H HEURISTICS

RESEARCH REPORT · OCTOBER 2026

The Global Variant Matrix

Recognizing Recurring Human Variants Across Continents, Populations, and Ancestral Groups

Africa · Europe · Asia · The Americas · Oceania · Middle East

Hunter Hughes

H Heuristics · Proposed flagship research report

Contents

About This Report **5**
A proposed idea, not an established finding 5
How to read this report 5
Key terms in one minute 6
Three ways to read this report 7
Evidence grades 7
A note on the figures and tables 7
How to cite 7
Executive Summary **7**
The central thesis 8
What the evidence says 8
The matrix in one picture 9
What the report proposes 9
1. The Central Hypothesis **10**
1.1 The phenomenon 10
1.2 The Global Variant Matrix hypothesis 10
1.3 The important distinction 11
1.4 Three levels of analysis 12
1.5 What the hypothesis does not claim 12
1.6 Relation to earlier H Heuristics work 12
1.7 Why the question matters 12
2. The Global Variant Matrix: Regions and Sampling Frames **13**
2.1 Why use regions at all? 13
2.2 Sampling rules 14
2.3 The regions in brief 14
2.4 Nested sampling: from region to person 15
2.5 The modern world adds a complication 15
3. From Observation to Evidence **16**
3.1 The power and the trap of anecdote 16
3.2 Three places where look-alikes are noticed 16
3.3 Why the matrix needs more than individual examples 17
3.4 Turning an impression into a study 17
3.5 A worked example 17
4. Why Look-Alikes Exist: The Science of Recurrence **18**
4.1 Many small genetic effects 18
4.2 Development narrows the possibilities 18
4.3 Counting look-alikes: the role of dimensions 18
4.4 Typical faces recur most 19
4.5 What real look-alikes show 20
4.6 Why this points to a global pattern 20
5. Human Variation Across Populations: What the Science Says **21**
5.1 One species, one origin, a gradual spread 21
5.2 Most variation lies within populations 21
5.3 The honest caveat: groups can still be told apart 22
5.4 A mosaic of traits 22
5.5 The same outcome by different routes 23
5.6 A two-layer picture of appearance 24
5.7 A caution about whose data we have 24
5.8 Summary for the matrix 24
6. The Matrix Formalized **25**
6.1 A definition 25
6.2 What determines each cell 26
6.3 A simulated world 26
6.4 Reading the matrix: four patterns 27
6.5 From matrix to variant 27
6.6 Why two people look alike: the hierarchy of explanation 27
6.7 Rules for reading a matrix result 28
7. Admixture, Migration and Bridge Populations **29**
7.1 Humans have always mixed 29
7.2 How mixing shapes individual faces 29
7.3 Bridge populations 30
7.4 Two routes to cross-region resemblance 31
7.5 Ancestry, residence and diaspora 31
7.6 Predictions for bridge populations 31
8. Perception and Culture: Who Is Looking? **32**
8.1 Resemblance is judged, not just measured 32
8.2 The other-group effect 32
8.3 Matching unfamiliar faces is hard for everyone 33

8.4 Machines see differently, and unevenly....33
8.5 Culture shapes what a look is....34
8.6 Design implications....34
9. Presentation, the Internet and the Visibility Problem....35
9.1 Presentation: the face we are shown....35
9.2 The internet as a variant amplifier....36
9.3 Attention is concentrated....36
9.4 Alpha variants....37
9.5 The feedback loop....38
9.6 What this means for the matrix....38
9.7 Summary....38
10. Variant Archetypes and the Path to a Catalogue....39
10.1 Why a catalogue?....39
10.2 Four working classes of variant....39
10.3 Variants are neighborhoods, not boxes....40
10.4 Naming and labeling rules....40
10.5 The catalogue pipeline....41
10.6 The variant card....41
10.7 A worked example of a variant card....42
10.8 Retiring entries....42
11. Research Design: How to Test the Matrix....43
11.1 Overview....43
11.2 Study 0: protocol and governance....44
11.3 Study A: reanalysis of existing 3D cohorts....44
11.4 Study B: the cross-region look-alike audit....44
11.5 Study C: perception experiments....45
11.6 Study D: visibility and discovery....45
11.7 Study E: genetic attribution....46
11.8 Risks to the programme....46
11.9 Quality and governance....46
12. Predictions, Falsification and Alternative Explanations....47
12.1 What the thesis predicts....47
12.2 What would show the thesis is wrong....47
12.3 Alternative explanations....48
12.4 Limitations....48
12.5 The evidence map....49
12.6 How the verdict would update....49
13. Ethics, Governance and the Boundary Against Typology....50
13.1 The historical shadow....50
13.2 Five boundaries....50
13.3 Race, ancestry and the language of regions....51
13.4 Privacy, consent and biometric data....51
13.5 Misuse and misreading....52
13.6 Rules for communication....52
13.7 Governance....52
13.8 A commitment....52
14. Applications and Implications....53
14.1 Facial recognition and security....53
14.2 Forensics and DNA phenotyping....53
14.3 Science of the face....53
14.4 Media, culture and online life....53
14.5 Identity and everyday life....54
14.6 Teaching human variation....54
14.7 Who might use this work....55
14.8 What the matrix should not be used for....55
14.9 Open questions for the field....55
15. Conclusion: A Recurring Cast....56
15.1 Summary of the argument....56
15.2 What is established, what is supported, what is proposed....56
15.3 The verdict....56
15.4 Next steps....57
15.5 Five things to take away....57
15.6 Closing....57
Appendix A. Glossary....58
Appendix B. Observation Card and Sampling Worksheet....60
B.1 The observation card....60
B.2 Standard imaging checklist....61
B.3 Regional sampling worksheet....61
Appendix C. References....62
H Heuristics....62

List of Figures

Figure 1. The structure of the report....6
Figure 2. From pair to variant to matrix....11
Figure 3. Why look-alikes are both rare and common....19
Figure 4. How genetic variation is divided....22
Figure 5. A simulated global variant matrix....26
Figure 6. A hypothesized decomposition of resemblance....28
Figure 7. Ancestry mosaics in unmixed and mixed populations....30
Figure 8. Same people, different resolution....33
Figure 9. Styling can turn an ordinary pair into a perceived look-alike....35
Figure 10. Attention is concentrated, and discovery follows attention....37
Figure 11. Variants as neighborhoods in a continuous space....40
Figure 12. The catalogue pipeline....41
Figure 13. An indicative 36-month programme....43
Figure 14. How many people and pairs are needed....45
Figure 15. Evidence map....49
Figure 16. The five boundaries of the Global Variant Matrix....50

List of Tables

Table 1. Key terms at a glance....6
Table 2. Evidence grades used in this report....7
Table 3. Summary of principal findings and confidence levels....9
Table 4. Three levels of analysis in the Global Variant Matrix....12
Table 5. The seven sampling regions of the Global Variant Matrix....13
Table 6. Four levels of sampling in the matrix....15
Table 7. Where variants are noticed, and what distorts each source....16
Table 8. Five ingredients of recurrence and the evidence behind each....20
Table 9. Different visible traits follow different geographic patterns....23
Table 10. Examples of convergent outcomes in different populations....23
Table 11. What Section 5 means for the matrix....24
Table 12. Terms used to define the matrix....25
Table 13. Four patterns the matrix could show and what each would mean....27
Table 14. Examples of bridge populations and their ancestral sources....30
Table 15. Two routes to cross-region look-alikes....31
Table 16. How perception problems affect the matrix and what to do about them....34
Table 17. Three different meanings of "common"....37
Table 18. Working classes of human variant....39
Table 19. Fields of a variant card....41
Table 20. Illustrative variant card (placeholder values)....42
Table 21. The five studies at a glance....43
Table 22. Main risks to the programme and how to manage them....46
Table 23. Eight predictions of the Global Variant Matrix....47
Table 24. Alternative explanations and how the framework tells them apart....48
Table 25. How results would change the grades....49
Table 26. Foreseeable misuses and safeguards....52
Table 27. Seven questions to ask about a look-alike story....54
Table 28. A teaching outline built on the matrix....54
Table 29. Who might use the matrix and how....55
Table 30. Observation card....60
Table 31. Standard imaging checklist for catalogue entries....61
Table 32. Regional sampling worksheet....61

About This Report

This report examines whether the human variant phenomenon, better known as the doppelgänger phenomenon, is limited to particular populations or is spread across all of humanity. Its central thesis is simple to state. People who strongly resemble one another, even though they are not close relatives, may turn up on every continent and in every ancestral group. If so, those resemblances are not scattered accidents. Taken together they may form a *global variant matrix*: a structured pattern of recurring human appearances that runs across national and continental borders.

The report builds on two earlier H Heuristics publications. The first is the essay The Human Variant (or Doppelgänger) Phenomenon, which introduced the word *variant* as a working label for a recognizable visual archetype, and raised the ideas of alpha variants and a Pareto distribution of visibility. The second is the research report The Human Variant Hypothesis, which argued that a shared genetic and developmental architecture channels human appearance into a structured space, so that similar configurations of traits arise again and again. That report asked *why* variants exist. This one asks *where* they are, and whether the whole world fits a single connected pattern.

The report is written for curious general readers, students and researchers from other fields. It uses bachelor's-level vocabulary and defines technical terms where they first appear; a glossary is in Appendix A. It is a proposal and a synthesis. It reviews published science, organizes it around a new idea, and sets out how that idea could be tested. It does not present new measurements of faces or genomes. The simulations in the figures are teaching models built for this report, and each caption says so.

A proposed idea, not an established finding

H Heuristics labels this report a *proposed flagship research report*. That label matters. The words *variant* and *matrix* are proposed analytical terms, not established biological classifications. Nothing in this report says that humanity is divided into a fixed set of physical types. The opposite is closer to the aim: to test whether resemblance moves freely across the lines that people usually use to sort one another.

How to read this report

Each section starts on a new page and can be read on its own. Sections 1 to 5 set out the hypothesis, the regions used for sampling, the science of why resemblance recurs, and what is known about variation between and within populations. Sections 6 to 9 build the matrix itself and examine the forces that could distort it: mixing between populations, human perception, styling, and the way the internet decides which faces we see. Sections 10 to 15 turn the idea into a research programme with a catalogue, studies, predictions that could fail, ethical limits, practical uses and a verdict. Figure 1 shows the layout.

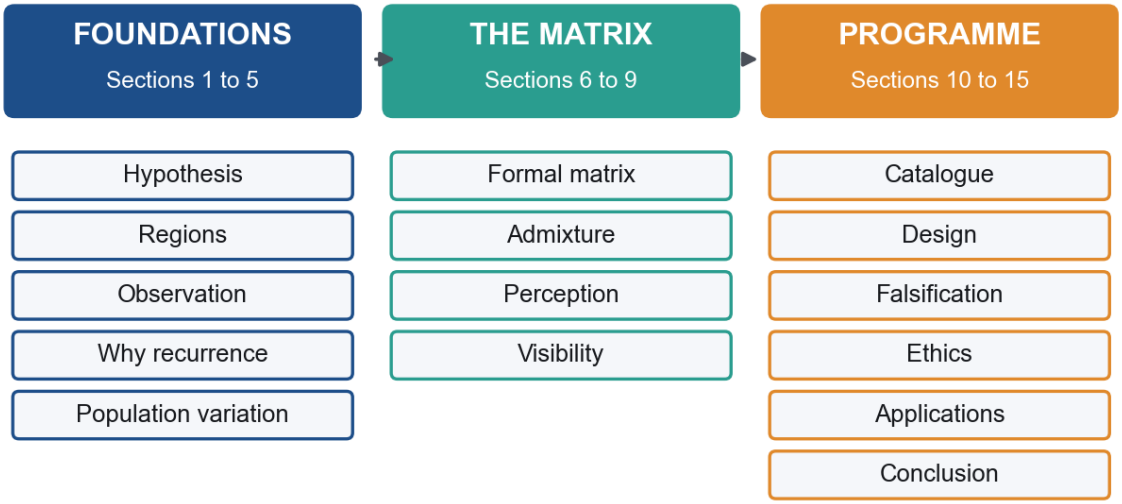

Figure 1. The structure of the report. Foundations (Sections 1 to 5) state the idea and its science. The matrix (Sections 6 to 9) formalizes it and tests it against mixing, perception and visibility. The programme (Sections 10 to 15) turns it into studies, limits and uses.
Source: H Heuristics. Illustrative.

Blue underlined text is a hyperlink. Links to scientific papers go to a DOI record, a permanent identifier that leads to the publisher's page. Every DOI cited in this report was checked against its registered title before publication, and the full reference list is in Appendix C.

Key terms in one minute

A handful of words carry most of the argument. Each is defined again where it first matters, and Appendix A gives the full glossary.

Table 1. Key terms at a glance

Term	Plain meaning	What it is not
Doppelgänger	A person who closely resembles an unrelated person	A sign of relatedness or a shared identity
Variant	A recurring, recognizable configuration of appearance that many people approximate	A kind of person, a race or a subspecies
Global variant matrix	A grid showing how often people in one region or ancestry group have look-alikes in another	A map of human types
Phenotype	The traits of a person that can be observed, such as face shape or skin tone	A prediction of ability or character
Ancestry	The mix of inherited DNA that traces to different parts of the world, measured as shares	A label that sorts people into boxes
Admixture	The mixing of ancestry from populations that were previously separate	A dilution of a pure type

Source: H Heuristics.

Three ways to read this report

- **General reader.** Read the Executive Summary, then Sections 1, 4, 8 and 14. These give the idea, the science behind it, the role of perception and the practical meaning.
- **Student.** Add Sections 5 to 7 for population variation and the formal matrix, and use the glossary and evidence grades to follow the argument.
- **Researcher.** Go straight to Sections 6 and 10 to 13, which define the matrix, the catalogue, the five studies, the falsification tests and the ethical limits.

Evidence grades

Claims are graded on a four-level scale, so that readers can tell settled science from a hunch. The same scale is used throughout the report and summarized in Figure 15.

Table 2. Evidence grades used in this report

Grade	Meaning	Example from this report
A: Established	Confirmed repeatedly by independent studies, or derived mathematically.	Most genetic variation lies within populations, not between them
B: Well supported	Strong evidence from more than one line of work, with some gaps.	Unrelated look-alikes share more common DNA variants than random pairs do
C: Plausible	A credible mechanism and partial evidence. Needs direct testing.	Cross-region look-alikes are common when comparing facial shape alone
D: Speculative	A hypothesis offered for testing and labeled as such.	Alpha variants are over-represented among highly visible people

The grade describes how well we know a claim, not how important it is.

A note on the figures and tables

The report contains 16 figures and a set of tables. Some figures are conceptual diagrams. Others are *illustrative simulations*: small computer models built for this report to show how an idea behaves, such as how the expected number of look-alikes changes as more facial features are compared, or how mixing between populations could affect the matrix. The numbers in these models were chosen to demonstrate a principle and are not measurements of real people. Where the text quotes a number from a simulation, it says so. One figure, on how genetic variation is divided within and between populations, redraws values reported in classic published studies.

How to cite

Hughes, H. (2026). *The Global Variant Matrix: Recognizing Recurring Human Variants Across Continents, Populations, and Ancestral Groups*. H Heuristics Research Report. hheuristics.substack.com

© 2026 Hunter Hughes. This report is independent research and is not medical, genetic, forensic or legal advice.

Executive Summary

> **The question.** Do people who strongly resemble one another turn up everywhere, and do those resemblances fit one connected pattern across the world?

Most people have been told that they look like someone else, and many have found a stranger online who looks uncannily like a friend. These cases are usually treated as curiosities, tied to a single family, town or ethnic group. This report treats them differently. It proposes that recognizable human appearances recur throughout humanity, across continents, ancestral populations and societies, and that these recurrences form a *global variant matrix*.

The central thesis

The human variant phenomenon may not be geographically isolated or exclusive to particular populations. Recognizable patterns of human resemblance may occur throughout humanity, forming a global matrix of recurring human variants. The report develops this thesis in five linked claims:

- Similar-looking individuals can emerge independently across geographically distant populations.
- Recurring facial configurations may be identifiable across different continents.
- Certain recognizable appearance patterns may occur across multiple ancestral backgrounds.
- Population-level differences in physical characteristics coexist with substantial individual variation and overlap.
- Migration, genetic admixture, shared ancestry and recurring developmental processes all shape how human appearances are distributed.

The important distinction is that the matrix does not represent a fixed collection of genetically distinct human types. It is a framework for studying recurring resemblance patterns.

What the evidence says

The report reaches a moderate verdict. The science that exists supports the *ingredients* of the thesis strongly, and the thesis itself is plausible but untested at global scale.

- **Faces are built from many small genetic effects.** Roughly two hundred regions of the genome influence normal facial shape, each with a small effect (White and colleagues, 2021). Many different genetic combinations can therefore produce similar faces.
- **Most human variation is shared across groups.** About 85 to 95 percent of the genetic variation measured in classic studies lies *within* populations (Lewontin, 1972; Rosenberg and colleagues, 2002). Overlap between groups is the rule.
- **Look-alikes carry a genetic signal.** In a 2022 study, about half of the most similar unrelated look-alike pairs shared more common DNA variants than random pairs (Joshi and colleagues, 2022). The sample was small and mostly European, so the result needs replication in other populations.
- **The same trait can evolve more than once.** Light skin evolved partly through different genes in Europeans and East Asians (Norton and colleagues, 2007), and lactase persistence arose through separate mutations in Europe and Africa (Tishkoff and colleagues, 2007). Independent routes to similar outcomes are a normal feature of human biology.

- **What counts as a look-alike depends on what is compared.** A simple model in this report shows that comparing more facial features makes look-alikes exponentially rarer, and that coarser perception makes them far more common (Figure 3, Figure 8).

The matrix in one picture

A global variant matrix is a table of how often people in one region or ancestral group have a look-alike in another. The diagonal shows resemblance within a region. The off-diagonal cells show resemblance across regions. In an illustrative simulation (Figure 5), cross-region look-alikes are almost as common as within-region ones when only facial shape is compared. When strongly regional surface traits such as skin tone and hair form are added, cross-region matches fall sharply for distant regions but stay high between neighbors and for mixed-ancestry populations. This pattern is a prediction of the framework, not a finding about real populations.

