

Journal: www.joqq.info

Originally Published: Volume 9, Number 1 (Fall 2021)

Reference Number: 91.002

WOULD GENEALOMICS BE AN APPROPRIATE TERM TO DESIGNATE FAMILY TREE RESEARCH BASED ON GENOME-WIDE DATA?

Author(s): *Clarice Sampaio Alho, PhD.; Marcio Dorn, PhD; Eduardo Avila, PhD.*

Would GENEALOMICS be an appropriate term to designate family tree research based on genome-wide data?

Alho, Clarice Sampaio^{1,2}; Dorn, Marcio^{2,3}; Avila, Eduardo^{1,2,4,*}

¹ *Laboratory of Human and Forensic Genetics, School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil.*

² *National Institute of Science and Technology – Forensic Sciences, Porto Alegre, Brazil.*

³ *Laboratory of Structural Bioinformatics and Computational Biology, Informatics Department, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.*

⁴ *Technical-Scientific Section, Rio Grande do Sul State Regional Division, Brazilian Federal Police, Porto Alegre, Brazil.*

* *Corresponding author: e.avila@edu.pucrs.br*

ABSTRACT:

Current abundance of available human individual genomic data has allowed the advent of genealogy applications based on DNA information. Family trees and distant kinship relations can be assessed using new proposed genetic data and analyses methods. This work proposes and discusses the feasibility and pertinence of employing the term “*Genealomics*” as a comprehensive term including genetic data and processing methods focused on DNA-based genealogy applications. More than just a trendy adoption of another “-omics” derived term, this concept is useful to delineate the range of genomic information and technical procedures that can be used with genealogic purposes, including in a forensic context.

KEYWORDS:

genealogy; genetics; investigative genetic genealogy; kinship; genealomics

Genealogy is understood as the research based in documental (primary or secondary) sources to demonstrate kinship and relationships among members in a family tree, pedigree, or lineage. Genealogy research is commonly based on historical reports, official documents and pictures, biographies, civil or military records regarding life events (birth, marriage, and death certificates, for instance), verbal statements, judicial records (as inventories, testaments, and other legal contracts), historic or biographical reports published in old newspapers and almanacs (personal announcements and other public proclamations). Obituaries also comprise a rich source of information, since they usually contain several biographical facts about the deceased and family or relatives, medical records and other sources.

Current genealogical research focuses not only on reconstruction of family trees or biogeographical ancestry determination due to personal interest or curiosity. Instead, such studies are also performed following court request, considering its legal implications, via Forensic Genealogy. Examples of genealogy applications in a forensic context are very common, including efforts to establish, grant or determine citizenship status of an applicant, admitting relatives or finding heirs, or demonstrate close or distant kinship relationships. Since most legal systems ensure inheritance rights to children or other relatives after someone dies, applications in civil or family law are largely widespread. Application of genetic and molecular biology tools to assist in this investigation is also a current practice: genetic analyses of biological samples are extensively employed to determine kinship in court cases, to determine paternity or other biological relationship between individuals. Paternity determination, identification of human remains and disasters victims, investigation of missing persons, human trafficking and illegal adoption practices, and a wide range of other types of criminal activities have been investigated using forensic genetics methods and techniques.

When compared to traditional genealogists, geneticists, correspondingly, build, propose, compare or confirm pedigrees through Genetic Genealogy, essentially based on DNA data, to recognize the genetic inheritance from a particular parent when a child is born. Law enforcement has applied the term *Investigative Genetic Genealogy*, or *IGG*, when DNA profiles from a crime scene or from unidentified human remains are used for human identification purposes by comparing the known genealogy of possible close relatives [1]; *IGG* is also known as *Forensic Genetic Genealogy*, to describe DNA-based relative combined matching.

