遺伝情報による顔貌認識

法数学勉強会

2019/10/19

統計遺伝学分野

山田　亮

nature COMMUNICATIONS

Article | OPEN | Published: 11 June 2019

Facial recognition from DNA using face-to-DNA classifiers

Dzemila Sero, Arslan Zaidi, Jiarui Li, Julie D. White, Tomás B. González Zarzar, Mary L. Marazita, Seth M. Weinberg, Paul Suetens, Dirk Vandermeulen, Jennifer K. Wagner, Mark D. Shriver & Peter Claes

Nature Communications **10**, Article number: 2557 (2019) | Download Citation

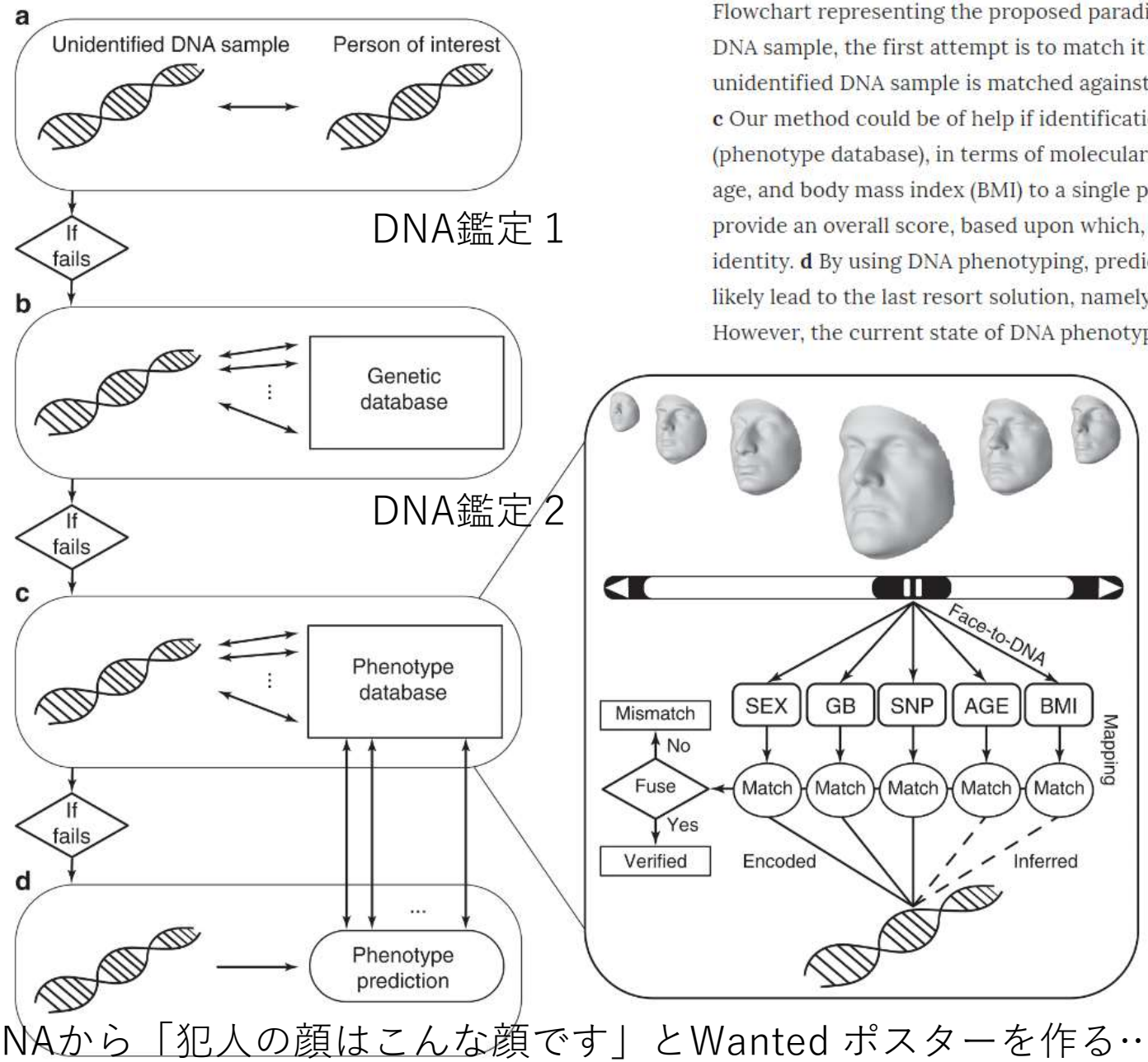
Abstract

Facial recognition from DNA refers to the identification or verification of unidentified biological material against facial images with known identity. One approach to establish the identity of unidentified biological material is to predict the face from DNA, and subsequently to match against facial images. However, DNA phenotyping of the human face remains challenging. Here, another proof of concept to biometric authentication is established by using multiple face-to-DNA classifiers, each classifying given faces by a DNA-encoded aspect (sex, genomic background, individual genetic loci), or by a DNA-inferred aspect (BMI, age). Face-to-DNA classifiers on distinct DNA aspects are fused into one matching score for any given face against DNA. In a globally diverse, and subsequently in a homogeneous cohort, we demonstrate preliminary, but substantial true (83%, 80%) over false (17%, 20%) matching in verification mode. Consequences of future efforts include forensic applications, necessitating careful consideration of ethical and legal implications for privacy in genomic databases.

- DNA情報から、予想顔貌を描き、それに似た人を探す、という手もあるが、ちょっと難しい
- 多人数の登録済み顔貌情報とDNA情報の関係から、「このDNAからこの顔貌は『アリ』か『ナシ』か」の判定ができると、一歩前進だ

Fig. 1

From: Facial recognition from DNA using face-to-DNA classifiers



Flowchart representing the proposed paradigm in the context of existing DNA investigative tools. **a** Given an unidentified DNA sample, the first attempt is to match it to the DNA of a person of interest. **b** If the matching fails, then the same unidentified DNA sample is matched against DNA profiles of persons with known identity enrolled in the genetic database. **c** Our method could be of help if identification fails again. Each face-to-DNA classifier matches a face, in a gallery of faces (phenotype database), in terms of molecular features including sex, genomic background (GB), individual genetic loci (SNP), age, and body mass index (BMI) to a single probe DNA. Multiple, one per aspect, matching scores are fused together to provide an overall score, based upon which, it becomes possible to verify or reject a DNA profile against a face with known identity. **d** By using DNA phenotyping, predicted phenotypes could as well be matched against given phenotypes, but more likely lead to the last resort solution, namely showing it to the public and hoping that someone recognizes the individual. However, the current state of DNA phenotyping has not achieved this ability yet

DNA鑑定 1

DNA鑑定 2

顔貌とDNAとのわかっているデータベースを使って
知りたいDNAから生じそうな顔貌をスコア付きで選び出す
「こんな顔かも」「あんな顔かも」という情報として提示される

DNAから「犯人の顔はこんな顔です」とWanted ポスターを作る…ちょっとまだ無理

顔貌 vs. (遺伝(SNP,民族性),性,年齢,BMI)