Table 3. Summary of principal findings and confidence levels

Finding	Basis	Grade
Facial form is highly polygenic and largely regulatory	Large 3D-imaging genetic studies of facial shape	A
Most genetic variation lies within populations, and groups overlap	Classic and genome-wide apportionment studies	A
Perceived resemblance depends on the observer and the image	Face-perception and algorithm-testing research	A
Unrelated look-alikes share common genetic variants	One multi-omic study of verified pairs	B
Typical faces have more look-alikes than distinctive ones	Face-space theory and probability models	B
Cross-region look-alikes are common in facial shape	Model-based, with no direct global test yet	C
Admixed populations bridge otherwise distant regions of the matrix	Admixture genetics and simulation	C
The internet inflates apparent recurrence	Network science and discovery bias	C
Alpha variants exist as a visibility pattern	Proposed in the H Heuristics essay; untested	D

Grades follow Table 1. A global test of the matrix has not been done.

What the report proposes

Section 10 proposes a *variant catalogue*, in which each variant is defined before examples are collected, measured under standard conditions and graded for the strength of its evidence. Section 11 describes five linked studies to test the matrix, from reanalysis of existing 3D face databases to a cross-region look-alike audit. Section 12 lists predictions that could prove the framework wrong. Section 13 sets firm ethical limits: variants are not types or races, appearance does not reveal character, and resemblance is not identity. Section 14 explains why the work matters for facial recognition, forensics, media and everyday life. Section 15 gives the verdict and next steps.

> **In short.** Humanity may contain a recurring cast of appearances that no border, ocean or ancestry label contains. The task is to find out whether that is true, and to do it without turning people into types.

01 CENTRAL HYPOTHESIS

1. The Central Hypothesis

1.1 The phenomenon

Everyone has had the experience. Someone says, "You look just like a person I know." A stranger on a train turns out to share a friend's jawline. A photo on social media seems to show a classmate who lives on another continent. These moments are usually told as funny stories. They are rarely treated as data.

In everyday language such a person is a *doppelgänger*, a German word meaning "double walker." It usually means one person who looks strikingly like another. H Heuristics uses a broader word, *variant*, as a working label for a recognizable visual archetype: a pattern of appearance that can be observed, compared and possibly measured. The word is chosen to shift attention from a pair of people to a pattern. Two look-alikes are a pair. A recurring appearance seen in many people is a variant. The Human Variant Hypothesis defines a human variant as a recurring configuration of observable traits, not a kind of person, and this report keeps that definition.

The starting point for this report is an observation made over time, in everyday life and online. The author of H Heuristics has noticed people who seem to reproduce the same recognizable appearance in African, Asian, Hispanic, Arab and European settings, and in others as well. The impression is that the doppelgänger phenomenon is not tied to any one region or ancestry. This report takes that impression seriously, states it as a hypothesis, and asks what it would take to test it.

1.2 The Global Variant Matrix hypothesis

> **Central thesis.** Recognizable patterns of human resemblance may occur throughout humanity, across continents, ancestral populations and societies, forming a global matrix of recurring human variants.

The word *matrix* is chosen on purpose. A matrix is a grid. Picture the world's populations along the rows and again along the columns. Each cell asks one question: how often does a person from the row population have a close look-alike in the column population? A world in which look-alikes were confined to families and villages would produce a grid that was bright along the diagonal and empty elsewhere. A world in which resemblance was spread evenly would produce a uniform grid. The hypothesis is that the real world sits somewhere between the two, with structure but not isolation. Figure 2 shows the idea in three steps.

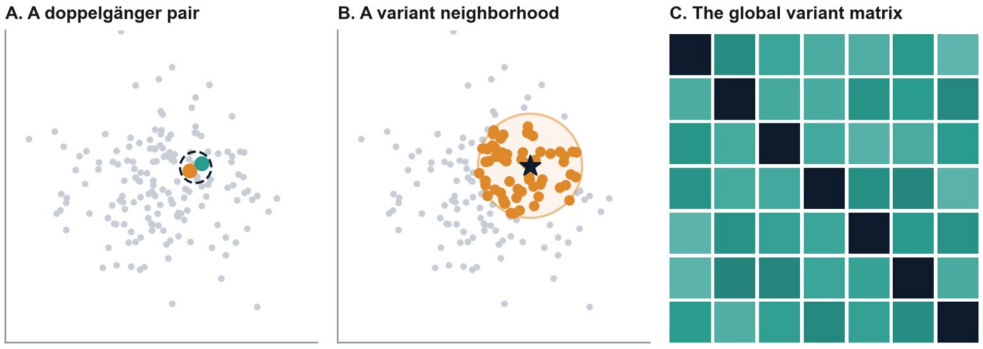

Figure 2. From pair to variant to matrix. (A) A doppelgänger pair is two people close together in a space of facial features. (B) A variant is the neighborhood around a reference point: everyone who resembles that reference to a similar degree. (C) The global variant matrix records how often such neighborhoods are shared between the world's populations. Dark cells on the diagonal show resemblance within a population; lighter cells show resemblance across populations.

Source: H Heuristics. Conceptual illustration; points and cell values are simulated.

The hypothesis is built from five claims.

1. **Independent emergence.** Similar-looking individuals can emerge independently across geographically distant populations.
2. **Cross-continental recurrence.** Recurring facial configurations may be identifiable across different continents.
3. **Cross-ancestry patterns.** Certain recognizable appearance patterns may occur across multiple ancestral backgrounds.
4. **Overlap and difference together.** Population-level differences in physical characteristics coexist with substantial individual variation and overlap.
5. **Shared processes.** Migration, genetic admixture, shared ancestry and recurring developmental processes contribute to the distribution of human appearances.

The fourth claim deserves emphasis. The matrix does not deny that populations differ on average in some visible traits. Skin tone, for example, differs strongly between regions, and this report treats that as established science (Section 5). The claim is that differences between group averages sit alongside very large overlap between individuals, so that a person at one end of a group can resemble someone in another group more than they resemble many of their own neighbors.

1.3 The important distinction

The matrix would not necessarily represent a fixed collection of genetically distinct human types. It would represent a framework for studying recurring resemblance patterns. This distinction runs through the whole report. A *type* is a box with a boundary: you are in it or you are not, and membership is assumed to say something deep about you. A *variant*, as used here, is a neighborhood in a continuous space: a region where many people happen to sit, with no sharp edge and no claim about what lies beneath the surface. The variant idea is meant to be the opposite of sorting people into types; Section 13 returns to why.

1.4 Three levels of analysis

The thesis works at three levels, each larger than the one before. Most popular writing about look-alikes stays at the first level. The report's contribution is to ask whether the second and third levels exist.

Table 4. Three levels of analysis in the Global Variant Matrix

Level	Unit of study	Main question	Typical evidence
1. Doppelgänger	A pair of people	Do these two people look alike, and why?	Photographs, face-matching scores, DNA comparison
2. Variant	A recurring appearance seen in many people	Is there a repeated configuration, and how widely is it shared?	Standard images, 3D measurements, blind ratings
3. Matrix	All variants across all regions and ancestral groups	Is there a connected structure in how variants are distributed worldwide?	Cross-population samples, recurrence statistics, admixture data

The report treats Level 3 as a hypothesis. Levels 1 and 2 already have partial support in the literature.

1.5 What the hypothesis does not claim

The Global Variant Matrix hypothesis does not claim any of the following.

- **A fixed number of human types.** The number and shape of variants depends on how finely you look, as Section 4 shows.
- **Hidden lineages or special genes.** Resemblance does not require a shared ancestor, a secret bloodline, or one gene for each archetype.
- **That appearance predicts personality, ability or destiny.** It does not, and Section 13 sets this as a boundary.
- **That population differences are unimportant.** Some visible traits differ systematically between regions, and any honest study must measure and control for that.
- **That the internet is a fair sample of humanity.** What we notice online is shaped by algorithms and fame, and Section 9 treats this as a major source of error.

1.6 Relation to earlier H Heuristics work

This report sits within a small family of H Heuristics writings. The essay on the human variant phenomenon uses named reference points such as the *Hunter variant* and the *Lars variant*: labels for perceived configurations, each a reference person plus a recurring cast of people who reproduce aspects of that appearance. The labels do not establish a shared ancestor, a variant gene or a fixed personality. A related essay on a *European Variant Matrix* applied a similar lens to one region. This report generalizes the idea to the whole world. The Human Variant Hypothesis supplies the biology (why resemblance recurs); this report supplies the geography (where it recurs and how the pieces connect).

1.7 Why the question matters

If the thesis is right, scientists studying faces would need samples from every region, because a pattern seen in one population may recur in others. People building facial-recognition systems would have to expect look-alikes everywhere, and more false matches in some groups than others. If it is wrong, that is also worth knowing: a null result would still be useful.

02 REGIONS AND SAMPLING FRAMES

2. The Global Variant Matrix: Regions and Sampling Frames

2.1 Why use regions at all?

To study a global pattern, researchers must decide where to look. The report proposes seven major sampling regions: Africa, Europe, Asia, North America, South America, Oceania and the Middle East. Each is a place on the map where a group of populations lives or came from. The regions are research sampling regions, not biological race categories. Their only job is to make sure that a study does not draw all of its faces from one corner of the world.

Geography is a practical way to organize sampling, but it is a crude one. The largest regions contain hundreds of languages and cultures, and ancestry does not stop at the edge of a continent. Africa alone holds more genetic diversity than the rest of the world combined, in the sense that people outside Africa carry a subset of the variation found within it (Ramachandran and colleagues, 2005). The study should therefore examine populations *within* each region rather than treating a whole continent as one uniform group.

Table 5. The seven sampling regions of the Global Variant Matrix

Geographic scope	Research focus	Points to keep in mind
Africa	Recurring facial and morphological patterns across diverse African populations	Greatest genetic diversity of any region; very wide range of facial form, stature and pigmentation within the continent
Europe	Recurring appearances across Northern, Southern, Eastern and Western Europe	Genetic variation follows geography closely, yet most variation is shared across the continent
Asia	Variants across East, Southeast, Central and South Asia	Over half of humanity lives here; the four subregions differ in history and ancestry
North America	Indigenous, European-descended, African-descended and admixed populations	Shaped by older Indigenous history and by the movements of recent centuries
South America	Indigenous, European, African and mixed-ancestry populations	Many people carry ancestry from three or more continents in varying proportions
Oceania	Aboriginal Australian, Melanesian, Polynesian and other Pacific populations	Long, separate histories in Australia and New Guinea, and a recent island-by-island settlement of the Pacific
Middle East	Recurring appearances across Arab, Persian, Levantine and other populations	A crossroads between Africa, Europe and Asia, with deep gene flow in all directions

Regions are sampling frames, not biological races. The scope column follows the research proposal in the H Heuristics project brief.

2.2 Sampling rules

Five rules keep the regional frame honest.

1. **Sample populations, not continents.** Within each region, define sub-samples by place, language community or genetic ancestry, and report results for each.
2. **Use ancestry as a measured covariate, not a label.** Where genetic data are available, estimate each person's ancestry as a continuous mixture. Where they are not, record self-described background and place of residence separately.
3. **Record how each person entered the study.** A volunteer from a photography project, a celebrity and a randomly invited participant are three different samples (Section 9).
4. **Include admixed and bridge populations on purpose.** Populations with mixed ancestry are where the matrix is most likely to show cross-region links (Section 7).
5. **Match samples on age and sex.** Faces change with age and differ on average between the sexes, so unmatched samples produce false patterns.

2.3 The regions in brief

Africa

Modern humans arose in Africa, and African populations hold the deepest genetic diversity on Earth. People within the continent differ widely in body build, skin tone, facial form and hair, and the differences between, say, West, East, Central, Southern and Northern African populations can exceed differences between populations on other continents. A single "African" sample would hide most of that variety, so the framework calls for sampling across the continent. Studies of skin pigmentation in African populations have found a wide span of tones and a complex genetic basis (Crawford and colleagues, 2017). The facial genetics of African populations is still the least studied of the major regions, which is a gap this report returns to.

Europe

Genetic variation within Europe tracks geography so closely that a map of genetic differences can look like a map of the continent (Novembre and colleagues, 2008). Even so, the differences between European countries are small compared with the differences among individuals within them. European-ancestry people dominate current research on facial genetics, so European faces are the best documented. This is a reason for caution: a pattern measured mainly in Europeans may be assumed to be general when it is not.

Asia

Asia holds the majority of the world's people and several quite different population histories. East Asia, Southeast Asia, Central Asia and South Asia each have their own genetic structure. South Asia, for instance, contains many long-separated groups, which makes it far from uniform (Reich and colleagues, 2009). Central Asia sits on old crossroads where ancestry from East and West meets. Treating Asia as one block would be the most misleading shortcut of all.

North America and South America

The Americas were settled from northeast Asia many thousands of years ago, and Indigenous populations went through strong founder effects as they spread. Since 1492 the continents have also

received large migrations from Europe and Africa, and from Asia. Many people in the Americas carry ancestry from several sources. A national study of the United States found that, on average, African Americans carry roughly three-quarters African ancestry and a quarter European, and Latinos on average carry mostly European ancestry with substantial Indigenous American and smaller African components, with wide variation between individuals and regions (Bryc and colleagues, 2015). In Mexico, the genetic structure of the present population reflects distinct Indigenous regional ancestries (Moreno-Estrada and colleagues, 2014). Because the Americas contain both older Indigenous populations and recently mixed ones, they are central to the matrix.

Oceania

Australia and New Guinea were settled long ago by populations that then remained largely distinct, while the islands of the Pacific were settled much more recently, in a sequence of voyages that reached Polynesia within the last few thousand years (Skoglund and colleagues, 2016). The region therefore combines very old and very recent population histories, which makes it a valuable test of whether similar facial configurations can arise on different branches of the human family tree.

Middle East

The Middle East has long been a land bridge between Africa, Europe and Asia. Populations there carry ancestry from several directions, and genetic studies of the Levant show that cultural groups, not just geography, shape the pattern of relatedness (Haber and colleagues, 2013). Because it connects other regions, the Middle East is expected to show some of the strongest cross-region links in the matrix.

2.4 Nested sampling: from region to person

A single level of sampling is never enough. The framework asks for sampling at four nested levels, so that a pattern can be traced from the world down to individuals and back up again.

Table 6. Four levels of sampling in the matrix

Level	Example	What it controls
Region	Africa; Asia; Oceania	That no part of the world is left out
Subregion	East Africa; Southeast Asia; Polynesia	That diversity inside a region is visible
Community or sample	A city, a language community, a cohort	That sampling bias can be named and tested
Individual	One person with age, sex, ancestry and discovery route recorded	That every face enters the study with its context

Source: H Heuristics. Proposed sampling design.

2.5 The modern world adds a complication

Migration in recent centuries has moved people between regions faster than at any earlier time. A sample drawn from a large modern city may contain people whose ancestry belongs to several regions, and photographs shared online may show people living far from where their families began. The matrix therefore needs two kinds of region: *regions of ancestry*, where a person's family history traces to, and *regions of residence*, where the person lives and is photographed. A study of look-alikes that confuses the two will find patterns that come from where people moved, not from how appearances recur. Section 7 takes up this problem directly.

03 FROM OBSERVATION TO EVIDENCE

3. From Observation to Evidence

3.1 The power and the trap of anecdote

The idea behind this report began as an observation: look-alikes seem to show up in every part of the world. Observations like this are where most science begins. They are also easy to get wrong. The human mind is built to notice faces and to link them to faces it already knows. We are good at spotting a match and poor at counting the times we saw no match at all. Psychologists call the first tendency the *availability heuristic*: events that come easily to mind seem more common than they are (Tversky and Kahneman, 1973). A related habit, *confirmation bias*, leads people to notice evidence that fits an idea they already hold (Nickerson, 1998).

None of this means the observation is wrong. A person who believes look-alikes appear everywhere will notice every one that does, and forget the thousands of faces that resembled nobody. This section explains how an impression can be turned into evidence that survives these biases. The short version: an observation is a lead; a rate with a denominator, a standard image and a blind rating is evidence.

3.2 Three places where look-alikes are noticed

Most reported look-alikes come from three sources: ordinary life, the internet and the media. Each shows something real, and each is distorted in its own way.

Table 7. Where variants are noticed, and what distorts each source

Source	What it offers	Main distortions
Everyday life	A real sample of people in one's own city, workplace and travels; faces seen in motion, with expression	Limited to the places a person goes; familiar groups are seen in finer detail than unfamiliar ones; memory keeps the hits
The internet	Faces from almost every country, shared by billions of people; easy to search and compare	Algorithms choose what appears; popular and young faces dominate; photos are edited, filtered and posed
Media and fame	Well-known reference faces that millions can compare against	A few famous faces are compared with huge numbers of strangers, so matches are found more often; styling and casting are shaped to look alike

Source: H Heuristics. See Section 9 for the role of visibility.

The three sources together give a rich supply of leads. They cannot, on their own, show how common variants are in the whole population. The internet is, in the words of the H Heuristics essay, a powerful place to generate hypotheses but not a representative population survey. The point is to separate *discovery* from *proof*.

3.3 Why the matrix needs more than individual examples

A collection of striking pairs is not yet a matrix. A matrix is about *rates*: how often people in one place have a look-alike in another. A rate needs a denominator, which means the number of comparisons made, not only the number of matches found. Suppose someone finds ten African, ten Asian and ten European look-alikes of a reference face. That is interesting, but without knowing how many faces were searched, the number says nothing about how common the match is. Finding ten matches among a thousand faces and finding ten among a hundred million are very different facts.

This is why Section 6 gives the matrix a formal definition built on match rates, and why Section 11 proposes studies that record every comparison, not just the successes.

3.4 Turning an impression into a study

The H Heuristics essay lays out a four-step path from impression to evidence. This report adopts it and adds detail.

1. **Define each variant before collecting examples.** Write down in advance which features make someone count as a match for a given reference. If the definition is drawn after looking at examples, it will fit the examples and prove nothing.
2. **Select images without cherry-picking, and judge them blind.** Collect photographs by a fixed rule, show them under standard conditions to independent raters who do not know the hypothesis, and include control pairs that were not chosen for resembling each other.
3. **Test prominence directly.** If a variant is said to be over-represented among famous or highly visible people, compare public figures with a matched sample of less visible people, accounting for age, geography and occupation.
4. **Measure attention rather than assume a curve.** If attention is concentrated, measure how concentrated, instead of assuming a textbook 80/20 split.

