Rapid evolutionary innovation during an Archaean genetic expansion

Rapid evolutionary innovation during an Archean Genetic Expansion

Lawrence A. David[1] & Eric J. Alm[2,3#]

[1]Computational & Systems Biology Initiative, [2]Department of Biological Engineering, & [3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

[#]Corresponding author

A natural history of Precambrian life remains elusive because of the rarity of microbial fossils and biomarkers[1,2]. The composition of modern day genomes, however, may bear imprints of ancient biogeochemical events[3-6]. Here we use an explicit model of macroevolution including gene birth, transfer, duplication and loss events to map the evolutionary history of 3,983 gene families across the three domains of life onto a geologic timeline. Surprisingly, we find that a brief period of genetic innovation during the Archean eon, which coincides with a rapid diversification of bacterial lineages, gave rise to 27% of major modern gene families. A functional analysis of genes born during this Archean Expansion reveals they are likely to be involved in electron transport and respiratory pathways. Genes arising after this expansion show growing use of molecular oxygen ($p=3.4\times10^{-8}$) and redox-sensitive transition metals and compounds, which is consistent with an increasingly oxygenating biosphere.

Describing the emergence of life on our planet is one of the grand challenges of the Biological and Earth sciences. Yet the roughly three-billion-year history of life preceding the emergence of hard-shelled metazoans remains obscure[1]. To date, the best understood event in early Earth history is the Great Oxidation Event, which is believed to follow the invention of oxygenic photosynthesis by ancestors of modern cyanobacteria[7] (though the precise timeline remains controversial[2,8]). If DNA sequences from extant organisms bear an imprint of this event, they can be used to make and test predictions; e.g., genes that use molecular oxygen are more likely to appear in organisms emerging after the Great Oxidation Event. Transfer of genes across species, however, can obscure patterns of descent and disrupt our ability to correlate gene histories with the geochemical record[9]. Widely distributed genes, for example, may descend from a Last Universal Common Ancestor (LUCA) as widely believed to be the case for the translational machinery[10], or may have been dispersed by horizontal gene transfer (HGT)[11,12], as in the case of antibiotic resistance cassettes.

An Archean Expansion

We developed a new phylogenomic method, *AnGST* (Analyzer of Gene and Species Trees), that explicitly accounts for HGT by comparing individual gene phylogenies to the phylogeny of organisms (the "Tree of Life") and generated detailed evolutionary histories for 3,983 major gene families. Gene histories reveal dramatic changes in the rates of gene birth, duplication, loss, and HGT over geologic time scales (Figure 1) including a burst of *de novo* gene family birth between 3.33-2.85 Ga which we refer to as the Archean Expansion. This window gave rise to 26.8% of extant gene

families and coincides with a rapid bacterial cladogenesis (Supplementary Fig. 15). A spike in the rate of gene loss (~3.1 Ga) follows the expansion and may represent consolidation of newly evolved phenotypes, as ancestral genomes became specialized for emerging niches. After 2.85 Ga, the rates of both gene loss and gene transfer stabilize at roughly modern-day levels. The rates of *de novo* gene birth and duplication after the Archean Expansion appear to show opposite trends: *de novo* gene family birth rates decrease and duplication rates increase over time. The near absence of *de novo* birth in modern times likely reflects the fact that ORFan gene families, which are widespread across all major prokaryotic groups, are not considered in this study[13]. The excess of gene duplications and ORFans in modern genomes suggests that novel genes from both sources experience high turnover. Although we did not observe changes in the rate of HGT after the Archean Expansion, we did detect an overrepresentation of HGT from alpha-proteobacteria to ancient eukaryotes ($p=3.3\times10^{-7}$ Wilcoxon rank sum test) and from cyanobacteria to plants ($p=8.3\times10^{-6}$ Wilcoxon rank sum test). These patterns of HGT likely reflect the endosymbioses that gave rise to mitochondria and chloroplasts[14,15], and serve to validate our phylogenomic approach against known macroevolutionary events.

