REVIEW

What Is a Genome?

Aaron David Goldman[1]*, **Laura F. Landweber**[2,3]*

1 Department of Biology, Oberlin College, Oberlin, Ohio, United States of America, **2** Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, United States of America, **3** Departments of Biochemistry & Molecular Biophysics and Biological Sciences, Columbia University, New York, New York, United States of America

* agoldman@oberlin.edu (ADG); Laura.Landweber@columbia.edu (LFL)

Abstract

The genome is often described as the information repository of an organism. Whether millions or billions of letters of DNA, its transmission across generations confers the principal medium for inheritance of organismal traits. Several emerging areas of research demonstrate that this definition is an oversimplification. Here, we explore ways in which a deeper understanding of genomic diversity and cell physiology is challenging the concepts of physical permanence attached to the genome as well as its role as the sole information source for an organism.

Introduction

The term genome was coined in 1920 to describe "the haploid chromosome set, which, together with the pertinent protoplasm, specifies the material foundations of the species" [1]. The term did not catch on immediately (Fig 1). Though Mendelian genetics was rediscovered in 1900, and chromosomes were identified as the carriers of genetic information in 1902 [2], it was not known in 1920 whether the genetic information was carried by the DNA or protein component of the chromosomes [3]. Furthermore, the mechanism by which the cell copies

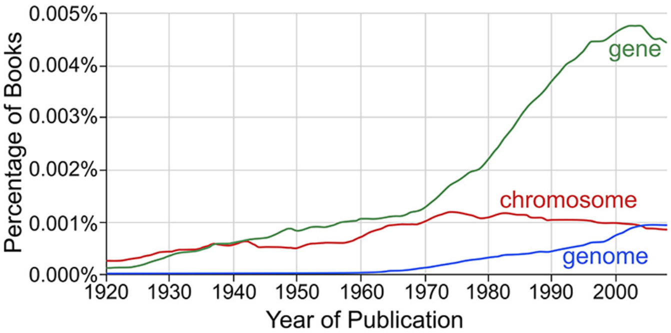

Fig 1. The change in usage of the term "genome" compared to related terms. A Google ngram [6] analysis shows the case-insensitive occurrences of the terms "gene," "genome," and "chromosome" in the corpus of books in English from 1920 to 2008. The data are smoothed by a three-year moving average. The term "genome" was coined in 1920 [1], and many sources, including the Oxford English Dictionary, attribute the word to a portmanteau of the words "gene" and "chromosome," although this etymology is disputed [1]. The term took decades to enter popular usage and only achieved its current level of usage by the turn of this century.

doi:10.1371/journal.pgen.1006181.g001

Citation: Goldman AD, Landweber LF (2016) What Is a Genome? PLoS Genet 12(7): e1006181. doi:10.1371/journal.pgen.1006181

Editor: W. Ford Doolittle, Dalhousie University, CANADA

Published: July 21, 2016

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

information into new cells [4] and converts that information into functions [5] was unknown for several decades after the term "genome" was coined.

Today, however, we are awash in genomic data. A recent release of the GenBank database [7], version 210.0 (released on October 15, 2015), contains over 621 billion base pairs from 2,557 eukaryal genomes, 432 archaeal genomes, and 7,474 bacterial genomes, as well as tens of thousands of viral genomes, organellar genomes, and plasmid sequences (http://www.ncbi.nlm.nih.gov/genome/browse/, on December 13, 2015). We also now have much broader and more detailed understandings of how the genome is expressed and how different biological and environmental factors contribute to that process. Even so, almost a century after coining the term, the standard definition of the genome remains very similar to its 1920 predecessor. For example, on its Genetics Home Reference website, the National Institutes of Health (NIH) definition reads: "An organism's complete set of DNA, including all of the genes, makes up the genome. Each genome contains all of the information needed to build and maintain that organism" (http://ghr.nlm.nih.gov/handbook/hgp/genome, on February 1, 2016).