A fifth step belongs to this report. **Always record the denominator.** Every study should state how many faces were examined, from which regions, and by what route they were found.

3.5 A worked example

A made-up example shows how this works. Imagine an observer who sees a person in a market in one city, a person in a news video from a second continent, and a person in a photo from a third, and feels that all three "have the same look." Three things follow.

- **Write the claim down.** Which features are shared? The set of the eyes, the length of the face, the jaw, the brow? Are hair and styling part of the claim, or are they set aside?
- **Standardize the images.** Find photos with similar lighting, a neutral expression, no glasses and a plain background, or use three-dimensional scans if they are available.
- **Compare with controls.** Show the three images, mixed with images of people not selected for resemblance, to a panel of raters from different backgrounds who do not know which are the "matches." See whether the raters agree, and whether the three stand out from the controls.

If the three stand out, the impression has passed its first real test. If not, it was probably a product of styling, mood or habit, which is useful to know too. Appendix B gives a one-page observation card for this process.

04 WHY LOOK-ALIKES EXIST

4. Why Look-Alikes Exist: The Science of Recurrence

4.1 Many small genetic effects

Before asking where variants occur, it is worth being clear about why they occur at all. The Human Variant Hypothesis gives the full argument; this section summarizes the parts that matter for a global pattern.

A face is not controlled by a few genes. Large genetic studies using three-dimensional face scans have linked roughly two hundred regions of the genome to differences in normal facial shape, each with a small effect (White and colleagues, 2021; Claes and colleagues, 2018). Traits built from many small effects have a useful property: very different genetic recipes can give nearly the same result. A cook can reach the same total sweetness from many different mixes of sugar, fruit and honey. In the same way, two people can arrive at a similar nose length or cheekbone width from quite different sets of inherited variants. Geneticists call this a many-to-one map from genes to traits.

A second feature of facial genetics matters for the matrix. Most of the signals sit in stretches of DNA that regulate when and how strongly genes switch on, not in the genes themselves. Regulatory variants tend to turn the dials of a shared developmental program rather than create new parts. The same dials exist in everyone, so people in different parts of the world vary along the same set of axes.

4.2 Development narrows the possibilities

A genome is a set of instructions, not a picture. Facial development is a long, carefully timed process in which migrating cells, growing tissue and signaling molecules build the face step by step. Biologists describe such systems as *canalized*: they are buffered, so that small differences in genes or environment are absorbed and development reaches similar end points. They are also *modular*: the eye region, nose, mouth and jaw develop partly independently, so one person may resemble another in the mid-face but not in the jaw (Klingenberg, 2008).

These features narrow the range of faces that are likely. Out of all conceivable faces, real ones occupy a small and structured region. When many people share a region of space, many of them end up close together. That is what a look-alike is.

4.3 Counting look-alikes: the role of dimensions

How many look-alikes should a person expect? The answer depends on a question that is easy to overlook: how many features are being compared, and how close is close enough? Imagine describing a

face with a set of independent measurements, such as eye spacing, nose width, jaw angle and so on. Say two faces count as look-alikes if they agree to within a set tolerance on every measurement. A simple model shows what happens as the number of measurements grows.

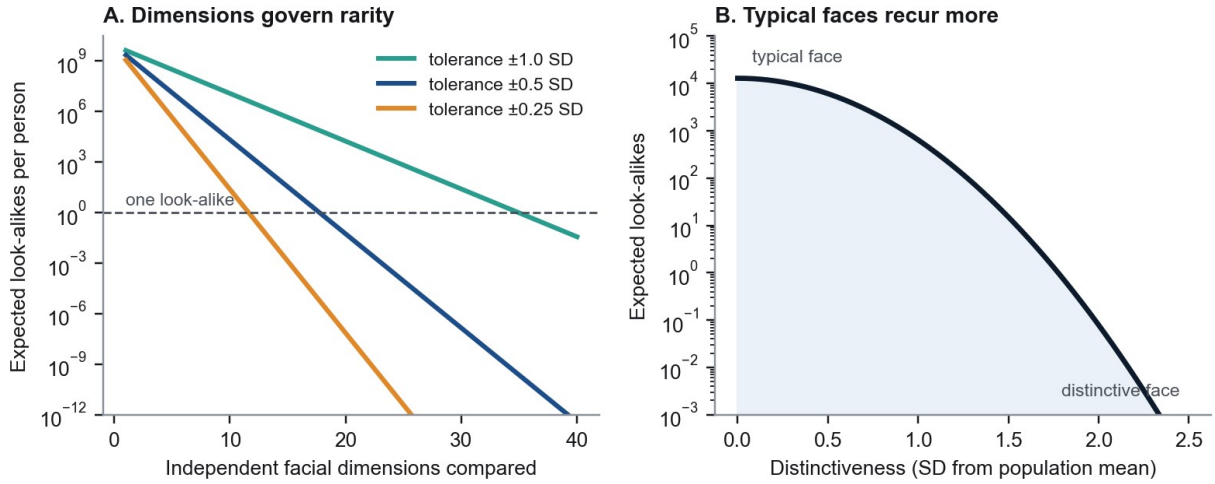

Figure 3. Why look-alikes are both rare and common. (A) Expected number of look-alikes per person in a population of 8.1 billion, as a function of the number of independent facial dimensions compared, for three levels of tolerance. Few dimensions or loose tolerance give everyone millions of look-alikes; many dimensions and tight tolerance give almost no one any. (B) For 12 dimensions at a tolerance of 0.6 standard deviations, people with typical faces (left) have far more expected look-alikes than people with distinctive faces (right).

Source: H Heuristics. Illustrative model: independent standard-normal dimensions; two faces match if they differ by less than the tolerance on every dimension.

The numbers in the model are striking. Comparing 8 dimensions at a loose tolerance of one standard deviation gives each person about 44 million expected look-alikes in a population of 8.1 billion. Comparing 12 dimensions at 0.6 standard deviations gives roughly 12,850. Comparing 20 dimensions at 0.5 standard deviations gives about 0.05, which means most people would have none. The question "Does everyone have a double?" therefore has no single answer. It depends on how finely the faces are compared.

This matches the best direct measurement. An analysis of faces in a large image database found that the chance of two people matching exactly on eight facial dimensions was below one in a trillion (Lucas and Henneberg, 2015). Exact duplication is effectively impossible, yet close resemblance on a smaller set of the features that people notice most is common.

4.4 Typical faces recur most

Panel B of the figure shows a second pattern. People whose faces sit near the middle of the population distribution have many more look-alikes than people whose faces are unusual. Psychologists have long described faces as points in a mental *face space* centered on an average face. Faces far from the center are distinctive and easy to recognize, while faces near the center are easier to confuse with each other (Valentine, 1991). It has also been shown that blended, average faces look familiar and are rated as attractive (Langlois and Roggman, 1990).

For the matrix, this means that variants are probably most common in the dense center of each population's distribution, and rarer toward its edges. It also means that a person with a distinctive face is less likely to appear in someone's list of doppelgängers, wherever they live.

4.5 What real look-alikes show

The most direct test of the science comes from a 2022 study of unrelated look-alikes. The researchers recruited pairs from a long-running photography project on people who resemble one another, scored each pair's similarity with three facial-recognition algorithms, and collected DNA, DNA methylation and microbiome data (Joshi and colleagues, 2022). Among the pairs rated highly similar by all three algorithms, a substantial subset clustered together on common genetic variants, sharing many more variants than random pairs did. The look-alikes were less alike in their epigenetic and microbiome profiles, which suggests the resemblance was linked mainly to inherited variation, not shared environment.

The study has limits that matter here. It was small. Its participants were recruited through a photography project and were mostly of European ancestry. The result therefore shows that unrelated look-alikes *can* share common genetic variants. It does not show how often this happens, and it does not test whether the same holds in African, Asian, Indigenous American, Oceanian or Middle Eastern populations. A global matrix is, in effect, a proposal to run that test everywhere.

Table 8. Five ingredients of recurrence and the evidence behind each

Ingredient	What it contributes	Key evidence	Grade
Many small genetic effects	Different genotypes converge on similar faces	Roughly 200 loci with small effects on face shape (White et al., 2021)	A
Shared regulatory dials	People vary along the same axes	Facial loci largely regulatory; consistent loci across groups (Xiong et al., 2019)	B
Canalization and modularity	Fewer effective dimensions; partial resemblance	Morphological integration research (Klingenberg, 2008)	A
Typicality	Most look-alikes sit near the population center	Face-space theory (Valentine, 1991)	B
Convergent genetic signal	Unrelated look-alikes share common variants	One multi-omic study (Joshi et al., 2022)	B

Grades follow Table 1. Full citations are in Appendix C.

4.6 Why this points to a global pattern

Nothing in the mechanism is limited to one region. The genetic architecture of the face is shared across the species. Genetic studies of facial shape have found loci with consistent effects in European and East Asian samples (Xiong and colleagues, 2019), and a study of admixed Latin Americans found loci influencing facial features there as well (Adhikari and colleagues, 2016). Developmental canalization and modularity are properties of the human species. The probability argument applies to any population large enough to contain many people.

Taken together, these points mean that recurring facial configurations are expected wherever there are many people. They do not prove that the *same* configurations recur across regions, which is the real claim of the matrix. To see whether they might, we need to look at how human variation is actually distributed across populations. That is the subject of the next section.

05 VARIATION ACROSS POPULATIONS

5. Human Variation Across Populations: What the Science Says

5.1 One species, one origin, a gradual spread

Every living human descends from populations that lived in Africa. As people spread across the planet, small groups left and founded new populations, each carrying only part of the variation of the group it came from. The result is a gradual loss of genetic diversity with distance from Africa, a pattern found in a survey of human populations and consistent with a serial founder effect (Ramachandran and colleagues, 2005; Henn and colleagues, 2012). Genetic differences between people also follow geography in a smooth way. Worldwide patterns of genetic variation form gradients, not sharp breaks (Li and colleagues, 2008).

Two consequences matter for the matrix. First, populations that share recent ancestry share more of the genetic variants that shape appearance, so neighbors tend to look more alike on average than distant populations do. Second, because the gradients are smooth and the founding groups were drawn from a common pool, very large amounts of variation are shared across the whole species.

5.2 Most variation lies within populations

The best-known finding in human population genetics is that most genetic variation is found *within* populations, not between them. In a classic 1972 analysis of protein and blood-group markers, Richard Lewontin estimated that about 85 percent of the variation lay within local populations, with the rest divided between populations within a major group and between major groups (Lewontin, 1972). Genome-wide studies using hundreds of markers found the same broad pattern: roughly 93 to 95 percent of the variation measured lay within populations (Rosenberg and colleagues, 2002). Figure 4 shows the two analyses side by side.

This is the first pillar of the matrix. If most variation lies within populations, any two populations have a great deal in common, and a person at one end of a group may have more in common with a person from another group than with the person at the other end of their own.

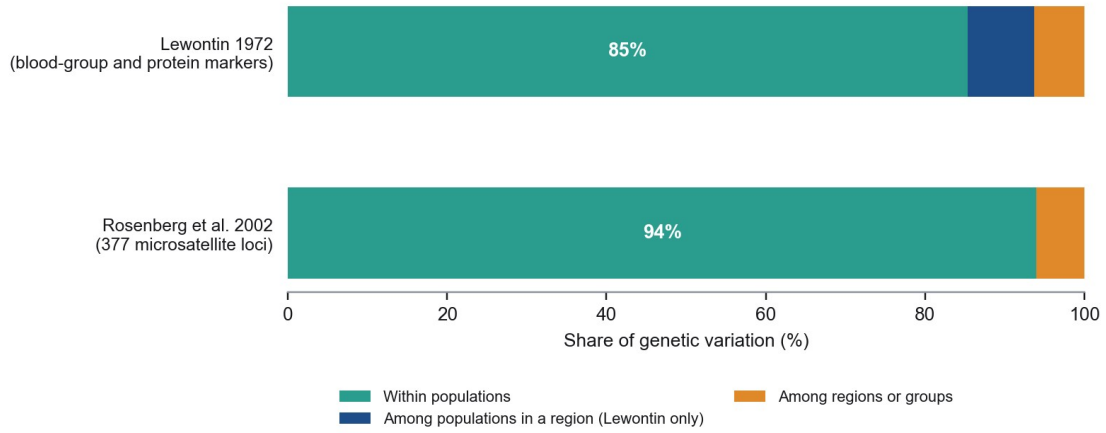

Figure 4. How genetic variation is divided. In both classic analyses most variation is found within populations. Values are approximate and rounded. In the Rosenberg study the small remainder was split between populations within regions and between regions.

Source: Lewontin (1972); Rosenberg et al. (2002). Redrawn by H Heuristics from the reported percentages.

5.3 The honest caveat: groups can still be told apart

A fair account has to include a second finding that points the other way. When researchers use many genetic markers at once, individuals can be assigned to broad geographic clusters with high accuracy (Rosenberg and colleagues, 2002). Small average differences on thousands of markers add up. A study of genetic similarities within and between populations found that with only a few markers, pairs of people from different populations were often more alike than pairs from the same population, but that the picture changed as the number of markers grew (Witherspoon and colleagues, 2007). Later theoretical work has shown that a high share of variation within populations does not by itself rule out clear differences between populations once many traits are combined (Edge and Rosenberg, 2015).

The lesson is the same one found in the look-alike model in Section 4: results depend on the number of dimensions compared. With few dimensions, overlap dominates. With many, small differences accumulate and groups separate. For the matrix this is a central point. Cross-region look-alikes should be common when only a few features are compared and rarer when many features, especially strongly regional ones, are compared together. The matrix is therefore not one grid but several, one for each choice of features.

5.4 A mosaic of traits

Visible traits do not all follow one geographic pattern. Some are strongly differentiated between regions and others are not. This mosaic is the reason that perceived resemblance across groups depends so heavily on which traits an observer attends to.

Table 9. Different visible traits follow different geographic patterns

Trait	Geographic pattern	Reason, as far as known	Effect on cross-region look-alikes
Skin pigmentation	Strongly differentiated between regions; tracks ultraviolet radiation	Natural selection linked to sunlight (Jablonski and Chaplin, 2000)	A strong regional signal; lowers cross-region matches when included
Craniofacial shape	Most variation within regions (Relethford, 2002)	Mostly neutral variation plus shared developmental architecture	Cross-region matches common
Nose shape	Modest correlation of some features with climate	Possible weak climate adaptation (Zaidi et al., 2017)	Small regional tendency
Hair form and color	Differentiated between regions	Genetic differences and some selection	Regional signal in appearance
Body build and stature	Varies within and between regions; strongly shaped by nutrition	Genes plus environment	Weak and context dependent

Source: H Heuristics synthesis of the cited studies. Descriptions are qualitative.

The mosaic has a practical meaning. Skin tone, hair form and eye shape are among the first things an observer sees, yet they account for only a small share of overall variation in facial form. As a result, two people may seem unlike each other on the traits that catch the eye while matching closely on the traits that define facial structure, such as the proportions of the jaw, brow and mid-face. The framework proposed here keeps these two layers apart.

5.5 The same outcome by different routes

A second pillar of the matrix is that similar traits have evolved more than once, in different populations, through different genetic routes. Three well-studied examples are shown below. In each case unrelated populations reached a similar result independently.

Table 10. Examples of convergent outcomes in different populations

Trait	Populations	What differed	Source
Light skin	Europeans and East Asians	Partly different genes underlie lighter skin in the two groups	Norton et al., 2007
Lactase persistence (digesting milk as an adult)	Europeans and several African pastoral groups	Different mutations in the same regulatory region, arising independently	Tishkoff et al., 2007
Living at high altitude	Tibetans and Andeans	Different physiological routes to coping with low oxygen	Beall, 2007

Full citations are in Appendix C.

Independent routes to similar outcomes are not a curiosity. They are a normal feature of human biology. If distant populations can converge on a similar skin tone, digestive ability or response to altitude, there is nothing odd about the idea that they could converge on similar facial configurations, which come from far more genetic ingredients. Many of those ingredients, such as the face-shape loci described in Section 4, are shared across the species. The three traits in the table are not facial traits, and they should not be mistaken for evidence of face resemblance. They show that convergence is possible and common, which makes cross-population look-alikes a reasonable expectation.

5.6 A two-layer picture of appearance

Putting these findings together gives a simple picture used in the rest of the report. Appearance has at least two layers.

- **A shape layer.** The geometry of the face and head: proportions, the arrangement of the eyes, nose, mouth and jaw. This layer is controlled by many small effects, mostly shared across populations. Variation is mainly within regions, and cross-region overlap is high.
- **A surface layer.** Skin tone, hair form and color, and some details of eye and nose shape. Several of these traits show strong regional differences, shaped partly by natural selection and ancestry.

Cross-population look-alikes are expected to be most common when resemblance is judged on the shape layer. When the surface layer is added, matches between distant regions become rarer, while matches between neighboring or mixed populations remain common. This is a prediction, and Section 6 turns it into a simulation.

5.7 A caution about whose data we have

Most large genetic studies have been done on people of European ancestry. In one count, the great majority of participants in genome-wide association studies were of European descent (Popejoy and Fullerton, 2016), and researchers have repeatedly called for broader sampling (Sirugo and colleagues, 2019). Methods built on one population can work worse on others (Martin and colleagues, 2019). For the matrix, this is a serious limit. We know the least about facial genetics in the regions where, according to the diversity argument above, the most variation may lie. Any claim about a *global* pattern is provisional until that gap closes.

5.8 Summary for the matrix

Table 11. What Section 5 means for the matrix

Finding	Consequence for the matrix	Grade
Humans share one origin and spread gradually	Neighboring populations resemble each other more than distant ones do; expect a gradient	A
Most variation lies within populations	Large overlap between regions in any single trait	A
Many small differences can still separate broad groups	The matrix is several grids, one per trait set; more features lower cross-region matches	A
Visible traits form a mosaic	Keep a shape layer and a surface layer apart	A
Similar traits arise by different routes	Independent convergence on similar faces is plausible	B
Research samples are uneven	A global claim is provisional until all regions are sampled	A

Source: H Heuristics synthesis of Section 5. Grades follow Table 1.

06 THE MATRIX FORMALIZED

6. The Matrix Formalized

6.1 A definition

To turn the idea of a matrix into something testable, the report gives it a definition. Take *k* sampling populations, such as the regions in Section 2 or any finer set of communities. Pick a set of traits *T* to compare, such as facial shape alone or shape together with skin tone and hair form. Pick a tolerance *ε* that says how close two people must be on those traits to count as a match. The *match rate* between populations *i* and *j* is the probability that a randomly chosen person from *i* and a randomly chosen person from *j* match.

$$M_{ij}(T, \varepsilon) = \Pr\left[d_T(x, y) < \varepsilon \mid x \in P_i, y \in P_j \right] \quad (1)$$

Here *d* measures the distance between two people on the chosen traits. The matrix is the *k* by *k* table of these rates. Because populations differ in how varied they are, raw rates are hard to compare, so the report divides each row by its within-population rate, called the *recurrence ratio*.

$$R_{ij}(T, \varepsilon) = M_{ij}(T, \varepsilon) \; / \; M_{ii}(T, \varepsilon) \quad (2)$$

A recurrence ratio of 1 means that a person in *i* is as likely to find a look-alike in *j* as in their own population. A ratio near 0 means look-alikes across the two populations are essentially absent. The average of the off-diagonal ratios is a single summary of how connected the world is.

Table 12. Terms used to define the matrix

Term	Meaning	Plain-language gloss
Trait set (T)	The traits compared: facial shape, proportions, surface traits, or all together	What counts as looking alike
Tolerance (ε)	How close two people must be on every trait to count as a match	How strict the judge is
Match rate (M)	Chance that a random pair, one from each population, match	How often a match turns up
Recurrence ratio (R)	Cross-population match rate divided by the within-population rate	How cross-group resemblance compares to resemblance at home
Connectivity	Average of all off-diagonal recurrence ratios	How linked the world's populations are
Variant reach	Number of regions in which a given variant neighborhood has members	How widely one recurring look is found

Source: H Heuristics. These are proposed working definitions.

6.2 What determines each cell

Three things set the value of any cell in the matrix. Understanding them is the key to reading the results without being misled.

- **The trait set.** Comparing facial shape gives a different matrix from comparing shape together with skin tone, hair form and eye shape. Section 5 predicts that the first is more uniform than the second.
- **The tolerance.** A loose tolerance produces many matches in almost every cell. A tight one produces few. Results should always be reported across a range of tolerances, not at a single convenient cutoff.
- **Population structure.** Shared ancestry raises match rates between populations that are close on the map. Admixture links populations that would otherwise be far apart. Both are real effects that the matrix is designed to show, not to remove.

6.3 A simulated world

To see how the matrix behaves, the report builds a small imaginary world. Seven populations stand for the seven regions and sit at positions on a map. Each person is described by 14 facial-shape features, correlated in pairs to mimic developmental modules, and by three surface traits standing for skin tone, hair form and eye shape. Differences between population averages grow with distance on the map, as in real isolation by distance. Shape features differ only slightly between regions. Surface traits differ strongly. One population, standing for South America, is built as a mixture of three source populations. The numbers in this model were chosen to demonstrate how the logic works. They are not measurements and should not be read as describing any real population.

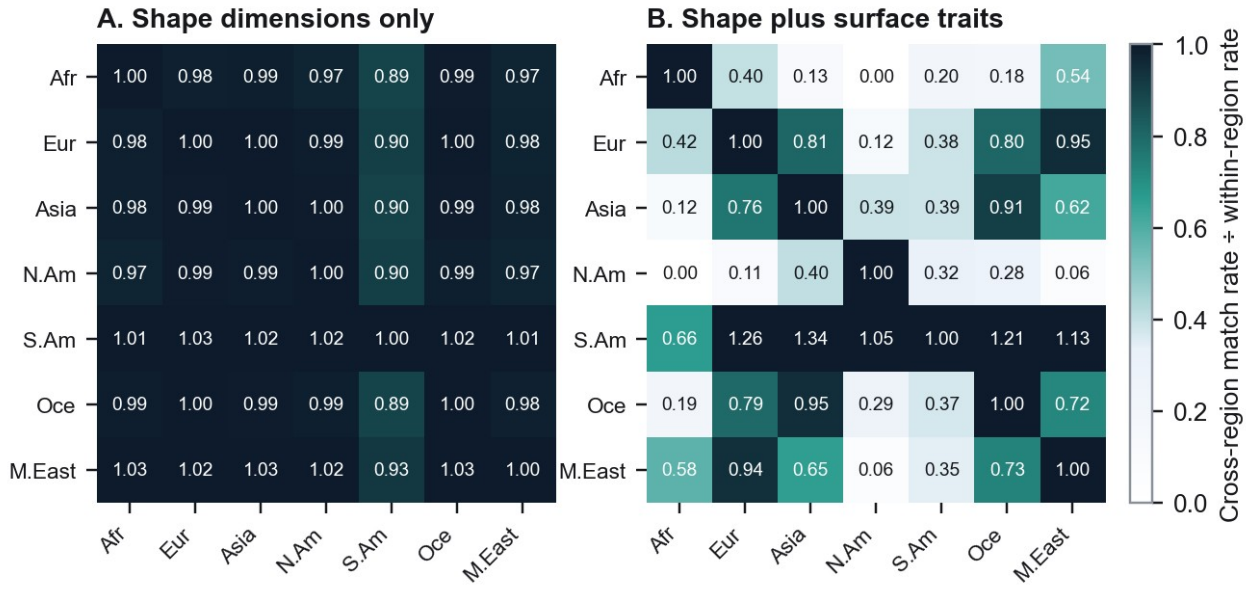

Figure 5. A simulated global variant matrix. Each cell shows the recurrence ratio: the cross-region match rate divided by the within-region rate for the row population. (A) Comparing facial-shape features only, almost every cell is close to 1. (B) Adding strongly regional surface traits makes distant regions rarer in each other's match lists, while neighbors and the mixed-ancestry population stay linked. Afr, Africa; Eur, Europe; N.Am, North America; S.Am, South America; Oce, Oceania; M.East, Middle East.

Source: H Heuristics. Illustrative simulation with invented parameters; not a measurement of real populations.

The results fall out of the assumptions, which is the point of a teaching model. With shape features alone, the average off-diagonal recurrence ratio is 0.98, with a range from 0.89 to 1.03. Cross-region look-alikes are as easy to find as local ones. When surface traits are added, the average falls to 0.54 and the range widens from nearly zero to above one. A pair of distant regions, such as the simulated Africa and Asia, drops from 0.99 to 0.13. Neighbors stay linked: the simulated Europe and Middle East pair is at 0.95. And the mixed population has ratios above 1 for several regions, such as 1.26 for the simulated Europe. A value above 1 is not a mistake. A mixed population is more varied than its parents, so its members have fewer look-alikes at home than the parent populations have in them. Admixed populations act as bridges, a point developed in Section 7.

6.4 Reading the matrix: four patterns

Real data could take many forms. Four broad patterns would mean different things for the thesis.

Table 13. Four patterns the matrix could show and what each would mean

Pattern	What it looks like	Meaning	Effect on the thesis
Uniform	All cells close to 1	Resemblance is spread evenly across regions	Strongly supports it
Gradient	Ratios fall smoothly with distance on the map	Shared ancestry matters but nothing is isolated	Supports it, with a geographic layer
Bridged	Gradient, plus high cells for admixed populations	Migration and mixing connect distant regions	Supports it and shows how
Block	High on the diagonal, near zero elsewhere	Look-alikes are confined to regions	Refutes it

Source: H Heuristics. Block-diagonal results would falsify the central thesis (Section 12).

The report's prediction is that real data will look like a *bridged gradient* for the full trait set and be close to *uniform* for facial shape alone. That is a specific, testable claim. If shape alone also shows a block pattern, the thesis fails.

6.5 From matrix to variant

A matrix of regions answers a coarse question. The finer question is whether *particular variants*, such as the neighborhood around a given reference person, are found in many regions. The report calls the number of regions in which a variant neighborhood has members its *reach*. A variant with a reach of one is local. A variant with a reach of six or seven is global. The catalogue proposed in Section 10 records the reach of each variant, along with how it was measured and how strong the evidence is.

6.6 Why two people look alike: the hierarchy of explanation

A high cell in the matrix does not say *why* two people match. The Human Variant Hypothesis separates the causes of resemblance into channels: shared descent (kinship), independent arrival at similar trait values (convergent configuration), similar development and environment, presentation (hair, grooming, pose, photography), and perception and sampling. The matrix adds a sixth, *ancestry-associated resemblance*, the baseline tendency of populations with shared history to resemble each other more. Figure 6 shows a hypothesized split for five kinds of pair, including the cross-region look-alikes at the heart of this report.

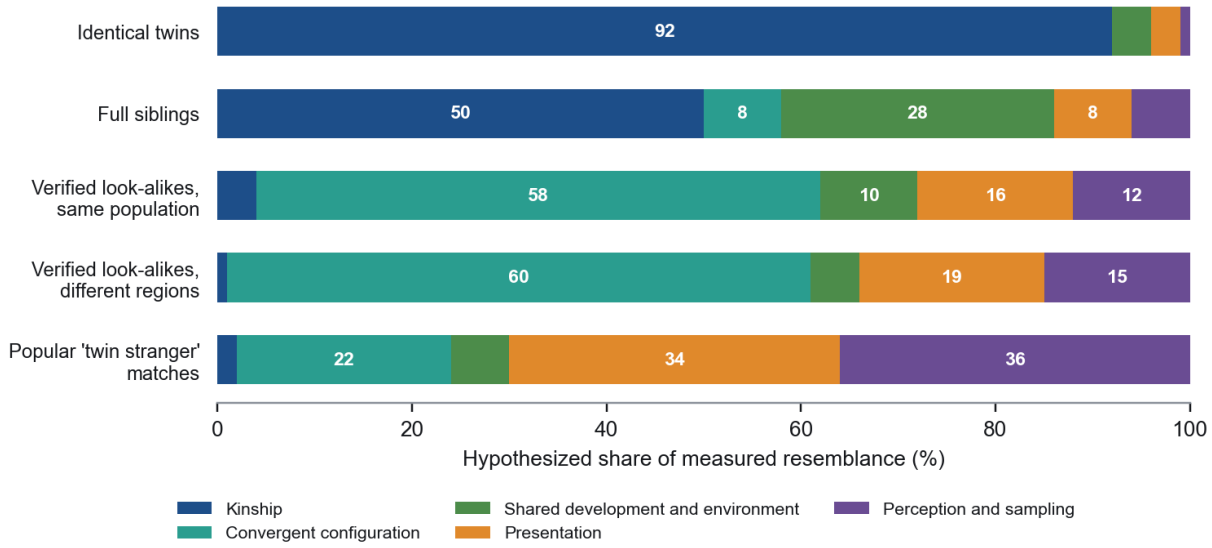

Figure 6. A hypothesized decomposition of resemblance. Shares are conceptual estimates for illustration, not measurements. Identical twins resemble each other almost entirely through shared genes. Siblings resemble each other through a mix of kinship and shared environment. Verified unrelated look-alikes, within or across regions, are expected to resemble each other mainly through convergent configuration, while popular matches found online are expected to owe more to styling, perception and selective discovery.

Source: H Heuristics. Hypothesized shares for illustration; extends the decomposition in the Human Variant Hypothesis report.

The figure expresses a simple expectation. The more rigorously a pair has been verified, the more its resemblance should come from convergent configuration. The more casually a pair was found, the more it should come from presentation and perception. That expectation is itself testable, and the studies in Section 11 are designed to test it.

6.7 Rules for reading a matrix result

1. **Always say which trait set.** A matrix of facial shape and a matrix of shape plus skin tone are different objects.
2. **Always say which tolerance.** Report results across several, not one.
3. **Check the diagonal first.** Within-population rates tell you how varied each population is.
4. **Look for pattern, not single cells.** One high cell may be chance; a gradient or a bridge is a pattern.
5. **Never turn a cell into a label.** A high or low ratio describes a sample, not a kind of person.

07 ADMIXTURE AND BRIDGE POPULATIONS

7. Admixture, Migration and Bridge Populations

7.1 Humans have always mixed

The idea of cleanly separate populations is a simplification. People have moved, traded, conquered and married across distances for as long as there is evidence. A large analysis of 95 populations across the world found signs of mixing between differently descended groups within the last few thousand years in most of them (Hellenthal and colleagues, 2014). Mixing is not an exception that needs special explanation. It is the normal condition of human populations.

For the matrix, this matters twice over. First, mixing is one of the processes that *links* regions: a population that carries ancestry from two or three regions will share traits with each of them. Second, mixing makes a study of resemblance harder to interpret, because a cross-region look-alike might be explained by recent shared ancestry rather than by independent convergence. The report keeps both possibilities in view.

7.2 How mixing shapes individual faces

When two populations mix, their descendants do not become a uniform blend. Each person inherits a patchwork of stretches of DNA from different ancestral sources, in proportions that differ from person to person, even among siblings. In a mixed population, one person may carry mostly one source, another mostly a second, and a third an even mix of three. The faces of mixed-ancestry people therefore show a great deal of individual variety. They also draw on trait variants from several source populations, which can produce combinations that are rare in any single parent population.

A real example comes from the United States. A study of more than 160,000 people found that, on average, African Americans carry about three-quarters African ancestry and about a quarter European ancestry. Latinos, on average, carry mostly European ancestry with a substantial Indigenous American share and a smaller African share. The averages hide wide differences: individuals and regions vary a great deal around them (Bryc and colleagues, 2015). Similar mosaics are found across Latin America and the Caribbean, where the mix of sources differs from country to country and region to region (Moreno-Estrada and colleagues, 2014).

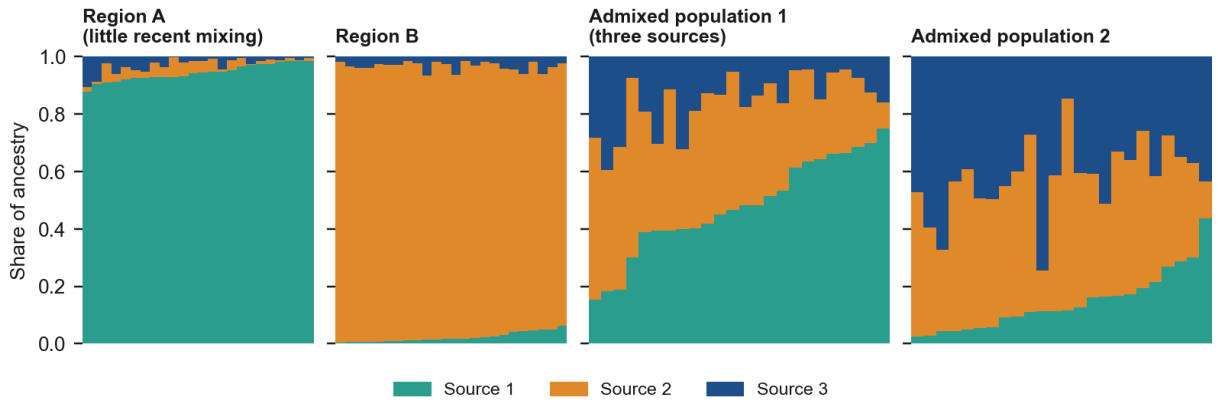

Figure 7. Ancestry mosaics in unmixed and mixed populations. Each thin bar is one person, and the colors show what share of that person's ancestry comes from each of three source populations. In the two unmixed regions, nearly everyone carries one source. In the mixed populations, people carry different shares of two or three sources, so individual faces draw on a range of inherited variants.

Source: H Heuristics. Illustrative simulation; proportions are invented to show the pattern.

7.3 Bridge populations

The report uses the term *bridge population* for a group that carries ancestry from two or more regions and is therefore expected to show recurrence with each of them. Several such populations are well documented.

Table 14. Examples of bridge populations and their ancestral sources

Population or region	Ancestral sources involved	Why it is a bridge in the matrix
Latin America	Indigenous American, European and African sources in varying proportions	Links the Americas, Europe and Africa within one population
African Americans and Caribbean populations	Mainly West and Central African ancestry, with European and some Indigenous American ancestry	Links Africa, Europe and the Americas
Middle East and North Africa	Near Eastern, African, European and Asian sources over thousands of years	Sits between three continents
Central Asia	Western and Eastern Eurasian ancestry in varying proportions	Links Europe, the Middle East and East Asia
Horn of Africa	African ancestry with a substantial West Eurasian component of ancient date (Pickrell et al., 2014)	Links Africa and the Middle East
Pacific Islands	Ancestry related to Asian populations and to Papuan populations (Skoglund et al., 2016)	Links Asia and Oceania
South Asia	Long-standing mixtures of several ancestral groups (Reich et al., 2009)	Links West Eurasia, East Asia and local sources

Descriptions are summaries of the cited work and omit much detail. Full citations are in Appendix C.

Bridge populations are where the matrix is most likely to show the connections the thesis predicts. If look-alikes cross regions anywhere, it should be here. That is not a trivial result, since it would show that the human appearance space is connected by real, recent ancestry. But it is also not the strongest form of the thesis. The strongest form says that look-alikes cross regions even between populations with *no* recent mixing, such as unmixed groups in distant parts of the world.

7.4 Two routes to cross-region resemblance

A cross-region match can arise in two quite different ways, and the matrix has to tell them apart.

Table 15. Two routes to cross-region look-alikes

	Route 1: Shared recent ancestry	Route 2: Independent convergence
How it works	Two people share trait variants because their populations exchanged ancestry in the past few thousand years	Two people reach similar trait values through different genetic combinations, with no shared recent ancestry
Where most likely	Neighbors, bridge populations, diaspora communities	Distant, unmixed populations
What it would show	The matrix is connected by gene flow	The matrix reflects deep structure in how faces are built
How to test	Estimate local ancestry at face-shape genes; compare with ancestry-matched control pairs	Look for matches in pairs with no detectable shared ancestry; test for excess sharing of face-associated variants

Source: H Heuristics. Both routes are consistent with the thesis; the second is the stronger claim.

Both routes would count as support for a global matrix, but they would mean different things. Route 1 would show that the world's appearances are linked through the long history of human contact. Route 2 would show something deeper: that the structure of the human face channels independent populations toward similar configurations. The tests are different, and the research design in Section 11 includes both.

7.5 Ancestry, residence and diaspora

Two more cautions apply. First, a person's face cannot be used to read off their ancestry. Ancestry proportions are statistical summaries of inherited DNA and do not map neatly onto how a face looks. Two siblings can differ visibly while sharing the same average ancestry. Second, people move. Large diaspora communities now live far from where their families began, and a photograph may show a person whose ancestry and residence are on different continents. The sampling rules in Section 2 therefore record ancestry and place of residence as separate facts.

