Human Genomics

EDITORIAL

Open Access

Why are keratins important?

Jeffrey Nicholas Fisk[1*] and Daniel W. Nebert[2,3*]

The human fertilized egg (*zygote*) is probably the most implausible cell on Earth, because it is equipped with all the genes needed to create a human body—comprising ~ 37 trillion cells and representing more than 200 distinctly different cell-types. This complexity emerges from how and when the regulatory genes become activated to express the essential structural genes at the right time during embryogenesis, fetogenesis, postpartum, and all the way to adulthood.

How does that 1-cell zygote hold itself together? How do the nucleus and nucleolus remain intact from the cytoplasm, and how do all the cytoplasmic organelles persist as distinct subcellular bodies? Likewise, how do all the zygote's successor cells hold themselves together, despite their diversification into > 200 cell-types that are localized uniquely into dozens of specific tissues? The answer, in part, lies in (regulatory) genes involved in signaling and adhesion, and formation of filaments and fibrils and their (structural) gene products (proteins). If these types of *Animalia* proteins had not evolved (with orgins as early as the the first eukaryote) to hold cells together and keep cells organized within a particular tissue, life on Earth would be drastically different.

One large subset of these genes responsible for keeping everything in their place is the Intermediate Filament (IntFil) gene superfamily. When we were first invited to join as coauthors on the Ho et al. [1] project, we knew nothing about IntFil genes and their proteins, nor why anyone would want to study them.

IntFils arose during early metazoan evolution to provide mechanical support for plasma membranes that are connected and interact with other cells and the extracellular matrix. IntFils are ubiquitous structural components that comprise, in a cell type-specific manner, the cytoskeleton infrastructure in all animal tissues. All IntFil proteins show a distinctly organized extended α-helical conformation, which is predisposed to form two-stranded coiled coils that reflect the basic building blocks of highly flexible, stress-resistant cytoskeletal filaments. In this issue, Ho et al. [1] studied the evolutionary history of IntFil genes. Although IntFils are divided into six types, the coauthors focused on the type I "acidic" and type II "basic" keratin genes—which are much larger in number and evolutionarily emerged more recently than the other four types.

The first keratin gene appeared in sponge, three keratin genes are found in arthropods, and then more rapid increases in keratin genes occurred in lungfish and amphibian genomes, concomitant with the sea animal-to land animal transition which occurred 440 to 410 million years ago. The human genome has 27 of 28 type I keratin genes clustered at chromosome (Chr) 17q21.2, and all 26 type II keratin genes clustered at Chr 12q13.13. The mouse genome has 27 of 28 type I keratin genes clustered on Chr 11, and all 26 type II clustered on Chr 15; all the mouse keratin genes are syntenic with the human keratin genes. On the other hand, the zebrafish genome has 18 type I keratin genes scattered on five chromosomes and three type II keratin genes on two chromosomes. The two clusters ("evolutionary blooms") of type I and type II keratin genes, each located along a chromosomal segment, have been found in all seven nonhuman mammalian genomes that have been examined to date, but not in fish genomes [1].

*Correspondence: jeffrey.fisk@yale.edu; nebertdw@ucmail.uc.edu
[1] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA
[2] Departments of Pediatrics and Molecular and Developmental Biology, Cincinnati Children's Research Center, Cincinnati, OH 45229, USA
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. 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 in a credit line to the data.

Screening 259 species and subspecies in 20 phyla of animals, from jellyfish to human, Ho et al. [1] examined various features found in the type I and type II keratin proteins. They found evidence that some genes appear to have arisen in an early species, disappeared in a later species, and then, on occasion, reappeared and apparently were repurposed to provide for new features in more recently diverged species.

To create the maximum-likelihood trees, Ho et al. [1] aligned sequences in *MAFFT* [2] using the *L-INS-i* local pair algorithm [3] with 10,000 iterative alignment steps. Evolutionary models were determined, using *ModelFinder* [4] as implemented in *IQTREE* [5], and using *Bayesian Information Criteria* [6] to select the optimal model and gamma rate categories [7]. Subsequently, they used, in successive steps, construction of maximum likelihood phylogenetic trees [8], and further optimization using a *hill-climbing nearest-neighbor interchange* [9] protocol.

To make the cross-species trees, Ho et al. [1] used the interactive Fast-Fourier Transform method in MAFFT to build multiple sequence alignments, evolutionary relationships were estimated by *Markov-chain Monte Carlo* [10] in the Bayesian Phylogenetics program and sampling every 1,000 generations in parallel using *the BEAGLE library* [11], following which the within-chain and between-chain variance *potential scale reduction factor* [12] was used to evaluate sufficient sampling. Finally, the sampled posteriors from the two independent executions were combined to generate a *maximum clade-credibility tree* [13]—summarizing the posterior distribution of estimated evolutionary relationships and branch lengths.