In biosciences, the suffix “-ome” (from the Greek –ωμα) is applied to form nouns in the sense of referring to a complete whole of a class of substances or data for a species or an individual, in which their constituents and interrelations are collectively and simultaneously considered [2-3]. Such definition usually

includes datasets comprising substantial volume of information [4]. Many “omes”, from the original Genome, first coined in 1920 [5-6], have been adopted by scientists as a term for large-scale analyses and connections in a specific field [7]. From the first sequenced human genome with over three billion base-pairs, current technologic advances led to thousands of complete genome data, which are now widely available. This number is increasing proportionally to sequencing and bioinformatics technology improvements, consolidating the new field called Genomics in order to distinguish and typify organisms, structures or systems from a wide genome data. The radical “-omics” is a derived neologism applied to studies that collectively identify, map and characterize big-data of biological molecules involved in structures, profiles, pathways, and composition of cells, tissues, organisms, species, or populations.

Expansion of massively parallel DNA sequencing facilities have provided large volumes of genomic investigations. A vast amount of genomic information from volunteers, obtained from publicly available research datasets or tested in Direct to Consumer-DTC services, is now accessible. Genealogical exploratory approaches using the genomic information from these collaborators have been the means to investigate kinship links among individuals. Investigators and researchers have been able to successfully identify consanguinity connections, applying methods to recognize identical-by-descent (IBD) DNA segments. Such approaches have been adopted to detect familial matches sharing DNA segments, even when the biological relationship is relatively distant [1]. In the same way as Phylogenomics studies procedures and techniques designed to propose fully resolved phylogenetic trees, methods employing

exhaustive analyzes in large-scale databases with simultaneous comparison of large numbers of highly dense or full-length human genetic profiles, may be required to allow recovering of genealogical relatedness correspondences [8].

Family search methods based on genome-wide big data require both the massive amount of information contained in full length DNA sequences databases and several specific-tailored computational approaches to precisely calculate the degree of relatedness between the contributors. As it intersects methods from the larger fields of genealogy, genomics and bioinformatics, we introduce the term *Genealomics* to refer to such strategy. *Genealomics* is an interdisciplinary field that draws information by computationally comparing entire genomes or large genetic variant datasets (containing thousands to millions DNA polymorphisms) to establish and clarify their kinship or common ancestry relationships; the term could be applied in multiple ways to represent any analysis involving genomic full-data or large-scale microarray genotype data and family reconstruction informatics procedures to predict or suggest personal biological interconnections. Although not exhaustive, the list of potential applications of *Genealomics*-based methods includes the mapping and identification of endogamy levels and pedigree collapse, which can be estimated for a single individual or for a group of people belonging to a local, regional population. Other applications might include reconstruction of historical genomes, belonging to early settlers of specific areas or other people of interest, by combining IBD fragments of a large number of relatively distant descendants. Such applications, of course, would depend on study and availability of across-population level genetic datasets, which might be hard to obtain or have limited access due to ethical or privacy concerns.

Coining a new term, often consisting of a combination of other previously defined words, is reasonable in terminology definitions when it emerges from an accepted, coherent, and well-known logic; we assume the *Genealomics* definition can be useful in optimizing and promoting a supportive keyword for genomic genealogists. Considering it comprises a subfield of Genetic Genealogy, and a particularly technical one, adoption of specific language can successfully drive conceptual structure and compartmentalization. The endorsement of distinguish jargon and terminology by the Genealogy community can serve to highlight a particular subdiscipline deserving of independent focus to establish methods and practices either unique to or adapted for that subfield. Adoption of this concept and terminology may also include the analysis and evaluation of documental (as provided by traditional, not DNA-related genealogical methods) support for kinship or biological relatedness. This consideration is especially relevant when considering the construction of genomic reference databases for a particular population, when clustering methods can be employed to narrow down possible family lines related to a particular sample. Legal and ethical concerns regarding this particular approach must also be discussed in order to verify how the *Genealomics* concept can be adopted in law enforcement and legal cases. Even though some criticisms have been published about the extensive use of “-omics”-derived terms, it is acknowledged that specific nomenclature regarding subsets of genomic applications can be useful in some cases [9-10]. The specificity of analytical methods and genetic information employed in DNA-based genealogical approaches may be considered sufficient to justify the adoption of a particular definition. A concise, direct terminology might help a deeper understanding and broader dissemination of technical concepts associated with these applications to scientific and legal communities or even to the general public, as traditional or genetic genealogists and clients from DTC companies.

