging from 4 to 103 °C. A total of 36,721 sequenced genomes for these

species were downloaded, indicating multiple genomes for each species on average. However, the number of genomes per species was highly skewed, with great redundancy for model organisms and pathogens (Fig S1C). To avoid having these relatively few species dominate the analysis, features were averaged by species and all regressions were done by species rather than by genome.

Individual genome derived features correlate with OGT

Based on the reverse genomics principle that an organism's adaptation to its environment is reflected within its genome, we hypothesize that a species' OGT could be predicted based on characteristics of its genome and genome derived sequences. This hypothesis was supported by previous noted correlations between OGT and individual features of the genomic [5,6,26], tRNA [6,7], rRNA [6–8], open reading frame [7,9,10,27,28], and proteomic or protein sequences [10–14,29–35]. These features are quantifiable properties of the sequence, such as G+C content, length, and nucleotide or amino acid fraction. Of the features calculated, 42 were found in this work to be correlated with OGT in the present dataset by the Pearson correlation coefficient with $|r| > 0.3$ (Fig 1, Table S1). However, these correlations to OGT were often weak and therefore insufficient for the calculation of a species' growth temperature. Furthermore, there was a strong association among many features (Fig S2). We therefore decided to consider them simultaneously, using multiple linear regression, with features added individually to minimize multicollinearity. We started by classifying features based on the source sequences (genomic, tRNA, rRNA, open reading frames, and proteome). Multiple linear regressions were calculated, progressively increasing the number of feature classes used in the regression.

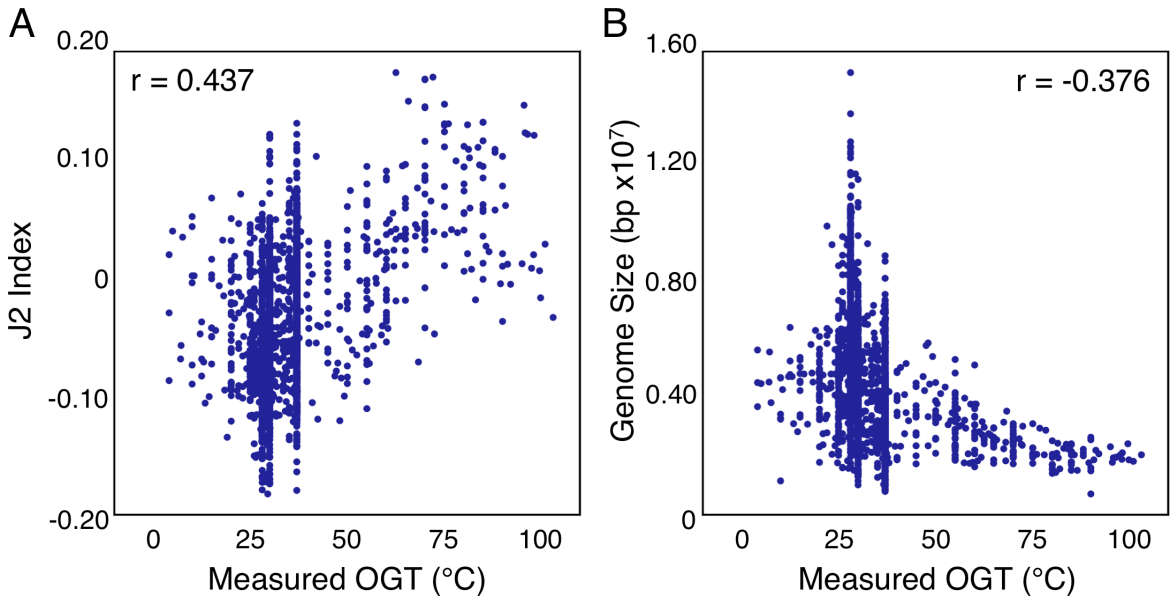


Figure 1. Individual genome derived features correlate weakly with the originating species' OGT. Measure optimal growth temperature for each species versus J2 index of genomic dinucleotide fractions (A) and total genome size (B).

A regression using only genomic sequence based features is weakly predictive of OGT

The genomic sequence provides information about the nucleotide content, nucleotide order, and chromosomal structure of an organism's hereditable genetic material. In the absence of any other knowledge, this sequence still reflects adaptations to the particular thermal environment of the organism. For example, total genome size has been shown to be negatively correlated with a species' OGT [26]. Accordingly, it has been proposed that the reduced time and energy of genomic replication offers selective advantages at higher temperatures. Additionally, the necessity of maintaining genomic structure with increased temperature is thought to be reflected in a species' genomic dinucleotide fractions [36], which is quantified in the J2 index [5].

136 In the present dataset, individual nucleotide and dinucleotide fractions of the genome, the J2
137 index, the G+C content, and total size were calculated for each genome. Of these features, the
138 J2 index, genome size, and the CT and AG dinucleotide fractions correlated with OGT, but only
139 weakly. Using these poorly correlated and collinear input features for regression, the resulting
140 multiple linear regression is poor at predicting OGT with a root mean squared error (RMSE) of
141 9.86 °C (r = 0.469) (Fig S3).
142
143 **tRNA and rRNA sequences improve OGT prediction**
144 tRNA and rRNA are nucleic acids whose structure, and enzymatic activity in the case of rRNA,
145 are essential to cell viability. Therefore, the direct correlation of OGT to G+C content of tRNAs
146 [6,7] and rRNAs [8,37] is thought to reflect the necessary increase in base pair hydrogen
147 bonding needed to maintain the structure of these nucleic acids at elevated temperatures. While
148 a subset of the previously analyzed genomic sequence, we hypothesized that features derived
149 from these tRNA and rRNA sequences might be more strongly correlated with OGT. To this
150 end, we identified their sequences bioinformatically. tRNA and 16S rRNA sequences were
151 identified in 100% and 98% of the species respectively, reflecting the highly conserved nature of
152 these genes.
153
154 Using these identified tRNA and rRNA sequences, nucleotide fractions and G+C content were
155 calculated for each. All calculated features for tRNA and rRNA sequences were correlated with
156 OGT. Calculating a new linear regression with the OGT using tRNA features, in addition to
157 genomic features, improved accuracy (RMSE = 7.30 °C, r = 0.757) (Fig 2A). Similarly, a
158 regression calculated with rRNA and genomic features also improved accuracy (RMSE = 6.99
159 °C, r = 0.784) (Fig 2B). By using all available tRNA, rRNA, and genomic features, a still more
160 accurate linear regression was calculated (RMSE = 6.71 °C, r = 0.802) (Fig 2C).

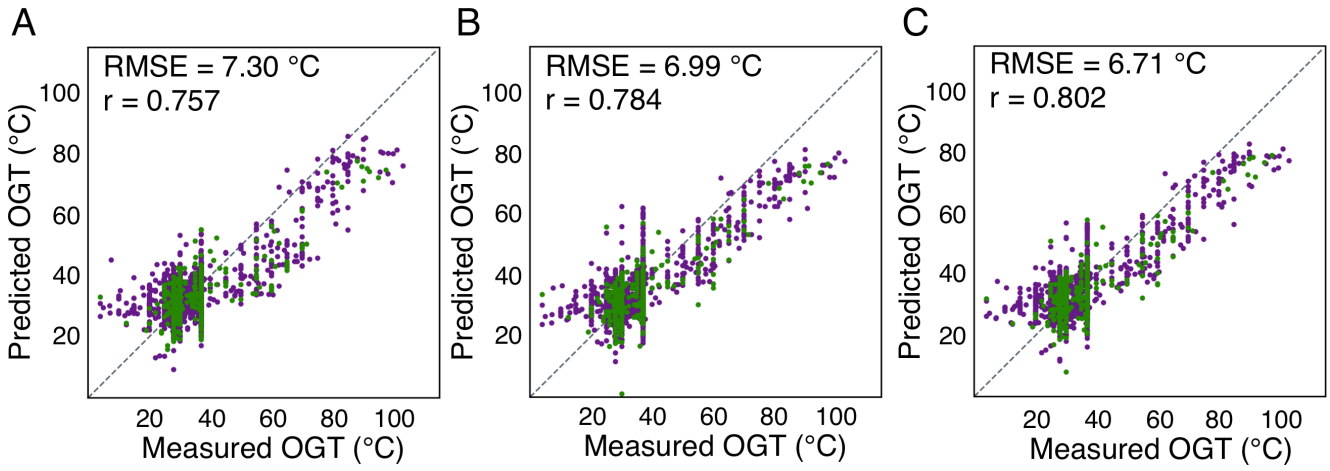
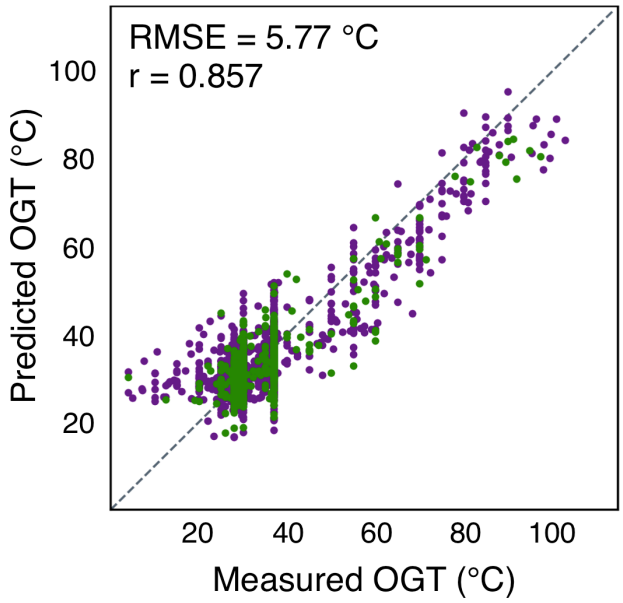


Figure 2. Using genomic and genic sequences improve OGT prediction accuracy. Predicted versus measured OGT for each species, using linear regressions with features derived from genomic and (A) tRNA, (B) rRNA, or (C) tRNA and rRNA sequences. Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

ORF sequences improve OGT prediction

As tRNA and rRNA features clearly improve the ability to predict a species' OGT, we examined if other gene sequences might also improve the regression. In particular open reading frames, which code for proteins but exclude the non-coding regions of the genome, were considered. We hypothesized that using coding regions alone would increase sensitivity to changes in OGT. Additionally, codon biases have previously been reported to correlate with OGT [13], likely reflecting both amino acid differences and the necessity of maintaining proper codon-anticodon pairing in differing thermal environments. Furthermore, the greater number of ORFs in a genome, relative to tRNAs and rRNAs, make the features of ORFs less sensitive to single gene

179 aberrations or mispredictions. Therefore, ORF derived features were hypothesized to more
180 sensitively and accurately report on the thermal environment than tRNA or rRNA sequences.
181
182 We identified ORFs within the genomic sequences bioinformatically. From these ORFs, a
183 number of derived features were calculated including nucleotide and dinucleotide fractions,
184 codon fractions, start and stop codon fractions, the coding ratio and fraction of the genome, the
185 ORF density of the genome, G+C and A+G content, and average length. Of these, nine were
186 found to be correlated with OGT. These include the A+G content, codon and dinucleotide
187 fractions, and the fraction of the alternative start codon TTG. These ORF derived features, in
188 addition to the genomic, tRNA, and rRNA features, were used to calculate a new multiple linear
189 regression with significantly improved accuracy (RMSE = 5.77 °C, r= 0.857) (Fig 3).
190



191
192
193 Figure 3. Open reading frame sequences further improve OGT prediction accuracy. Predicted
194 versus measured OGT for each species, using a linear regression with features derived from

195 genomic, tRNA, rRNA, and ORF sequences. Species used for regression and evaluation are
196 shown in purple and green, respectively. The dotted line indicates a perfect prediction.
197
198 **Including proteome features significantly improves OGT prediction**
199 While ORF feature correlation to OGT partially reflects the adaptation of the coding regions and
200 mRNAs to the thermodynamic environment, it has been suspected that this correlation also
201 reflected adaptations in each species' proteome to OGT. Temperature is known to correlate with
202 protein folding, biochemistry, and enzyme kinetics, all of which are essential to organismal
203 viability [10,14,32]. Based on these biological consequences, proteome derived features were
204 hypothesized to be especially sensitive to thermal environment. Therefore, the proteome was
205 translated from each species' ORFs, and features calculated from the proteome's primary
206 sequence. These features included amino acid fractions, the fraction of the proteome to be
207 charged or thermolablile, and the EK/QH, LK/Q, Polar/Charged, and Polar/Hydrophobic amino
208 acid ratios.
209
210 Supporting this hypothesis, proteome derived features were found to have the strongest
211 correlation to OGT (Table S1), with the greatest correlation being the fraction of the proteome
212 composed of the amino acids ILVWYGERKP [13]. The linear regression of OGT using proteome
213 features, in addition to previously described features, significantly improved accuracy (RMSE =
214 4.69 °C, r = 0.908). (Fig 4, Eq S1).
215

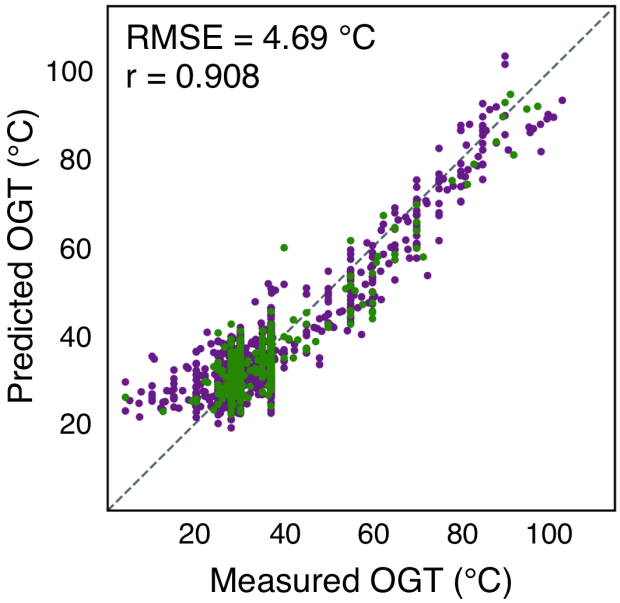


Figure 4. Proteome derived features significantly improve OGT prediction accuracy. Predicted versus measured OGTs for each species, using a linear regression with features derived from genomic, tRNA, rRNA, ORF, and proteome sequences. Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

Taxonomic clade specific regressions are the most accurate

The regressions described up to this point were all made using all prokaryotic species. However, we had noted that the number of individual features correlated with OGT was much higher in Archaea than Bacteria (Table S1). In addition, we hypothesized that the magnitude of the response of each feature to OGT may be distinct in each superkingdom.

Based on these distinctions, we tested whether superkingdom specific regressions would be more accurate than the regression of all prokaryotes (Fig. 5). Using the NCBI taxonomic assignment for each species, an Archaea-only regression dramatically improved accuracy for

these species (RMSE = 3.21 °C, r = 0.995) (Eq S2). However, the Bacteria-only regression only showed only a slight improvement (RMSE = 4.61 °C, r = 0.816) (Eq S3). This likely reflects bias of the general prokaryotic regression, due to the numerical majority of bacterial species and the greater diversity of bacterial species.

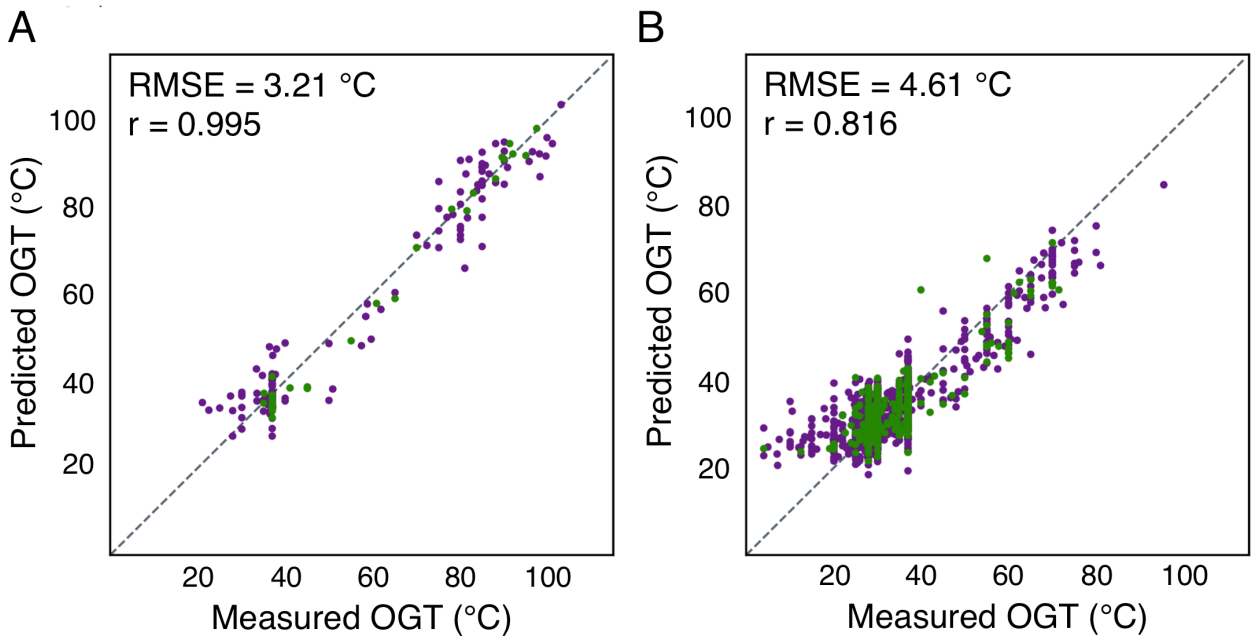


Figure 5. Taxon specific linear regressions are most accurate. Predicted versus measured OGT for each species using superkingdom specific linear regressions for Archaea (A) and Bacteria (B). Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

Addressing this diversity in bacteria, the taxonomic specific regression can be further improved when the data is separated by phylum or class. OGT regression was limited to clades where the number of species (N) was greater than 50 to ensure the significance of the regression. Of the individual phyla, the most accurate regressions are found in the *Firmicutes* (RMSE = 4.88 °C, r = 0.831), *Actinobacteria* (RMSE = 2.90 °C, r = 0.818), *Bacteroidetes* (RMSE = 1.58 °C, r =

249 0.964), and *Euryarchaeota* (RMSE = 4.00 °C, r = 0.985) (Fig S4). In contrast, the *Proteobacteria*
250 regression had much more weakly correlated predicted and reported OGTs (RMSE = 4.10 °C, r
251 = 0.569), though the small RMSE likely reflects the narrow OGT range of this phylum. Further
252 subdivision of the *Proteobacteria* into classes (Fig S5) resulted in significant correlation of the
253 *Betaproteobacteria* (RMSE = 2.94 °C, r = 0.789), and *Deltaproteobacteria* regressions (RMSE =
254 2.04 °C, r = 0.761). However, no correlation was found in regressions for the *Proteobacteria*
255 classes of *Alphaproteobacteria* or *Gammaproteobacteria*.
256
257 **Discussion**
258 Knowing an organism's optimal growth characteristics is central to addressing basic biological
259 questions about how organisms adapt to a particular environmental niche. Further, the
260 systematic study of adaption often requires the optimal growth conditions of the species of origin
261 for each species and gene or protein examined. Additionally, proteins from organisms adapted
262 to particular environmental niches are often particularly suited for structural biology [38–40] and
263 industrial applications [41,42].
264
265 However, if the growth characteristics of already sequenced organisms are uncharacterized, the
266 physiochemical properties of these genes that otherwise might be inferred are lost [20].
267 Consequently, this limits the use of these genomes in academic study and mining for
268 biotechnology applications. Exacerbating this issue, high throughput sequencing has enabled
269 rapid growth in the number of available genomic, metagenomic, and derived proteomic
270 sequences. This growth in genetic information is likely to outpace the laborious experimental
271 task of characterizing the growth conditions of each species, leading to an increasing number of
272 genomic sequences with unknown growth characteristics. This is already apparent by those