What evolutionary factors were responsible for the period of innovation marked by the Archean Expansion? While we cannot provide an unequivocal answer to this question using gene birth dates alone, we can ask whether the functions of genes born during this time suggest plausible hypotheses. In general, birth of metabolic genes is enriched during the expansion, especially those involved in energy production and coenzyme metabolism (Supplementary Table 2), but further inspection also reveals an

enrichment for metabolic gene family birth prior to the Archean Expansion. To focus on specific metabolic changes linked to the Archean Expansion we: (i) grouped genes according to the metabolites they used; and (ii) we directly compared the occurrence of these metabolites in genes born during the Archean Expansion to their abundance prior to the Archean Expansion. The results are striking: the Archean Expansion-specific metabolites (positive bars, Fig. 2 inset) include most of the compounds annotated as redox/e^- transfer (blue bars), with Fe-S-, Fe-, and O_2-binding gene families showing the most significant enrichment (False Discovery Rate < 5%, Fisher's exact test). Gene families that use ubiquinone and FAD (key metabolites in respiration pathways) are also enriched, albeit at slightly lower significance levels (False Discovery Rate < 10%). The ubiquitous NADH and NADPH are a notable exception to this trend and appear to have played a role early in life history. By contrast, enzymes linked to nucleotides (green bars) exhibited strong enrichment in genes of more ancient origin than the expansion.

The observed metabolite usage bias suggests that the Archean Expansion was associated with an expansion in microbial respiratory and electron transport capabilities. Proving this association to be causal is beyond the power of our phylogenomic model. Yet this hypothesis is appealing because more efficient energy conservation pathways could increase the total free energy budget available to the biosphere, possibly enabling the support of more complex ecosystems and a concomitant expansion of species and genetic diversity. We note, however, that while the use of oxygen as a terminal electron acceptor would have significantly increased biological energy budgets, oxygen-utilizing genes are only enriched toward the end of the expansion (Supplementary Fig. 10). Thus,

the earliest redox genes we identified as part of the expansion were likely to be used in anaerobic respiration or oxygenic/anoxygenic photosynthesis, and may have been co-opted later for use in aerobic respiration pathways.

Phylogenomic evidence for ancient changes in global redox potential

Our metabolic analysis supports an increasingly oxygenated biosphere following the Archean Expansion, as the fraction of proteins utilizing oxygen gradually increases from the expansion to the present day (Figure 2; $p=3.4\times10^{-8}$, two-sided Kolmogorov-Smirnov test). Further indirect evidence of rising oxygen levels comes from compounds whose availability is sensitive to global redox potential. We observe significant increases over time in the usage of the transition metals copper and molybdenum (Figure 2; False Discovery Rate < 5%, two-sided Kolmogorov-Smirnov test), which is in agreement with geochemical models of these metals' solubility in increasingly oxidizing oceans[5,6] and with molybdenum enrichments from black shales that suggest molybdenum began accumulating in the oceans only after the Archean eon[16]. Our prediction of a significant increase in nickel utilization accords with geochemical models that predict a ten-fold increase in dissolved nickel concentration between the Proterozoic and modern day[5], but conflicts with a recent analysis of banded iron formations that inferred monotonically decreasing maximum dissolved nickel concentrations from the Archean onwards[17]. The abundance of enzymes using oxidized forms of nitrogen (N_2O and NO_3) also grows significantly over time, with 1/3 of nitrate-binding gene families appearing at the beginning of the expansion and 3/4 of nitrous oxide-binding gene families appearing by the end of the expansion. The timing of these gene family births provides phylogenomic

evidence for an aerobic nitrogen cycle by the Late Archean[18].

One striking discrepancy between our phylogenomic patterns and geochemical predictions, however, is a modest but significant increase in iron-using genes over time (Figure 2; False Discovery Rate < 5%, two-sided Kolmogorov-Smirnov test). Declining iron solubility in oxygenated ocean surface waters and sulfide-mediated iron removal from anoxic deeper waters are thought to have reduced overall iron bioavailability during the Proterozoic[19]. If the abundance of iron-using genes tracks iron bioavailability, we would expect these genes to decrease in abundance following the Archean. The conflicting phylogenomic result may reflect the confounding effect of evolutionary inertia, whereby microbes could have found more success evolving a handful of metal-acquisition proteins (e.g. siderophores), rather than replacing a host of iron-binding proteins in the face of declining iron availability[5]. Alternatively, the insolubility of iron in modern oceans may be offset by large organic pools of reduced iron.