This bioinformatics analysis led Ho et al. [1] to conclude that type I KRT18 resembles most closely the ancestral precursor of all other type I keratins, and the type II KRT8 resembles most closely the ancestral precursor of all other type II keratins. It is suggested for other gene superfamilies—containing evolutionary blooms in which an ancestral ordering is difficult to resolve—that the comparative genomics approach used in this publication might be helpful in determining which is the earliest diverging gene in a cluster.

Lastly, comparative-genomics approaches on genes relevant to human health and disease can offer insight into the nature and etiology of specific disorders. Are there keratin gene variants known to cause human disease? Ho et al. [1] found that the ClinVar database currently lists 26 human disease-causing variants within the various domains of keratin proteins.

Authors' contributions
Both authors have read and approved the manuscript.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA. [2]Departments of Pediatrics and Molecular and Developmental Biology, Cincinnati Children's Research Center, Cincinnati, OH 45229, USA. [3]Department of Environmental Health and Center for Environmental Genetics, University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA.

Published online: 30 January 2022

References

1. Ho M, Thompson B, Fisk JN, et al. Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders. Hum Genom. 2022;16:1. https://doi.org/10.1186/s40246-021-00374-9.
2. In bioinformatics, *MAFFT* is a program that creates multiple sequence alignments of amino acid (or nucleotide sequences). MAFFT uses an algorithm based on progressive alignment, in which the sequences are clustered with help of the Fast Fourier Transform. https://medium.datadriveninvestor.com/mafft-a-novel-method-for-multiple-sequence-alignment-8cbc2710a037
3. L-INS-i is probably most accurate, and recommended for <200 sequences; it is an iterative refinement method incorporating local pairwise alignment information. https://mafft.cbrc.jp/alignment/software/manual/manual.html
4. Kalyaanamorthy, S., Minh, B.Q., Wong, T.K.F. et al., ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017; 14:587–589. *ModelFinder* is a fast model-selection method that greatly improves the accuracy of phylogenetic estimates by incorporating a model of rate heterogeneity across sites—not previously considered in this context—and by allowing concurrent searches of model space and tree space.
5. *IQ-TREE* takes, as input, a multiple-sequence alignment and will then reconstruct a maximum-likelihood evolutionary tree that is best explained by the input data. http://www.iqtree.org/doc/molevol
6. The Bayesian Information Criterion (BIC) is an index used in Bayesian statistics to choose between two or more alternative models. The BIC is also known as the *Schwarz information criterion* (SIC) or the *Schwarz-Bayesian information criteria*; it was published in a 1978 paper by Gideon E. Schwarz, and is closely related to the Akaike information criterion (AIC), which was formally published in 1974. https://www.statisticshowto.com/bayesian-information-criterion/
7. The gamma categories represent the number of bins in which you discretize a gamma distribution that describes best the rate heterogeneity—without over-parameterizing the model.
8. "Maximum likelihood" simply evaluates a hypothesis about evolutionary history in terms of the probability that the proposed model and the hypothesized history would give rise to the observed data set. A history with a higher probability of reaching the observed state is preferred to a history having a lower probability; obviously, this method searches for the tree with the highest probability or likelihood. http://www.deduveinstitute.be/~opperd/private/max_likeli.html
9. *Hill-climbing nearest-neighbor interchange* is a tree topology search strategy that attempts to improve the likelihood of a given tree, by performing the following operations: Each internal branch of a binary unrooted tree has four subtrees connected to it (a subtree may comprise a single node). NNI then exchanges those subtrees to obtain a new tree. There are only two exchanges, which lead to new unrooted binary trees; the procedure is repeated for each internal branch, until no further likelihood improvements can be obtained. https://www.cs.rice.edu/~ogilvie/comp571/2019/12/02/hill-climbing.html

10. *Markov-chain Monte Carlo* (MCMC) sampling provides a class of algorithms for systematic random sampling from high-dimensional probability distributions. Unlike simple Monte Carlo-sampling methods that are able to draw independent samples from the distribution, MCMC methods draw samples where the next sample is dependent on the existing sample, called a Markov Chain; this allows the algorithms to narrow in on the quantity that is being approximated from the distribution—even with a large number of random variables.
11. *BEAGLE* is a high-performance library that can perform the core calculations at the heart of most Bayesian and Maximum Likelihood phylogenetics packages. https://beagle-dev.github.io/
12. "*Potential scale reduction factor*" (PSRF) is an estimated factor by which the scale of the current distribution for the target distribution might be decreased, if the simulations were continued for an infinite number of iterations; each PSRF declines to 1 as the number of iterations approaches infinity. https://mc-stan.org/docs/2_18/reference-manual/notation-for-samples-chains-and-draws.html
13. Each clade within the tree is given a score, based on the fraction of times that it appears in the set of sampled posterior trees, and the product of these scores is then taken as the tree's score. The tree with the highest score is therefore assigned the maximum clade-credibility tree (MCCT). https://beast2.blogs.auckland.ac.nz/summarizing-posterior-trees/

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions