New Insights into Ethnic and Genomic Diversity

Professor Sung-Hou Kim

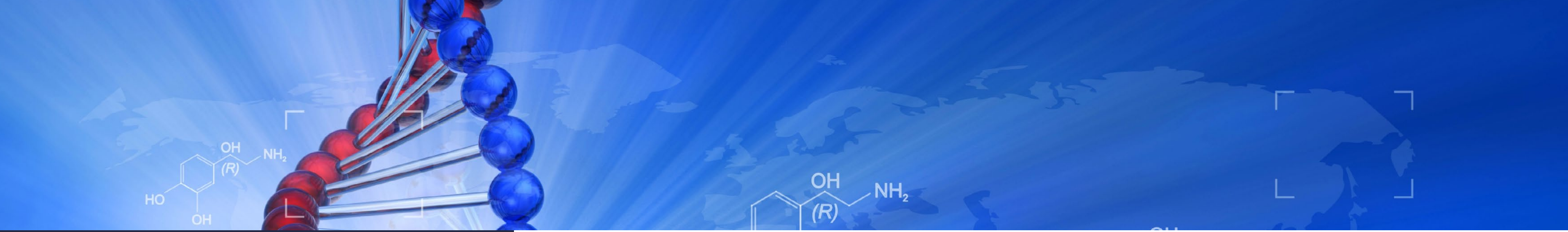
DECEMBER 2024

doi.org/10.33548/SCIENTIA1138

MEDICAL & HEALTH SCIENCES

LIFE SCIENCES & BIOLOGY

Scientia



New Insights into Ethnic and Genomic Diversity

Does our ethnic diversity translate to genomic diversity? New findings suggest that it might not and point instead to considerable genomic similarities across multiple ethnicities. Professor Sung-Hou Kim at the University of California, Berkeley, and his colleagues classified 164 ethnic groups into 14 genomic clusters spread across various geographical regions. Their findings reveal important new insights into our shared human genetic heritage.

Genetic Diversity: From Past to Present

How we, as human populations, diversified from our last common ancestor has been a topic of long debates and discussion. While there are many contrasting theories, it is now widely accepted that overall genomic – the collection of all genes of known and unknown functions – diversity of human species is very small (0.2%) and that a subgroup migrated from the African continent to other parts of the world shows a slightly lower genomic diversity among non-African groups. Understanding these genomic diversities is key to learning about our evolutionary history, identifying genetic links to health and diseases, and predicting our future adaptations.

The two landmark whole genome project initiatives, the 1000 Genomes Project and the Simons Genome Diversity Project (SGDP) have significantly advanced our understanding of human genetic diversity. The SGDP, published in 2016, provided a broader view of genetic diversity by deep sequencing 300 genomes from 142 diverse ethnic populations. In contrast, the 1000 Genomes Project, published earlier in 2015, sequenced over 2,500 individuals from 26 'populations' using a combination of low-coverage whole genome sequencing and dense genotyping to create a detailed map of human genetic variations of the populations. However, despite the availability of this large-scale genomic data, we still have much to learn about how to categorise human populations today.

The Need for Genome-Based Categorisation of Human Populations

Historically, the classification of human populations has relied heavily on physical, cultural, and societal characteristics, often intertwined with other non-genomic factors such as presumed ancestry, language, cultural history, religion, and socioeconomic status. These traditional categorisations have sparked heated debates due to their subjective and qualitative nature and the potential for misclassification or bias.

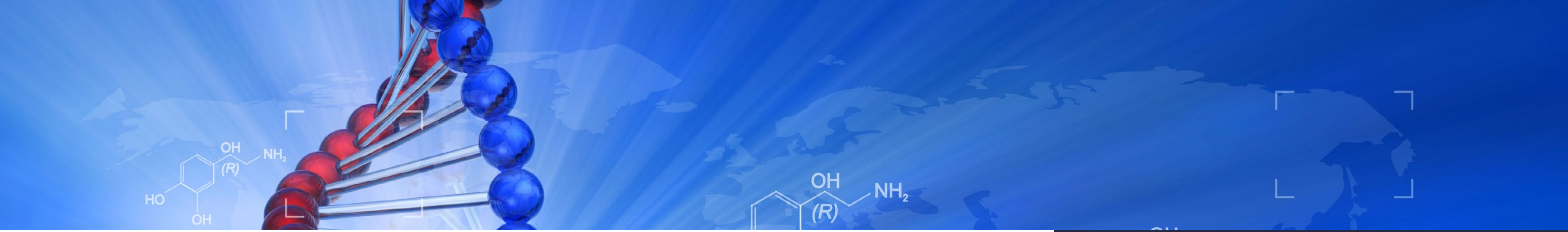
In the past decade, advances in genomics have provided a wealth of data that have the potential to revolutionise our understanding of human diversity. Unlike traditional methods, these genomic data offer objective and quantitative insights into the biological and genomic characteristics of populations. This has been particularly transformative in fields such as human genetics, health sciences, and medical practices, where understanding the genetic or genomic underpinnings of diseases and health conditions can lead to more effective treatments and interventions.

Thus, a better *whole-genome-based* classification system is urgently needed to bridge the gap between genomic data and traditional population categories, enabling a more objective correlation between genetics and health-related outcomes. Professor Sung-Hou Kim and his group at the University of California, Berkeley, has recently provided a comprehensive analysis of human genomic diversity, focusing on the extent of shared genomic material among different ethnic groups.

Genomic Demography

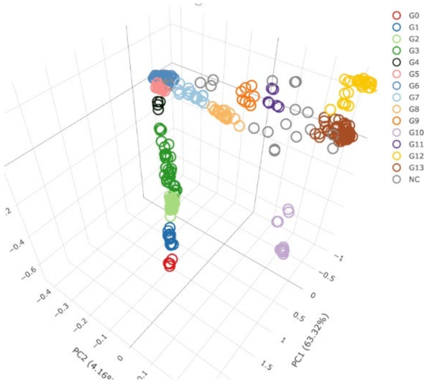
Professor Kim and his colleagues used the recently published SGDP datasets for the whole-genome-based grouping pattern. Although SGDP data is sampled based on ethnic diversity, it is currently the most genomically diverse dataset. Based on their genomic similarities and differences, the study identified 14 distinct genomic groups (GGs) across the world's populations.

These could be further classified into two main supergroups: one consisting of all five African GGs (GG0-GG4) and the other of all non-African GGs (GG5-GG13), which included 119 non-African ethnic groups (EGs) from SGDP data. The African GGs were linearly



connected but not well clustered as compared to the other group. The researchers suggest this was due to a limited availability of sequences representing the vast diversity of African EGs.

Importantly, each GG consisted of multiple EGs, suggesting that no direct correspondence exists between GGs and EGs and that ethnicity is not a genomic construct, i.e., it is a construct of social, cultural, mythical, and other non-genomic factors. However, many GGs represented distinct geographical or geological regions. Notably, members of GG6, GG12 and GG13 are geographically widespread today despite showing lower genomic variation. These groups include populations from several different geographical regions of Europe, the Americas, the far Middle East, East Asia, and Southeast Asia. However, the study suggests that GGs are defined by geological barriers, thus, each genome-based categorisation correlates better with respective environment of its geological region.



Amended from BJ Kim, JJ Choi, SH Kim, On whole-genome demography of world's ethnic groups and individual genomic identity, *Scientific Reports*, 2023, 13, 6316. DOI: https://doi.org/10.1038/s41598-023-32325-w

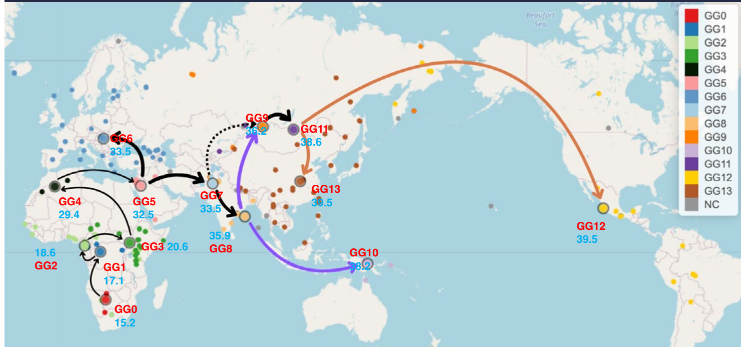
Emergence Order of Genomic Groups

Using a combination of different phylogenetic analyses, Professor Kim's group determined the emergence order of different GGs. For instance, the African GGs emerged sequentially from GG0 to GG4. However, the European GGs and the rest of non-African GGs emerged in a burst separately from the Middle Eastern GG5 during a narrow evolutionary window. The GG12 and GG13, which consist of the Americas and East and Southeast Asia, respectively, emerged more recently. Looking closely at these patterns of genomic divergence offers a new insight into the relationship between EGs and their genomic diversity. Interestingly, the emergence-order mapped on the current world map exclusively based on genomic divergence shows some similarities and differences to each of various maps proposed for the 'migration' of early humans based on various hypothesis combined with various non-genomic factors.

Individual Level Genomic Identity: An Astonishing 99.8%

Professor Kim and colleagues then compared the genomic identity between individuals in EGs from the SGDP dataset and population groups (PGs) derived from the 1000 Genomes Project database. One of the most significant findings was that, on average, 99.8% of genomic material (excluding sex chromosomes, which account about 1% of whole genome) is identical between any two individuals, regardless of their ethnic backgrounds or GGs. This highlights the extensive genomic commonality among all humans. It also emphasises that ethnic differences are relatively minor on a genomic scale. Together, these analyses showed that genomic variation is largely independent of traditional ethnic categorisations. Thus, the identification of GGs based on genomic data provides a more objective representation of the narrow

> "Genome-based categorisation has the potential to enhance the precision of medical research by allowing scientists to identify genetic variants associated with diseases and their prevalence across different genetic groups more accurately.



Amended from Supplementary Fig. S2B of BJ Kim, JJ Choi, SH Kim, On whole-genome demography of world's ethnic groups and individual genomic identity, *Scientific Reports*, 2023, 13, 6316. DOI: https://doi.org/10.1038/s41598-023-32325-w human genomic diversity and better inform future research in genetics, anthropology, sociology and other related fields.

Benefits of Genome-Based Categorisations

Genome-based categorisation has the potential to enhance the precision of medical research by allowing scientists to identify genetic or genomic variants associated with diseases and their prevalence across different ethnic or genomic groups more accurately. This can lead to the development of targeted therapies and personalised medicine, improving patient outcomes. Moreover, genomic data provides a more detailed and objective picture of human evolution and migration patterns for anthropology studies. Lastly, a genome-based approach removes the influence of social, cultural and race bias, leading to a more accurate and equitable understanding of human diversity.

Challenges and Further Considerations

The use of genomic data invariably raises important ethical questions concerning privacy, consent, and potential misuse. Therefore, any further studies on classifying genetic diversity among human populations must prioritise responsible data handling and the careful integration of traditional categories. It is crucial to ensure that the benefits of genome-based research are accessible to all populations, prompting equity in scientific advancements and medical applications. By addressing these ethical considerations, researchers can help safeguard individual rights and foster trust in genomic studies.

MEET THE RESEARCHER

Professor Sung-Hou Kim
Department of Chemistry and Centre for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

Professor Sung-Hou Kim is a member of the Chemistry Department and the Centre for Computational Biology at the University of California, Berkeley. He is also affiliated with the Division of Molecular Biophysics and Integrated Bioimaging at Lawrence Berkeley National Laboratory. Professor Kim obtained his PhD in Physical Chemistry from the University of Pittsburgh under the supervision of Professor GA Jeffrey. He then pursued postdoctoral research at the Massachusetts Institute of Technology under the supervision of Professor Alex Rich. His group at the University of California, Berkeley recently developed a method (based on the Natural Language Analysis model of Information Theory, commonly used in Artificial Intelligence field) to create a 'Whole-genome Tree of Life', showing the relationships among all living organisms, and is applying this to study genomic demography and human ethnic groups as well as viral population such as the COVID-19 virus. Professor Kim is a member of the US National Academy of Sciences and a Fellow of both the American Academy of Arts and Sciences and the American Association for the Advancement of Science.

CONTACT

sunghou@berkeley.edu
https://chemistry.berkeley.edu/faculty/chem/kim

University
of California
Berkeley

KEY COLLABORATORS

Byung-Ju Kim, Department of Chemistry and Centre for Computational Biology, University of California, Berkeley; Incheon National University, Incheon, South Korea

JaeJin Choi, Department of Chemistry and Centre for Computational Biology, University of California, Berkeley.

FUNDING

A gift grant to the University of California (41349-10805-44-CCSHK), Berkeley, CA, USA;

The Priority Research Centers Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2020R1A6A03041954)

FURTHER READING

BJ Kim, JJ Choi, SH Kim, On whole-genome demography of world's ethnic groups and individual genomic identity, *Scientific Reports*, 2023, 13, 6316. DOI: https://doi.org/10.1038/s41598-023-32325-w