

RESEARCH HIGHLIGHT

Open Access



# Adding resolution and dimensionality to comparative genomics: moving from reference genomes to clade genomics

Jeffrey Rogers<sup>1,2</sup>

## Abstract

The main goal and promise of comparative genomics has been to create a comprehensive catalog of genomic information and function across the phenomenal diversity of living systems. A recent study has demonstrated the evolutionary insights possible by generating high-quality whole-genome assemblies from multiple species of a clade.

## Making impossible things feasible

The first whole-genome reference assemblies were available for only the highest priority species (humans, mice, and *Drosophila*). Happily, advances in technologies for DNA sequencing and computational analysis have made it practical to generate new high-quality genome assemblies at an accelerating pace. Excellent genomic resources are now publicly available for hundreds of plant, invertebrate, and vertebrate species, to say nothing of the remarkable databases of microbial diversity.

Nevertheless, cost remains a constraint and as a result it has been commonplace for research communities to collectively choose one key species (or a small number of species) from a major clade for thorough analysis. This leaves investigators who are working on other closely related species to 'borrow' information from the selected keystone representative of a given taxonomic group. Fortunately, the continuing reductions in sequencing costs, and the development of new tools that generate more robust assemblies, are progressively removing this barrier. It is increasingly feasible to generate high-quality, highly contiguous whole-genome assemblies not only for individual representatives of clades but also in parallel for

multiple species within a group. This increasing granularity of comparative genomic information provides new opportunities for evolutionary analysis. The increased resolution of information facilitates the ability of investigators to estimate the timing of genomic innovations and to measure variation in rates of genomic evolution through time and across lineages. This is not possible on a sparsely annotated phylogenetic tree. These improved assemblies (with fewer gaps) can reveal patterns of evolutionary change that previously could not be discerned. For example, recent efforts have added the dimensions of segmental duplications, structural variants, and even haplotype phasing to the traditional genomic data sets of single, unphased nucleotide differences and small insertion/deletions [1].

## A new illustration from murid rodents

A valuable and informative example of this trend toward finer-scale phylogenetic resolution has recently been published by Thybert and colleagues [2]. In this study, the authors sequenced, assembled, and compared the genomes of two murid rodents, *Mus caroli* and *Mus pahari*. They chose these two relatives of the well-studied laboratory mouse (*Mus musculus*) because previous work had shown that *M. caroli* and *M. pahari* shared a common ancestor with *M. musculus* approximately 3 million years ago and 6 million years ago, respectively. Together with the rat (*Rattus norvegicus*), this constitutes a clade of four rodents with high-quality genome sequences that Thybert et al. explicitly present as a parallel to the hominid clade of humans, chimpanzees, gorillas, and orangutans. Using the newly included murid genomes, Thybert et al. have addressed various questions regarding patterns of genomic change in these two mammalian groups. When all we had available were the human, chimpanzee, and mouse genomes, one could ask questions regarding differences in gene content or genomic organization between representative primates and a representative rodent. But with

Correspondence: [jr13@bcm.edu](mailto:jr13@bcm.edu)

<sup>1</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA

<sup>2</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

clades of genomes within each lineage, Thybert and colleagues have compared patterns of evolutionary change at a much higher resolution and have begun addressing the temporal dynamics of specific aspects of genome evolution that have not been previously investigated in these lineages.

It has long been known that rodents exhibit a higher rate of large (multi-megabase) chromosomal rearrangements than do primates. New intra-family comparisons show that there have been periods of rapid rearrangement interspersed with periods of much slower change during murid evolution. New assemblies also show that other features of the murid genome evolve more rapidly than they do in primates, including single nucleotide changes (previously reported, but now examined across multiple rodent species), SINE elements, CTCF-binding motifs, and, particularly, LINE elements. The evidence suggests that the expansion of LINE elements led to various changes in genome function in the genus *Mus*. In one case, new LINE elements specific to murid rodents accumulated within the androgen-binding protein gene cluster, apparently contributing to gene copy number increases in murids, particularly *M. musculus*. This gene family has been associated with behavioral mate choice and the dynamics of a hybrid zone between two mouse subspecies [3]. New analyses demonstrate that LINE element expansion in murid rodents differs significantly from that in the parallel clade of hominids, which has interesting biological consequences.

Another example of the value of more extensive sampling of genomes within and across clades is the novel finding by Thybert and colleagues that, although the B2 SINE element has increased in number in all three *Mus* species, only *M. caroli* displays a novel CTCF-binding site within the B2 element. Consequently, only *M. caroli* has increased CTCF-binding site numbers due to SINE element proliferation. Given the potential functional impact of increased CTCF-binding sites, this has considerable implications for understanding the evolution of gene regulation across rodent lineages. Overall, Thybert et al. observed that murid genomes exhibit faster evolution than do primates across multiple domains (coding exons, introns, regulatory sequences, ancestral repeat elements, and even inter-chromosomal rearrangements), but that episodes of accelerated change can occur independently in different murid lineages and that some of these rates fluctuate over evolutionary timescales within a given lineage.

This analysis by Thybert et al. provides a welcome example of the insights to be gained by comparing multiple genomes across several major lineages (families, superfamilies, or orders among mammals, or equivalent phylogenetic diversity in other taxonomic groups). A 2014 analysis of 48 bird genomes was another clear illustration of the value of this approach [4]. Other recent studies of smaller data sets have also indicated how comparisons made

locally within a clade, when performed in parallel with comparisons across widely divergent lineages, can generate novel unanticipated results [5, 6].

### No longer just dreams

Several projects are now underway that promise to dramatically increase the number of species for which high-quality genome assemblies are available. Additional assemblies for a wider range of primates are in progress or have recently been published [7]. This is incrementally generating new clades of assemblies on additional branches of the primate evolutionary tree. At a considerably broader scale, the Vertebrate Genomes Project (VGP) is an international, multi-institution program designed to achieve impressive goals starting with the assembly of the genome of one species from each of approximately 260 vertebrate orders [8]. The VGP has established explicit recommended standards for the quality of new reference assemblies and plans to sequence many more species in its second and subsequent phases, reaching tens of thousands of new high-quality assemblies over several years. The insect i5k project is another such effort that intends to produce de novo assemblies for 1000 arthropods [9].

The continuing advances in genome analysis technology, involving methods for both data generation and computational analysis, are making truly revolutionary sequencing projects feasible. One can foresee a not-too-distant future in which almost gapless high-accuracy whole-genome assemblies are publicly available for tens of thousands of plant, invertebrate, and vertebrate species. Goals that were inconceivable even a few years ago are now under serious discussion [10]. A golden age of comparative genomics is coming. The recent paper by Thybert et al. [2] takes a welcome step towards the day when biologists will have ready access to extensive genomic information for entire clades of species.

### Funding

JR receives support from NIH grants R01-EY026045, R24-OD11173, UM1-HG008898, and P51-OD011106.

### Author's contribution

JR wrote the manuscript, and approved the final manuscript.

### Author's information

JR is an Associate Professor in the Human Genome Sequencing Center and Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas.

### Competing interests

The author declares that he has no competing interests.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 14 August 2018

#### References

1. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet.* 2018;19:329–46.
2. Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M, Janousek V, Akanni W, et al. Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Res.* 2018;28:448–59.
3. Bimova BV, Macholan M, Baird SJ, Munclinger P, Dufkova P, Laukaitis CM, Karn RC, Luzynski K, Tucker PK, Pialek J. Reinforcement selection acting on the European house mouse hybrid zone. *Mol Ecol.* 2011;20:2403–24.
4. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science.* 2014;346:1320–31.
5. Foote AD, Liu Y, Thomas GW, Vinar T, Alfoldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. Convergent evolution of the genomes of marine mammals. *Nature Genet.* 2015;47:272–5.
6. Meadows JRS, Lindblad-Toh K. Dissecting evolution and disease using comparative vertebrate genomics. *Nature Rev Genet.* 2017;17:624–36.
7. Palesch D, Bosinger SE, Tharp GK, Vanderford TH, Paiardini M, Chahroudi A, Johnson ZP, Kirchhoff F, Hahn BH, Norgren RB, et al. Sooty mangabey genome sequence provides insight into AIDS resistance in a natural SIV host. *Nature.* 2018;553:77–81.
8. Vertebrate Genomes Project plan. <https://rockefeller.edu/research/vertebrate-genomes-project/vertebrate-genomes-project-plan/>. Accessed 12 Jul 2018.
9. i5k: sequencing five thousand arthropod genomes. <http://i5k.github.io/>. Accessed 12 Jul 2018.
10. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. Earth BioGenome project: sequencing life for the future of life. *Proc Natl Acad Sci U S A.* 2018;115:4325–33.