- 顔貌に影響することが明らかな要因
  - 性別、年齢、BMI
  - 民族性
  - それらを差し引いて（共変量として）、顔貌影響多型を同定・選択

顔貌識別モデルの機械学習

- サポートベクターマシン
  - 性別、年齢、BMI、民族性（多数のSNPからの遺伝的民族指標）、SNP

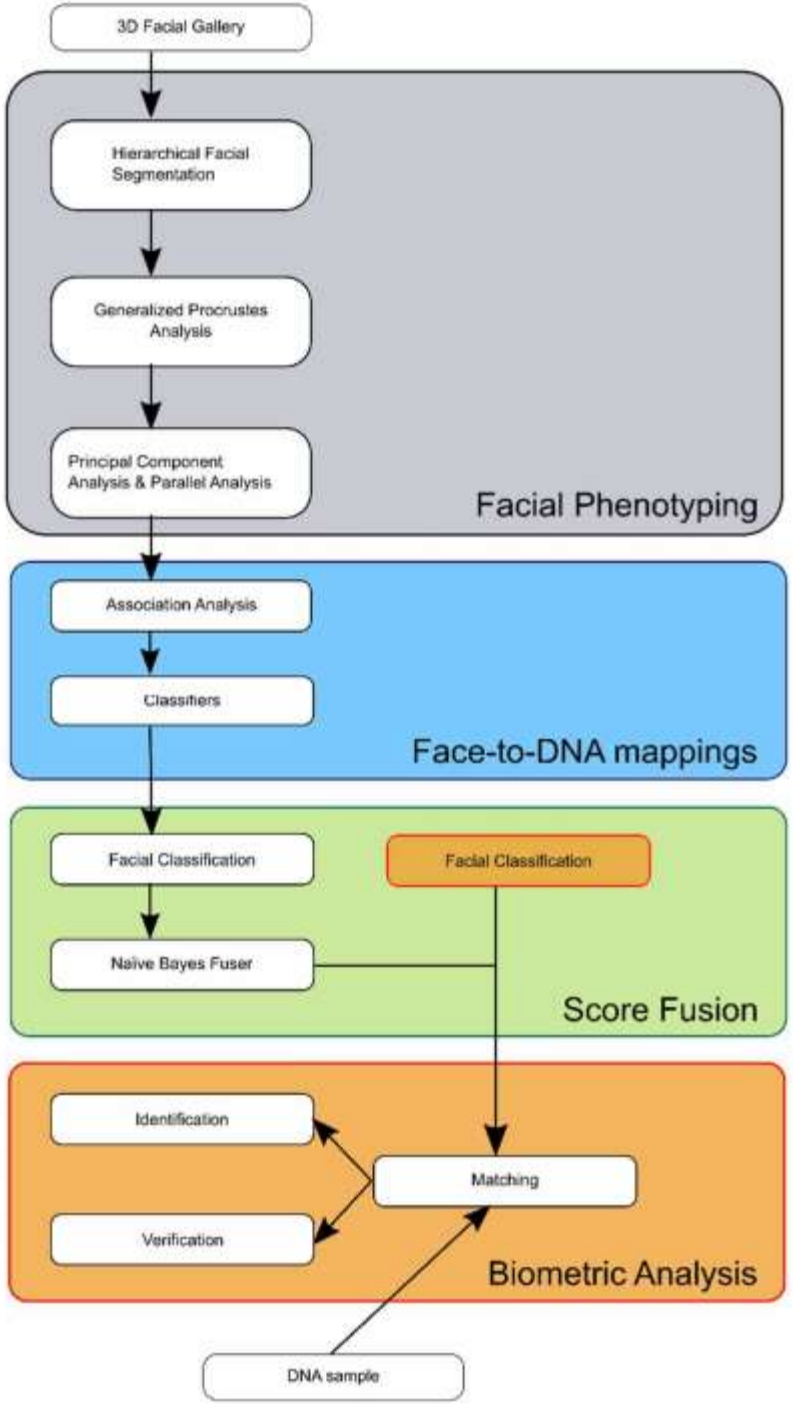
3D顔貌ギャラリー

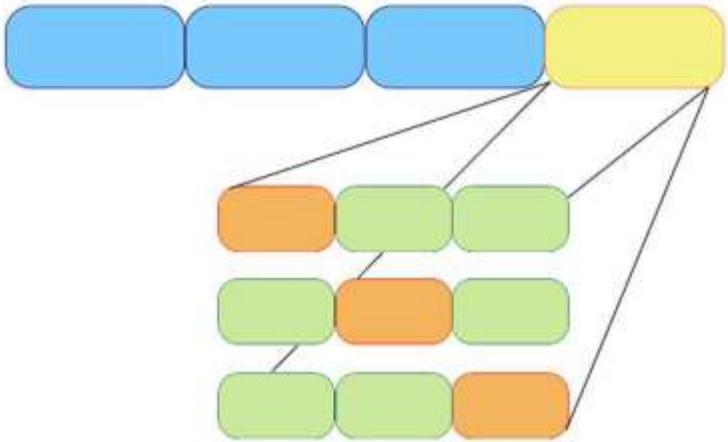
データ解析できるように特徴量に変える

個々の特徴量に関係する遺伝子多型を見つける

複数の特徴量の統合をして判定モデルを作る

クロスバリデーション





GLOBAL Cohort

= COMPLETE SET, 3,295
= TRAINING SET, 2,471 participants
= REMAINING SET, 824 participants
= VALIDATION SET, 549 participants
= TEST SET, 275 participants

民族差が大きく影響している場合

EURO Cohort

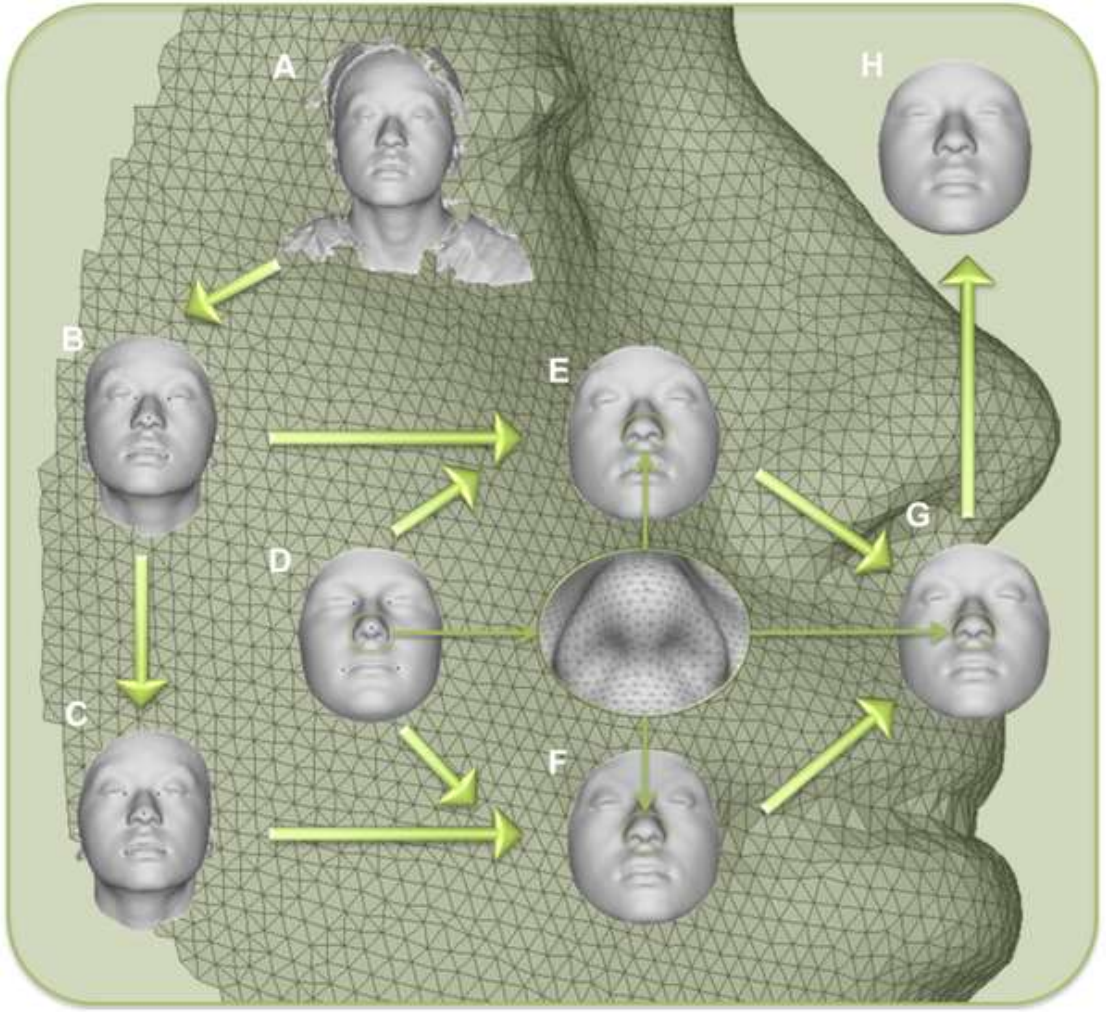
= COMPLETE SET, 3,542
= TRAINING SET, 2,656 participants
= REMAINING SET, 886 participants
= VALIDATION SET, 591 participants
= TEST SET, 295 participants

民族的に比較的均質な場合

Figure 1. Workflow for 3D face scan processing.

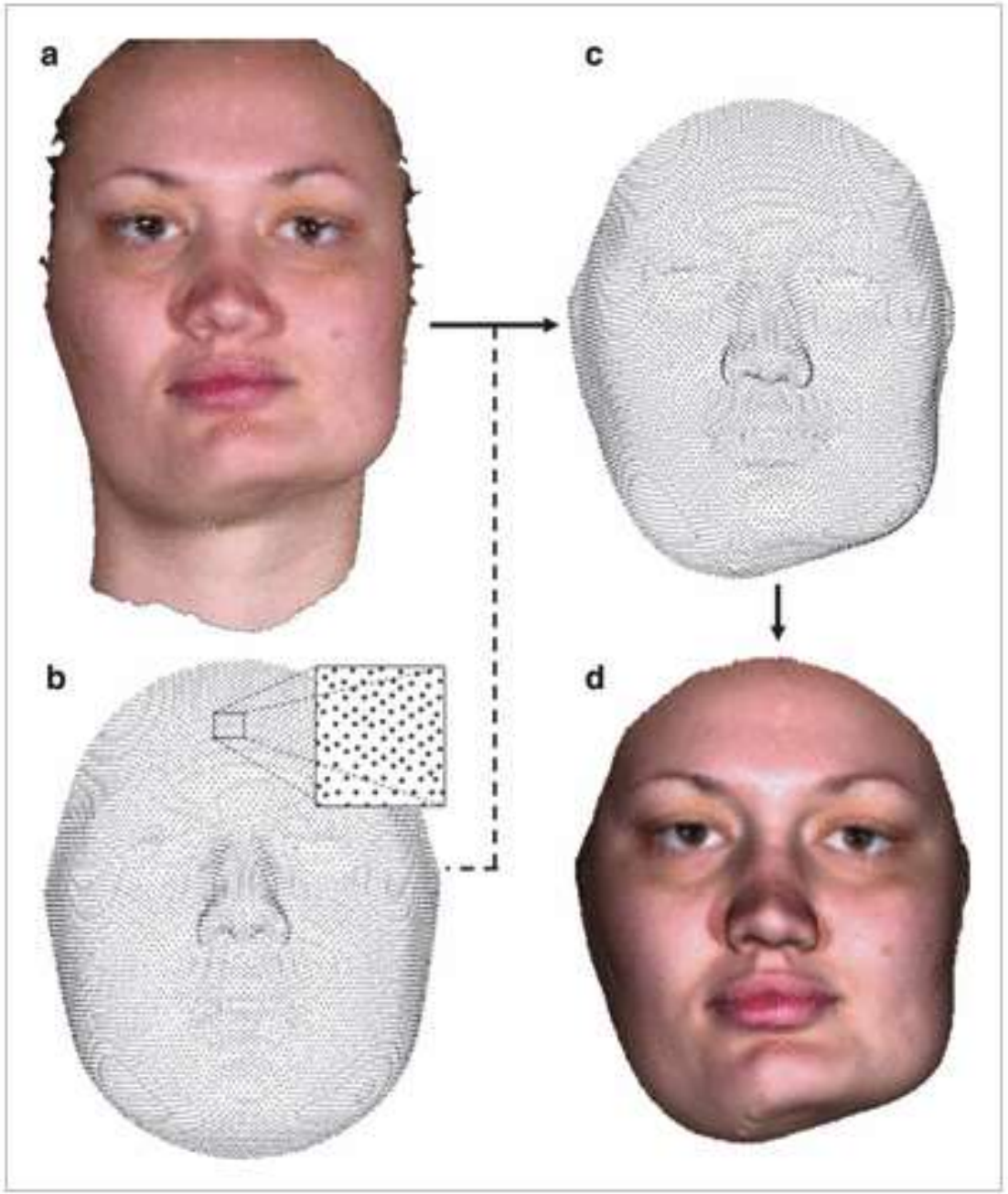
おおよそのアラインメントを決める
メッシュを張り付ける
左右の平均を取る

各頂点を、Quasi-Landmarksとする



Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, et al. (2014) Modeling 3D Facial Shape from DNA. PLOS Genetics 10(3): e1004224. https://doi.org/10.1371/journal.pgen.1004224
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004224

PLOS | GENETICS



igure 1

Open in figure viewer | PowerPoint

左右非対称問題は、
先行研究の手法で対
処する

Journal of Anatomy / Volume 219, Issue 4

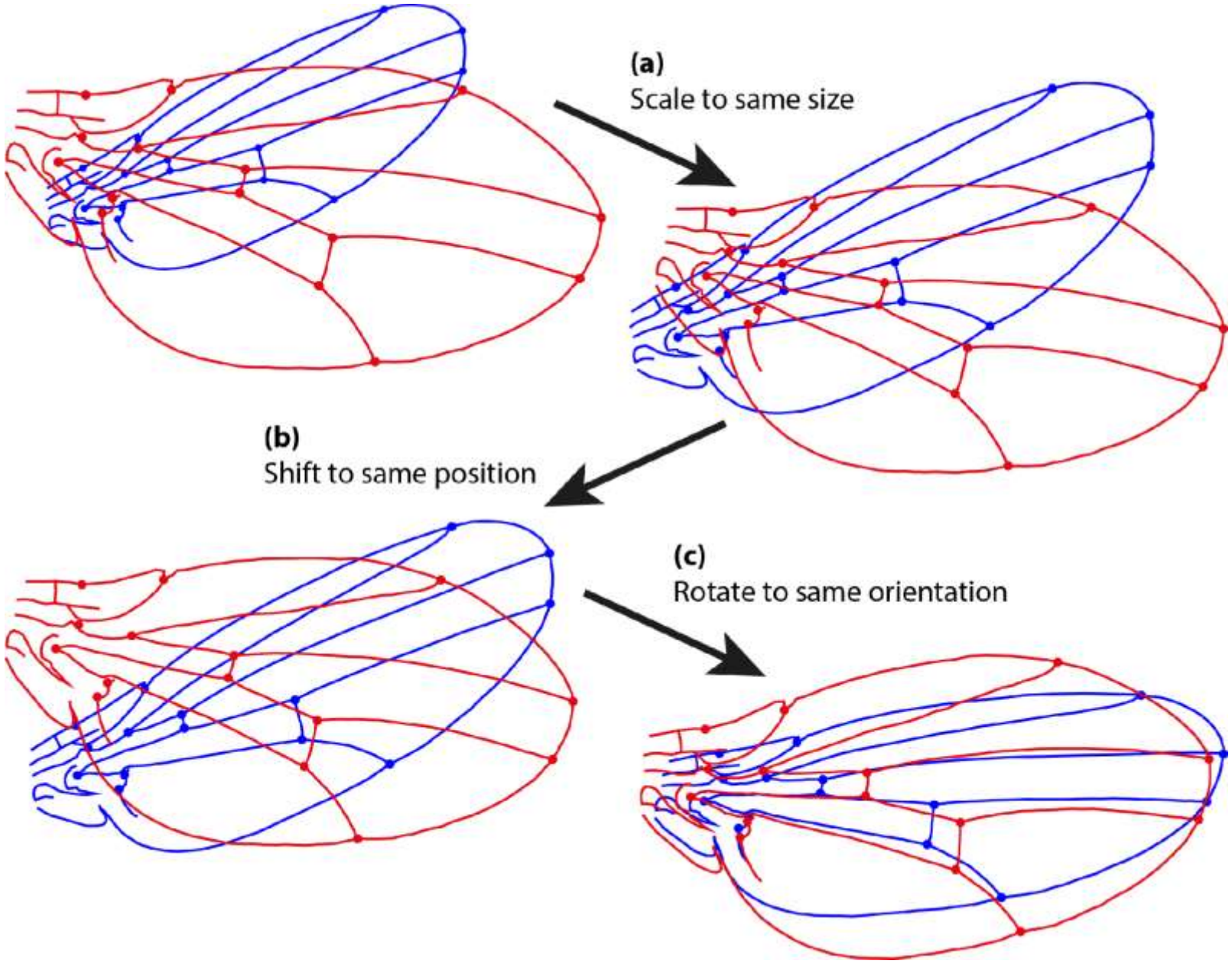
Open Access

Spatially-dense 3D facial asymmetry assessment in both typical and disordered growth