Furthermore, it is important to mention that traditional DNA-based methods for kinship determination can be highly improved by *Genealomic* analyses, where both a pipeline of high

computational complexity and very dense panels of DNA markers may lead, in particular cases (especially those including distant biological relatedness or where traditional techniques are insufficient), to more precise and accurate relationships evaluations; studies propose that a large number of nucleotide variations are necessary to determine siblings (at least tens of thousands polymorphisms), while most distant biologic kinship relations would require increasingly higher genetic variants genotyping [8]. Likewise, to return few people with high IBD, e.g., two or more individuals sharing over 100cM (centimorgans) DNA segments inherited in haplotype phase without recombination from a common ancestor, a long-range familial search has to be traced within databases containing full-genomic or highly dense records from a large number of samples, preferentially containing thousands of individuals [1].

Genealomics is a proposed *omics*-inspired term, motivated by the new genealogical proceedings relying on DNA sequence comparisons to infer relationships, where numerous and extensive computational tools are required. It differs from traditional kinship-resolving methods, since it is based on large amounts of genetic information, in contrast to a relatively smaller number of highly polymorphic genetic variants (as short tandem repeats). As such, it has the ability to propose or identify more distant biological relatedness. Therefore, to the scientific community and other interested parties, we suggested the term *Genealomics* could be broadly used to refer to approaches aiming at personal interconnection and kinship prediction or estimation, especially those employing large-scale DNA data and including comprehensive population databases. It consists basically of a subset of the complete genomic data (evaluated at both individual and population levels), particularly the genetic information necessary or useful to investigate or determine biological relatedness among distinct individuals. The concept also includes all analytical methods, approaches and computational systems used to handle and process the biological data. By comparing vast numbers of highly dense DNA sequences or entire genomes among many people, *Genealomics* can be used to propose kinship relationships and identify how closely related they are. It is also important to mention that, as a recent advance derived from the exponential growth in available genomic data, as well as modern progresses provided by bioinformatics research, the need for extensive research, development and improvement in both field-related methodologies and proper use (considering legal, ethical, and technical aspects of available personal genome-wide data) is essential. The contribution of *Genealomics*-derived methods to traditional and genetic genealogy has the potential to significantly improve family and pedigree studies, with implications on many fields as kinship mapping or forensic investigation. The genealogy community is challenged to explore the possibilities presented by this promising area.

ACKNOWLEDGEMENTS:

The authors would like to thank the reviewers and editors for their insightful contributions and suggestions to the present work.

- [1] Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: Current methods, knowledge and practice. *Forensic Sci Int: Genetics*. 2021;52: 102474. doi: 10.1016/j.fsigen.2021.102474
- [2] Yadav SP. The Wholeness in Suffix -omics, -omes, and the Word Om. *J Biomol Tech*. 2007; 18(5): 277 PMID: PMC2392988 PMID: 18166670
- [3] Kuska B. Beer, Bethesda, and biology: how "genomics" came into being. *J Natl Cancer Inst*. 1998;90(2):93. doi: 10.1093/jnci/90.2.93.
- [4] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights*. 2020; 14: 1177932219899051. doi: 10.1177/1177932219899051
- [5] Lederberg J, McCray AT. `Ome Sweet `Omics - A Genealogical Treasury of Words. *The Scientist*. 2001;15(7): 8. Gale Academic OneFile.
- [6] Goldman AD, Landweber LF. What Is a Genome? *PLoS Genet*. 2016;12(7): e1006181. doi: 10.1371/journal.pgen.1006181
- [7] Myers AJ. The age of the "ome": genome, transcriptome and proteome data set collection and analysis. *Brain Res Bull*. 2012;88(4):294-301. doi: 10.1016/j.brainresbull.2011.11.015
- [8] Erlich Y, Shor T, Pe'er O, Carmi S. Identity inference of genomic data using long-range familial searches. *Science*. 2018;362: 690–694. doi: 10.1126/science.aau4832
- [9] Petsko GA. No place like Ome. *Genome Biol*. 2002; 3: comment1010.1. doi: 10.1186/gb-2002-3-7- comment1010.
- [10] Eisen, JA. Badomics words and the power and peril of the ome-meme. *GigaScience*. 2012;1(1):6. doi: 10.1186/2047-217X-1-6.