273 organisms which have been 'unculturable' to date, but which have been sequenced by
274 metagenomics.
275
276 To satisfy the need for growth condition data when only genomic sequences are available, here
277 we demonstrate a novel reverse ecology tool to accurately predict the OGT using solely the
278 genomic sequence as input. Our method can predict the OGT for sequenced Archaea and
279 bacteria with an accuracy of 3.21 °C and 4.61 °C, respectively.
280
281 **OGT can be accurately predicted using only genome derived parameters**
282 Genome classification is clearly essential to the most accurate prediction of OGT. The programs
283 used for tRNA, rRNA, and ORF identification all require some level of taxonomic classification.
284 When applying the general prokaryotic regression, this is only requires the relatively simple
285 exclusion of eukaryotic samples prior to sequencing [43]. However, the most accurate OGT
286 regressions are taxon specific, and therefore genomic samples require further classification.
287 This assignment is routinely addressed *in silico*, using specialized bioinformatic tools which can
288 easily assign taxonomic clade to genomic material [44,45].
289
290 As a simple proof-of-concept, the prokaryotic genomes were also classified by superkingdom
291 using the best scoring 16S rRNA hidden Markov model in Barrnap (Fig S6). These regressions
292 were of similar accuracy to those using NCBI superkingdom assignments.
293
294 **Excluding genome size does not alter the regression accuracy**
295 While prokaryote genome size is strongly correlated with OGT, it is unique among all features
296 used here in requiring a complete genome for calculation. Therefore, this feature might not be
297 available in metagenomic samples, or otherwise incompletely assembled genomes. Excluding

298 this feature has only a minor impact on the regression for all prokaryotes (RMSE = 5.07 °C, r =
299 0.891), or the separate regressions for Bacteria (RMSE = 4.97 °C, r = 0.783), or Archaea
300 (RMSE = 3.21 °C, r = 0.995) (Fig S7).
301
302 **Psychrophiles are poorly fit**
303 While the final regressions of prokaryotes and Bacteria were generally accurate, species with
304 optimal growth temperatures less than approximately 25 °C are clearly poorly fit. This outcome
305 is unsurprising, as few psychrophilic sequences are present in the dataset (Fig. S1), and the
306 mechanisms of thermoadaptation to higher and lower temperatures are not equivalent [46].
307 Excluding those species with an OGT of less than 25 °C yields a slightly better general
308 prokaryotic regression (RMSE = 4.42 °C, r = 0.916) (Fig. S6). The archaeal regression was
309 slightly worse (RMSE = 3.12 °C, r = 0.993), while the bacterial regression improved (RMSE =
310 4.26 °C, r = 0.832), reflecting the known OGT ranges of each superkingdom.
311
312 **Improvements over comparable methods**
313 Our method significantly expands and improves upon the individual features previously
314 described to correlate with OGT. By studying a much larger set of genomes, a more precise
315 correlation between each feature and OGT can be calculated. Further, by using multiple
316 features, more accurate and predictive regression models have been calculated. Notably, our
317 method improves on previously reported analyses requiring particular genes being present in
318 the genome, thereby making the method more general in application [47]. Also, this method
319 quantitatively predicts an OGT rather than using classification (psychrophile, mesophile,
320 thermophile, or hyperthermophile). This improves on methods which predict OGT ranges [47–
321 50], where classification necessarily limited accuracy.
322

323 The most comparable method is reported by Zeldovich *et al.* calculating OGT from the proteome
324 as OGT = 937F – 335, where F is the sum of the proteome fraction for the amino acids
325 IVYWREL [13]. Using the current larger dataset, we calculate a lower correlation (r = 0.726) and
326 accuracy (RMSE = 10.5 °C) than previously reported. This is likely a consequence of more
327 genomic sequences being available, and our keeping of individual species separate rather than
328 averaging those with the same OGT. By considering more features derived from the source
329 organism's genome, the prokaryotic regression presented here clearly advances upon this
330 previous method improving in both correlation and accuracy. While we focus on growth
331 temperature, the same principle could be readily applied to other quantifiable characteristics of
332 an organism's optimal growth environment, such as pH, salinity, osmolarity, or oxygen
333 concentration.
334
335 **Application and validation**
336 Applying these regressions, we predicted OGTs for those species with a genomic sequence
337 available, but without a reported OGT in Sauer *et al.* (2015), using the most taxon specific linear
338 regression available. Only the *Betaproteobacteria* and *Deltaproteobacteria* classes of
339 *Proteobacteria* were predicted, excluding the *Alphaproteobacteria*, *Gammaproteobacteria*, and
340 other *Proteobacteria* due to the poor predictive values of those taxon specific regressions. In
341 total, 482 species' OGTs were predicted (Table S2). Of the species with newly predicted OGTs,
342 a more recent literature search revealed reported OGTs for 36 species [51–87]. The predicted
343 and measured OGTs were strongly correlated (RMSE = 6.94 °C, r = 0.857), validating the
344 predictive value of this method (Fig S9).
345
346 **Materials and Methods**
347 **Source data and sequence extraction**

348 Experimentally measured OGTs of various prokaryotic species were used as previously
349 published without modification [88]. Taxonomic assignments for each species were collected
350 from NCBI [89]. All available top level genome sequences for each species were downloaded
351 from Ensembl [90]. tRNA sequences were identified with tRNAScan-SE 1.3.1 [91] with general
352 settings. Ribosomal RNA genes were identified with Barrnap 0.8 [92] using superkingdom
353 specific hidden Markov models, and rRNA sequences extracted from the genome using
354 BEDtools 2.26.0 [93]. Open reading frame sequences were identified with GenemarkS 4.32 [94]
355 using the default settings. ORFs were also translated into protein sequences using the standard
356 genetic code. Features were calculated for each genome and derived proteome, ignoring
357 ambiguous nucleotides and amino acids. All calculated features were averaged by species.
358 Twenty percent of the species with available genomes were set aside as a test set and never
359 used for regression, only evaluation.
360

361 **Multiple linear regression**

362 Only individual features linearly correlated with OGT ($|r| > 0.3$) were used for multiple linear
363 regression. To minimize multicollinearity, the initial regression input feature set consisted of only
364 the feature most correlated with OGT. To this set all other correlated features were added
365 individually, and multiple linear regressions were calculated. If the correlation between
366 measured and predicted OGTs increased for any regression, the input feature which most
367 increased the correlation was added to the input set. This was repeated until the correlation did
368 not increase.
369

370 **Regression evaluation and prediction**

371 The test set was only used for evaluation of the multiple linear regressions, comparing the
372 calculated and measured OGTs. Regressions were evaluated by comparing the predicted and
373 reported OGT using the Pearson correlation coefficient and root mean square error.
374