> **Reminder.** A mixed-ancestry person is not a "blend" of two types. Each person is an individual mosaic of inherited segments, and no one carries a single pure type.

7.6 Predictions for bridge populations

The bridge idea is easy to state and easy to test. If it is right, four things should hold.

- **Wider reach.** Variants found in a bridge population should have members in more regions than variants found only in unmixed populations.
- **Gradient by ancestry.** Within a bridge population, a person's matches should lean toward the regions that supply more of their ancestry, although individual faces will not follow average proportions.
- **Weaker signal in unmixed pairs.** Between distant, unmixed populations, recurrence should be lower for the full trait set and close to the within-population rate for facial shape alone.
- **No special face for mixing.** Mixed-ancestry people should not form a distinct type. They should be as varied as the sources allow.

08 PERCEPTION AND CULTURE

8. Perception and Culture: Who Is Looking?

8.1 Resemblance is judged, not just measured

A look-alike is a judgment made by a perceiver, human or machine. Two faces can be measured as geometrically different and still seem the same, or the reverse. This section looks at how perception shapes what we count as a variant, and at why the observer's background has to be part of the study design.

Human faces are processed by specialized brain systems. In monkeys, recordings from face-selective brain cells suggest that facial identity may be coded along roughly fifty dimensions of shape and appearance (Chang and Tsao, 2017). The brain therefore does not store every detail. It compresses a face into a limited description, and two faces that match on that description feel alike.

8.2 The other-group effect

People are generally better at recognizing and telling apart faces from groups they see often. A large analysis of studies spanning decades found a reliable *own-race bias* in memory for faces: people recognized faces of their own group more accurately, and were more likely to wrongly think they had seen a new face from another group (Meissner and Brigham, 2001). Researchers explain this partly through experience. Face-space theory suggests that familiar faces are represented with more detail, while unfamiliar faces are represented more coarsely (Valentine, 1991).

The effect matters for the matrix because it changes the number of look-alikes that an observer perceives. A finer mental resolution means fewer faces count as matches. A coarser one means more do. Figure 8 shows the logic with the same counting model as Section 4.

This produces an unusual prediction. An observer will tend to find *more* look-alikes among groups they know less well, and *fewer* among groups they know well. Reports of cross-region look-alikes may therefore be partly inflated by coarse perception, and reports of look-alikes within one's own group may be undercounted. This is a reason to use many raters from many backgrounds, and to record each rater's background.

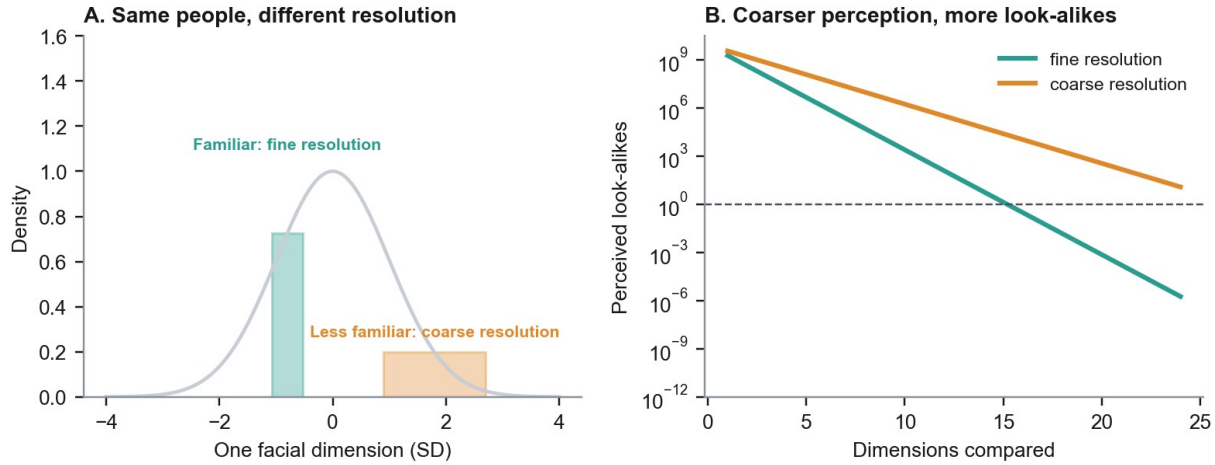

Figure 8. Same people, different resolution. (A) An observer with fine resolution for familiar faces treats only very close neighbors as matches. With coarse resolution for less familiar faces, the same tolerance covers a wider band, so more people look alike. (B) In the counting model, coarser perception raises the expected number of perceived look-alikes by orders of magnitude at every number of dimensions.

Source: H Heuristics. Illustrative model using the counting approach of Section 4; tolerances of 0.4 and 0.8 standard deviations.

8.3 Matching unfamiliar faces is hard for everyone

Even the basic job of deciding whether two photos show the same unfamiliar person is difficult. Different photographs of one person can be judged to show different people, and photographs of different people can be judged to show one (Jenkins and colleagues, 2011; Hancock and colleagues, 2000). People differ widely in the skill. So-called super-recognizers perform far above average (Russell and colleagues, 2009), and the best results in one comparison came from teams of trained examiners working with algorithms (Phillips and colleagues, 2018). Ordinary judgments of resemblance, including the viral "I found my twin" post, therefore have large error.

8.4 Machines see differently, and unevenly

Algorithms compress faces too. Early methods represented a face with a few dozen numbers, and modern deep-learning systems map each face to a vector of a few hundred numbers. These systems can be accurate overall and still behave unevenly across groups. A large government test of face-recognition algorithms found that false matches between different people occurred at rates that differed across demographic groups, in many cases by factors of ten to one hundred (Grother and colleagues, 2019). Similar disparities in commercial systems were documented in an earlier study of gender classification (Buolamwini and Gebru, 2018).

This has a direct implication for any study that uses algorithms to find look-alikes. An algorithm that was trained mostly on faces from one region may under-resolve faces from others, and so produce more false look-alikes there. The algorithm's score is a measure of *its* perception and cannot be treated as the truth about faces. Section 11 requires several algorithms, tested for fairness across regions, to be used alongside geometry and human raters.

8.5 Culture shapes what a look is

Culture reaches the face in many ways that have nothing to do with genes. Hairstyle, facial hair, grooming, makeup, eyewear, clothing, posture and expression all differ between societies and across time, and all affect whether two people seem to look alike. Beauty ideals shape which features are emphasized. Photographic conventions and filters shape how faces appear online. Two people from different continents who adopt the same haircut, glasses and expression will look more alike than they would otherwise, and this resemblance is real in perception but not in biology.

People also tend to judge faces on social dimensions. They form first impressions about trustworthiness, competence and other traits within a fraction of a second, but those impressions are not accurate guides to character (Todorov and colleagues, 2015). This is another reason for the boundary that Section 13 sets: a resemblance to a well-known person may bring along assumptions about personality that have no basis.

Finally, people use social categories to sort faces quickly. It is plausible that this habit causes observers to under-notice resemblances that cross a familiar category line, and to over-notice those that stay within it. The idea is untested (Grade D), but it is easy to test, and a design with raters from many backgrounds could do so.

8.6 Design implications

Table 16. How perception problems affect the matrix and what to do about them

Problem	Effect on the matrix	Remedy in the research design
Other-group effect	Inflates perceived look-alikes in less familiar groups; may deflate them in familiar ones	Raters from many backgrounds; record rater background; analyze by rater group
Unfamiliar-face matching error	Adds noise to every rating	Many raters per pair; use of controls; super-recognizer panels
Algorithmic bias	Uneven false-match rates across regions	Several algorithms; test fairness by region; report all
Cultural styling	Adds resemblance that is not biological	Standardized images; neutral styling; compare with and without styling
Social categorization	May hide resemblances that cross category lines	Blind rating of unlabeled images; test whether raters group by label

Source: H Heuristics. Proposed design safeguards.

09 PRESENTATION AND VISIBILITY

9. Presentation, the Internet and the Visibility Problem

9.1 Presentation: the face we are shown

What we see in a photograph is not a bare face. It is a face with a hairstyle, a facial expression, a camera angle, lighting, a filter and a background. All of these are choices, and many of them follow fashion. Two strangers who happen to wear the same cut, the same beard and the same glasses can look strikingly alike even when their bone structure differs. In a simple simulation, adding a shared style lifts perceived similarity by a fixed amount, which pushes many pairs over the line that people use to call someone a look-alike (Figure 9).

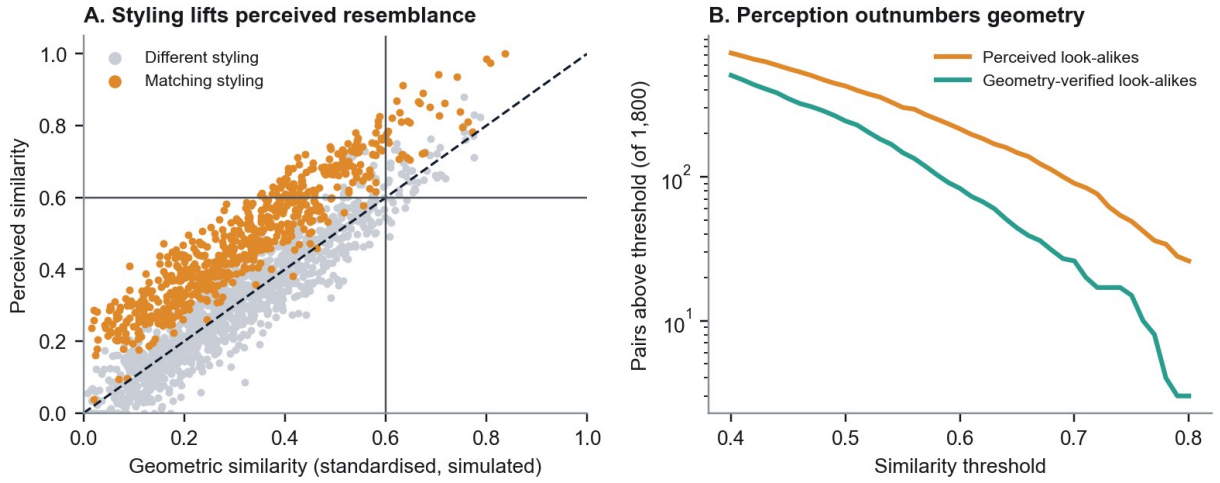

Figure 9. Styling can turn an ordinary pair into a perceived look-alike. (A) In a simulation of 1,800 random pairs, pairs that share a hairstyle, grooming or pose (orange) are perceived as more alike than their underlying geometry would predict. (B) At a similarity threshold of 0.6, 215 pairs count as perceived look-alikes but only 83 pass the same threshold on geometry alone.

Source: H Heuristics. Illustrative simulation; the size of the styling effect is assumed, not measured.

The simulation is a teaching model, and the numbers are not real. But the logic is general: whenever people judge resemblance from ordinary photos, perceived look-alikes outnumber geometric ones. This is *perceptual amplification*, a term from the Human Variant Hypothesis, and any matrix built from casual photographs will contain it.

9.2 The internet as a variant amplifier

The internet changes the scale and the filter. A person can now compare their face with billions of images, and applications and communities exist to find "twin strangers." This is a huge advantage as a source of leads, because it is the first time in history that anyone can search so many faces. It is also a powerful source of distortion.

- **Algorithms choose what we see.** Recommendation systems rank content by predicted engagement. The faces shown to a user are not a random sample of humanity.
- **Exposure makes a look feel common.** A face seen again and again becomes familiar, and familiarity makes the next resembling face stand out.
- **Posting is selective.** People share flattering, posed and edited images, and viral look-alike stories are those that someone chose to post.
- **Failures are invisible.** We see the posts where a match was found. We do not see the many comparisons where nothing matched.

The internet is therefore best used as an engine for generating hypotheses. It cannot, by itself, serve as a representative population survey.

9.3 Attention is concentrated

The key fact about online visibility is that it is very unequal. A few people receive enormous attention and most receive almost none. Such patterns are often described by a *Pareto distribution*, commonly summarized as the 80/20 rule, in which roughly 80 percent of an outcome comes from 20 percent of the sources. Heavy-tailed patterns of this kind are common in social and technical systems (Clauset and colleagues, 2009). One standard explanation is *preferential attachment*, in which things that are already popular become more popular because newcomers connect to them (Barabási and Albert, 1999). Experiments on artificial music markets showed that when people can see what others chose, inequality of popularity rises and which items succeed becomes hard to predict (Salganik and colleagues, 2006).

The Pareto pattern is not a universal law, and there is no reason to assume that attention follows an exact 80/20 split. The report therefore treats the shape of the curve as something to be measured. A simulation with a heavy-tailed distribution is shown in Figure 10.

Panel B of the figure below is the important result. In the model, every pair of people is equally likely to be a good match. Yet more than half of the matches that people *notice* involve a very visible person, simply because very visible people are compared with far more observers. This is *discovery bias*. It means that when most reported look-alikes involve famous people, the cause may be where attention is directed, not where resemblance lies.

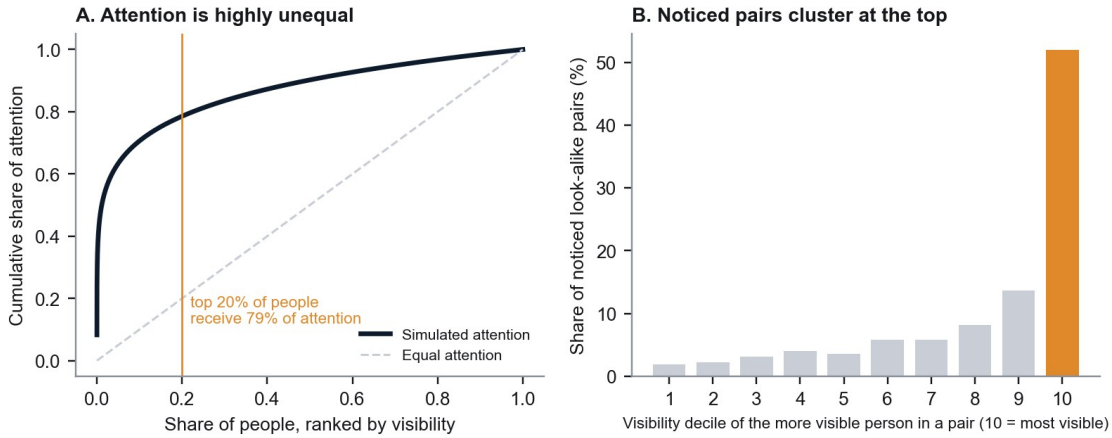

Figure 10. Attention is concentrated, and discovery follows attention. (A) In a simulated group of 100,000 people with Pareto-distributed visibility, the top 20 percent of people receive 79 percent of attention, and the top 1 percent receive 49 percent. (B) If look-alike pairs are noticed in proportion to how visible the more visible person is, then the most visible tenth of pairs accounts for 52 percent of all noticed pairs, and the less visible half accounts for only 15 percent, although every pair in the simulation was equally likely to be a look-alike.

Source: H Heuristics. Illustrative simulation: Pareto shape 1.16; discovery probability proportional to visibility, capped.

9.4 Alpha variants

The H Heuristics essay on the variant phenomenon proposed a provisional term for this pattern. An *alpha variant* is a variant archetype that is disproportionately represented among highly visible people. The word *alpha* describes prominence in the attention environment, not biological superiority. It is a hypothesis about visibility, not a ranking of human value.

The idea separates three claims that are easy to blur together.

Table 17. Three different meanings of "common"

Claim	What it says	How it would be tested
Common in the population	The variant is frequent among ordinary people in a region or worldwide	Random or population-based samples with standard imaging
Common among public figures	The variant is more frequent among prominent people than among a matched comparison group	Matched samples of public figures and less visible people, controlling for age, geography and occupation
Common in what an observer sees	The variant fills the content that reaches a particular person	Audit of the observer's feeds, searches and surroundings

Source: H Heuristics, after the essay on the Human Variant (or Doppelgänger) Phenomenon.

A variant could be common on screen without being common in the street, or common in the street without appearing on screen. Prominence can come from many sources: casting conventions, beauty standards, fashion cycles, professional networks, access to resources and platform recommendation. A genuine alpha variant would be one that is more common among prominent people than among a suitable comparison group *after* these factors are accounted for. That test has not been done, so the idea is graded D.

9.5 The feedback loop

Attention may also feed itself. A person becomes prominent. Their look becomes familiar. The familiarity makes people notice others who resemble them. Those people are shared and compared, and the archetype becomes more memorable still. A Pareto-like pattern may describe the distribution of attention far better than it describes the distribution of faces. Whether a feedback loop of this kind exists for appearance is an open question, and one the research design in Section 11 addresses with a dedicated study.

9.6 What this means for the matrix

Visibility is a serious threat to the matrix, because the data that exist about look-alikes come mostly from highly visible people and the places where they are most photographed. The report adopts four safeguards.

1. **Record the discovery route.** Every pair or set in the catalogue notes how it was found: a public appeal, a photography project, an algorithm applied to a consented dataset, a viral post, a celebrity comparison.
2. **Use visibility-independent samples as the core.** The matrix is estimated from population-based samples, with discovery-based samples used only to propose hypotheses.
3. **Treat visibility as a covariate.** Measure how widely each person is photographed or known, and test whether it explains the pattern.
4. **Report the denominator.** State how many comparisons were made as well as how many matches were found.

9.7 Summary

The internet is the best source of leads about look-alikes that has ever existed and the worst source of rates. Styling lifts perceived resemblance above geometric resemblance, and attention concentrates on a few faces, so that most of the pairs we notice involve the most visible people. Neither effect says anything about how common variants are in the population. The matrix therefore rests on population-based samples, and uses the internet to generate hypotheses and to study visibility itself.

> **Key point.** Noticing a look-alike is not evidence that look-alikes are common. It is evidence that someone looked.

10 ARCHETYPES AND CATALOGUING

10. Variant Archetypes and the Path to a Catalogue

10.1 Why a catalogue?

The matrix describes regions. The variants themselves are the units that move through it. To study variants across the world, researchers need agreed descriptions that can be checked, repeated and criticized. A *variant catalogue* is a register of such descriptions. Each entry says what the variant is, how it was measured, where it has been found, how strong the evidence is and what the limits are. The H Heuristics essay on the phenomenon called for a more disciplined variant catalogue as the natural next step. This section describes what one would contain.