Peter Claes, Mark Walters, Dirk Vandermeulen, John Gerald Clement

First published: 11 July 2011
https://doi.org/10.1111/j.1469-7580.2011.01411.x
Cited by: 42

Generalized Procrutes Analysis



ランドマークをうまく対応づける幾何統計手法
全ての顔貌について、回転・伸び縮みを入れて補正する

全ての顔貌について、各点の「平均位置」を決める

平均からのずれを数値化する

点のペアワイズでの関係性を数値化する

相関行列をもとに階層型クラスタリ

Fig. 3

From: Facial recognition from DNA using face-to-DNA classifiers

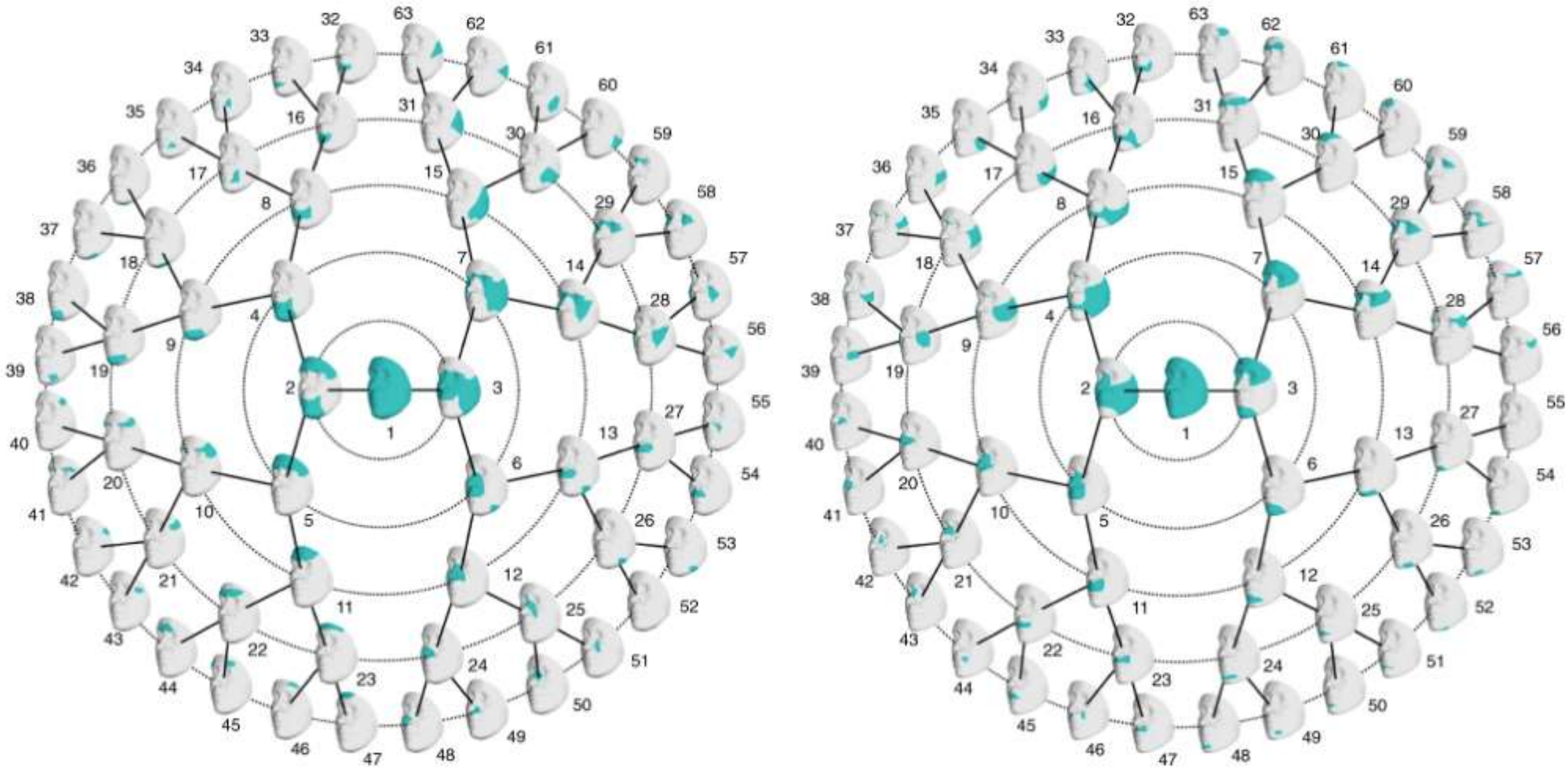
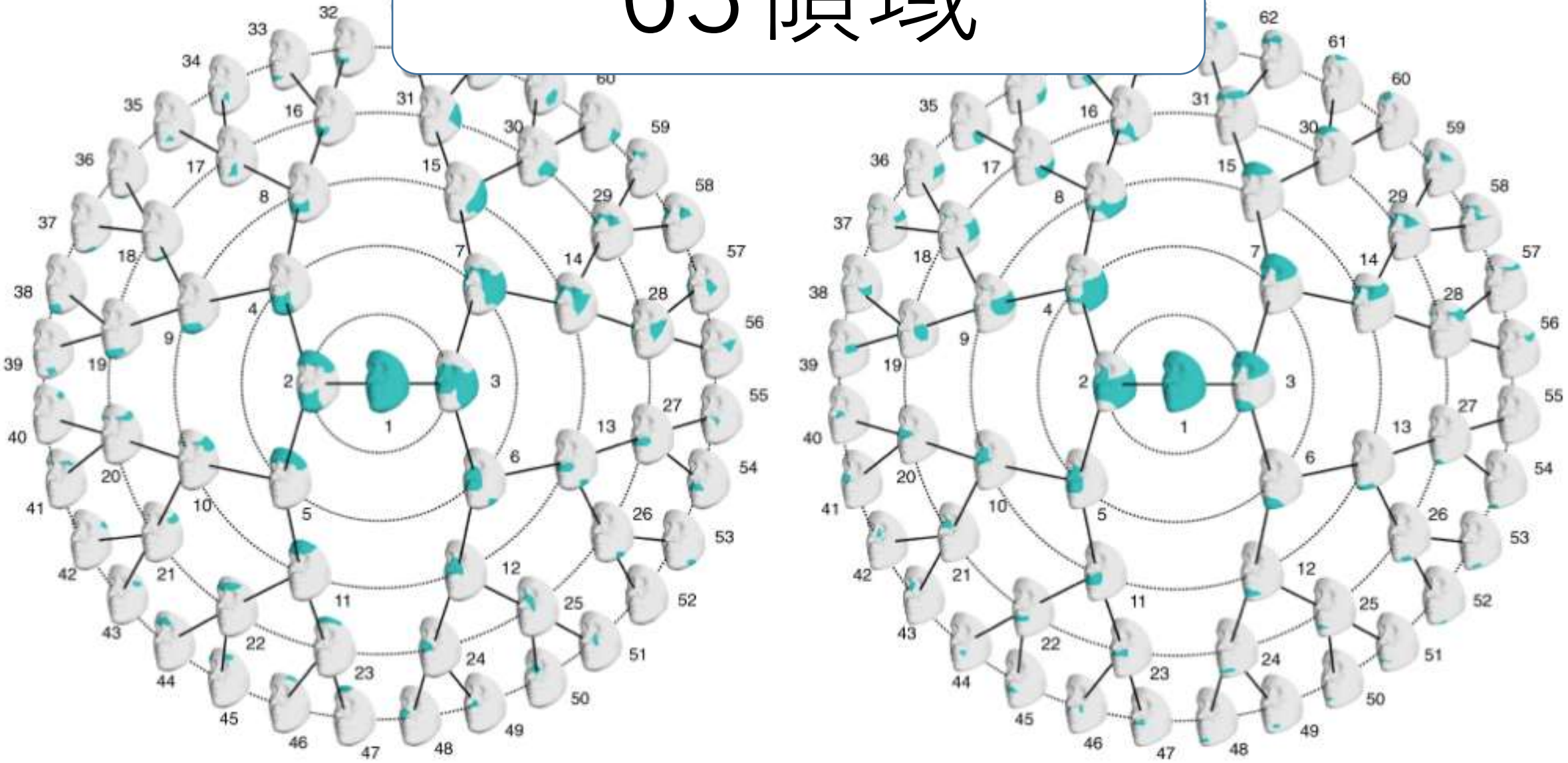
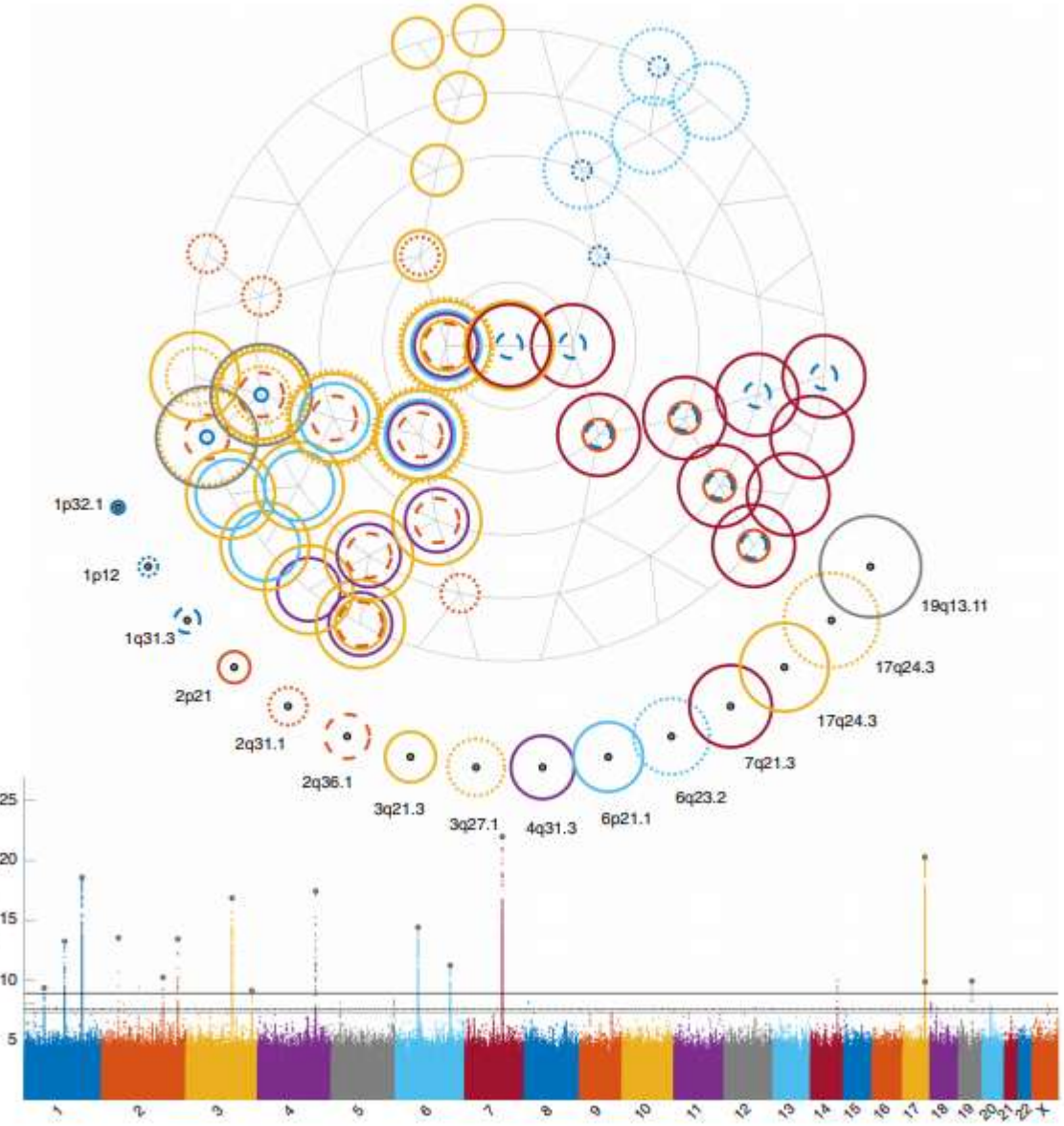