375 ***De novo* OGT prediction and validation**

376 All top level genomes in Ensembl Bacteria were downloaded for each species where there was
377 not a reported OGT in the Sauer *et al.* (2015) dataset. Taxonomic assignment and feature
378 calculation were preformed as described above. The most taxonomic specific regression
379 available, using genomic, tRNA, ORF, and proteome features, was used to predict the OGT for
380 each species. For these newly predicted species, Pubmed was searched using the binomial
381 name and "optimal growth" as keywords. From the returned publications, OGTs were manually
382 collected where available.
383

384 Analyses were carried out using custom Python scripts using Biopython 2.7.12 [95], NumPy
385 1.13.3 [96], SciPy 1.0.0, Scikit-learn 0.19.1 [97], and MatPlotLib 2.1.0 [98].
386

387 **Acknowledgements**

388 The authors thank Jennifer Marden for discussion and critical review of this manuscript.
389

390 This work was financially supported in part by an American Cancer Society Postdoctoral
391 Fellowship (16-A1-00-005739 to D.B.S), the Department of Defense (W81XWH-16-1-0153 to
392 D.B.S.), and NIH (R01-GM121994, R01-DK099023, and R01-GM093825 to D-N.W). This work
393 was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the
394 Peer Reviewed Cancer Research Program under Award No. W81XH-16-1-0153. Opinions,

interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

399 **References**

400 1. Li YF, Costello JC, Holloway AK, Hahn MW. "Reverse ecology" and the power of
401 population genomics. Evol Int J Org Evol. 2008;62: 2984–2994. doi:10.1111/j.1558-
402 5646.2008.00486.x

403 2. Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. Population resequencing
404 reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 2010;42: 260–
405 263. doi:10.1038/ng.515

406 3. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population
407 genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS
408 Genet. 2010;6: e1000862. doi:10.1371/journal.pgen.1000862

409 4. Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, et al. Population genomics and
410 local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci U S A.
411 2011;108: 2831–2836. doi:10.1073/pnas.1014971108

412 5. Kawashima T, Amano N, Koike H, Makino S, Higuchi S, Kawashima-Ohya Y, et al.
413 Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma*
414 *volcanium*. Proc Natl Acad Sci U S A. 2000;97: 14257–14262.
415 doi:10.1073/pnas.97.26.14257

416 6. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary
417 structures, and optimal growth temperature in prokaryotes. J Mol Evol. 1997;44: 632–636.

418 7. Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high
419 temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci. 2001;268: 493–
420 497. doi:10.1098/rspb.2000.1397

421 8. Khachane AN, Timmis KN, dos Santos VAPM. Uracil content of 16S rRNA of thermophilic
422 and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures.
423 Nucleic Acids Res. 2005;33: 4016–4022. doi:10.1093/nar/gki714

424 9. Lynn DJ, Singer GAC, Hickey DA. Synonymous codon usage is subject to selection in
425 thermophilic bacteria. Nucleic Acids Res. 2002;30: 4272–4277.

426 10. Singer GAC, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon
427 usage, amino acid composition and nucleotide content. Gene. 2003;317: 39–47.

428 11. Lobry JR, Chessel D. Internal correspondence analysis of codon and amino-acid usage in
429 thermophilic bacteria. J Appl Genet. 2003;44: 235–261.

430 12. Tekaia F, Yeramian E, Dujon B. Amino acid composition of genomes, lifestyles of
431 organisms, and evolutionary trends: a global picture with correspondence analysis. Gene.
432 2002;297: 51–60.

433 13. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of
434 thermophilic adaptation. PLoS Comput Biol. 2007;3: e5. doi:10.1371/journal.pcbi.0030005

435 14. Suhre K, Claverie J-M. Genomic correlates of hyperthermostability, an update. J Biol Chem.
436 2003;278: 17198–17202. doi:10.1074/jbc.M301327200

437 15. Nguyen V, Wilson C, Hoemberger M, Stiller JB, Agafonov RV, Kutter S, et al. Evolutionary
438 drivers of thermoadaptation in enzyme catalysis. Science. 2017;355: 289–294.
439 doi:10.1126/science.aah3717

440 16. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer
441 thermostability on a cold shock protein. Nat Struct Biol. 2000;7: 380–383.
442 doi:10.1038/75151

443 17. Elliott RP. Temperature-Gradient Incubator for Determining the Temperature Range of
444 Growth of Microorganisms. J Bacteriol. 1963;85: 889–894.

445 18. Honglin Z, Yongjun L, Haitao S. Determination of thermograms of bacterial growth and
446 study of optimum growth temperature. Thermochim Acta. 1993;216: 19–23.
447 doi:10.1016/0040-6031(93)80377-M

448 19. Stewart EJ. Growing unculturable bacteria. J Bacteriol. 2012;194: 4151–4160.
449 doi:10.1128/JB.00345-12

450 20. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide
451 to metagenomics. Microbiol Mol Biol Rev MMBR. 2008;72: 557–578, Table of Contents.
452 doi:10.1128/MMBR.00009-08

453 21. Rose M, Landman D, Quale J. Are community environmental surfaces near hospitals
454 reservoirs for gram-negative nosocomial pathogens? Am J Infect Control. 2014;42: 346–348.
455 doi:10.1016/j.ajic.2013.12.025

456 22. Cangelosi GA, Meschke JS. Dead or alive: molecular assessment of microbial viability. Appl
457 Environ Microbiol. 2014;80: 5884–5891. doi:10.1128/AEM.01763-14

458 23. Hearing J, Hunter E, Rodgers L, Gething MJ, Sambrook J. Isolation of Chinese hamster
459 ovary cell lines temperature conditional for the cell-surface expression of integral membrane
460 glycoproteins. J Cell Biol. 1989;108: 339–353.

461 24. Hashimoto H, Moritani N, Saito TR. Comparative study on circadian rhythms of body
462 temperature, heart rate, and locomotor activity in three species hamsters. Exp Anim.
463 2004;53: 43–46.

464 25. Wang Q, Cen Z, Zhao J. The survival mechanisms of thermophiles at high temperatures: an
465 angle of omics. Physiol Bethesda Md. 2015;30: 97–106. doi:10.1152/physiol.00066.2013

466 26. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in bacteria
467 are negatively correlated, suggesting genomic streamlining during thermal adaptation.
468 Genome Biol Evol. 2013;5: 966–977. doi:10.1093/gbe/evt050

469 27. Li W, Zou H, Tao M. Sequences downstream of the start codon and their relations to G + C
470 content and optimal growth temperature in prokaryotic genomes. Antonie Van
471 Leeuwenhoek. 2007;92: 417–427. doi:10.1007/s10482-007-9170-6

472 28. Zheng H, Wu H. Gene-centric association analysis for the correlation between the guanine-
473 cytosine content levels and temperature range conditions of prokaryotic species. BMC
474 Bioinformatics. 2010;11 Suppl 11: S7. doi:10.1186/1471-2105-11-S11-S7

475 29. Burra PV, Kalmar L, Tompa P. Reduction in structural disorder and functional complexity in
476 the thermal adaptation of prokaryotes. PloS One. 2010;5: e12069.
477 doi:10.1371/journal.pone.0012069

478 30. Robinson-Rechavi M, Alibés A, Godzik A. Contribution of electrostatic interactions,
479 compactness and quaternary structure to protein thermostability: lessons from structural
480 genomics of Thermotoga maritima. J Mol Biol. 2006;356: 547–557.
481 doi:10.1016/j.jmb.2005.11.065

482 31. Puigbò P, Pasamontes A, Garcia-Vallve S. Gaining and losing the thermophilic adaptation in
483 prokaryotes. Trends Genet TIG. 2008;24: 10–14. doi:10.1016/j.tig.2007.10.005

484 32. Cambillau C, Claverie JM. Structural and genomic correlates of hyperthermostability. J Biol
485 Chem. 2000;275: 32383–32386. doi:10.1074/jbc.C000497200

486 33. Saelensminde G, Halskau Ø, Helland R, Willassen N-P, Jonassen I. Structure-dependent
487 relationships between growth temperature of prokaryotes and the amino acid frequency in
488 their proteins. Extrem Life Extreme Cond. 2007;11: 585–596. doi:10.1007/s00792-007-
489 0072-3

490 34. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid
491 compositions deduced from their genomes. Nucleic Acids Res. 2001;29: 1608–1615.

492 35. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal adaptation
493 analyzed by comparison of protein sequences from mesophilic and extremely thermophilic
494 *Methanococcus* species. Proc Natl Acad Sci U S A. 1999;96: 3578–3583.

495 36. Amano N, Ohfuku Y, Suzuki M. Genomes and DNA conformation. Biol Chem. 1997;378:
496 1397–1404.

497 37. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life
498 forms. Science. 1999;283: 220–221.

499 38. Yernool D, Boudker O, Jin Y, Gouaux E. Structure of a glutamate transporter homologue
500 from *Pyrococcus horikoshii*. Nature. 2004;431: 811–818. doi:10.1038/nature03018

501 39. Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R. Crystal structure and
502 mechanism of a calcium-gated potassium channel. Nature. 2002;417: 515–522.
503 doi:10.1038/417515a

504 40. Karpowich NK, Wang D-N. Assembly and mechanism of a group II ECF transporter. Proc
505 Natl Acad Sci U S A. 2013;110: 2534–2539. doi:10.1073/pnas.1217361110

506 41. Acharya S, Chaudhary A. Bioprospecting thermophiles for cellulase production: a review.
507 Braz J Microbiol Publ Braz Soc Microbiol. 2012;43: 844–856. doi:10.1590/S1517-
508 83822012000300001

509 42. Koskinen PEP, Lay C-H, Beck SR, Tolvanen KES, Kaksonen AH, Örlygsson J, et al.
510 Bioprospecting Thermophilic Microorganisms from Icelandic Hot Springs for Hydrogen and
511 Ethanol Production. Energy Fuels. 2008;22: 134–140. doi:10.1021/ef700275w

512 43. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al.
513 Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304: 66–74.
514 doi:10.1126/science.1093857

515 44. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact
516 alignments. Genome Biol. 2014;15: R46. doi:10.1186/gb-2014-15-3-r46

517 45. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification
518 of metagenomic sequences. Genome Res. 2016;26: 1721–1729. doi:10.1101/gr.210641.116

519 46. Yang L-L, Tang S-K, Huang Y, Zhi X-Y. Low Temperature Adaptation Is Not the Opposite
520 Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition.
521 Genome Biol Evol. 2015;7: 3426–3433. doi:10.1093/gbe/evv232

522 47. Jensen DB, Vesth TC, Hallin PF, Pedersen AG, Ussery DW. Bayesian prediction of bacterial
523 growth temperature range based on genome sequences. BMC Genomics. 2012;13 Suppl 7:
524 S3. doi:10.1186/1471-2164-13-S7-S3

525 48. Taylor TJ, Vaisman II. Discrimination of thermophilic and mesophilic proteins. BMC Struct
526 Biol. 2010;10 Suppl 1: S5. doi:10.1186/1472-6807-10-S1-S5

527 49. Li Y, Middaugh CR, Fang J. A novel scoring function for discriminating hyperthermophilic
528 and mesophilic proteins with application to predicting relative thermostability of protein
529 mutants. BMC Bioinformatics. 2010;11: 62. doi:10.1186/1471-2105-11-62

530 50. Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. J
531 Microbiol Methods. 2011;84: 67–70. doi:10.1016/j.mimet.2010.10.013

532 51. Ogg CD, Patel BKC. *Thermotalea metallivorans* gen. nov., sp. nov., a thermophilic,
533 anaerobic bacterium from the Great Artesian Basin of Australia aquifer. Int J Syst Evol
534 Microbiol. 2009;59: 964–971. doi:10.1099/ijs.0.004218-0

535 52. Zhao W, Zeng X, Xiao X. *Thermococcus eurythermalis* sp. nov., a conditional piezophilic,
536 hyperthermophilic archaeon with a wide temperature range for growth, isolated from an oil-
537 immersed chimney in the Guaymas Basin. Int J Syst Evol Microbiol. 2015;65: 30–35.
538 doi:10.1099/ijs.0.067942-0

539 53. Puente-Sánchez F, Sánchez-Román M, Amils R, Parro V. *Tessaracoccus lapidicaptus* sp.
540 nov., an actinobacterium isolated from the deep subsurface of the Iberian pyrite belt. Int J
541 Syst Evol Microbiol. 2014;64: 3546–3552. doi:10.1099/ijs.0.060038-0

542 54. Debnath R, Saikia R, Sarma RK, Yadav A, Bora TC, Handique PJ. Psychrotolerant
543 antifungal *Streptomyces* isolated from Tawang, India and the shift in chitinase gene family.
544 Extrem Life Extreme Cond. 2013;17: 1045–1059. doi:10.1007/s00792-013-0587-8

545 55. Chen Z, Feng D, Zhang B, Wang Q, Luo Y, Dong X. Proteomic insights into the temperature
546 responses of a cold-adaptive archaeon *Methanolobus psychrophilus* R15. Extrem Life
547 Extreme Cond. 2015;19: 249–259. doi:10.1007/s00792-014-0709-y

548 56. Pivovarova TA, Kondrat'eva TF, Batrakov SG, Esipov SE, Sheĭchenko VI, Bykova SA, et
549 al. Phenotypic features of *Ferroplasma acidiphilum* strains Yt and Y-2. Mikrobiologiia.
550 2002;71: 809–818.