10.2 Four working classes of variant

The Human Variant Hypothesis proposed four working classes, which this report adopts and extends with a measure of global reach.

Table 18. Working classes of human variant

Class	What recurs	Main measurement	Main confounders
Morphological	Broad facial geometry: jaw, cheekbones, face length, feature placement	Aligned 3D landmarks or dense surface distances	Age, body composition, expression
Proportional	Relative proportions and overall structure, regardless of size	Scale-free ratios and angles	Choice of measure, growth stage
Composite	A distinctive combination of face, hair and body traits	Geometry plus appearance and perceptual ratings	Hairstyle, grooming, photography, rater bias
Ancestry-associated	Statistical tendencies shaped by shared population history	Genetic ancestry estimates as covariates	Sampling design, social categories

Source: Human Variant Hypothesis report (H Heuristics, 2026), Table 3, with the matrix's use added.

The *composite* class is what most people mean by a doppelgänger, and it is the class most affected by styling and photography. The *morphological* and *proportional* classes are the best candidates for a global shape-layer matrix. The *ancestry-associated* class is not a type of variant at all. It is a baseline, the statistical tendency of populations with shared history to resemble each other, against which the other classes are measured.

10.3 Variants are neighborhoods, not boxes

A variant has a center, a *reference point*, and a neighborhood around it. People near the reference point resemble it to a high degree. People farther out resemble it less and merge into the neighborhoods of other reference points. There is no edge, and the neighborhoods of different variants overlap. Figure 11 shows the picture in two dimensions.

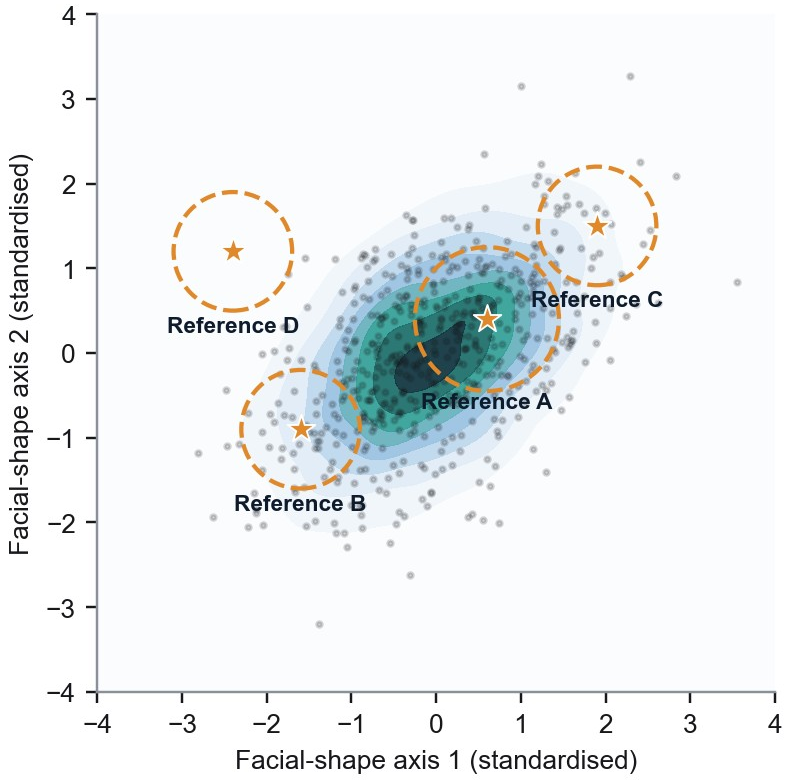

Figure 11. Variants as neighborhoods in a continuous space. The blue shading shows the density of people in a simulated two-dimensional space of facial-shape features, with most people near the center. Each orange star is a reference point, and the dashed circle shows its neighborhood. Neighborhoods near the center contain many people and overlap, while those at the edge contain few. There are no sharp boundaries between variants.
Source: H Heuristics. Illustrative simulation of 3,200 people in two standardized shape dimensions. Reference points are generic.

The picture also explains why variants near the center of the distribution are more widely shared. They sit where the density is highest, as the counting model of Section 4 predicts. A variant's reach in the matrix is a direct reflection of this density.

10.4 Naming and labeling rules

The names given to variants carry weight, and a careless name can slide into the typing of people that this report rejects. Four rules are proposed.

1. **Name by reference point or code, not by region or ancestry.** A variant is given an identifier, such as a code, or named after its reference point. It is never named after a continent, a nation or a racial label.
2. **Describe where a variant has been observed as a fact about a sample.** A variant card may say "observed in samples from five regions." It may not say "the variant of" a region.

3. **State that the label names a perceived configuration only.** In the words of the H Heuristics essay, labels give a name to a perceived configuration; they do not establish a shared ancestor, a unique variant gene or a fixed personality.
4. **Get consent before naming a living person as a reference point.** Authors may use themselves as references. Anyone else must agree.

The Hunter and Lars variants, already described in earlier H Heuristics work, would be the first entries in the catalogue, since they already have working descriptions. They would be treated like any other entry, subject to the same definitions, measures and grading.

10.5 The catalogue pipeline

Entries move through six steps from nomination to publication.

Figure 12. The catalogue pipeline. Candidate pairs and sets are nominated from many channels and then standardized, measured, rated blind, attributed to their likely causes, and finally entered in the catalogue with a grade and a statement of limits.
Source: H Heuristics. Proposed process.

Step 1, *nominate*, welcomes leads from everywhere, including the internet, but records the route. Step 2, *standardize*, creates comparable images or scans with neutral expression, restrained hair, no eyewear and controlled pose and lighting. Step 3, *measure*, produces geometric, proportional and modular measures. Step 4, *rate*, collects blind judgments from diverse human raters and from several algorithms. Step 5, *attribute*, asks how much of the resemblance is explained by kinship, ancestry, visibility, presentation and perception. Step 6, *catalogue*, writes up the entry.

10.6 The variant card

Each catalogue entry takes the form of a *variant card*, a short standard record. The fields below keep every entry comparable and make its weaknesses visible.

Table 19. Fields of a variant card

Field	What it records
Identifier and class	Code, working name, and which of the four classes applies
Reference definition	The traits that define the center of the neighborhood, written before examples were collected
Measurement method	Imaging protocol, landmarks or surface method, rating protocol and algorithms used

Field	What it records
Neighborhood tolerance	How close a person must be to count as a member, and results across several tolerances
Sample by region	Numbers of people examined and numbers matching, by region and subregion, with denominators
Reach and recurrence ratios	Number of regions with members, and recurrence ratios between regions
Discovery routes	How each member was found, and whether the variant is also found in visibility-independent samples
Attribution	Estimated shares of kinship, ancestry, convergence, presentation and perception
Evidence grade	A to D, with a plain-language reason
Limitations and ethics	Known weaknesses; consent status; the five boundaries restated

Source: H Heuristics. Proposed template.

10.7 A worked example of a variant card

To make the template concrete, the card below is filled in for an invented candidate. Every value is a placeholder chosen to show how the fields work. It does not describe a real variant or any real person.

Table 20. Illustrative variant card (placeholder values)

Field	Illustrative entry
Identifier and class	GVM-000 (example only); morphological
Reference definition	Pre-registered description of mid-face proportions and jaw width, written before any examples were collected
Measurement method	Standardized frontal images; 3D landmarks where available; blind ratings by raters from five regions; three algorithms
Neighborhood tolerance	Results shown at three tolerances; main result at the middle one
Sample by region	Example: 400 imaged per region in 7 regions; numbers matching and total comparisons listed for each
Reach and recurrence ratios	Example: members in 6 of 7 regions; recurrence ratios between 0.7 and 1.1 for shape only
Discovery routes	Population-based sample plus nominated pairs, each tagged by route
Attribution	Example shares of kinship, convergence, presentation and perception, with intervals
Evidence grade	C: plausible, awaiting replication
Limitations and ethics	Rater background effect noted; consent obtained; boundaries restated

Source: H Heuristics. Placeholder values for illustration only; no real data.

10.8 Retiring entries

A catalogue that only grows is not science. Entries that fail replication, that turn out to depend on styling, or that cannot be found in a population-based sample would be marked *retired* and kept in the record with the reason. The H Heuristics essay states that a null result would still be useful. A catalogue that records its failures is more trustworthy than one that records only its successes.

11 RESEARCH DESIGN

11. Research Design: How to Test the Matrix

11.1 Overview

A good idea earns its keep by surviving tests. This section proposes a programme of five linked studies, preceded by a preparatory phase, that together could confirm or reject the Global Variant Matrix. The studies are arranged so that the cheapest and least intrusive come first, and the expensive ones are only started if the early results justify them. Figure 13 shows an indicative 36-month plan.

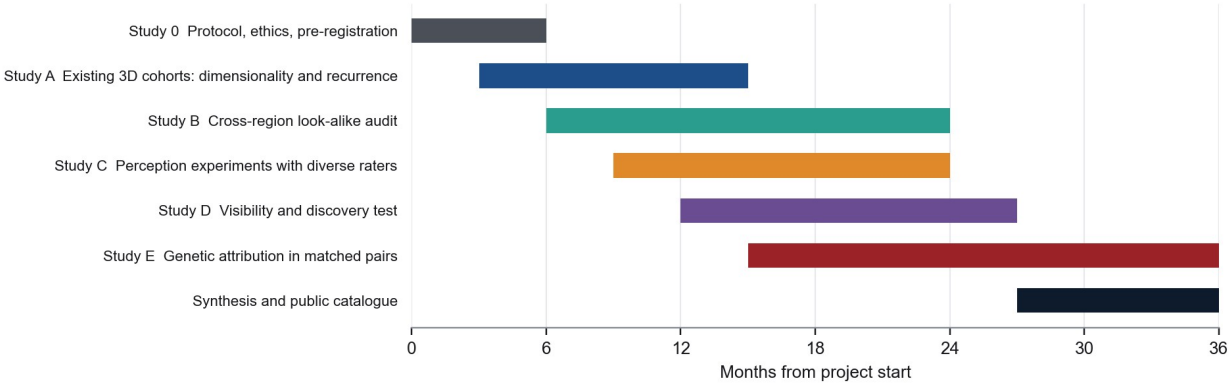

Figure 13. An indicative 36-month programme. Study 0 sets up protocols, ethics approval and pre-registration. Studies A, B, C and D overlap so that early results can steer later ones. Study E, the genetic attribution, starts once matched pairs exist. The final phase pulls the results into a public catalogue.
Source: H Heuristics. Indicative plan; timelines and sample sizes would be refined in a full proposal.

Table 21. The five studies at a glance

Study	Question	Main data	Would support the thesis	Would count against it
A	How many effective dimensions does facial shape have, and where do people crowd together?	Existing 3D face and genetic cohorts	Low effective dimensionality; crowding near the center in every population	Dimensionality near the number of traits
B	How often do people in one region have a look-alike in another?	Standardized images from all seven regions, with matched controls	Shape-only recurrence ratios near 1; bridged gradient overall	Block-diagonal matrix, even for shape
C	How does the observer change what counts as a look-alike?	Ratings by diverse humans and by several algorithms	Perception exceeds geometry, with a measurable observer effect	No observer dependence

Study	Question	Main data	Would support the thesis	Would count against it
D	Is visibility driving what we notice?	Discovery routes; public figures versus matched samples	Visibility explains noticed pairs better than population match rates	Visibility explains nothing
E	Do cross-region look-alikes share face-associated genetic variants?	Genotypes of matched look-alike and control pairs	Excess sharing beyond ancestry and kinship	No excess after adjustment

Source: H Heuristics. Proposed programme.

11.2 Study 0: protocol and governance

Before any data are collected, the programme writes down its definitions, metrics, thresholds and analysis plans and files them publicly (*pre-registration*). It sets up independent ethics review, a data protection plan and an advisory group with expertise in genetics, developmental biology, perception science, ethics and community representation. It also fixes the catalogue rules in Section 10, including the naming rules. This phase costs little and prevents most of the avoidable errors.

11.3 Study A: reanalysis of existing 3D cohorts

Research groups around the world have collected 3D face scans linked to genetic data for studies of normal facial variation and craniofacial disorders, and shared resources such as FaceBase show the value of well-governed craniofacial data. Study A would use such data, within the terms of consent, to estimate how many independent directions of variation facial shape really has, and where in the space people crowd together, separately for each population in the data. It requires no new recruitment. It tests the foundations of the thesis: whether the space of faces is as constrained, and as shared across populations, as the argument in Section 4 assumes.

11.4 Study B: the cross-region look-alike audit

Study B is the core test. It would collect standardized images of people from each of the seven regions, with sub-samples from within each region, by a protocol that does not depend on fame or on whether anyone thought the person looked like someone else. For every pair of regions, the study would compute the match rate under several trait sets and tolerances, and build the matrices of Section 6, with confidence intervals. Each match would be checked by geometry, by blind raters and by algorithms.

Sample size is the hard part, because look-alike pairs are rare by design. If a random pair qualifies as a match with probability p, then the number of people needed to find k matching pairs grows as the square root of $2k$ divided by p. Figure 14 shows how this works.

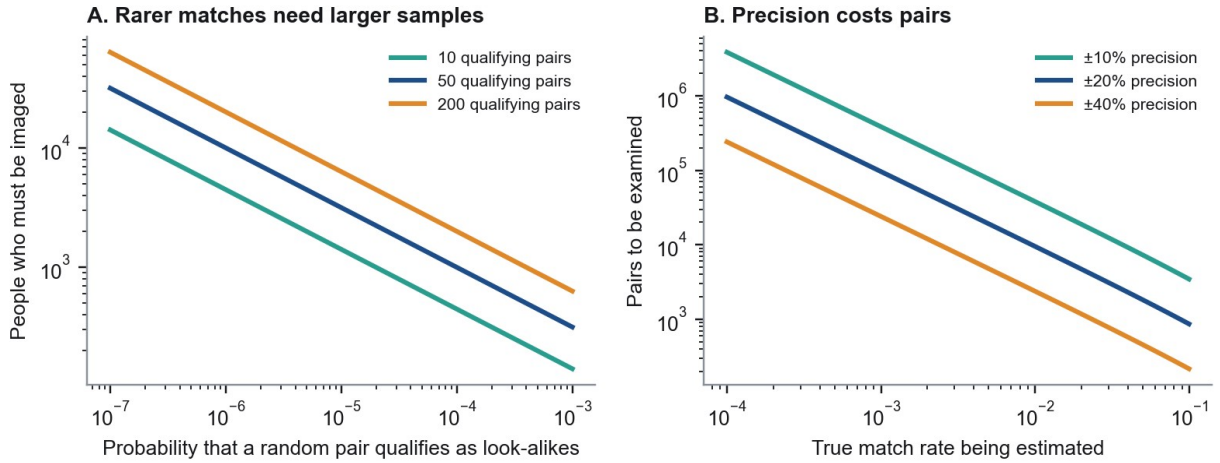

Figure 14. How many people and pairs are needed. (A) The number of people who must be imaged to find 10, 50 or 200 qualifying pairs, as a function of how rare a match is. Because the number of pairs grows with the square of the sample, large samples are efficient. (B) The number of pairs that must be examined to estimate a match rate to within 10, 20 or 40 percent at 95 percent confidence.

Source: H Heuristics. Calculation: people needed = sqrt(2k/p); pairs needed = 1.96^2 (1-p) / (p c^2) for relative precision c.

The figures are workable. If a qualifying match has a probability of one in a hundred thousand, about 3,200 people would give some 50 matching pairs. If the probability is one in a million, about 10,000 people would be needed. Estimating a match rate of one in a thousand to within 10 percent would take about 384,000 pairs, which a sample of fewer than 900 people supplies. Because an image sample compares every person with every other, even modest samples contain a very large number of pairs. The cost lies in standardized imaging across seven regions, not in the arithmetic.

11.5 Study C: perception experiments

Study C would show the same image pairs to raters from many regions and backgrounds and to several face-recognition algorithms, and would compare their judgments with the geometric measures. It would estimate the size of perceptual amplification, test the other-group effect from Section 8, and measure whether raters sort faces by social category when asked to rate resemblance. Each rater's background would be recorded, and raters would be unaware of the study's hypothesis. Because it needs no genetic data, Study C can start early and is relatively cheap.

11.6 Study D: visibility and discovery

Study D asks whether visibility explains what we notice. It has two parts. In the first, a registry of reported look-alike pairs records the discovery route of each one, so that the share arising from fame, virality, photography projects and algorithmic screening can be measured. In the second, public figures in selected occupations and countries are compared with a matched sample of less visible people to test whether any variant is over-represented among the visible, after adjusting for age, geography and occupation. This is the test of the alpha variant idea. Attention itself is measured, for example with search or view counts, instead of being assumed to follow a particular curve.

11.7 Study E: genetic attribution

Study E asks why cross-region matches occur. For verified look-alike pairs and matched control pairs, it would estimate genome-wide relatedness, ancestry similarity and similarity in genetic scores for facial shape. Mixed statistical models would then test whether look-alikes share more face-associated variation than controls after adjusting for kinship and ancestry. Local ancestry at face-shape genes would help to separate Route 1, shared recent ancestry, from Route 2, independent convergence (Section 7). Because the 2022 study that first reported this signal drew mostly on one ancestry group, Study E must sample widely. It is also the study with the heaviest ethical load, and it would go ahead only with explicit consent from both members of every pair.

11.8 Risks to the programme

Table 22. Main risks to the programme and how to manage them

Risk	Why it matters	Management
Unrepresentative samples	A global claim needs data from every region	Staged sampling; partnerships with regional institutions; coverage worksheet (Appendix B)
Measurement differences between sites	Different cameras or protocols create false differences	One imaging protocol; calibration charts; repeat images
Over-interpretation	A grid can look more orderly than the data justify	Pre-registration; null models; blind rating; report all results
Misuse of findings	Results could be read as a map of races or used for surveillance	Naming rules; five boundaries; controlled data access
Participant harm or loss of privacy	Faces and genomes are sensitive and identifying	Specific consent; consent from both pair members; no re-identification
Null result	The matrix may not exist	Treat as a finding; publish with equal prominence

Source: H Heuristics.