Fig. 3

From: Facial recognition from

63領域





ARTICLES
https://doi.org/10.1038/s41588-018-0057-4
nature genetics

Genome-wide mapping of global-to-local genetic effects on human facial shape

Peter Claes[1,2,3*], Jasmien Roosenboom[4], Julie D. White[5], Tomek Swigut[6], Dzemila Sero[1,2], Jiarui Li[1,2], Myoung Keun Lee[4], Arslan Zaidi[5], Brooke C. Mattern[5], Corey Liebowitz[5], Laurel Pearson[5], Tomás González[5], Elizabeth J. Leslie[4], Jenna C. Carlson[7], Ekaterina Orlova[8], Paul Suetens[1,2], Dirk Vandermeulen[1,2], Eleanor Feingold[7,8], Mary L. Marazita[4,8], John R. Shaffer[8], Joanna Wysocka[6,9*], Mark D. Shriver[5*] and Seth M. Weinberg[4,10*]

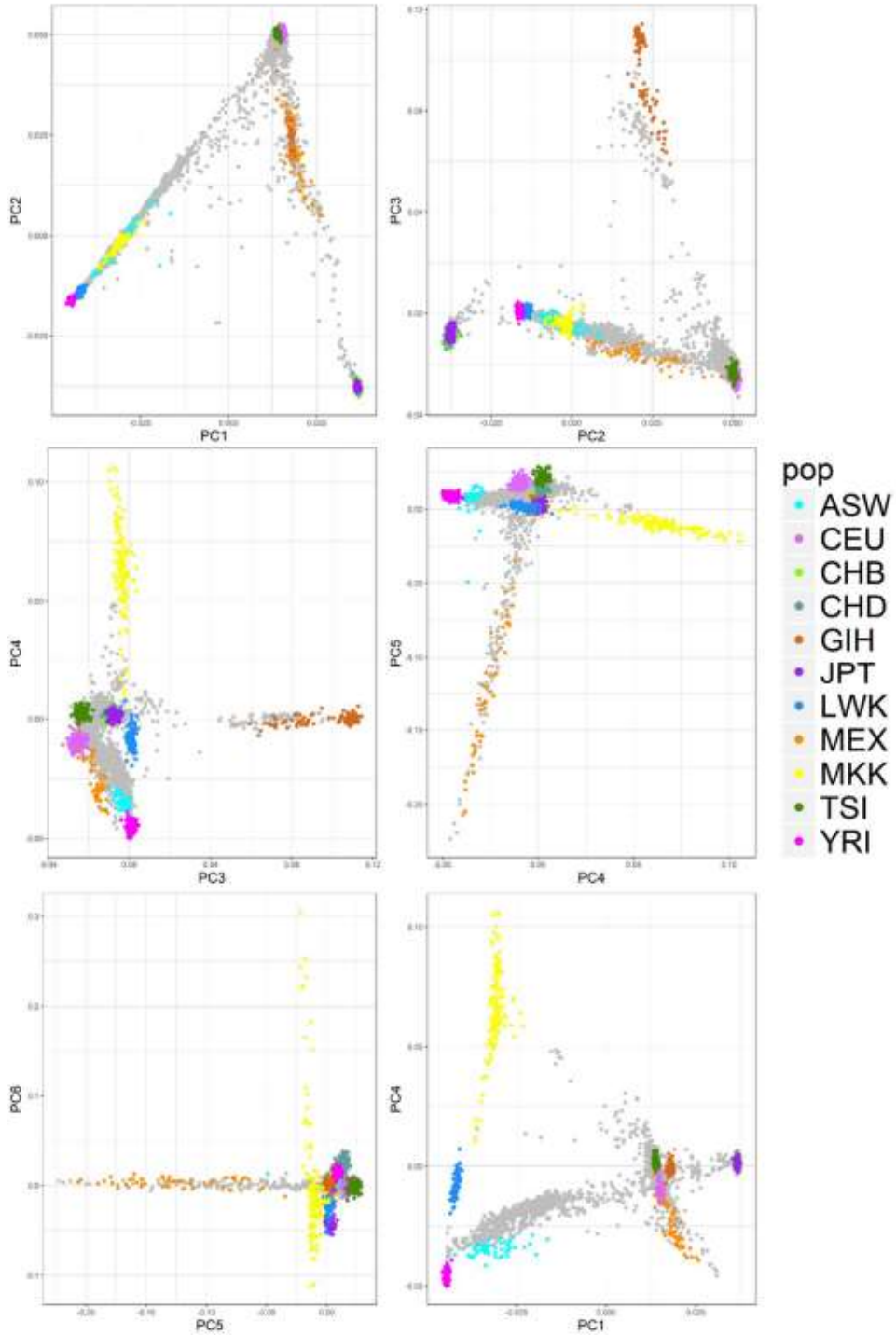
Genome-wide association scans of complex multipartite traits like the human face typically use preselected phenotypic measures. Here we report a data-driven approach to phenotyping facial shape at multiple levels of organization, allowing for an open-ended description of facial variation while preserving statistical power. In a sample of 2,329 persons of European ancestry, we identified 38 loci, 15 of which replicated in an independent European sample ($n=1,719$). Four loci were completely new. For the others, additional support ($n=9$) or pleiotropic effects ($n=2$) were found in the literature, but the results reported here were further refined. All 15 replicated loci highlighted distinctive patterns of global-to-local genetic effects on facial shape and showed enrichment for active chromatin elements in human cranial neural crest cells, suggesting an early developmental origin of the facial variation captured. These results have implications for studies of facial genetics and other complex morphological traits.

SNP
顔のどこを説明するか？

Fig. 2 | Fifteen replicating loci. Top, a polar dendrogram structurally corresponding to the polar dendrogram displaying the facial segments in Fig. 1; concentric circles represent loci reaching genome-wide significance for each segment. Bottom, Manhattan plot of all 63 facial segments combined illustrating the chromosomal position of the associated loci. The bottom (dotted) horizontal line represents the genome-wide significance threshold ($P=5\times10^{-8}$), the middle (dashed) horizontal line represents the FDRd significance threshold ($P=2.82\times10^{-8}$) and the top (solid) horizontal line represents study-wide significance ($P=1.28\times10^{-9}$). The 15 loci are ordered by chromosome number, increasing from left to right, and are colored according to corresponding chromosome.

顔貌 vs. (遺伝(SNP,民族性),性,年齢,BMI)

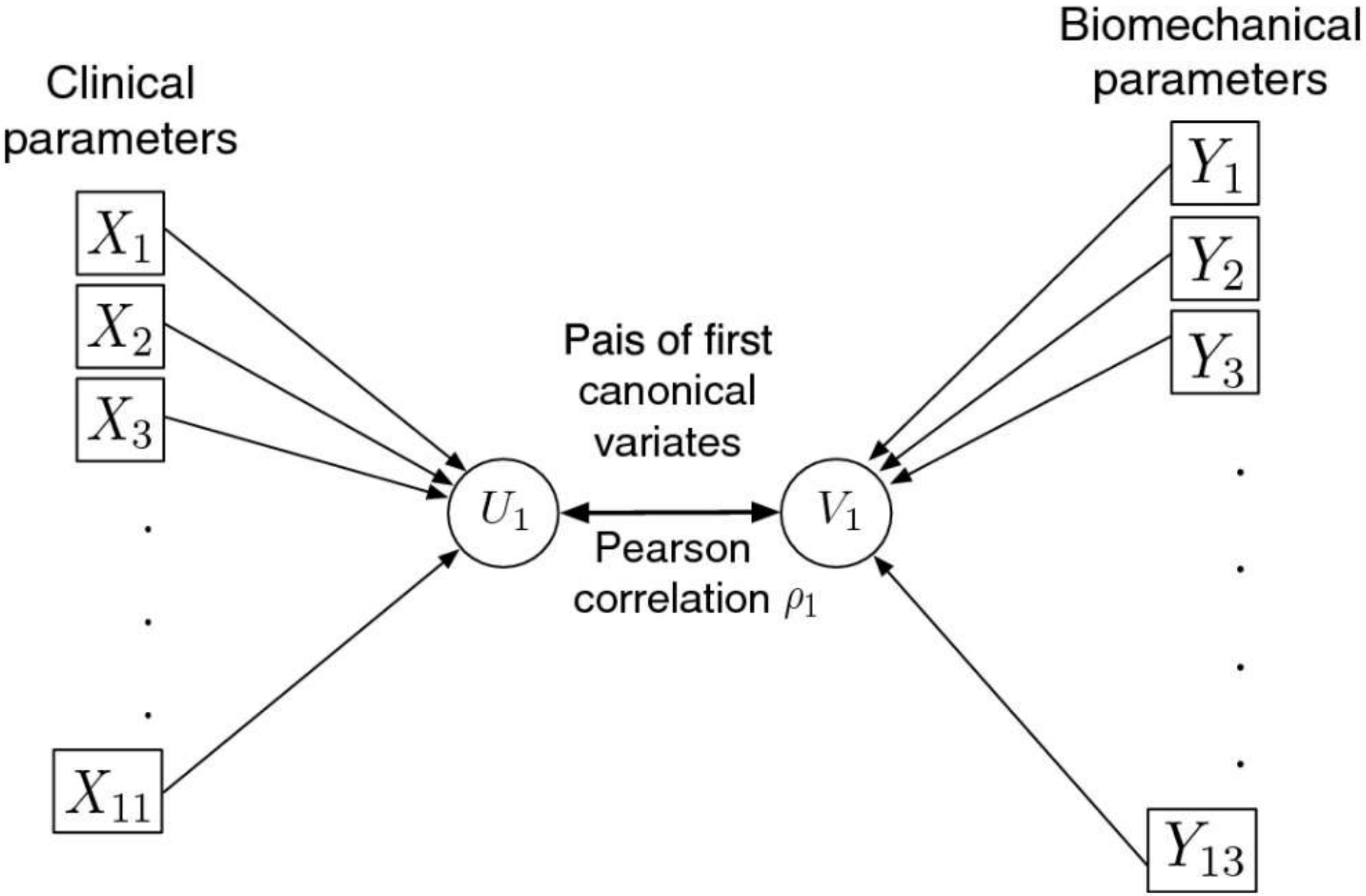
- 顔のどの部分を、(遺伝(SNP,民族性),性,年齢,BMI)の組がどのように説明するかを解析
- 顔貌側：多変数
- 遺伝要因等：多変数
- 多変数 vs. 多変数　～　行列 vs. 行列
  - 正準相関解析
    - 顔貌に影響するSNP/遺伝子を同定



- 多数のSNP
- 個人間距離
- 主成分分析

- 多民族顔貌コホートの場合
  - 987 成分
  - 個別SNPの寄与は考えない
- EUROコホートの場合
  - 4成分で民族性を決め
  - 顔貌関連SNPを検出