Our chronologies of oxygen and redox-sensitive metal and compound utilization suggest ancient increases in oxygen bioavailability, as well as an Archean biosphere with some of the basic genetic components required for oxygenic photosynthesis and respiration. These results are consistent with recent biomarker-based evidence for oxygenesis preceding the Paleoproterozoic by hundreds of millions of years[20]. Still, a precise timeline for the origins of oxygenesis is beyond the resolution of our phylogenomic model. In the results described above, we estimated lineage divergence times using *PhyloBayes*[21], which enabled us to explicitly account for uncertainty in the timing of inferred events (Supplementary Figure 13). An alternative model of

evolutionary rates[22] dated the rapid bacterial cladogenesis to 2.75-2.5 Ga (in contrast to 3.33-2.85 for *PhyloBayes*), but still finds evidence for an Archean Expansion (Supplementary Fig. 9) characterized by the emergence of electron transport genes. Uncertainty or errors in the reference tree present a further challenge for phylogenomic analysis, and while they do not obscure our ability to detect the large-scale events described here, they may prevent analysis at greater temporal resolution and limit the precision with which we can date important events, such as the birth of oxygenesis. Future studies that benchmark biomarker and other geochemical data against the predicted age of associated gene families can be used to refine the Tree of Life, ultimately yielding an abundant and reliable source of Precambrian fossils: modern-day genomes.

Methods summary

We developed a new phylogenomic method, *AnGST* (Analyzer of Gene and Species Trees), to account for gene transfer, duplication, loss, and *de novo* birth by comparing individual gene phylogenies to a previously described reference phylogeny. We refer to this process as tree reconciliation and provide a detailed description of the *AnGST* algorithm in the Supplementary Information. Unlike some previous methods[23-25], *AnGST* uses the topology of the gene family tree rather than just its presence/absence across genomes and can infer the direction of gene transfer in addition to gene duplication, birth, and loss events. *AnGST* also accounts for uncertainty in gene trees by incorporating reconciliation into the tree-building process: the tree that minimizes the macroevolutionary cost function, but is still supported by the sequence data, is chosen as the best gene tree. We reconciled gene families against a published

Tree of Life[26]. To assess our method's sensitivity to the reference tree topology, we reconciled gene families against 30 alternative reference trees rooted on either the Bacterial, Archaeal, or Eukaryotic branches. Inferred gene family birth ages were consistent across the ensemble of reference trees, and the Archean Expansion was a uniformly observed feature (Supplementary Figs. 8 & 9). A conservative set of eight temporal constraints was selected from the geochemical and paleontological literature (Supplementary Fig. 7 and Supplementary Table 1), and the *PhyloBayes* software package was used to infer a range of divergence times for each ancestral lineage on the reference tree[21]. We did not apply temporal constraints to lineage ages on the gene trees.

REFERENCES

1. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083-1091 (2001).
2. Rasmussen, B., Fletcher, I. R., Brocks, J. J. & Kilburn, M. R. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**, 1101-1104 (2008).
3. Dupont, C. L., Yang, S., Palenik, B. & Bourne, P. E. Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc Natl Acad Sci USA* **103**, 17822-17827 (2006).
4. Dupont, C. L., Butcher, A., Valas, R. E., Bourne, P. E. & Caetano-Anollés, G. History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc Natl Acad Sci USA* **107**, 10567-10572 (2010).
5. Saito, M. A., Sigman, D. M. & Morel, F. M. M. The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean-Proterozoic boundary? *Inorganica Chimica Acta* **356**, 308-318 (2003).
6. Zerkle, A. L., House, C. H. & Brantley, S. L. Biogeochemical signatures through time as inferred from whole microbial genomes. *American Journal of Science* **305**, 467-502 (2005).
7. De Marais, D. J. Evolution. When did photosynthesis emerge on Earth? *Science* **289**, 1703-1705 (2000).
8. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033-1036 (1999).
9. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution* **19**, 2226 (2002).
10. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801-3806 (1999).
11. Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond, B, Biol Sci* **364**, 2241-2251 (2009).
12. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).
13. Fischer, D. & Eisenberg, D. Finding families for genomic ORFans. *Bioinformatics* **15**, 759-762 (1999).
14. Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. Mitochondrial origins. *Proc Natl Acad Sci USA* **82**, 4443-4447 (1985).
15. Giovannoni, S. J. et al. Evolutionary relationships among cyanobacteria and green chloroplasts. *J Bacteriol* **170**, 3584-3592 (1988).
16. Scott, C. et al. Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature* **452**, 456-459 (2008).
17. Konhauser, K. O. et al. Oceanic nickel depletion and a methanogen famine before the Great Oxidation Event. *Nature* **458**, 750-753 (2009).
18. Garvin, J., Buick, R., Anbar, A. D., Arnold, G. L. & Kaufman, A. J. Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. *Science* **323**, 1045-1048 (2009).
19. Canfield, D. E. A new model for Proterozoic ocean chemistry. *Nature* **396**, 450-453 (1998).

20. Waldbauer, J. R., Sherman, L. S., Sumner, D. Y. & Summons, R. E. Late Archean molecular fossils from the Transvaal Supergroup record the antiquity of microbial diversity and aerobiosis. *Precambrian Research* **169**, 28-47 (2009).
21. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**, 1095-1109 (2004).
22. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-302 (2003).
23. Alm, E., Huang, K. & Arkin, A. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* **2**, e143 (2006).
24. Kunin, V. & Ouzounis, C. A. The balance of driving forces during genome evolution in prokaryotes. *Genome Res* **13**, 1589-1594 (2003).
25. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**, 17-25 (2002).
26. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287 (2006).
27. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
28. Alm, E. J. et al. The MicrobesOnline Web site for comparative genomics. *Genome Res* **15**, 1015-1022 (2005).

CAPTIONS

Figure 1. Rates of macroevolutionary events over time.

The figure shows the average rates of gene birth (red), duplication (blue), HGT (green), and loss (yellow) *per lineage* (events / 10 Ma / lineage). Events that increase gene count are plotted to the right, and gene loss events are shown to the left. Genes already present at the LUCA are not included in the analysis of birth rates because the time over which those genes formed is not known. The Archean Expansion was also detected when 30 alternative chronograms were considered (Supplementary Fig. 9). Inset: metabolites or classes of metabolites ordered according to the number of gene families that use them that were born during the Archean Expansion compared to the number born before the expansion, plotted on Log (base 2) scale. Metabolites whose enrichments are statistically significant at a False Discovery Rate <10% or < 5% (Fisher's Exact Test) are identified using one or two asterisks, respectively. Bars are colored by functional annotation or compound type (functional annotations were assigned manually). Metabolites were obtained from the KEGG database release 51.0[27] and associated with COGs using the MicrobesOnline September 2008 database[28]. Metabolites associated with fewer than 20 COGs or sharing more than 2/3 of gene families with other included metabolites are omitted.

Figure 2. Genome utilization of redox-sensitive compounds over time. The first bar illustrates a gradual increase in the fraction of enzymes that bind molecular oxygen

predicted to be present over Earth history ($p=3.4\times10^{-8}$, two-sided Kolmogorov-Smirnov test). Colors indicate abundance normalized to present-day values. The lower four panels group transition metals, nitrogen compounds, sulfur compounds, and C1 compounds. The fraction of each group's associated genes that bind a given compound, normalized to present-day fractions, is shown over time using a color gradient. Enclosed boxes show raw fractional values at three time points: 3.5 Ga (left); 2.5 Ga (middle); and the present day (right). For example, 18.9% of transition metal-binding genes are predicted to have bound Mn at 2.5 Ga, a value 1.26 times the size of the modern day percentage of 15.0%. Values within parentheses give the overall number of gene families in each group. To determine which compounds showed divergent genome utilization over time, the timing of copy number changes for each compound's associated genes was compared to a background model derived from all other compounds. Compounds whose utilization significantly differs from the background model are marked with an asterisk (False Discovery Rate < 5%, two-sided Kolmogorov-Smirnov test). Nitrite and nitric oxide are not shown due to their COG-binding similarity to nitrate and nitrous oxide, respectively.