551 57. Dsouza M, Taylor MW, Ryan J, MacKenzie A, Lagutin K, Anderson RF, et al. *Paenibacillus*
552 *darwinianus* sp. nov., isolated from gamma-irradiated Antarctic soil. Int J Syst Evol
553 Microbiol. 2014;64: 1406–1411. doi:10.1099/ijs.0.056697-0

554 58. Sukweenadhi J, Kim Y-J, Lee KJ, Koh S-C, Hoang V-A, Nguyen N-L, et al. *Paenibacillus*
555 *yonginensis* sp. nov., a potential plant growth promoting bacterium isolated from humus soil
556 of Yongin forest. Antonie Van Leeuwenhoek. 2014;106: 935–945. doi:10.1007/s10482-014-
557 0263-8

558 59. Kwon YM, Yang S-H, Kwon KK, Kim S-J. *Nonlabens antarcticus* sp. nov., a psychrophilic
559 bacterium isolated from glacier ice, and emended descriptions of *Nonlabens marinus* Park et
560 al. 2012 and *Nonlabens agnitus* Yi and Chun 2012. Int J Syst Evol Microbiol. 2014;64: 400–
561 405. doi:10.1099/ijs.0.056606-0

562 60. Stieglmeier M, Klingl A, Alves RJE, Rittmann SK-MR, Melcher M, Leisch N, et al.
563 *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-
564 oxidizing archaeon from soil and a member of the archaeal phylum *Thaumarchaeota*. Int J
565 Syst Evol Microbiol. 2014;64: 2738–2752. doi:10.1099/ijs.0.063172-0

566 61. Cui H-L, Tohty D, Liu H-C, Liu S-J, Oren A, Zhou P-J. *Natronorubrum sulfidifaciens* sp.
567 nov., an extremely haloalkaliphilic archaeon isolated from Aiding salt lake in Xin-Jiang,
568 China. Int J Syst Evol Microbiol. 2007;57: 738–740. doi:10.1099/ijs.0.64651-0

569 62. Itoh T, Yamaguchi T, Zhou P, Takashina T. *Natronolimnobius baerhuensis* gen. nov., sp.
570 nov. and *Natronolimnobius innermongolicus* sp. nov., novel haloalkaliphilic archaea isolated
571 from soda lakes in Inner Mongolia, China. Extrem Life Extreme Cond. 2005;9: 111–116.
572 doi:10.1007/s00792-004-0426-z

63. Xin H, Itoh T, Zhou P, Suzuki K, Nakase T. *Natronobacterium nitratireducens* sp. nov., a aloalkaliphilic archaeon isolated from a soda lake in China. Int J Syst Evol Microbiol. 2001;51: 1825–1829. doi:10.1099/00207713-51-5-1825

64. Kern T, Fischer MA, Deppenmeier U, Schmitz RA, Rother M. *Methanosarcina flavescens* sp. nov., a methanogenic archaeon isolated from a full-scale anaerobic digester. Int J Syst Evol Microbiol. 2016;66: 1533–1538. doi:10.1099/ijsem.0.000894

65. Sun L, Toyonaga M, Ohashi A, Tourlousse DM, Matsuura N, Meng X-Y, et al. *Lentimicrobium saccharophilum* gen. nov., sp. nov., a strictly anaerobic bacterium representing a new family in the phylum *Bacteroidetes*, and proposal of *Lentimicrobiaceae* fam. nov. Int J Syst Evol Microbiol. 2016;66: 2635–2642. doi:10.1099/ijsem.0.001103

66. Baek K, Choi A, Kang I, Lee K, Cho J-C. *Kordia antarctica* sp. nov., isolated from Antarctic seawater. Int J Syst Evol Microbiol. 2013;63: 3617–3622. doi:10.1099/ijs.0.052738-0

67. Surendra V, Bhawana P, Suresh K, Srinivas TNR, Kumar PA. *Imtechella halotolerans* gen. nov., sp. nov., a member of the family *Flavobacteriaceae* isolated from estuarine water. Int J Syst Evol Microbiol. 2012;62: 2624–2630. doi:10.1099/ijs.0.038356-0

68. Birkenbihl RP, Neef K, Prangishvili D, Kemper B. Holliday junction resolving enzymes of archaeal viruses SIRV1 and SIRV2. J Mol Biol. 2001;309: 1067–1076. doi:10.1006/jmbi.2001.4761

69. Castillo AM, Gutiérrez MC, Kamekura M, Ma Y, Cowan DA, Jones BE, et al. *Halovivax asiaticus* gen. nov., sp. nov., a novel extremely halophilic archaeon isolated from Inner Mongolia, China. Int J Syst Evol Microbiol. 2006;56: 765–770. doi:10.1099/ijs.0.63954-0

70. Gutiérrez MC, Castillo AM, Kamekura M, Ventosa A. *Haloterrigena salina* sp. nov., an extremely halophilic archaeon isolated from a salt lake. Int J Syst Evol Microbiol. 2008;58: 2880–2884. doi:10.1099/ijs.0.2008/001602-0

71. Cui H-L, Tohty D, Zhou P-J, Liu S-J. *Haloterrigena longa* sp. nov. and *Haloterrigena limicola* sp. nov., extremely halophilic archaea isolated from a salt lake. Int J Syst Evol Microbiol. 2006;56: 1837–1840. doi:10.1099/ijs.0.64372-0

72. Gutiérrez MC, Castillo AM, Pagaling E, Heaphy S, Kamekura M, Xue Y, et al. *Halorubrum kocurii* sp. nov., an archaeon isolated from a saline lake. Int J Syst Evol Microbiol. 2008;58: 2031–2035. doi:10.1099/ijs.0.65840-0

73. Hong H, Kim S-J, Min U-G, Lee Y-J, Kim S-G, Jung M-Y, et al. *Geosporobacter ferrireducens* sp. nov., an anaerobic iron-reducing bacterium isolated from an oil-contaminated site. Antonie Van Leeuwenhoek. 2015;107: 971–977. doi:10.1007/s10482-015-0389-3

74. Söderholm H, Derman Y, Lindström M, Korkeala H. Functional csdA is needed for effective adaptation and initiation of growth of *Clostridium botulinum* ATCC 3502 at suboptimal temperature. Int J Food Microbiol. 2015;208: 51–57. doi:10.1016/j.ijfoodmicro.2015.05.013

610 75. Davidova IA, Wawrik B, Callaghan AV, Duncan K, Marks CR, Suflita JM.
611 *Dethiosulfatarculus sandiegensis* gen. nov., sp. nov., isolated from a methanogenic paraffin-
612 degrading enrichment culture and emended description of the family *Desulfarculaceae*. Int J
613 Syst Evol Microbiol. 2016;66: 1242–1248. doi:10.1099/ijsem.0.000864

614 76. Abin CA, Hollibaugh JT. *Desulfuribacillus stibiiarsenatis* sp. nov., an obligately anaerobic,
615 dissimilatory antimonate- and arsenate-reducing bacterium isolated from anoxic sediments,
616 and emended description of the genus *Desulfuribacillus*. Int J Syst Evol Microbiol. 2017;67:
617 1011–1017. doi:10.1099/ijsem.0.001732

618 77. An TT, Picardal FW. *Desulfocarbo indianensis* gen. nov., sp. nov., a benzoate-oxidizing,
619 sulfate-reducing bacterium isolated from water extracted from a coal bed. Int J Syst Evol
620 Microbiol. 2014;64: 2907–2914. doi:10.1099/ijs.0.064873-0

621 78. Hahnke S, Langer T, Koeck DE, Klocke M. Description of *Proteiniphilum*
622 *saccharofermentanssp*. nov., *Petrimonas mucosasp*. nov. and *Fermentimonas caenicolagen*.
623 nov., sp. nov., isolated from mesophilic laboratory-scale biogas reactors, and emended
624 description of the genus *Proteiniphilum*. Int J Syst Evol Microbiol. 2016;66: 1466–1475.
625 doi:10.1099/ijsem.0.000902

626 79. Hahnke S, Striesow J, Elvert M, Mollar XP, Klocke M. *Clostridium bornimense* sp. nov.,
627 isolated from a mesophilic, two-phase, laboratory-scale biogas reactor. Int J Syst Evol
628 Microbiol. 2014;64: 2792–2797. doi:10.1099/ijs.0.059691-0

629 80. Xu Y, Zhou P, Tian X. Characterization of two novel haloalkaliphilic archaea
630 *Natronorubrum bangense* gen. nov., sp. nov. and *Natronorubrum tibetense* gen. nov., sp.
631 nov. Int J Syst Bacteriol. 1999;49 Pt 1: 261–266. doi:10.1099/00207713-49-1-261

632 81. Yang S-H, Seo H-S, Woo J-H, Oh H-M, Jang H, Lee J-H, et al. *Carboxylicivirga* gen. nov.
633 in the family *Marinilabiliaceae* with two novel species, *Carboxylicivirga mesophila* sp. nov.
634 and *Carboxylicivirga taeanensis* sp. nov., and reclassification of *Cytophaga fermentans* as
635 *Saccharicrinis fermentans* gen. nov., comb. nov. Int J Syst Evol Microbiol. 2014;64: 1351–
636 1358. doi:10.1099/ijs.0.053462-0

637 82. Lee G-H, Rhee M-S, Chang D-H, Kwon KK, Bae KS, Yang S-H, et al. *Bacillus*
638 *solimangrovi* sp. nov., isolated from mangrove soil. Int J Syst Evol Microbiol. 2014;64:
639 1622–1628. doi:10.1099/ijs.0.058230-0

640 83. Dunlap CA, Kwon S-W, Rooney AP, Kim S-J. *Bacillus paralicheniformis* sp. nov., isolated
641 from fermented soybean paste. Int J Syst Evol Microbiol. 2015;65: 3487–3492.
642 doi:10.1099/ijsem.0.000441

643 84. Dunlap CA, Saunders LP, Schisler DA, Leathers TD, Naeem N, Cohan FM, et al. *Bacillus*
644 *nakamurai* sp. nov., a black-pigment-producing strain. Int J Syst Evol Microbiol. 2016;66:
645 2987–2991. doi:10.1099/ijsem.0.001135

646 85. Kim S-J, Dunlap CA, Kwon S-W, Rooney AP. *Bacillus glycinifermentans* sp. nov., isolated
647 from fermented soybean paste. Int J Syst Evol Microbiol. 2015;65: 3586–3590.
648 doi:10.1099/ijsem.0.000462

649 86. Shi W, Takano T, Liu S. *Anditalea andensis* gen. nov., sp. nov., an alkaliphilic, halotolerant
650 bacterium isolated from extreme alkali-saline soil. Antonie Van Leeuwenhoek. 2012;102:
651 703–710. doi:10.1007/s10482-012-9770-7

652 87. Chu Y, Zhu Y, Chen Y, Li W, Zhang Z, Liu D, et al. aKMT Catalyzes Extensive Protein
653 Lysine Methylation in the Hyperthermophilic *Archaeon Sulfolobus* islandicus but is
654 Dispensable for the Growth of the Organism. Mol Cell Proteomics MCP. 2016;15: 2908–
655 2923. doi:10.1074/mcp.M115.057778

656 88. Sauer DB, Karpowich NK, Song JM, Wang D-N. Rapid Bioinformatic Identification of
657 Thermostabilizing Mutations. Biophys J. 2015;109: 1420–1428.
658 doi:10.1016/j.bpj.2015.07.026

659 89. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al.
660 GenBank. Nucleic Acids Res. 2017;45: D37–D42. doi:10.1093/nar/gkw1070

661 90. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, et al. Ensembl
662 Genomes 2016: more genomes, more complexity. Nucleic Acids Res. 2016;44: D574-580.
663 doi:10.1093/nar/gkv1209

664 91. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA
665 genes in genomic sequence. Nucleic Acids Res. 1997;25: 955–964.

666 92. Seemann T. Barrnap [Internet]. Available: https://github.com/tseemann/barrnap

667 93. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
668 Bioinforma Oxf Engl. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

669 94. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction
670 of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory
671 regions. Nucleic Acids Res. 2001;29: 2607–2618.

672 95. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely
673 available Python tools for computational molecular biology and bioinformatics. Bioinforma
674 Oxf Engl. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

675 96. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient
676 Numerical Computation. Comput Sci Eng. 2011;13: 22–30. doi:10.1109/MCSE.2011.37

677 97. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
678 Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

679 98. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95.
680 doi:10.1109/MCSE.2007.55

681 **Supporting Information Captions**
682
683 Figure S1. The genomes available are dominated by mesophiles, bacteria, and repetitively
684 sequenced organisms.
685
686 Figure S2. Features are often highly associated.
687
688 Figure S3. Using only genomic sequence features is poorly predictive of OGT.
689
690 Figure S4. Phylum specific regressions are often strongly predictive.
691
692 Figure S5. Class specific regressions can be strongly predictive.
693
694 Figure S6. Bioinformatic classification allows for accurate OGT prediction.
695
696 Figure S7. Genome size is not necessary for OGT prediction accuracy.
697
698 Figure S8. Excluding psychrophiles improves OGT prediction.
699
700 Figure S9. OGT prediction validated using previously unknown species-OGT values.
701
702 Equation S1. Features and coefficients for the prediction of the OGT for a prokaryote.
703
704 Equation S2. Features and coefficients for the prediction of the OGT for an Archaea.
705
706 Equation S3. Features and coefficients for the prediction of the OGT for a Bacterium.
707
708 Table S1. Correlation of features to OGT.
709
710 Table S2. *De novo* predicted OGT for species without a measured OGT in Sauer *et al.* 2015