With a greater understanding of genomic content, diversity, and expression, we can now reassess our basic understanding of the genome and its role in the cell. For example, closer scrutiny of the NIH definition reveals that its two halves are mutually exclusive; that is, the "complete set of DNA" cannot be "all of the information needed to build and maintain (an) organism." Of course, this was probably meant to be a simplified definition for both scientists and nonscientists. While it is useful to continue thinking of the genome as a physical entity encoding the information required to maintain and replicate an organism, our present understanding shows that this definition is incomplete.

Examples of Physical Transience in Genomes

Many diverse genetic systems challenge the material definition of the genome as "the complete set of *chromosomes*" [1] or "an organism's complete set of *DNA*" (http://ghr.nlm.nih.gov/handbook/hgp/genome). Perhaps the most familiar and straightforward example of a genome's physical impermanence occurs in the retroviral infection cycle. Upon infection, retroviruses convert their single-stranded RNA genomes into double-stranded DNA. These intermediate DNA copies of the genome are integrated into the host cell and, thus, no longer constitute a separate physical entity from the host's genome. As an integrated DNA sequence, transcription into mRNA can both express retroviral genes and also reconstitute the original single-stranded (ss)RNA genome. Other types of viruses share similar features. Many temperate phages and viruses integrate into the host's genome, removing themselves and lysing the host cell only after certain conditions are met. The hepadnaviruses, including Hepatitis B, infect the cell as double-stranded DNA, but are transcribed after infection into single-stranded RNA and subsequently follow a similar course as the retroviruses, wherein they are reverse transcribed back into DNA [8].

The chemical conversions of these genomes between different nucleic acids offer cogent examples that challenge our assumption of the physical permanence of genomes. It is tempting to explain this physical transience as another eccentric quirk of viruses. Many viruses, after all, do not have genomes composed of double-stranded DNA, a feature that already flouts the NIH definition given earlier. But an equally cogent example of the physical impermanence of a genome is found in the eukaryotic genus *Oxytricha* [9–11], a group of ciliates that are distantly related to *Tetrahymena* and *Paramecium* [12].

Like other ciliates, *Oxytricha* possesses two distinct versions of its genome, a germline version and a somatic version. *Oxytricha*'s germline genome is an archive of approximately 1 Gb of DNA sequence containing approximately one-quarter million embedded gene segments.

These DNA pieces assemble following sexual recombination to form the somatic, expressed chromosomes (Fig 2). Thousands of these gene segments are present within the germline chromosomes in a scrambled order or reverse orientation, such that their reassembly requires translocation and/or inversion with respect to one another [13]. The resulting somatic genome, containing protein-coding sequences in the correct order, contains just 5%–10% the original sequence of the germline genome. This somatic genome resides on over 16,000 unique "nanochromosomes" that typically bear single genes and have an average size of just 3.2 kb [14]. These nanochromosomes also exist in high copy number, averaging approximately 2,000 copies per unique chromosome [14,15].

Much of the information required to reproduce the somatic genome derives from RNA rather than DNA. Long, RNA-cached copies of somatic chromosomes from the previous generation provide templates to guide chromosome rearrangement [16]. Germline transposases participate in the whole process, probably by facilitating DNA cleavage events [17,18] that allow genomic regions to rearrange in the order according to the RNA templates [16]. Experimental introduction of long artificial RNAs can reprogram a developing *Oxytricha* cell to follow the order of gene segments specified by the artificial RNA templates, rather than the wild-type chromosome.

RNA performs other essential roles in building *Oxytricha*'s somatic genome. Millions of small, 27-nt piRNAs, which also derive from the previous generation's somatic genome, mark and protect the retained DNA regions in the new zygotic germline that assemble (according to the RNA template) to form the new somatic genome [19,20]. In addition, the relative abundance of the long template RNAs also establishes chromosome copy number in the daughter cells [17]. Because these RNA templates derive from the previous generation's somatic genome, this means that both the genomic sequence and chromosome ploidy are inherited from the old somatic nucleus to the new somatic nucleus through information transfer from DNA to RNA and back again to DNA.