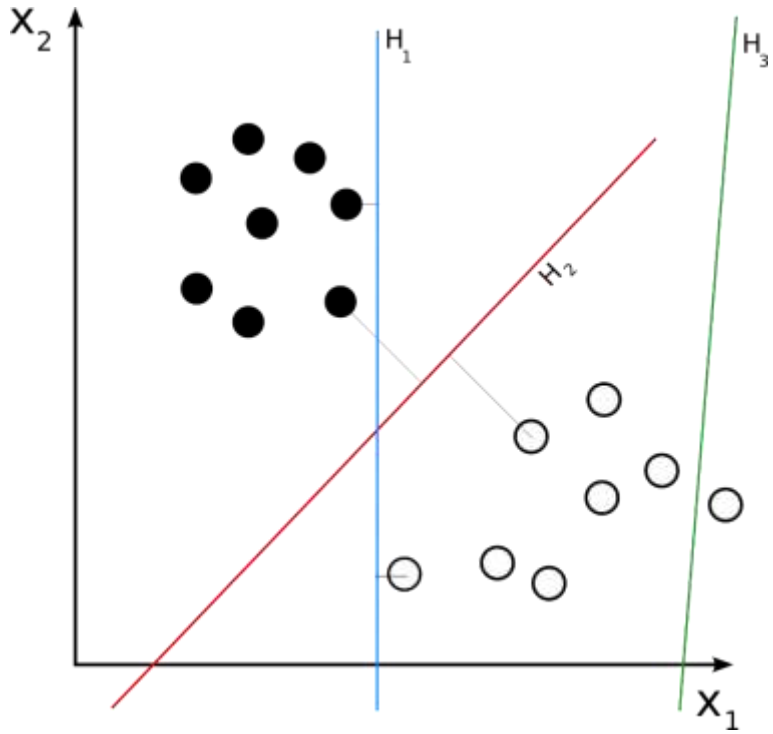
- 主成分分析(PCA)は１つの行列を回転して、説明軸を選ぶ
- 正準相関解析は、２つの行列を両方とも回して、相関係数が高い軸が選ばれるように説明軸を選ぶ

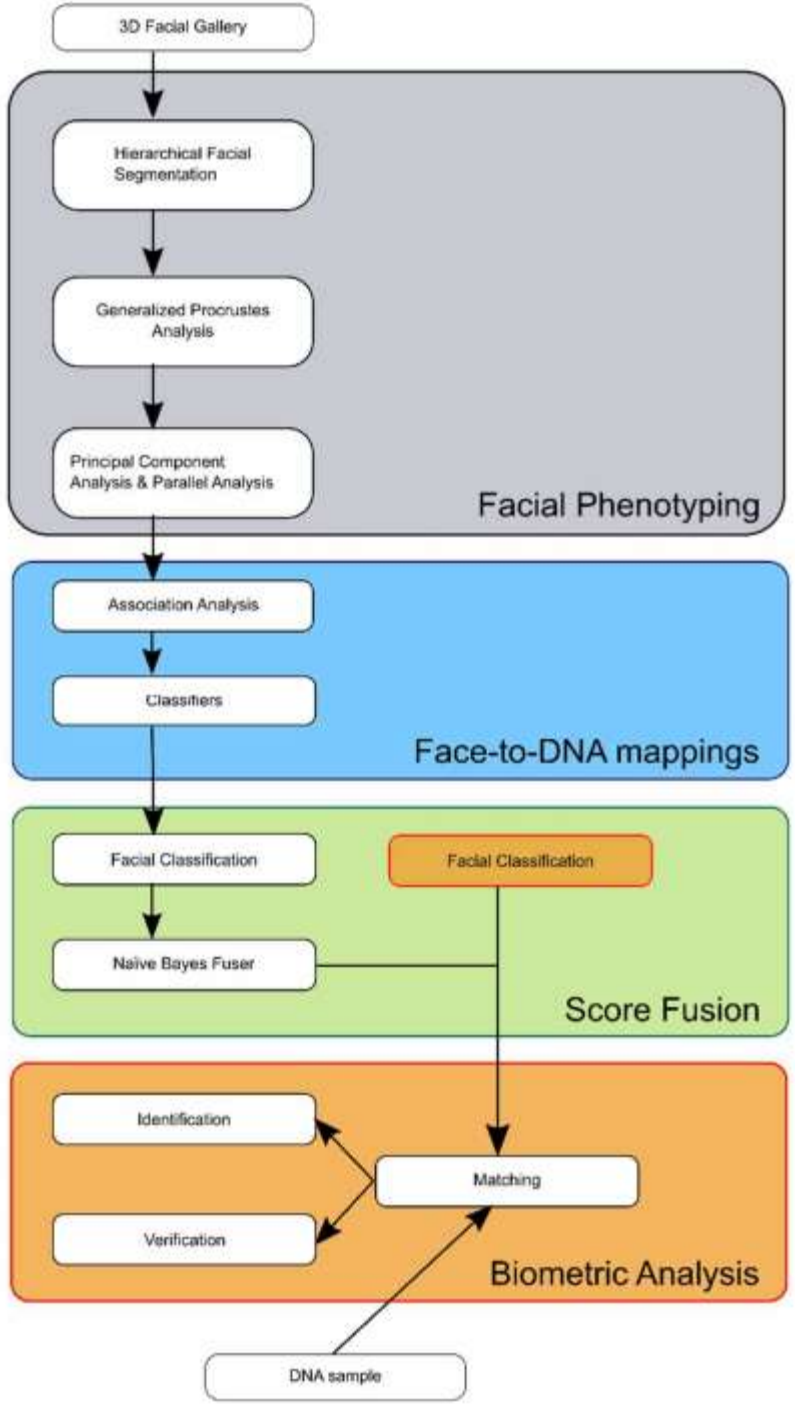


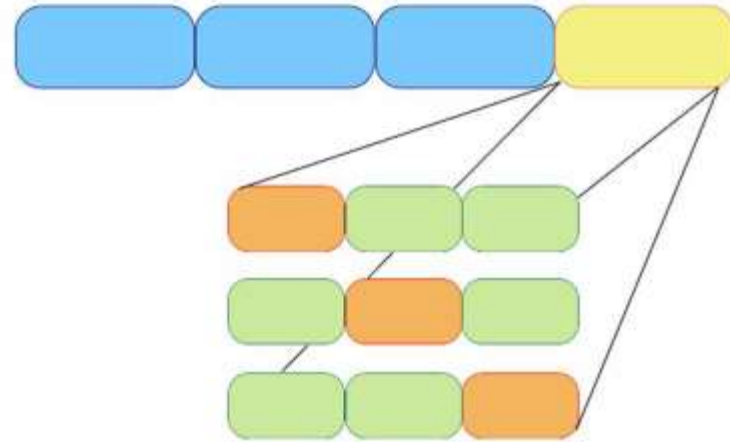
https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-019-0676-8

顔貌識別モデルの機械学習

- サポートベクターマシン
  - 性別、年齢、BMI、民族性（多数のSNPからの遺伝的民族指標）、SNP
  - 各指標を2値化して、2^kクラスに分類
  - 指標組み合わせでの、適切な閾値探索
  - 指標によるクラス分類の精度を評価