11.9 Quality and governance

- **Pre-registration.** Hypotheses, metrics, thresholds and analysis plans are filed before data access.
- **Replication.** Every primary finding is replicated in at least one held-out sample.
- **Sensitivity analyses.** Results are shown across tolerance levels, imaging methods and null models, not at one convenient setting.
- **Open methods.** Analysis code and summary statistics are released. Individual-level images and genomes stay in controlled-access storage.
- **Diverse sampling.** Sampling must reach every region. A global claim cannot be tested with a one-region sample.

12 PREDICTIONS AND FALSIFICATION

12. Predictions, Falsification and Alternative Explanations

12.1 What the thesis predicts

A hypothesis that can explain every possible result explains nothing. The Global Variant Matrix makes specific claims that could fail. The table lists eight predictions, tied to the studies in Section 11.

Table 23. Eight predictions of the Global Variant Matrix

ID	Prediction	Study	Would disconfirm
M1	With facial shape alone, recurrence ratios between regions are close to 1 for most pairs of regions	B	A block-diagonal matrix for shape
M2	Adding strongly regional surface traits lowers recurrence between distant regions more than between neighbors	B	No change, or the reverse
M3	Admixed populations show higher cross-region recurrence than unmixed populations	B, E	No difference
M4	Variants near the center of the distribution have wider reach than variants at the edge	A, B	No relation between density and reach
M5	Unrelated cross-region look-alikes share more face-associated variants than matched controls, after adjustment for ancestry	E	No excess sharing
M6	Perceived resemblance exceeds geometric resemblance, and the excess depends on rater background	C	Perfect agreement with no observer effect
M7	Visibility predicts which pairs are noticed but not population-based match rates	D	Visibility has no effect on discovery
M8	If alpha variants exist, they are over-represented among prominent people relative to a matched comparison group	D	No over-representation

Thresholds such as "close to 1" must be pre-registered; they are conventions, not natural boundaries.

12.2 What would show the thesis is wrong

The thesis would be seriously weakened by any of the following.

- **A block-diagonal matrix for facial shape.** If look-alikes are confined within regions even when only shape is compared, the central claim fails.
- **High effective dimensionality.** If facial shape turns out to have nearly as many independent dimensions as measured traits, then cross-population look-alikes are too rare to form a matrix.

- **No genetic excess in matched pairs.** If cross-region look-alikes share no more face-associated variants than controls, their resemblance is better explained by presentation and perception.
- **Complete dependence on styling.** If resemblance disappears when images are standardized, the matrix is a matrix of fashion.
- **No population-based signal.** If cross-region matches turn up only in discovery-based samples and not in population-based ones, they are a product of selection.

Any of these would be a result worth publishing. The H Heuristics essay makes the same point: a null result would still be useful.

12.3 Alternative explanations

Table 24. Alternative explanations and how the framework tells them apart

Alternative	Claim	Distinguishing test
Pure chance	Cross-region look-alikes are simply the expected tail of very large populations	Compare recurrence with an independent-trait null; the framework predicts more recurrence than the null for constrained traits
Hidden shared ancestry	Matches are undetected relatives or recent shared ancestors	Estimate genome-wide relatedness; restrict to pairs with no detectable recent kinship
Population structure	Matches reflect ancestry clusters only	Ancestry-matched controls and null models within ancestry strata
Presentation	Resemblance comes from hair, grooming, pose and photography	Standardized imaging; measure the drop when styling is removed
Perceptual artifact	Resemblance is in the eye of the beholder	Compare geometry with ratings from diverse raters and algorithms
Selective discovery	Only striking or famous pairs are reported	Record the discovery route; use population-based samples
Algorithmic bias	Machine scores reflect training data, not faces	Use several algorithms; test fairness by region; always compare with geometry

Source: H Heuristics; extends the alternatives table of the Human Variant Hypothesis report.

12.4 Limitations

Several limits would remain even in the best-designed programme.

- **Measurement dependence.** Every resemblance metric reflects choices about landmarks, alignment, scale and weighting. Reporting several metrics helps but no measure is neutral.
- **Coverage.** Facial genetics is far better known for some populations than others. Findings in one region may not generalize.
- **Age.** Faces change over a lifetime. Two people may look alike at 25 and less so at 60. Age must be treated as part of the measurement.
- **Causal inference.** Even a strong association between shared variants and resemblance does not show which variants cause it.
- **The risk of over-reading patterns.** The word *matrix* implies more order than may exist. Human minds find patterns even in random data. This is why pre-registration, null models and blind rating are central to the design.

12.5 The evidence map

Figure 15 gathers the report's claims by grade, from those that are firmly established to those that are speculative.

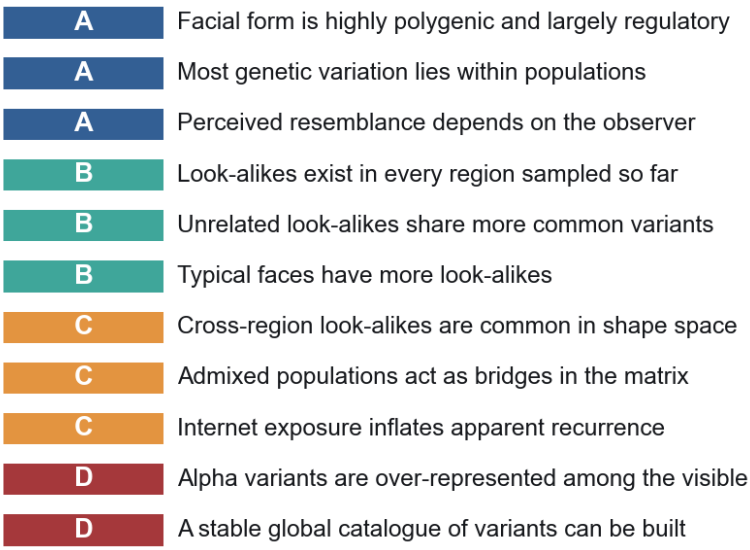

Figure 15. Evidence map. The three claims graded A are well established in the wider literature. The claims graded B have good but limited support. The cross-region, bridge and internet claims are plausible and graded C because they have not been directly tested. The alpha-variant and catalogue claims are speculative.
Source: H Heuristics. Grades as defined in Table 1.

> **Verdict so far.** The ingredients of the thesis are well supported. The thesis itself is a hypothesis awaiting its first global test.

12.6 How the verdict would update

The report commits in advance to how its own grade would change with the evidence.

Table 25. How results would change the grades

If the studies find	Then the thesis would be	Grade of cross-region claim
Shape-only recurrence ratios near 1, a bridged gradient for the full set, and excess genetic sharing	Supported, in its strong form	Rises to B or A
Shape-only ratios near 1 but no genetic excess	Supported as a pattern; the cause is open (perception or sampling)	Stays C
A gradient with low ratios between unmixed distant regions	Supported in a weak form: resemblance spreads mainly through contact	Stays C, with a narrower claim
A block-diagonal matrix even for shape	Rejected	Falls to D or is withdrawn

Source: H Heuristics. Thresholds to be pre-registered.

13 ETHICS AND BOUNDARIES

13. Ethics, Governance and the Boundary Against Typology

13.1 The historical shadow

Any project that sorts, compares or catalogues human faces walks in the footprints of earlier projects that did great harm. In the 1870s, Francis Galton pioneered composite photography, blending many portraits into one to reveal a supposedly typical face (Galton, 1879). Galton was also a founder of eugenics. Nineteenth-century physiognomy and criminal anthropology claimed to read character, criminality and worth from the shape of a face. Twentieth-century racial typology sorted people into supposed biological types and ranked them. These programs were scientifically wrong and caused profound suffering.

The danger has not gone away. Modern machine-learning work has revived attempts to infer criminality, sexuality or personality from facial photographs, which critics have described as physiognomy in algorithmic form (Agüera y Arcas and colleagues, 2017). Research on human variants touches the same raw material, faces and resemblance. It must therefore be built so that these errors are made impossible, not merely unlikely.

13.2 Five boundaries

The Human Variant Hypothesis sets out five boundaries. This report adopts them as binding rules for the matrix, and Figure 16 shows them.

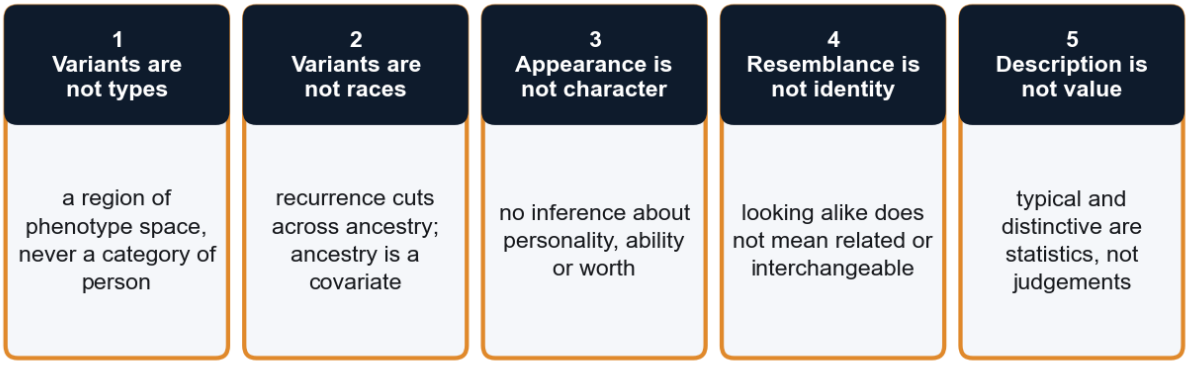

Figure 16. The five boundaries of the Global Variant Matrix. Each is a rule for how results may be described and used.

Source: Human Variant Hypothesis report (H Heuristics, 2026), Section 11, applied to the matrix.

1. **Variants are not types.** A human variant is a region of phenotype space, never a category of person. The matrix describes where people cluster, not what people are.
2. **Variants are not races.** Recurring configurations cut across ancestry. Regions are sampling frames only, and ancestry is a continuous measured covariate, never a box into which people are sorted.
3. **Appearance is not character.** The framework makes no claims, and permits no inferences, about personality, ability, behavior or worth from appearance.
4. **Resemblance is not identity.** Looking alike does not make two people interchangeable, related, or the same in any social or legal sense.
5. **Description is not value.** Typical and distinctive configurations are statistical descriptions, not judgments of beauty, normality or desirability. Alpha describes visibility, never superiority.

13.3 Race, ancestry and the language of regions

The relationship between a global matrix and race needs the most care. Race is a broad and socially important classification, not a neat set of separate biological boxes. Two landmark reports make the point. A 2023 report from the US National Academies recommends that genetics research move away from racial typology in its population descriptors and use more precise terms (National Academies, 2023). The American Association of Biological Anthropologists has issued a statement that race, as commonly understood, does not describe biological reality and that racial categories are social constructions with real consequences (AABA, 2019).

The matrix uses regions for a narrow purpose, to make sure that sampling covers the world. It uses ancestry as a measured number, not a label. It reports results for populations defined by specific samples and avoids grand names for whole continents. Where a regional word appears in this report, such as African or Asian, it refers to a sample from a place or to a described population, not to a biological category. Anyone using the matrix is asked to keep to the same practice.

13.4 Privacy, consent and biometric data

Facial images and genetic data are among the most sensitive kinds of personal data. Under the European Union's General Data Protection Regulation, biometric data used for unique identification and genetic data are special categories that need heightened protection (GDPR, Article 9). Several US states also regulate biometric identifiers. Research on look-alikes adds a specific risk: data showing that two people look alike could be used to impersonate one of them or to defeat a biometric lock.

Required safeguards include the following.

- **Specific, informed consent** that covers imaging, genetic analysis, perceptual studies and any sharing of data, with the right to withdraw.
- **Consent from both members of any pair** before pair-level data are analyzed or published.
- **Controlled-access storage.** No identifiable images or individual-level genomes in open repositories.
- **A ban on re-identification,** and no use of research data for surveillance, policing or commercial face matching.
- **No scraping of private individuals' images.** Photographs of private people found online are not to be collected, analyzed or republished without their agreement, however striking the resemblance.

13.5 Misuse and misreading

Three kinds of misuse are foreseeable, and each has a remedy built into the report.

Table 26. Foreseeable misuses and safeguards

Misuse	How it could happen	Safeguard
Reading the matrix as a map of races	A grid of regions looks like a grid of groups; variants could be named after peoples	Naming rules (Section 10.4); regions defined as sampling frames; boundaries 1 and 2 restated in every publication
Reading alpha as hierarchy	Alpha could be taken to mean superior or dominant	Always define alpha as visibility only; never rank groups; boundary 5
Inferring traits from faces	A resemblance to a famous person could carry assumed personality	Boundary 3; reference to evidence that first impressions of faces are unreliable guides to character (Todorov et al., 2015)
Using look-alike data to impersonate or surveil	Pair data could help defeat identity checks or support tracking	Controlled access; consent from both members; no policing or commercial use

Source: H Heuristics. Citation: Todorov et al. (2015), see Appendix C.

13.6 Rules for communication

How results are described matters as much as what they are. Every public communication from the programme should state the five boundaries in a short form, avoid language that implies types, ranks or essences, such as "kinds of faces" or "facial races," and use population descriptors consistent with current guidance. Participant and community representatives should read communications before release. Where a result is a null or negative one, it should be given the same prominence as a positive one.

13.7 Governance

The programme would be overseen by an independent ethics board and an advisory group that includes members of the communities being sampled. Its pre-registered protocols would be public. Variants and findings that fail replication would be retired in the open. A programme that invites scrutiny, publishes its failures and states its limits is the best protection against the use of its work for purposes it rejects.

13.8 A commitment

> **Commitment.** H Heuristics commits to publishing this programme's results whether they support the thesis or not, to describing every finding within the five boundaries, and to withdrawing or correcting any claim that the evidence does not support.

14 APPLICATIONS

14. Applications and Implications

14.1 Facial recognition and security

If look-alikes are common and spread across populations, then any system that identifies people from faces will produce false matches between strangers, and those errors will not fall evenly across groups. The US government's large test of face-recognition algorithms found that many systems made false matches at very different rates for different demographic groups (Grother and colleagues, 2019). The matrix offers a way to estimate, for each population, how many look-alikes a person is likely to have at a given strictness, which is exactly the quantity a security designer needs. Used in this way, the research would help reduce harm: it would support human review of high-stakes matches, corroborating evidence, and honest reporting of error rates by group.

14.2 Forensics and DNA phenotyping

Researchers have shown that facial shape can be partly predicted from DNA (Claes and colleagues, 2014), and that predicted faces can sometimes be matched to photographs in small test sets (Sero and colleagues, 2019). The ability of such methods to identify an individual from a genome has been strongly debated (Lippert and colleagues, 2017). The matrix strengthens a key caution. Because many different genetic combinations converge on similar faces, and because perceived resemblance is inflated by perception, a predicted face will resemble many people in a large population. Forensic use risks wrongly implicating a look-alike. If the thesis is correct, the number of look-alikes in the world is not small, and a face predicted from DNA can never single out one person with certainty.

14.3 Science of the face

The programme would widen the base of facial science. Cross-population samples would show which genetic effects on the face are shared across the species and which are not. They would help to test theories of developmental constraint in humans. They would extend facial genetics to populations it has largely missed. And they would offer a way to separate the contributions of ancestry, convergence and presentation to how faces look, a question that anthropology and developmental biology both care about.

14.4 Media, culture and online life

For ordinary readers, the most useful result may be a better way to read look-alike stories. Viral posts about twin strangers and celebrity doubles are everywhere. The framework in this report suggests seven questions that any reader can ask.

Table 27. Seven questions to ask about a look-alike story

Question	Why it matters
How was the pair found, and how many faces were searched?	Without the denominator, a match says little about how common matches are
Are the people styled alike?	Hair, glasses, makeup, pose and filters create resemblance that is not in the face
Is one of the two famous?	Fame multiplies the number of observers and so the number of matches noticed
Do the photos have matching lighting and angle?	Different photos of one person can look like different people
Who is judging?	Raters are better at telling apart familiar groups than unfamiliar ones
Does the claim go beyond looks?	A resemblance says nothing about relatedness, character or ability
Has it been checked blind?	A pair that stands out only to those who know the claim is not yet evidence

Source: H Heuristics.

14.5 Identity and everyday life

The idea of a global matrix has a human side. If appearances recur across all populations, then the sense that a stranger looks like a friend from far away is not a mistake or a trick of memory. It is a trace of the shared architecture of the species, as the Human Variant Hypothesis puts it. People who share a look with others across the world need not share anything else. The report hopes that this might, modestly, soften the sense of difference between groups. Resemblance is evidence that humanity is more continuous than the borders drawn on it suggest. It is not evidence that anyone is the same.

14.6 Teaching human variation

The topic is also a good vehicle for teaching. Look-alikes make the central ideas of human variation vivid: overlap, continuity and the difference between individuals and averages. A short teaching outline follows.

Table 28. A teaching outline built on the matrix

Idea	Classroom activity	Takeaway
Overlap	Plot a classmate measurement for several groups and see how distributions overlap	Group averages differ less than individuals do
Dimensions	Match a face on one, three and ten features and count the matches	More features mean fewer matches
Perception	Rate the same pairs with and without matching hairstyles	Styling changes what looks alike
Selection	Collect "found" look-alikes and ask what was left out	Reported matches are not a random sample
Ethics	Discuss the history of typing people by their faces	Why labels and rankings do harm

Source: H Heuristics. Suggested activities for secondary or undergraduate teaching.

14.7 Who might use this work

Table 29. Who might use the matrix and how

User	Use	Caution
Face and genetics researchers	Plan diverse samples; test shared and regional effects on facial shape	Do not generalize from one population
Developers of face recognition	Estimate look-alike risk by population; set match thresholds	Test fairness by region; keep human review
Forensic scientists	Understand how many people a predicted face could fit	Never treat a predicted face as an identification
Journalists and fact-checkers	Judge viral look-alike claims with the seven questions	A resemblance is not a relationship
Teachers and students	Teach overlap, continuity and the ethics of classification	Avoid labeling groups

Source: H Heuristics.

14.8 What the matrix should not be used for

The report states the limits plainly. The matrix is not a tool for estimating anyone's ancestry from their face, for judging who belongs to which group, for inferring traits, for ranking people or populations, or for surveillance. It is a framework for studying resemblance. Uses that stray beyond that are outside its scope and would contradict the boundaries in Section 13.