These examples of physical transience in genomes show that a genome's chemical composition and stability are not necessarily fixed requirements at all times in every organism. Synthetic biologists have further demonstrated this point through the chemical synthesis of viral [21,22] and bacterial [23] genomes. Prior to the chemical synthesis of these DNA chromosomes, the genomes existed in a purely informational state as nucleotide sequences in a computer file. In these cases, the genome of the virus or cell is not transferred from one type of nucleic acid to another, but from a physical DNA molecule to a non-physical nucleotide sequence and back again to a physical DNA molecule. Though this example is not a naturally occurring phenomenon, it provides a straightforward demonstration that the information content of the genome is more important than its physical permanence. Therefore, the concept of informational supremacy that is used to define genomes, e.g., "all of the information needed to build and maintain that organism," also deserves further scrutiny.

Extra-Genomic Information

Information is both an essential concept that underpins our understanding of a genome's function and a notoriously difficult concept to define. The genome contains information, but so do other constituents of the cell. A typical and uncontroversial view is that the genome carries information but requires the presence of proteins, ribosomal RNAs, and transfer RNAs in the cell for the meaningful conversion of genomic information to molecular function. Indeed, the construction of synthetic genomes mentioned earlier required transplantation of the chemically synthesized genome into a pre-existing cell [23]. Evidence for heritable information beyond the genome has also been known since the 1960s [24]. A greater understanding of