GLOBAL Cohort

= COMPLETE SET, 3,295

= TRAINING SET, 2,471 participants

= REMAINING SET, 824 participants

= VALIDATION SET, 549 participants

= TEST SET, 275 participants

EURO Cohort

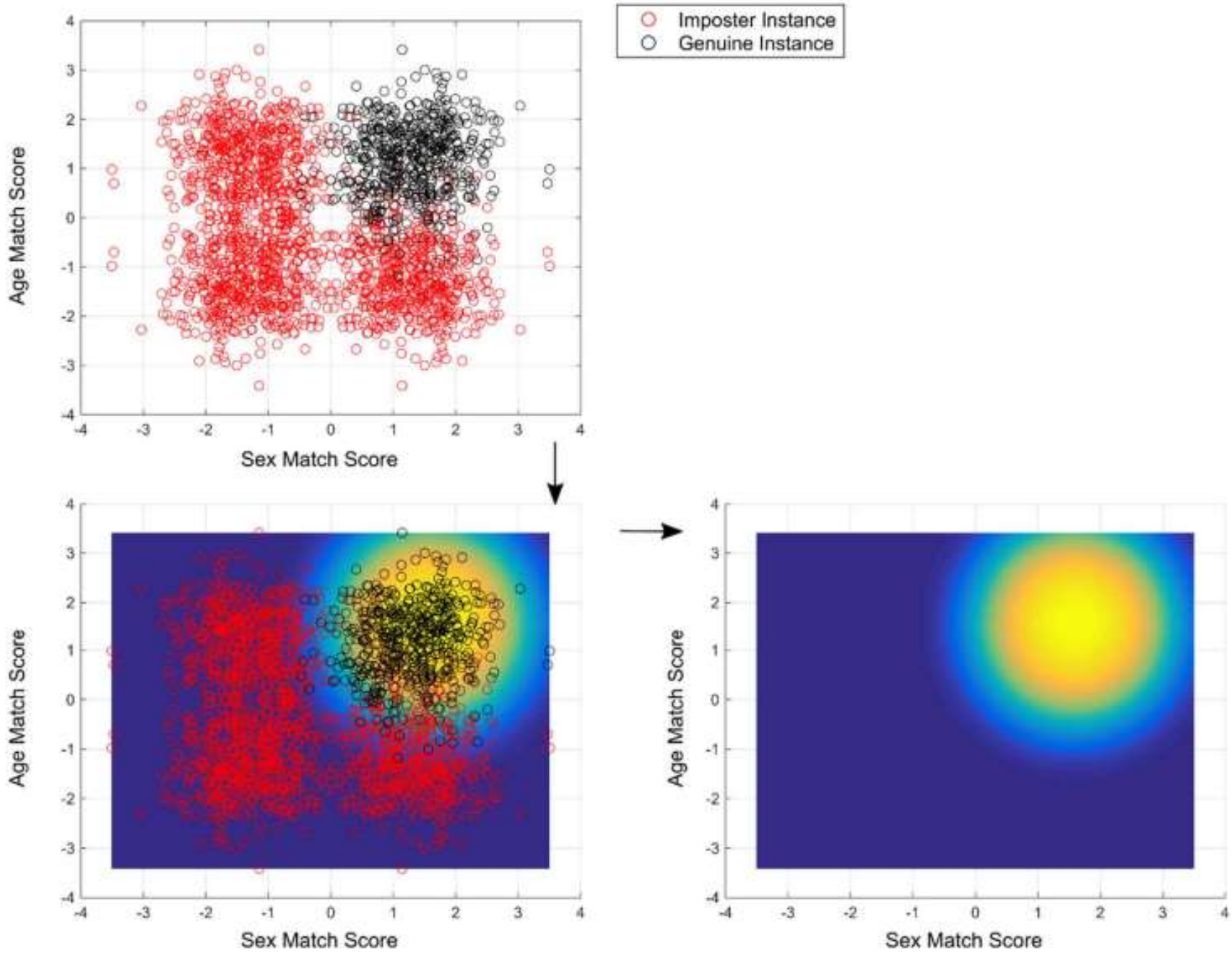
= COMPLETE SET, 3,542

= TRAINING SET, 2,656 participants

= REMAINING SET, 886 participants

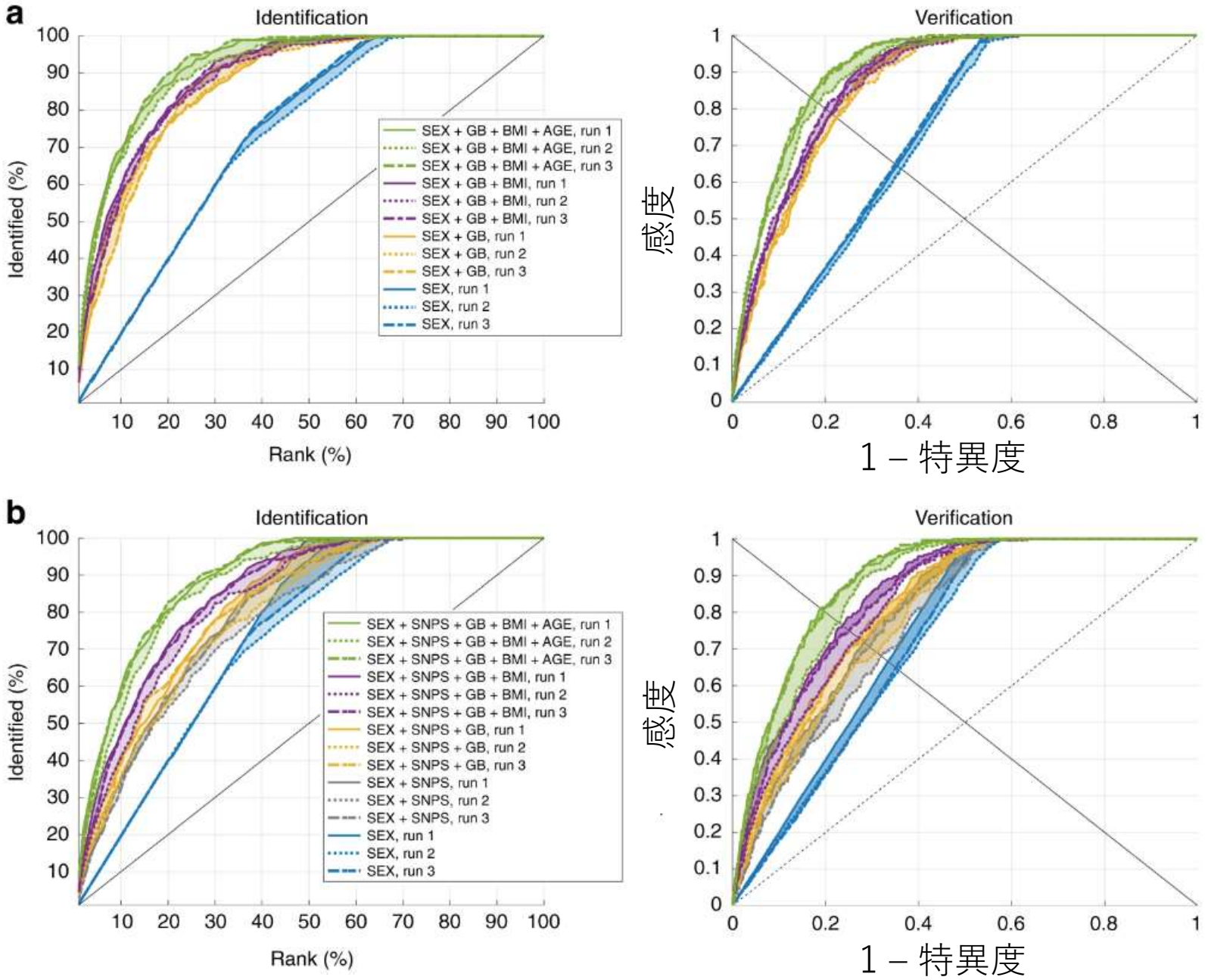
= VALIDATION SET, 591 participants

= TEST SET, 295 participants


Imposter Instance
Genuine Instance
Age Match Score
Sex Match Score
Age Match Score
Sex Match Score
Age Match Score
Sex Match Score

分類器のパフォーマンス

From: Facial recognition from DNA using face-to-DNA classifiers



まとめ

- 顔貌を決める遺伝子・遺伝子多型がある
- 遺伝情報から民族性も抽出できる。多民族集団の場合は、その民族性も顔貌判別情報として使える
- 感度・特異度が、まあまあ、というレベル（個人の特定、というのは、超高感度・超高特異度だが）

補助

Genotype Imputation

Genotype data in the PITT and PSU samples, separately, were imputed using the 1000 Genomes Project Phase 3 reference panel[29]. SHAPEIT2 was used for pre-phasing haplotypes and IMPUTE2 was used to impute nearly 35 M variants. SNP-level (INFO score > 0.5) and genotype per participant-level (genotype probability > 0.9) filters were used to omit poorly-imputed variants. A further reduction of SNPs with MAF <5% was performed, before aligning allele encodings in both datasets with the 1000 Genomes Project and to merge them into a single cohort of European ancestry.

Population Structure

Population structure was assessed using PCA of approximately 100 K autosomal SNPs chosen for call rate (>95%), MAF (>5%), and LD (pairwise r2 < 0.1 across variants in a sliding window of 10 Mb). Tests of genetic association between the first 20 PCs and all SNPs confirmed that PCs did not represent local variation at specific genetic loci. Based on Supplementary Fig. 2, four PCs were sufficient to capture population structure within this European-derived cohort towards the purpose of a GWAS.