14.9 Open questions for the field

- How many independent dimensions does facial shape have within each module, and does the answer differ by population?
- How large is the other-group effect for look-alike judgments, and does it vary with a rater's life history?
- Are the same variants found in every region once styling is removed?
- Do cross-region look-alikes share face-associated variants beyond what ancestry explains?
- How much of the internet's apparent recurrence survives a visibility-matched control?

15 CONCLUSION

15. Conclusion: A Recurring Cast

15.1 Summary of the argument

The report began with an experience nearly everyone has had, meeting a stranger who looks like someone else, and a more specific impression: that these look-alikes turn up in every part of the world. It proposed that the impression might reflect a real pattern, a *global variant matrix*, in which recognizable human appearances recur across continents, populations and ancestral groups.

The case was built in layers. Faces are made from many small genetic effects acting through a shared developmental program, so very different genotypes converge on similar faces (Section 4). Most human variation lies within populations and groups overlap heavily, though small differences on many traits can still separate broad geographic groups (Section 5). The same trait can arise by different genetic routes in different peoples (Section 5). With these ingredients, a formal matrix of cross-population match rates can be defined and its behavior predicted, with facial shape expected to be broadly shared and surface traits adding a regional signal (Section 6). Mixing between populations creates bridges (Section 7). Perception, styling and the internet change what we notice and amplify some faces far more than others (Sections 8 and 9). A variant catalogue, five studies and a set of failure conditions would turn the idea into science (Sections 10 to 12), under firm ethical limits (Section 13).

15.2 What is established, what is supported, what is proposed

- **Well established (A).** Facial form is highly polygenic and largely regulatory. Most genetic variation lies within populations. Perceived resemblance depends on the observer and the image.
- **Supported but preliminary (B).** Unrelated look-alikes share common genetic variants, as shown in one small study. Typical faces have more look-alikes.
- **Plausible and untested (C).** Cross-region look-alikes are common in facial shape. Admixed populations bridge distant regions. The internet inflates apparent recurrence.
- **Speculative (D).** Alpha variants exist as a visibility pattern. A stable global catalogue of variants can be built.

15.3 The verdict

The report's verdict is a moderate one. The ingredients of the Global Variant Matrix are well supported by existing science. The thesis that those ingredients assemble into a connected global structure is plausible, specific and testable, and has not yet been tested. It would be wrong to say the matrix exists. It would be equally wrong to say it does not. The honest statement is that the first global test is still to be done, and that the programme in Sections 10 and 11 is designed to do it.

Closing thought. Humanity may contain a recurring cast of appearances that no border contains. The cast may be recurring. The people, their lives and their choices remain distinct.

15.4 Next steps

1. **Write the protocol.** Draft and file the pre-registration, ethics plan and catalogue rules (Study 0).
2. **Reanalyze what exists.** Use existing diverse 3D face and genetic datasets to estimate effective dimensionality and recurrence by population (Study A).
3. **Run the perception and visibility studies early.** They are cheap, need no genetic data and answer the most pressing objections (Studies C and D).
4. **Pilot the audit.** Run a small, standardized, population-based cross-region audit in three or four regions to test logistics and refine tolerances (Study B).
5. **Convene the advisory group.** Bring together genetics, developmental biology, perception science, ethics and community representatives to review the programme.
6. **Start the catalogue.** Enter the Hunter and Lars variants as first reference points, graded and limited like any other.

15.5 Five things to take away

1. **Look-alikes are expected, not mysterious.** With billions of people and a shared way of building faces, close resemblance between strangers is a statistical certainty.
2. **How many look-alikes depends on how closely you look.** More features compared means fewer matches.
3. **Observers, styling and the internet shape what we notice.** Perceived resemblance outnumbers measured resemblance.
4. **A global matrix is a hypothesis.** It is plausible and testable, and it has not yet been tested everywhere.
5. **Resemblance is not identity, ancestry or character.** The five boundaries apply to every finding.

15.6 Closing

A doppelgänger is a surprise. A variant is a pattern. A matrix is a hypothesis about how the patterns connect. The three words describe three steps from a curiosity to a question that can be answered. If the answer is yes, then the human face is a shared inheritance, and the sense of recognizing a stranger is a small glimpse of how much humanity has in common. If the answer is no, we will have learned where resemblance stops. Either way, the question is worth asking, and asking it carefully is the whole point of this report.

16 APPENDICES

Appendix A. Glossary

Working definitions of key terms as used in this report.

Admixture. The mixing of ancestry from two or more previously separate populations.

Alpha variant. A provisional term for a variant archetype that is disproportionately represented among highly visible people. It describes prominence, not biological superiority.

Ancestry-associated resemblance. The statistical tendency of populations with shared history to resemble each other more than distant populations do. A baseline, not a type.

Availability heuristic. The habit of judging how common something is by how easily examples come to mind.

Bridge population. A population that carries ancestry from two or more regions and is expected to show recurrence with each.

Canalization. The buffering of development so that it reaches similar outcomes despite small genetic or environmental differences.

Composite variant. A variant class in which a distinctive combination of facial, hair and bodily traits recurs.

Confirmation bias. The tendency to notice and remember evidence that fits an idea one already holds.

Connectivity. The average of all off-diagonal recurrence ratios in the matrix; a summary of how linked populations are.

Convergent configuration. Independent arrival of unrelated people at similar trait values through different genetic combinations.

Denominator. The total number of comparisons made, needed to turn a count of matches into a rate.

Discovery bias. Error that arises because pairs are noticed in proportion to visibility rather than at random.

Doppelgänger. A person who closely resembles another, unrelated person.

Effective dimensionality. How many independent directions of variation a set of traits really has, after allowing for correlations among them.

Face space. A psychological model in which faces are points in a multidimensional space centered on an average face.

Heritability. The share of variation in a trait, in a given population and environment, associated with genetic differences.

Match rate (M). The probability that a random pair, one person from each of two populations, qualify as look-alikes.

Matrix, global variant. A table of match rates or recurrence ratios between the world's populations. A proposed analytical framework, not a classification of people.

Modularity. Organization of traits into sets that vary together internally and only weakly with other sets, such as the nose region or jaw.

Morphological variant. A variant class based on broad facial geometry.

Own-race bias (other-group effect). The tendency to recognize and tell apart faces of familiar groups more accurately than those of less familiar groups.

Pareto distribution. A heavy-tailed pattern in which a small share of sources accounts for most of an outcome, commonly summarized as the 80/20 rule.

Perceptual amplification. Inflation of perceived resemblance relative to geometric similarity, caused by styling, perception and presentation.

Phenotype. The observable traits of an organism.

Polygenic trait. A trait influenced by many genetic variants, each usually of small effect.

Pre-registration. Publicly filing a study's hypotheses and analysis plan before collecting or examining data.

Preferential attachment. A process by which popular items gain further popularity because newcomers attach to them.

Proportional variant. A variant class based on relative proportions rather than absolute size.

Recurrence ratio (R). Cross-population match rate divided by the within-population match rate of the row population.

Reference point. The central configuration around which a variant neighborhood is defined.

Serial founder effect. Loss of genetic diversity as small groups repeatedly split off from larger ones during the spread of a species.

Tolerance (ε). How close two people must be on every compared trait to count as a match.

Variant. A proposed analytical term for a recurring phenotypic configuration, a recognizable visual archetype. It is not a type, race or category of person.

Variant card. The standard record of a catalogue entry: definition, method, samples, reach, attribution, grade and limits.

Variant catalogue. A register of variant cards.

Variant reach. The number of regions in which a variant neighborhood has members.

Appendix B. Observation Card and Sampling Worksheet

The observation card below turns an everyday impression into a record that can be used later. One card is filled out for each candidate look-alike pair or set. Its purpose is to capture the facts a researcher needs and the biases that are easy to forget.

B.1 The observation card

Table 30. Observation card

Field	What to record	Example entries
Date and observer	Who noticed the resemblance and when	Initials, date
Where it was noticed	In person, online, in the media; platform if online	Street; video platform; news photo
Discovery route	How the pair or set was found	Chance meeting; search; viral post; friend's tip
Reference point	The person or description everyone is compared with	Code or consented name
Claimed shared features	The specific features said to match, written before looking further	Eye set; jaw; face length; brow
Styling similarity	Hair, beard, glasses, makeup, expression, pose, filter	Same cut; both bearded; both smiling
Image quality	Resolution, lighting, angle, date of photo	Frontal, soft light, recent
Apparent age and sex of each	As best known; self-reported if available	Age band; sex
Region of ancestry and residence	Recorded separately, as stated by the person where possible	Region of family origin; country of residence
Visibility of each person	Private, locally known, widely known	Private; regional; global
Comparisons made	How many faces were looked at to find this one (denominator)	About 200; unknown
Controls	Non-matching faces shown alongside, if any	Five unrelated faces
Consent status	Whether each person agreed to be recorded	Yes; no; not applicable (public figure)
Observer's own confidence	How strong the resemblance feels, 1 to 5	3

Source: H Heuristics. Proposed template. Do not record or share images of private individuals without their agreement.

B.2 Standard imaging checklist

Table 31. Standard imaging checklist for catalogue entries

Item	Standard
Expression	Neutral, mouth closed
Hair	Restrained away from the face and ears
Eyewear and makeup	None
Pose	Frontal, head level, eyes at the camera
Lighting	Even and diffuse; same setup for every person
Background	Plain
Distance and lens	Fixed distance; same camera and lens across sites
Colour calibration	Standard color chart in frame
3D scan, where available	Face and head, same protocol across sites
Repeat images	A subsample imaged twice to estimate measurement error

Source: H Heuristics. Follows the phenotyping protocol of the Human Variant Hypothesis report.

B.3 Regional sampling worksheet

For each study, a worksheet is kept for every region and subregion, so that gaps in coverage are visible early.

Table 32. Regional sampling worksheet

Region	Subregions covered	Number imaged	Source of sample	Ancestry data	Gaps
Africa					
Europe					
Asia					
North America					
South America					
Oceania					
Middle East					

Blank worksheet to be completed for each study.

Appendix C. References

The list below gives full details for every primary source cited by DOI in the text, generated from DOI registry metadata and checked against the registered titles. Web sources and project documents follow.

Adhikari, K. et al. (2016). A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nature Communications 7, 10815*. doi:10.1038/ncomms10815

Barabási, A. L. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science 286, 509-512*. doi:10.1126/science.286.5439.509

Beall, C. M. (2007). Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proceedings of the National Academy of Sciences 104, 8655-8660*. doi:10.1073/pnas.0701985104

Bryc, K. et al. (2015). The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics 96, 37-53*. doi:10.1016/j.ajhg.2014.11.010

Chang, L. & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell 169, 1013-1028.e14*. doi:10.1016/j.cell.2017.05.011

Claes, P. et al. (2014). Modeling 3D Facial Shape from DNA. *PLoS Genetics 10, e1004224*. doi:10.1371/journal.pgen.1004224

Claes, P. et al. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics 50, 414-423*. doi:10.1038/s41588-018-0057-4

Clauset, A. et al. (2009). Power-Law Distributions in Empirical Data. *SIAM Review 51, 661-703*. doi:10.1137/070710111

Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al. (2023). Using Population Descriptors in Genetics and Genomics Research. **. doi:10.17226/26902

Crawford, N. G. et al. (2017). Loci associated with skin pigmentation identified in African populations. *Science 358, eaan8433*. doi:10.1126/science.aan8433

Edge, M. D. & Rosenberg, N. A. (2015). Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 52, 32-45*. doi:10.1016/j.shpsc.2014.12.005

Galton, F. (1879). Composite Portraits, Made by Combining Those of Many Different Persons Into a Single Resultant Figure. *The Journal of the Anthropological Institute of Great Britain and Ireland 8, 132*. doi:10.2307/2841021

Grother, P. et al. (2019). Face recognition vendor test part 3:. **. doi:10.6028/NIST.IR.8280

Haber, M. et al. (2013). Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture. *PLoS Genetics 9, e1003316*. doi:10.1371/journal.pgen.1003316

Hancock, P. J. et al. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences 4, 330-337*. doi:10.1016/S1364-6613(00)01519-9

Hellenthal, G. et al. (2014). A Genetic Atlas of Human Admixture History. *Science 343, 747-751*. doi:10.1126/science.1243518

Henn, B. M. et al. (2012). The great human expansion. *Proceedings of the National Academy of Sciences 109, 17758-17764*. doi:10.1073/pnas.1212380109

Jablonski, N. G. & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution 39, 57-106*. doi:10.1006/jhev.2000.0403

Jenkins, R. et al. (2011). Variability in photos of the same face. *Cognition 121, 313-323*. doi:10.1016/j.cognition.2011.08.001

Joshi, R. S. et al. (2022). Look-alike humans identified by facial recognition algorithms show genetic similarities. *Cell Reports 40, 111257*. doi:10.1016/j.celrep.2022.111257

Klingenberg, C. P. (2008). Morphological Integration and Developmental Modularity. *Annual Review of Ecology, Evolution, and Systematics 39, 115-132*. doi:10.1146/annurev.ecolsys.37.091305.110054

Langlois, J. H. & Roggman, L. A. (1990). Attractive Faces Are Only Average. *Psychological Science 1, 115-121*. doi:10.1111/j.1467-9280.1990.tb00079.x

Lewontin, R. C. (1972). The Apportionment of Human Diversity. *Evolutionary Biology, 381-398*. doi:10.1007/978-1-4684-9063-3_14

Li, J. Z. et al. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science 319, 1100-1104*. doi:10.1126/science.1153717

Lippert, C. et al. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences 114, 10166-10171*. doi:10.1073/pnas.1711125114

Lucas, T. & Henneberg, M. (2015). Are human faces unique? A metric approach to finding single individuals without duplicates in large samples. *Forensic Science International 257, 514.e1-514.e6*. doi:10.1016/j.forsciint.2015.09.003

Martin, A. R. et al. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics 51, 584-591*. doi:10.1038/s41588-019-0379-x

Meissner, C. A. & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law 7, 3-35*. doi:10.1037/1076-8971.7.1.3

Moreno-Estrada, A. et al. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science 344, 1280-1285*. doi:10.1126/science.1251688

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology 2, 175-220*. doi:10.1037/1089-2680.2.2.175

Norton, H. L. et al. (2006). Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Molecular Biology and Evolution 24, 710-722*. doi:10.1093/molbev/msl203

Novembre, J. et al. (2008). Genes mirror geography within Europe. *Nature 456, 98-101*. doi:10.1038/nature07331

Phillips, P. J. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences 115, 6171-6176*. doi:10.1073/pnas.1721355115

Pickrell, J. K. et al. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences 111, 2632-2637*. doi:10.1073/pnas.1313787111

Popejoy, A. B. & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature 538, 161-164*. doi:10.1038/538161a

Ramachandran, S. et al. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences 102, 15942-15947*. doi:10.1073/pnas.0507611102

Reich, D. et al. (2009). Reconstructing Indian population history. *Nature 461, 489-494*. doi:10.1038/nature08365

Relethford, J. H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *American Journal of Physical Anthropology 118, 393-398*. doi:10.1002/ajpa.10079

Rosenberg, N. A. et al. (2002). Genetic Structure of Human Populations. *Science 298, 2381-2385*. doi:10.1126/science.1078311

Russell, R. et al. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review 16, 252-257*. doi:10.3758/PBR.16.2.252

Salganik, M. J. et al. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science 311, 854-856*. doi:10.1126/science.1121066

Sero, D. et al. (2019). Facial recognition from DNA using face-to-DNA classifiers. *Nature Communications 10, 2557*. doi:10.1038/s41467-019-10617-y

Sirugo, G. et al. (2019). The Missing Diversity in Human Genetic Studies. *Cell 177, 26-31*. doi:10.1016/j.cell.2019.02.048

Skoglund, P. et al. (2016). Genomic insights into the peopling of the Southwest Pacific. *Nature 538, 510-513*. doi:10.1038/nature19844

Tishkoff, S. A. et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics 39, 31-40*. doi:10.1038/ng1946

Todorov, A. et al. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology 66, 519-545*. doi:10.1146/annurev-psych-113011-143831

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology 5, 207-232*. doi:10.1016/0010-0285(73)90033-9

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A 43, 161-204*. doi:10.1080/14640749108400966

White, J. D. et al. (2021). Insights into the genetic architecture of the human face. *Nature Genetics 53, 45-53*. doi:10.1038/s41588-020-00741-7

Witherspoon, D. J. et al. (2007). Genetic Similarities Within and Between Human Populations. *Genetics 176, 351-359*. doi:10.1534/genetics.106.067355

Xiong, Z. et al. (2019). Novel genetic loci affecting facial shape variation in humans. *eLife 8, e49898*. doi:10.7554/eLife.49898

Zaidi, A. A. et al. (2017). Investigating the case of human nose shape and climate adaptation. *PLOS Genetics 13, e1006616*. doi:10.1371/journal.pgen.1006616

Web sources and project documents

American Association of Biological Anthropologists. (2019). *AABA statement on race and racism*. bioanth.org/about/aaba-statement-on-race-racism/

Agüera y Arcas, B., Mitchell, M., & Todorov, A. (2017). *Physiognomy's new clothes*. Medium. medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91. proceedings.mlr.press/v81/buolamwini18a.html

GDPR. (2016). Article 9: Processing of special categories of personal data. Regulation (EU) 2016/679. gdpr-info.eu/art-9-gdpr/

Hughes, H. (2026). *The Human Variant (or Doppelgänger) Phenomenon*. H Heuristics. hheuristics.substack.com/p/the-variant-or-doppelganger-phenomenon

Hughes, H. (2026). *The Human Variant Hypothesis: DNA, Creation Structures, and the Recurrence of Human Form* (H Heuristics Research Report HH-2026-30). researchreports.hheuristics.com/reports/human-variant-hypothesis/

H Heuristics

Navigating a changing world.

H Heuristics is an independent publication by Hunter Hughes exploring the ideas, structures and patterns that shape a changing world, from science and technology to society and human experience.

This report is part of an ongoing series of long-form research essays. It is a proposed research programme and conceptual synthesis. It presents no new empirical data on faces or genomes, and it is intended to stimulate rigorous inquiry and discussion. It is not medical, genetic, forensic or legal advice.

Read more and subscribe at hheuristics.substack.com.