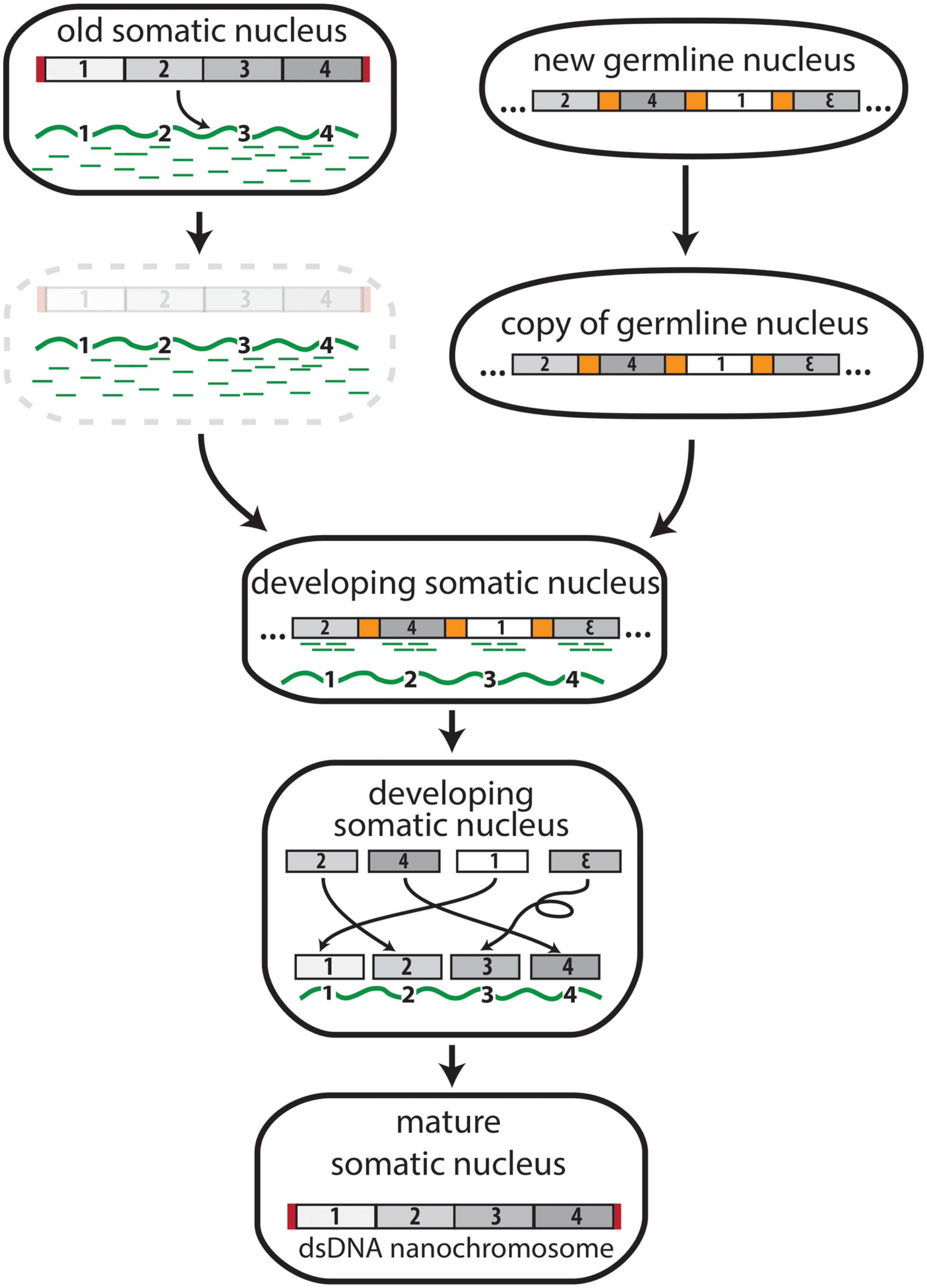

Fig 2. The transfer of genomic information from DNA to RNA in *Oxytricha trifallax*. The physical transition of genomic information from DNA to RNA and back to DNA occurs after mating in the ciliate, *Oxytricha trifallax*. RNA templates (wavy green line) and piRNAs (green dashes) derive from RNA transcripts of the previous generation's somatic DNA nanochromosomes before the old somatic nucleus degrades. A mitotic copy of the new, zygotic germline genome provides precursor DNA segments (numbers 1–4) that are retained in the developing somatic nucleus through piRNA associations and rearranged according to the inherited RNA templates. This step sometimes reorders

or inverts precursor segments to build the mature DNA molecule. The number of copies of each new nanochromosome is also influenced by the concentration of RNA templates supplied by the previous somatic genome during development. Red rectangles represent telomeres added to the ends of somatic chromosomes. Only one representative nanochromosome (of over 16,000 in *Oxytricha*) is shown for simplicity, and it derives from a representative locus containing 4 scrambled precursor segments in the germline genome.

doi:10.1371/journal.pgen.1006181.g002

molecular biology has revealed that extra-genomic sources of information are not only required to read the genome but can influence the information encoded within the genome [25].

Epigenetic control of gene regulation provides a subtler—but in many ways more cogent—example of extra-genomic information. DNA methylation [26,27], histone modification encoding chromatin [28,29], and certain proteins (e.g., [30,31]) and noncoding RNAs [32,33], including *Oxytricha's* noncoding RNAs described in the previous section [17,18,20], all offer platforms that permit information transfer across generations, while seeming to bypass the DNA genome. It has not yet been shown whether epigenetic information can persist over scales of evolutionary time, but it is clear that many if not most genomes have evolved a capacity for epigenetic control. This makes such genomes sensitive to external information that they do not encode, which, in turn, should influence their ability to adapt to changing environments while, in some cases, preserving the ability to revert to the former wild-type genome. This is epitomized by the genome duality in *Oxytricha*, in which millions of small and long noncoding RNAs sculpt and decrypt the information in its somatic epigenome, while the germline genome provides a more stable archive.

A second example of extra-genomic information has come by way of genome-wide association studies, which have identified correlations between many phenotypic traits and genetic variants [34]. In doing so, such studies have also revealed the so-called "missing heritability" problem, that genetic variation does not always account for 100% of the measured heritability, let alone the observed phenotypic variance, in many complex traits. In many cases, this missing heritability can be explained as a lack of statistical power due to low phenotypic impact of the genetic variation or low frequency in the population [35]. The missing heritability can also be explained, however, by a gene–environment interaction, such that the genes may only encode a trait that is expressed under certain environmental conditions [36,37]. In this example, genomes do not necessarily encode all of the information of the cell, but rather a set of potential states that may be realized through interaction with different environments.

As these examples demonstrate, the way in which the information content of the genome becomes realized as functions and phenotypes depends on other cellular constituents as well as the environment. The ability of genomes to be affected by this external information is, itself, encoded on the genome. In this way, genomes are not a sole source of cellular information, but rather a more expansive archive of possible states that can be generated through interactions with internal and external factors.

Conclusion

Many biologists already know that the genome is not always best defined as "all of the information needed to build and maintain" a cell or an organism. While this definition is useful in the context of an online glossary for the public, it is, by necessity, an oversimplification. But if a genome is not a complete set of DNA containing all of the information needed to build and maintain the organism, then what is it?

We have demonstrated through examples from retroviruses, the microbial eukaryote *Oxytricha*, and synthetic biology that the genome can change its physical character while still maintaining the necessary information encoded within it. We also describe examples in which non-genomic factors can alter the way in which the information within the genome translates to molecular functions and phenotypes. These examples suggest a more expansive definition of the genome as an informational entity, often but not always manifest as DNA, encoding a broad set of functional possibilities that, together with other sources of information, produce and maintain the organism. Whether or not even this definition stands up to future discoveries remains to be seen.

Acknowledgments

We thank Ford Doolittle and Susan Rosenberg for organizing this series of papers on "How Microbes 'Jeopardize' the Modern Synthesis."

References

1. Lederberg J, McCray AT. 'Ome Sweet 'Omics: A Genealogical Treasury of Words. *The Scientist*. 2001; 15:8.
2. Sutton WS. On the morphology of the chromosome group in Brachystola magna. *Biol*. *Bull*. 1902; 4:24–39
3. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. *J Exp Med*, 1944; 79:137–159. PMID: 19871359
4. Watson JD, Crick FHC. Genetical Implications of the structure of Deoxyribonucleic Acid. *Nature*. 1953; 171:964–967. PMID: 13063483
5. Crick FHC. On protein synthesis. *Symp Soc Exp Biol*. 1958; 12:138–163. PMID: 13580867
6. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, et al. Quantitative analysis of culture using millions of digitized books. *Science*. 2011; 331:176–182. doi: 10.1126/science.1199644 PMID: 21163965
7. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2005; 33:D34–D38. PMID: 15608212
8. Nassal M, Schaller H. Hepatitis B virus replication. *Trends Microbiol*. 1993; 1:221–228. PMID: 8137119
9. Nowacki M, Shetty K, Landweber LF. RNA-Mediated Epigenetic Programming of Genome Rearrangements. *Annu Rev Genomics Hum Genet*. 2011; 12:367–389. doi: 10.1146/annurev-genom-082410-101420 PMID: 21801022
10. Goldman AD, Landweber LF. *Oxytricha* as a modern analog of ancient genome evolution. *Trends Genet*. 2012; 28:382–388. doi: 10.1016/j.tig.2012.03.010 PMID: 22622227
11. Bracht JR, Fang W, Goldman AD, Dolzhenko E, Stein EM, Landweber LF. Genomes on the edge: programmed genome instability in ciliates. *Cell*. 2013; 152:406–416. doi: 10.1016/j.cell.2013.01.005 PMID: 23374338
12. Zoller SD, Hammersmith RL, Swart EC, Higgins BP, Doak TG, et al. Characterization and taxonomic validity of the ciliate *Oxytricha trifallax* (Class Spirotrichea) based on multiple gene sequences: limitations in identifying genera solely by morphology. *Protist*. 2012; 163:643–657
13. Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, et al. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*. 2014; 158:1187–98. doi: 10.1016/j.cell.2014.07.034 PMID: 25171416
14. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, et al. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol*. 2013; 11:e1001473. doi: 10.1371/journal.pbio.1001473 PMID: 23382650
15. Prescott DM. The DNA of ciliated protozoa. *Microbiol Mol Biol Rev*. 1994; 58:233–267.
16. Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, Landweber LF. RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature*. 2008; 451:153–158. PMID: 18046331
17. Nowacki M, Haye JE, Fang W, Vijayan V, Landweber LF. RNA-mediated epigenetic regulation of DNA copy number. *Proc Natl Acad Sci U S A*, 2010; 107:22140–22144. doi: 10.1073/pnas.1012236107 PMID: 21078984
18. Vogt A, Goldman AD, Mochizuki K, Landweber LF. Transposon domestication versus mutualism in ciliate genome rearrangements. *PLoS Genet*. 2013; 9:e1003659. doi: 10.1371/journal.pgen.1003659 PMID: 23935529

19. Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. *Cell*. 2012; 151:1243–1255. doi: 10.1016/j.cell.2012.10.045 PMID: 23217708

20. Zahler AM, Neeb ZT, Lin A, Katzman S. Mating of the stichotrichous ciliate *Oxytricha trifallax* induces production of a class of 27 nt small RNAs derived from the parental macronucleus. *PLoS ONE*, 2012; 7: e42371. doi: 10.1371/journal.pone.0042371 PMID: 22900016

21. Cello J, Paul AV, Wimmer E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*. 2002; 297:1016–1018. PMID: 12114528

22. Smith HO, Hutchison CA 3rd, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci U S A*. 2003; 100:15440–5. PMID: 14657399

23. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 2010; 329:52–56. doi: 10.1126/science.1190719 PMID: 20488990

24. Nanney DL. Corticotype transmission in *Tetrahymena*. *Genetics*. 1966; 54:955–968. PMID: 5972435

25. Walker SI. Top-down causation and the rise of information in the Emergence of Life. *Information*. 2014; 5:424–439.

26. Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*. 1975; 14:9–25. PMID: 1093816

27. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003; 33:245–254. PMID: 12610534

28. D'Urso A, Brickner JH. Mechanisms of epigenetic memory. *Trends Genet*. 2014; 30:230–236. doi: 10.1016/j.tig.2014.04.004 PMID: 24780085

29. Siklenka K, Erkek S, Godmann M, Lambrot R, McGraw S, et al. Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science*. 2015; 350:aab2006 doi: 10.1126/science.aab2006 PMID: 26449473

30. Zordan R, Miller M, Galgoczy D, Tuch B, Johnson A. Interlocking transcriptional feedback loops control white-opaque switching in *Candida albicans*. *PLoS Biol*. 2007; 5:1–11.

31. Zacharioudakis I, Gligoris T, Tzamarias D. A yeast catabolic enzyme controls transcriptional memory. *Curr Biol*. 2007; 17:2041–2046. PMID: 17997309

32. Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I, Cuzin F. RNA-mediated non-Mendelian inheritance of an epigenetic change in the mouse. *Nature*. 2006; 441:469–474. PMID: 16724059

33. Rodgers AB, Morgan CP, Leu NA, Bale TL. Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. *Proc Natl Acad Sci U S A*. 2015; 112:13699–13704. doi: 10.1073/pnas.1508347112 PMID: 26483456

34. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*. 2012; 8: e1002822. doi: 10.1371/journal.pcbi.1002822 PMID: 23300413

35. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet*. 2012; 13:135–145. doi: 10.1038/nrg3118 PMID: 22251874

36. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. *PLoS Biol*. 2008; 6: e83. doi: 10.1371/journal.pbio.0060083 PMID: 18416601

37. Manuck SB, McCaffery JM. Gene-environment interaction. *Annu Rev Psychol*. 2014; 65:41–70. doi: 10.1146/annurev-psych-010213-115100 PMID: 24405358