

# 1 **GestaltMatcher Database - A global reference for facial** 2 **phenotypic variability in rare human diseases**

3 Hellen Lesmann<sup>1,2\*</sup>, Alexander Hustinx<sup>2\*</sup>, Shahida Moosa<sup>3</sup>, Hannah Klinkhammer<sup>2,4</sup>,  
4 Elaine Marchi<sup>5</sup>, Pilar Caro<sup>6</sup>, Ibrahim M. Abdelrazek<sup>7</sup>, Jean Tori Pantel<sup>8,9</sup>, Merle ten  
5 Hagen<sup>2</sup>, Meow-Keong Thong<sup>10</sup>, Rifhan Azwani Binti Mazlan<sup>10</sup>, Sok Kun Tae<sup>10</sup>, Tom  
6 Kamphans<sup>11</sup>, Wolfgang Meiswinkel<sup>11</sup>, Jing-Mei Li<sup>2</sup>, Behnam Javanmardi<sup>2</sup>, Alexej  
7 Knaus<sup>2</sup>, Annette Uwineza<sup>12</sup>, Cordula Knopp<sup>13</sup>, Tinatin Tkemaladze<sup>14,15</sup>, Miriam  
8 Elbracht<sup>13</sup>, Larissa Mattern<sup>13</sup>, Rami Abou Jamra<sup>16</sup>, Clara Velmans<sup>17</sup>, Vincent  
9 Strehlow<sup>16</sup>, Maureen Jacob<sup>18</sup>, Angela Peron<sup>19,20</sup>, Cristina Dias<sup>21,22,23,24</sup>, Beatriz  
10 Carvalho Nunes<sup>25</sup>, Thainá Vilella<sup>25</sup>, Isabel Furquim Pinheiro<sup>26</sup>, Chong Ae Kim<sup>26</sup>,  
11 Maria Isabel Melaragno<sup>25</sup>, Hannah Weiland<sup>2</sup>, Sophia Kaptain<sup>2</sup>, Karolina  
12 Chwiałkowska<sup>27,28</sup>, Mirosław Kwasniewski<sup>28,27</sup>, Ramy Saad<sup>22,29</sup>, Sarah Wiethoff<sup>30</sup>,  
13 Himanshu Goel<sup>31</sup>, Clara Tang<sup>32</sup>, Anna Hau<sup>33</sup>, Tahsin Stefan Barakat<sup>34</sup>, Przemysław  
14 Panek<sup>35</sup>, Amira Nabil<sup>7</sup>, Julia Suh<sup>13</sup>, Frederik Braun<sup>36</sup>, Israel Gomy<sup>37</sup>, Luisa  
15 Averdunk<sup>38</sup>, Ekanem Ekure<sup>39</sup>, Gaber Bergant<sup>40</sup>, Borut Peterlin<sup>41</sup>, Claudio Graziano<sup>42</sup>,  
16 Nagwa Gaboon<sup>43,44</sup>, Moisés Fiesco-Roa<sup>45,46</sup>, Alessandro Mauro Spinelli<sup>47</sup>, Nina-  
17 Maria Wilpert<sup>48,49,50</sup>, Prasit Phowthongkum<sup>51,52</sup>, Nergis Güzel<sup>13</sup>, Tobias B. Haack<sup>53</sup>,  
18 Rana Bitar<sup>54,55</sup>, Andreas Tzschach<sup>56</sup>, Agusti Rodriguez-Palmero<sup>57</sup>, Theresa Brunet<sup>18</sup>,  
19 Sabine Rudnik-Schöneborn<sup>58</sup>, Silvina Noemi Contreras-Capetillo<sup>59</sup>, Ava Oberlack<sup>18</sup>,  
20 Carole Samango-Sprouse<sup>60,61,62</sup>, Teresa Sadeghin<sup>63</sup>, Margaret Olaya<sup>63</sup>, Konrad  
21 Platzer<sup>16</sup>, Artem Borovikov<sup>64</sup>, Franziska Schnabel<sup>16</sup>, Lara Heuft<sup>16</sup>, Vera Herrmann<sup>16</sup>,  
22 Renske Oegema<sup>65</sup>, Nour Elkhateeb<sup>66</sup>, Sheetal Kumar<sup>1</sup>, Katalin Komlosi<sup>56</sup>,  
23 Khoushoua Mohamed<sup>7</sup>, Silvia Kalantari<sup>67</sup>, Fabio Sirchia<sup>67,68</sup>, Antonio F. Martinez-  
24 Monseny<sup>69</sup>, Matthias Höller<sup>56</sup>, Louiza Toutouna<sup>56</sup>, Amal Mohamed<sup>7</sup>, Amaia Las-  
25 Aranzasti<sup>70,71</sup>, John A. Sayer<sup>72,73</sup>, Nadja Ehmke<sup>74</sup>, Magdalena Danyel<sup>74</sup>, Henrike  
26 Sczakiel<sup>74</sup>, Sarina Schwartzmann<sup>74</sup>, Felix Boschann<sup>74</sup>, Max Zhao<sup>74</sup>, Ronja Adam<sup>74</sup>,  
27 Lara Einicke<sup>74</sup>, Denise Horn<sup>74</sup>, Kee Seang Chew<sup>75</sup>, Choy Chen KAM<sup>75</sup>, Miray  
28 Karakoyun<sup>76</sup>, Ben Pode-Shakked<sup>77,78</sup>, Aviva Eliyahu<sup>79,80,81</sup>, Rachel Rock<sup>82,83</sup>, Teresa  
29 Carrion<sup>84</sup>, Odelia Chorin<sup>85</sup>, Yuri A. Zarate<sup>86,87</sup>, Marcelo Martinez Conti<sup>88</sup>, Mert  
30 Karakaya<sup>17</sup>, Moon Ley Tung<sup>89,90</sup>, Bharatendu Chandra<sup>89,90</sup>, Arjan Bouman<sup>34</sup>, Aime  
31 Lumaka<sup>91</sup>, Naveed Wasif<sup>92,93</sup>, Marwan Shinawi<sup>94</sup>, Patrick R. Blackburn<sup>95</sup>, Tianyun  
32 Wang<sup>96,97,98</sup>, Tim Niehues<sup>99</sup>, Axel Schmidt<sup>1</sup>, Regina Rita Roth<sup>100</sup>, Dagmar  
33 Wieczorek<sup>100</sup>, Ping Hu<sup>101</sup>, Rebekah L. Waikel<sup>101</sup>, Suzanna E. Ledgister Hanchard<sup>101</sup>,  
34 Gehad Elmakkawy<sup>7</sup>, Sylvia Safwat<sup>7</sup>, Frédéric Ebstein<sup>102,103</sup>, Elke Krüger<sup>104</sup>,  
35 Sébastien Küry<sup>102,103</sup>, Stéphane Bézieau<sup>102,103</sup>, Annabelle Arlt<sup>2</sup>, Eric Olinger<sup>105</sup>, Felix  
36 Marbach<sup>6</sup>, Dong Li<sup>106</sup>, Lucie Dupuis<sup>107</sup>, Roberto Mendoza-Londono<sup>107</sup>, Sofia  
37 Douzgou Houge<sup>108</sup>, Denisa Weis<sup>109</sup>, Brian Hon-Yin Chung<sup>110,111</sup>, Christopher C.Y.  
38 Mak<sup>111</sup>, Hülya Kayserili<sup>112</sup>, Nursel Elcioglu<sup>113</sup>, Ayca Aykut<sup>114</sup>, Peli Özlem Şimşek-  
39 Kiper<sup>115</sup>, Nina Bögershausen<sup>116</sup>, Bernd Wollnik<sup>116,117,118</sup>, Heidi Beate Bentzen<sup>119,120</sup>,  
40 Ingo Kurth<sup>13</sup>, Christian Netzer<sup>17</sup>, Aleksandra Jezela-Stanek<sup>35</sup>, Koen Devriendt<sup>121</sup>,  
41 Karen W. Gripp<sup>122</sup>, Martin Mücke<sup>8,9</sup>, Alain Verloes<sup>123</sup>, Christian P. Schaaf<sup>6</sup>,

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

42 Christoffer Nellåker<sup>124</sup>, Benjamin D. Solomon<sup>101</sup>, Markus M. Nöthen<sup>1</sup>, Ebtesam  
43 Abdalla<sup>7</sup>, Gholson J. Lyon<sup>125,126,127</sup>, Peter M. Krawitz<sup>2</sup>, Tzung-Chien Hsieh<sup>2#</sup>

44  
45

46 <sup>1</sup>Institute of Human Genetics, University of Bonn, Bonn, NRW, Germany, <sup>2</sup>Institute  
47 for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, NRW,  
48 Germany, <sup>3</sup>Division of Molecular Biology and Human Genetics, Stellenbosch  
49 University and Medical Genetics, Tygerberg Hospital, Stellenbosch, South Africa,  
50 <sup>4</sup>Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn,  
51 Bonn, NRW, Germany, <sup>5</sup>New York State Institute for Basic Research in  
52 Developmental Disabilities, New York State, Albany, New York, USA, <sup>6</sup>Institute of  
53 Human Genetics, Heidelberg University, Heidelberg, Baden-Württemberg, Germany,  
54 <sup>7</sup>Department of Human Genetics, Medical Research Institute, Alexandria University,  
55 Alexandria, Alexandria, Egypt, <sup>8</sup>Institute for Digitalization and General Medicine,  
56 University Hospital RWTH Aachen, Aachen, NRW, Germany, <sup>9</sup>Centre for Rare  
57 Diseases Aachen (ZSEA), University Hospital RWTH Aachen, Aachen, NRW,  
58 Germany, <sup>10</sup>Department of Paediatrics, Faculty of Medicine, University of Malaya,  
59 50603 Kuala Lumpur, Malaysia, <sup>11</sup>GeneTalk GmbH, Bonn, NRW, Germany,  
60 <sup>12</sup>College of Medicine and Health Sciences, University of Rwanda, and University  
61 Teaching Hospital of Kigali, Kigali, Rwanda, <sup>13</sup>Institute for Human Genetics and  
62 Genomic Medicine, Medical Faculty, RWTH Aachen University, Aachen, NRW,  
63 Germany, <sup>14</sup>Department of Molecular and Medical Genetics, Tbilisi State Medical  
64 University, Tbilisi, Georgia, <sup>15</sup>Givi Zhvania Pediatric Academic Clinic, Tbilisi State  
65 Medical University, Georgia, <sup>16</sup>Institute of Human Genetics, University of Leipzig  
66 Medical Center, Leipzig, Germany, <sup>17</sup>Institute of Human Genetics, University of  
67 Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, NRW,  
68 Germany, <sup>18</sup>Institute of Human Genetics, Klinikum rechts der Isar, Technical  
69 University of Munich, School of Medicine and Health, Munich, Germany, <sup>19</sup>Medical  
70 Genetics, Meyer Children's Hospital IRCCS, Firenze, Italy, <sup>20</sup>Department of  
71 Experimental and Clinical Biomedical Sciences "Mario Serio", Università degli Studi  
72 di Firenze, Italy, <sup>21</sup>Department of Medical Genetics, Guy's and St. Thomas' NHS  
73 Foundation Trust, London, UK, <sup>22</sup>North East Thames Regional Genetics Service,  
74 Great Ormond Street Hospital for Children, Great Ormond Street, London, UK,  
75 <sup>23</sup>Neural Stem Cell Biology Laboratory, The Francis Crick Institute, UK, <sup>24</sup>Department  
76 of Medical & Molecular Genetics, School of Basic and Medical Biosciences, Faculty  
77 of Life Sciences & Medicine, King's College London, UK, <sup>25</sup>Genetics Division,  
78 Department of Morphology and Genetics, Universidade Federal de São Paulo, São  
79 Paulo, Brazil, <sup>26</sup>Genetics Unit, Instituto da Criança, Universidade de São Paulo, São  
80 Paulo, Brazil, <sup>27</sup>Centre for Bioinformatics and Data Analysis, Medical University of  
81 Bialystok, Bialystok, Poland, <sup>28</sup>IMAGENE.ME SA, Bialystok, Poland, <sup>29</sup>Department of  
82 Genetics and Genomic Medicine, UCL Institute of Child Health, London UK,  
83 <sup>30</sup>Department of Neurology with Institute of Translational Neurology, University  
84 Hospital Münster, Münster, NRW, Germany, <sup>31</sup>School of Medicine and Public Health,  
85 University of Newcastle, Callaghan NSW, Australia, <sup>32</sup>Kabuki Syndrome Foundation,

86 Northbrook, IL, USA, <sup>33</sup>Hunter Genetics, Hunter New England Health Service,  
87 Newcastle, Australia, <sup>34</sup>Department of Clinical Genetics, Erasmus MC University  
88 Medical Center, Rotterdam, The Netherlands, <sup>35</sup>Department of Genetics and Clinical  
89 Immunology, National Institute of Tuberculosis and Lung Diseases, Warsaw, Poland,  
90 <sup>36</sup>Institute of Human Genetics, University Hospital Essen, University Duisburg-Essen,  
91 Essen, NRW, Germany, <sup>37</sup>Department of Genetics, Faculdade de Medicina de  
92 Ribeirão Preto, Universidade de São Paulo, Sao Paulo, Brazil, <sup>38</sup>Department of  
93 General Pediatrics and Neonatology, University Children's Hospital, Heinrich-Heine-  
94 University, Medical Faculty, Düsseldorf, Germany, <sup>39</sup>Department of Paediatrics,  
95 College of Medicine, University of Lagos, Lagos, Nigeria, <sup>40</sup>Clinical Institute of  
96 Genomic Medicine, University Medical Centre Ljubljana, Ljubljana, Slovenia,  
97 <sup>41</sup>Clinical Institute of Genomic Medicine, University Medical Centre Ljubljana,  
98 <sup>42</sup>Medical Genetics Unit, Ausl Romagna, Cesena, Italy, <sup>43</sup>Medical Genetics Center,  
99 Faculty of Medicine, Ain Shams University, Cairo, Egypt, <sup>44</sup>Medical Genetics  
100 Department, Armed Forces College of Medicine, Cairo, Egypt, <sup>45</sup>Programa de  
101 Maestría y Doctorado en Ciencias Médicas, Odontológicas y de la Salud,  
102 Universidad Nacional Autónoma de México, México City, Mexico, <sup>46</sup>Laboratorio de  
103 Citogenética, Instituto Nacional de Pediatría, México City, Mexico, <sup>47</sup>Institute for  
104 Maternal and Child Health, IRCCS Burlo Garofolo, Trieste, Italy, <sup>48</sup>NeuroCure Cluster  
105 of Excellence; Charité–Universitätsmedizin Berlin, corporate member of Freie  
106 Universität Berlin and Humboldt-Universität zu Berlin, D-10117 Berlin, Germany,  
107 <sup>49</sup>Department of Neuropediatrics, Charité–Universitätsmedizin Berlin, corporate  
108 member of Freie Universität Berlin and Humboldt-Universität zu Berlin, D-13353  
109 Berlin, Germany, <sup>50</sup>Berlin Institute of Health at Charité - Universitätsmedizin Berlin,  
110 BIH Biomedical Innovation Academy, BIH Charité Junior Clinician Scientist Program,  
111 D-10117 Berlin, German, <sup>51</sup>Excellence Center for Genomics and Precision Medicine,  
112 King Chulalongkorn Memorial Hospital, the Thai Red Cross Society, Bangkok,  
113 Thailand, <sup>52</sup>Division of Medical Genetics and Genomics, Department of Medicine,  
114 Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand, <sup>53</sup>Institute of  
115 Medical Genetics and Applied Genomics, University of Tübingen, Tübingen,  
116 Germany, <sup>54</sup>Pediatric Gastroenterology Department, Sheikh Khalifa Medical City,  
117 Abu Dhabi, United Arab Emirates, <sup>55</sup>Khalifa University, Abu Dhabi, United Arab  
118 Emirates, <sup>56</sup>Institute of Human Genetics, Medical Center - University of Freiburg,  
119 Faculty of Medicine, University of Freiburg, Freiburg, Germany, <sup>57</sup>Paediatric  
120 Neurology Unit, Department of Pediatrics, Hospital Universitari Germans Trias i  
121 Pujol, Universitat Autònoma de Barcelona, Barcelona, Spain, <sup>58</sup>Institute of Human  
122 Genetics, Medical University Innsbruck, Innsbruck, Austria, <sup>59</sup>Universidad Autónoma  
123 de Yucatan (University Autonomus of Yucatan), Merida, Yucatan, Mexico,  
124 <sup>60</sup>Department of Pediatrics, George Washington University, 2121 I St. NW,  
125 Washington D.C. 2005, <sup>61</sup>Department of Human and Molecular Genetics, Florida  
126 International University, 11200 SW 8th Street, AHC2 Miami, Florida 22199,  
127 <sup>62</sup>Department of Research, The Focus Foundation, 820 W. Central Ave. #190,  
128 Davidsonville, MD 21035, <sup>63</sup>Department of Research, The Focus Foundation, 2772  
129 Rutland Road P.O. Box 190, Davidsonville, MD 21035, <sup>64</sup>Research Centre for

130 Medical Genetics (RCMG), Moscow, Russia, <sup>65</sup>Department of Genetics, University  
131 Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands, <sup>66</sup>Department  
132 of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust,  
133 Cambridge, UK, <sup>67</sup>Department of Molecular Medicine, University of Pavia, Pavia,  
134 Italy, <sup>68</sup>Medical Genetics Unit, IRCCS San Matteo Foundation, Pavia, Italy,  
135 <sup>69</sup>Department of Clinical Genetics, SJD Barcelona Children's Hospital, Esplugues del  
136 Llobregat (Barcelona), Spain, <sup>70</sup>Medicine Genetics Group, Vall d'Hebron Institut de  
137 Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Hospital  
138 Universitari, Barcelona, Catalunya, Spain, <sup>71</sup>Department of Clinical and Molecular  
139 Genetics, Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Hospital  
140 Universitari, Barcelona, Catalunya, Spain, <sup>72</sup>Biosciences Institute, Newcastle  
141 University, Central Parkway, Newcastle upon Tyne, UK, <sup>73</sup>Renal Services, The  
142 Newcastle Upon Tyne NHS Hospitals Foundation Trust, Freeman Road, Newcastle  
143 Upon Tyne, UK, <sup>74</sup>Institute of Medical Genetics and Human Genetics, Charité-  
144 Universitätsmedizin Berlin, Humboldt-Universität zu Berlin and Berlin Institute of  
145 Health, Berlin, Germany, <sup>75</sup>Department of Paediatrics, Faculty of Medicine,  
146 University Malaya, 59100 Kuala Lumpur, Malaysia, <sup>76</sup>Ege University, Faculty of  
147 Medicine, Department of Pediatric Gastroenterology Hepatology and Nutrition, Izmir,  
148 Turkey, <sup>77</sup>The Institute of Rare Diseases, Edmond and Lily Safra Children's Hospital,  
149 Sheba Medical Center, Ramat Gan, Israel, <sup>78</sup>The faculty of Medical and Health  
150 Sciences, Tel-Aviv University, Tel-Aviv, Israel, <sup>79</sup>The Danek Gertner Institute of  
151 Human Genetics, Sheba Medical Center, Tel-Hashomer, Israel, <sup>80</sup>Sackler Faculty of  
152 Medicine, Tel-Aviv University, Tel-Aviv, Israel, <sup>81</sup>The Mina and Everard Goodman  
153 Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, <sup>82</sup>Metabolic  
154 Diseases Clinic, Edmond and Lily Safra Children's Hospital, Sheba Medical Center,  
155 <sup>83</sup>National Newborn Screening Program, Public Health Services, Ministry of Health  
156 Tel-Hashomer, Israel, <sup>84</sup>Rare diseases Unit, Pediatric Department, Hospital  
157 Universitari Son Espases, Palma de Mallorca, Spain, <sup>85</sup>The Institute of Rare  
158 Diseases, Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Tel-  
159 Hashomer, Israel, <sup>86</sup>Department of Pediatrics, Section of Genetics and Metabolism,  
160 University of Arkansas for Medical Sciences and Arkansas Children's Hospital, Little  
161 Rock, AR, USA, <sup>87</sup>Division of Genetics and Metabolism, University of Kentucky,  
162 Lexington, KY, USA, <sup>88</sup>Project Director AI for Health at Foundation 29, Foundation  
163 29, Madrid, Spain, <sup>89</sup>University of Iowa Roy J and Lucille A Carver College of  
164 Medicine, Iowa City, IA 52242, USA, <sup>90</sup>Division of Medical Genetics and Genomics,  
165 Stead Family Department of Pediatrics, University of Iowa Hospitals and Clinics,  
166 Iowa City, IA 52242, USA, <sup>91</sup>Center for Human Genetics, Faculty of Medicine,  
167 University of Kinshasa, Kinshasa, DR Congo, <sup>92</sup>Institute of Human Genetics,  
168 University of Ulm, Ulm, Baden-Württemberg, Germany, <sup>93</sup>University Hospital  
169 Schleswig-Holstein, Campus Kiel, Kiel, Germany, <sup>94</sup>Division of Genetics and  
170 Genomic Medicine, Department of Pediatrics, Washington University School of  
171 Medicine, St. Louis, MO, USA, <sup>95</sup>Department of Pathology, St. Jude Children's  
172 Research Hospital, Memphis, Tennessee 38105, USA, <sup>96</sup>Department of Medical  
173 Genetics, Center for Medical Genetics, Peking University Health Science Center,



174 Beijing 100191, China, <sup>97</sup>Neuroscience Research Institute, Peking University; Key  
175 Laboratory for Neuroscience, Ministry of Education of China & National Health  
176 Commission of China, Beijing 100191, China, <sup>98</sup>Autism Research Center, Peking  
177 University Health Science Center, Beijing 100191, China, <sup>99</sup>Department of Pediatrics,  
178 Helios Klinik Krefeld, Krefeld 47805, Germany, <sup>100</sup>Institute of Human Genetics,  
179 Medical Faculty, University Hospital Düsseldorf, Heinrich Heine University  
180 Düsseldorf, Germany, <sup>101</sup>Medical Genomics Unit, Medical Genetics Branch, National  
181 Human Genome Research Institute, Bethesda, USA, <sup>102</sup>Nantes Université, CHU  
182 Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France, <sup>103</sup>Nantes  
183 Université, CHU Nantes, Service de Génétique Médicale, F-44000 Nantes, France,  
184 <sup>104</sup>Institute for Medical Biochemistry and Molecular Biology, University of Greifswald,  
185 Greifswald, Greifswald, Germany, <sup>105</sup>Center for Human Genetics, Cliniques  
186 Universitaires Saint-Luc, Brussels, Belgium, <sup>106</sup>Division of Human Genetics,  
187 Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, <sup>107</sup>Department  
188 to Paediatrics, Division of Clinical and Metabolic Genetics, The Hospital of Sick  
189 Children, Toronto, Ontario, Canada, <sup>108</sup>Department of Medical Genetics, Haukeland  
190 University Hospital, Bergen, Norway, <sup>109</sup>Institute for Medical Genetics, Kepler  
191 University Hospital, Linz, Austria, <sup>110</sup>Hong Kong Genome Institute, Hong Kong,  
192 China, <sup>111</sup>Department of Paediatrics and Adolescent Medicine, The University of  
193 Hong Kong, Hong Kong, China, <sup>112</sup>Medical Genetics Department, Koç University  
194 School of Medicine (KUSoM), 34010, Istanbul, Türkiye, <sup>113</sup>Department of Pediatric  
195 Genetics, Marmara University School of Medicine, Istanbul, Türkiye, <sup>114</sup>Department  
196 of Medical Genetics, Ege University Faculty of Medicine, Izmir, Türkiye, <sup>115</sup>Hacettepe  
197 University Faculty of Medicine, Department of Pediatric Genetics, Ankara, Türkiye,  
198 <sup>116</sup>Institut of Human Genetics, University Medical Center Göttingen, Göttingen,  
199 Germany, <sup>117</sup>Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines  
200 to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen,  
201 Germany, <sup>118</sup>German Center for Cardiovascular Research (DZHK), Partner Site  
202 Göttingen, Göttingen, Germany, <sup>119</sup>Centre for Medical Ethics, Faculty of Medicine,  
203 University of Oslo, Oslo, Norway, <sup>120</sup>Cancer Registry of Norway, Norwegian Institute  
204 of Public Health, Oslo, Norway, <sup>121</sup>Center for Human Genetics, KU Leuven, Leuven,  
205 Belgium, <sup>122</sup>Division of Medical Genetics, A.I. du Pont Hospital for Children/Nemours,  
206 USA, Wilmington, Delaware, USA, <sup>123</sup>Department of Clinical Genetics, Robert-Debré  
207 Hospital, Paris, France, <sup>124</sup>Big Data Institute, Li Ka Shing Centre for Health  
208 Information and Discovery, Nuffield Department of Women's & Reproductive Health,  
209 University of Oxford, Oxford, UK, <sup>125</sup>Department of Human Genetics, New York State  
210 Institute for Basic Research in Developmental Disabilities, Staten Island, New York,  
211 United States of America, <sup>126</sup>George A. Jervis Clinic, New York State Institute for  
212 Basic Research in Developmental Disabilities, Staten Island, New York, United  
213 States of America, <sup>127</sup>Biology PhD Program, The Graduate Center, The City  
214 University of New York, New York, United States of America  
215

216 \*These authors contributed equally

217 #Corresponding author

## 218 Abstract

219 The most important factor that complicates the work of dysmorphologists is the  
220 significant phenotypic variability of the human face. Next-Generation Phenotyping  
221 (NGP) tools that assist clinicians with recognizing characteristic syndromic patterns  
222 are particularly challenged when confronted with patients from populations different  
223 from their training data. To that end, we systematically analyzed the impact of genetic  
224 ancestry on facial dysmorphism. For that purpose, we established the GestaltMatcher  
225 Database (GMDB) as a reference dataset for medical images of patients with rare  
226 genetic disorders from around the world. We collected 10,980 frontal facial images –  
227 more than a quarter previously unpublished - from 8,346 patients, representing 581  
228 rare disorders. Although the predominant ancestry is still European (67%), data from  
229 underrepresented populations have been increased considerably via global  
230 collaborations (19% Asian and 7% African). This includes previously unpublished  
231 reports for more than 40% of the African patients. The NGP analysis on this diverse  
232 dataset revealed characteristic performance differences depending on the  
233 composition of training and test sets corresponding to genetic relatedness. For clinical  
234 use of NGP, incorporating non-European patients resulted in a profound enhancement  
235 of GestaltMatcher performance. The top-5 accuracy rate increased by +11.29%.  
236 Importantly, this improvement in delineating the correct disorder from a facial portrait  
237 was achieved without decreasing the performance on European patients. By design,  
238 GMDB complies with the FAIR principles by rendering the curated medical data  
239 findable, accessible, interoperable, and reusable. This means GMDB can also serve  
240 as data for training and benchmarking. In summary, our study on facial dysmorphism  
241 on a global sample revealed a considerable cross ancestral phenotypic variability  
242 confounding NGP that should be counteracted by international efforts for increasing  
243 data diversity. GMDB will serve as a vital reference database for clinicians and a  
244 transparent training set for advancing NGP technology.

## 245 Introduction

246 Facial dysmorphism is one of the most complex and informative clinical features in  
247 syndromic disorders, and is therefore often crucial in terms of establishing a diagnosis

248 in rare genetic diseases<sup>1,2</sup>. However, the recognition of dysmorphic patterns, is a  
249 challenging endeavour, and relies on the skills, knowledge, and experience of the  
250 examiner. In certain syndromes, in particular those that are ultra-rare, variability in  
251 facial features can pose challenges even for highly experienced clinicians<sup>3</sup>. Facial  
252 features can also vary according to sex, age, and ancestry, which further complicates  
253 the recognition of a specific dysmorphic pattern<sup>4–6</sup>.

254 Ancestry plays a particularly significant role since considerable inter-ancestral  
255 variability exists in facial gestalt<sup>7</sup>. Thus, facial features that are common in certain  
256 ancestral groups may be considered dysmorphic in others. For example, while  
257 upslanting palpebral fissures are common in healthy Asians, they may be perceived  
258 as dysmorphic in other populations<sup>8</sup>. Previous studies have also highlighted  
259 differences in facial gestalt between different ancestries in common dysmorphic  
260 genetic syndromes such as Down Syndrome, 22q11.2 deletion syndrome, Noonan  
261 syndrome, and Williams–Beuren syndrome<sup>4,9,10</sup>. Furthermore, Lumaka et al. have  
262 demonstrated that this variability can influence the assessor, with European clinicians  
263 failing to recognize dysmorphic features in individuals of African ancestry<sup>11</sup>. This is a  
264 growing problem as globalization and migration increasingly blur ancestral and cultural  
265 boundaries, and geography is no longer a key determining factor in mating patterns<sup>12</sup>.  
266 Hence, in diverse populations, such as those with admixed ancestries, the challenge  
267 of accurately diagnosing rare diseases becomes even more pronounced since new  
268 phenotypes can evolve via admixture<sup>13</sup>.

269 Ancestry also has a significant impact on the detection of rare dysmorphic disorders  
270 via artificial intelligence (AI)<sup>11</sup> because in most healthcare datasets, non-European  
271 ancestries are underrepresented<sup>14</sup>. Many next-generation phenotyping (NGP)  
272 approaches that predict disorders on the basis of facial image analysis, such as  
273 GestaltMatcher<sup>15</sup>, have demonstrated high accuracy in patients from the ancestries in  
274 which they were predominantly trained and validated, i.e., European and North  
275 American<sup>15–19</sup>.

276 Since the significantly higher birth rates in non-European regions account for 80% of  
277 the global population and 90% of all annual births (Figure 1a)<sup>20</sup>, action is required to  
278 include non-European patients currently considered to be underrepresented. So far,  
279 few studies exist about the performance of NGP tools where the ancestry composition

280 of individuals in the training and test set differs. Literature suggests that AIs trained on  
281 individuals of European ancestry perform better on a test set of Asian rather than  
282 African ancestry<sup>21-24</sup> that may be explained by their closer genetic relatedness<sup>25</sup>. This  
283 raises the question of whether AIs need to be trained for different ancestries or whether  
284 a similar performance can be achieved by sufficiently increasing the ancestral diversity  
285 in the joint training set. The latter is indicated by a study conducted on Down syndrome  
286 patients of African ancestry<sup>11</sup>. However, comparing these studies is difficult since they  
287 were not performed on data compliant with FAIR principles that are findable,  
288 accessible, interoperable, and reusable, meaning the results cannot be reproduced.

289 The motivation of our work is therefore threefold: 1) scientific, because we wanted to  
290 study the effect of inter- and intra-ancestral phenotypic variability on NGP, such as  
291 GestaltMatcher, in a systematic manner; 2) clinical, because more diverse training  
292 data can presumably increase the performance of NGP on non-European ancestries;  
293 and 3) societal, because so far underrepresented populations would benefit from  
294 potential performance improvements.

295 To achieve these goals, we aimed for a FAIR database with an increased number of  
296 patients of non-European ancestry with respect to comparable databases<sup>20,26,27</sup>.  
297 Therefore, we established the GestaltMatcher Database (GMDB) as a community-  
298 driven online framework that facilitates acquiring patient consent and incentivizes data  
299 sharing, acknowledging contributions from clinician-scientists as citeable micro-  
300 publications (Figure 2)<sup>28-31</sup>. Through this framework, we established global  
301 collaborations, enabling the collection of a wide range of data from various ancestries.

302 GMDB is the first database for medical imaging data of patients with rare genetic  
303 disorders from diverse ancestries that is compliant with the FAIR principles<sup>32</sup>. By its  
304 machine-readable design, GMDB also enables systematic analyses of the influence  
305 of genetic background on NGP performance, which we will report in this study.

## 306 Results:

### 307 Overview of FAIR data in GMDB

308 Retrospective data from curated publications, along with data provided by clinicians or  
309 patients, were made available as FAIR cases in the GMDB (Figure 3, Supplementary  
310 Figures 1 and 2)<sup>33</sup>. At the time of the data freeze for this paper on April 6<sup>th</sup> 2024, we



311 curated the GMDB-FAIR dataset consisting of 10,980 portrait images (Supplementary  
312 Figure 3) of 8,346 patients with 581 genetic disorders, including patients curated from  
313 2,224 scientific publications. 2,312 unpublished images were contributed by 138  
314 clinicians from 106 institutions (indicated by location markers in Figure 1a), including  
315 novel cases from GMDB micro-publications (micro-publication section in  
316 Supplementary Note). For the portrait data, which is the scope of this study, in terms  
317 of sex, the data distribution is relatively balanced (Figure 4a). However, age is biased  
318 toward patients aged below 10 years (Figure 4b). Figure 4c shows a two-dimensional  
319 representation of Human Phenotype Ontology<sup>34</sup> (HPO)-defined symptom groups in  
320 GMDB via Uniform Manifold Approximation and Projection (UMAP). While GMDB  
321 incorporates cases from all HPO-defined symptom groups across the disease  
322 landscape, the HPO-defined symptom group ‘facial dysmorphism’ is enriched in  
323 GMDB. Since each individual can be attributed to several HPO-defined symptom  
324 groups according to their features, facial dysmorphism was also present in the other  
325 HPO-defined symptom groups, as shown in the heatmap.

### 326 **Underrepresented populations benefited from micro-publication case reports in** 327 **GMDB**

328 Through our international collaborations (Figure 1a), the representation of non-  
329 European ancestral groups is 19% for Asian, 7% for African, and 7% for Others. 67%  
330 comprises individuals of European descent (Figure 1b). Moreover, the ancestry  
331 distribution varies among different disorders. Some disorders, such as Williams-  
332 Beuren syndrome, Hyperphosphatasia with impaired intellectual development  
333 syndrome, and Cohen syndrome, have relatively diverse and balanced ancestral  
334 distributions (Supplementary Figure 4).

335 Notably, the proportion of African ancestry was strongly increased by means of GMDB  
336 micro-publications which account for 40% of the individuals with African ancestry  
337 (Figure 4d). In terms of specific sub-ancestries (Figure 4e), more than 80% of cases  
338 with sub-Saharan ancestry and over 20% of cases with North African, Native American,  
339 and Latin American ancestries were obtained through GMDB micro-publications.

## 340 **Performance disparities in underrepresented populations**

341 We analyzed the performance of GestaltMatcher on the test set of 882 images of 275  
342 disorders with different ancestries that have not been used for the training of  
343 GestaltMatcher. Performance is measured as a top-k accuracy (as described in  
344 Methods). We report the top-1 to top-30 accuracies in Table 1. When considering top-  
345 1 accuracy, the 'Others' group demonstrated the highest performance at 73.91%,  
346 followed by the African group at 62.07%, the Asian group at 53.54%, and the European  
347 group at 55.45%. The African group achieved the highest top-5 accuracy (82.76%),  
348 the Asian group attained the highest top-10 accuracy (85.04%), while the European  
349 group only achieved 75.14% and 82.60% for top-5 and top-10 accuracies, respectively.  
350 However, the European group contains more than 50% of the testing images (523 out  
351 of 882), covering many more disorders than the other ancestry groups. That includes  
352 ultra-rare disorders known to achieve lower performances<sup>19</sup>.

353 To fairly compare the European group to another non-European ancestry, we only  
354 looked at the disorders that were present in both ancestry groups. In Table 2, when  
355 comparing the African and European groups on the six overlapping disorders, the  
356 European group outperformed the African group by achieving +16.96% top-1 accuracy  
357 and +11.17% top-10 accuracy. The European group also exhibited higher accuracies  
358 compared to the Asian group, with a top-1 accuracy of +6.92% and a top-10 accuracy  
359 of +4.15%. However, the European and 'Others' groups achieved relatively  
360 comparable results. The 'Others' group had a higher top-1 accuracy, while the  
361 European group performed better on the top-10 accuracy.

362 We further reported the performance of sex and age groups in Table 1. The distribution  
363 of testing images was relatively balanced across different groups, and no significant  
364 performance gap was observed between males and females. However, the under-  
365 one-year-old group exhibited the lowest performance, while the five- to ten-year-old  
366 group demonstrated notably higher top-5 and top-10 accuracies.

## 367 **Diverse ancestry data enhance prediction accuracy for underrepresented** 368 **populations**

369 To investigate the impact of incorporating ancestry-diverse data on the overall  
370 performance of GestaltMatcher across ancestries, we designed two sets of ancestry  
371 analysis experiments. First, we investigated the expansion of the training set of

372 GestaltMatcher (as described in Methods), including either European only (EU + EU\*)  
373 or European and non-European (EU + non-EU) patients. We measured a top-1  
374 accuracy averaged over all ancestral groups of 49.65% for the European only training  
375 set (EU + EU\*) and 66.90% for the diverse training set (EU + non-EU) (Figure 5a).  
376 Similarly, top-5 accuracy of the European training set was 69.95%, and when we  
377 trained on the diverse set, the top-5 accuracy increased to 81.24%. Notably, the  
378 evaluation performance on images of patients with European ancestry showed only a  
379 marginal performance dropdown. Specifically, the top-1 accuracy decreased by 3.82%  
380 and the top-5 accuracy by 3.61% when the dataset was augmented with 50% more  
381 non-European images. Meanwhile, the top-1 and top-5 performance increased notably  
382 for almost every other ancestral group. Figure 5a and Table 3 show further per-  
383 ancestry performances.

384 The training of GestaltMatcher results in a clinical face phenotype space that can be  
385 populated by additional cases, which we refer to as the gallery set (as described in  
386 Methods). We next investigated the influence of expanding the gallery with ancestry-  
387 diverse data by gradually raising the proportion of included non-European data from  
388 10% to 100%. Figure 5b shows that the top-1 accuracy of the non-European groups  
389 was clearly increased when we added more non-European data in the gallery.  
390 However, the top-1 accuracy of the European group did not change even when we  
391 added 100% of the non-European data into the gallery.

### 392 **GMDB-FAIR dataset drives the advancement of NGP technology**

393 GMDB-FAIR dataset is the first dataset that can be shared with the research  
394 community to train and benchmark their NGP approaches. After the first publication of  
395 the GestaltMatcher approach in 2022, for which we initially started the collection of our  
396 FAIR data, many researchers have utilized GMDB-FAIR to develop different NGP  
397 approaches. Hustinx et al.<sup>19</sup>, Sumer et al.<sup>35</sup>, and Campbell et al.<sup>36</sup> improved the  
398 prediction accuracy of their models significantly by utilizing different loss functions,  
399 network architectures, and data augmentation. Recently, Wu et al. proposed  
400 combining a large language model with facial image analysis to streamline the rare  
401 disorder diagnosis<sup>37</sup>. Furthermore, running facial analysis with an on-premise solution  
402 is possible using the FAIR data set to further prioritize genomic variants<sup>38</sup>.

403 Moreover, the GMDB-FAIR dataset can be taken as a validatable control cohort to  
404 facilitate the delineation of the facial phenotype of disorders. GestaltMatcher can  
405 detect clusters and assess whether, for example, cases with an identical variant or  
406 pathogenic variants in the same gene share a similar facial phenotype. For example,  
407 Ebstein et al. showed that facial dysmorphism was heterogeneous among the entire  
408 *PSMC3* patient cohort, but facial similarities were found in patients sharing the same  
409 pathogenic variants<sup>39</sup>. To date, 15 publications have analyzed the facial phenotype of  
410 the cohort with the GMDB-FAIR dataset and GestaltMatcher<sup>39-53</sup>. All results can be  
411 reproduced in the research platform of GMDB, which we introduce in the Methods  
412 section (Figure 2c, Figure 3c and Supplementary Note).

## 413 Discussion

414 GMDB is a modern, searchable reference and publication medium encompassing  
415 diverse populations that is designed for both clinicians and computer scientists  
416 engaged in NGP development. The ultimate goal of this study is to drive research in  
417 rare genetic disorders to understand the phenotypic variability among ancestries  
418 systematically and improve support for underrepresented populations.

419 GMDB stands out as the sole database compliant with FAIR principles, distinguished  
420 by its extensive collection of facial images covering diverse populations. This was  
421 mainly possible through the contributions and crowd-sourced annotations by our  
422 global collaborators. To increase motivation for data submission in the future, every  
423 case in the database has the potential to become a citable micro-publication with a  
424 Digital Object Identifier (DOI)<sup>54</sup>. Furthermore, future micro-publications could be  
425 indexed in reputable scientific indexing services, such as PubMed, as is the case for  
426 some existing micro-publication communication platforms<sup>55</sup>. Active patient  
427 involvement and the ability to access, upload and delete their data enhance patient  
428 autonomy and facilitate the acquisition of longitudinal patient data, further enriching  
429 GMDB's repository of facial images. Similar to other natural history study data, the  
430 longitudinal image and associated phenotypic meta data add significant value to the  
431 understanding of disease progression in patients with facial dysmorphism<sup>56</sup>. Moreover,  
432 micro-publication encourages the recruitment of patients from underrepresented  
433 populations. For example, more than 40% of all images obtained for Africans had been  
434 previously unpublished. These micro-publications from unpublished images of

435 patients with underrepresented ancestries underscored the importance of GMDB  
436 since they cannot be found in any medical journals.

437 The diverse ancestry data in GMDB further enabled us to investigate the  
438 GestaltMatcher performance differences among ancestral groups systematically. In  
439 Table 2, the performance disparities in the Asian and African groups were observed  
440 when compared to the European group. The “Others” group showed a comparable or  
441 even higher performance than the European group. The reason could be that Latin  
442 Americans in the ‘Others’ group show relatively similar facial phenotypes to the  
443 Europeans.

444 Our findings indicate that increasing the ancestral diversity in FAIR databases will  
445 particularly benefit populations currently regarded as underprivileged. We investigated  
446 how the top-1 and top-5 accuracies for the different ancestries changed when equally  
447 sized groups of European or non-European patients were added to the training set.  
448 Overall, the top-5 accuracy for non-European ancestral groups increased significantly  
449 when the training set was expanded with non-Europeans (+11.29%). When the  
450 training data were extended from only Europeans to Europeans and non-Europeans,  
451 only a marginal change in the performance of the European group was observed.  
452 Including more non-European patients in the gallery can also improve non-European  
453 groups' performances dramatically while European performance remains roughly the  
454 same (Figure 5b). The results indicate that recruiting non-European patients to support  
455 the underrepresented populations is more effective than recruiting more European  
456 patients, which often leads to models' extreme bias toward European ancestry.

457 The GMDB-FAIR dataset offers a transparent AI training set, which is crucial for the  
458 NGP development because all FAIR data are available to the clinical and scientific  
459 community. This transparency, combined with the increased representativeness of the  
460 training set, helps minimise the risk of algorithmic bias, which is key for ensuring  
461 respect for the fundamental right to non-discrimination<sup>57</sup>. The high quality of the GMDB  
462 data allows researchers to train, validate, and test AI in a manner that aligns with the  
463 expectation in the EU AI Act and the EU Medical Device Regulation<sup>58</sup>. Finally, the  
464 controlled access and consent options as described in the Methods section not only  
465 ensures respect for the fundamental right to protection of personal data<sup>57</sup> and EU  
466 General Data Protection Regulation (GDPR)<sup>59</sup> compliance, but it also enabled the



467 creation of a more diverse, representative, and larger data set as people are more  
468 comfortable with sharing health and genetic data, including images, under controlled  
469 conditions and responsible data governance than in open access publications and  
470 repositories. By this, the GMDB-FAIR dataset falls in line with other large public  
471 datasets, such as ImageNet<sup>60</sup> for object classification or Labeled Faces in the Wild  
472 (LFW)<sup>61</sup> for face verification, which have been fundamental for deep-learning  
473 technology driving computer vision over the last decade. GMDB-FAIR has been used  
474 to develop many NGP approaches<sup>19,35–37</sup> for predicting rare disorders after the first  
475 usage in GestaltMatcher in 2022. Moreover, GMDB-FAIR data can be used in the  
476 research platform (Supplementary Note) to validate the results shown in the published  
477 works<sup>39–53</sup> that provides transparency to the researcher using GestaltMatcher and the  
478 probability to extend the existing research with the user's additional data.

479 Due to variability in facial phenotypes secondary to ancestry, diverse reference image  
480 databases are crucial in order to enable clinicians to learn about the phenotypic  
481 variability in facial dysmorphism within a given disorder. While efforts have been made  
482 to create an atlas of human malformations that addresses the issue of ancestral  
483 diversity, this remains limited to only a few disorders<sup>20</sup>. With GMDB-FAIR, we created  
484 a large-scale dataset that can be searched for disorders or genes of interest in the  
485 GMDB gallery view (Figure 2c, Figure 3b), which provides clinicians with a  
486 comprehensive selection of patient images from different ancestries at a glance,  
487 thereby eliminating the need for extensive literature searches. In addition, it facilitates  
488 facial phenotype comparisons within a given disorder among different ancestries  
489 (Supplementary Note). GMDB also represents a valuable teaching tool for training  
490 students and residents to recognize disorders based on facial features.

491 To conclude, GMDB is a medical imaging database for rare disorders that  
492 encompasses diverse populations. The FAIR data will serve as reference material for  
493 clinicians that facilitates learning about facial dysmorphism across ancestries, and as  
494 a transparent training and benchmarking dataset for advancing the NGP approach.  
495 While we show improved performance for the underrepresented populations, it is  
496 important to point out that the performance is far from the optimum that can be  
497 achieved by collecting more diverse data. We envision that the gap between the  
498 European ancestral group and the underrepresented ancestries can be mitigated by

499 micro-publications in the future, and this will result in substantially improved support  
500 for underrepresented populations.

## 501 **Methods**

### 502 **Implementation of the online GMDB platform**

503 The online platform was built using Ruby on Rails in order to allow users to input  
504 images and other patient data. A database was set up using MySQL to store the  
505 patient data. GMDB is hosted physically in the University Hospital of Bonn and is  
506 maintained by Arbeitsgemeinschaft für Gen-Diagnostik e.V. (AGD), which is a non-  
507 profit organization for genomic research. The service is funded by membership fees  
508 of the AGD and donations from the Eva-Luise und Horst Köhler Foundation and the  
509 Wirtgen Foundation.

### 510 **Image data and meta data stored in GMDB**

511 An entry in GMDB consists of a medical image such as a portrait, X-ray, or fundoscopy  
512 and machine-readable meta information containing: 1) demographic data (including  
513 sex, age, and ancestry); 2) the molecularly confirmed diagnosis (OMIM index<sup>62</sup>); 3) the  
514 disease-causing mutation reported in Human Genome Variation Society format<sup>63</sup>  
515 (HGVS) or International System for Human Cytogenomic Nomenclature<sup>64</sup> (ISCN) with  
516 test method and zygosity; and 4) the clinical feature encoded in HPO terminology<sup>34</sup>  
517 (Figure 2b). When submitting data, clinicians are also asked to state their expert  
518 opinion concerning the distinctiveness of a phenotype: They are asked to score  
519 whether the medical imaging data was supportive (1), important (2), or key (3) in  
520 establishing the clinical diagnosis. Computer scientists can use this information to  
521 interpret the performance of their AI<sup>15</sup>.

### 522 **Digital consent form and patient-centered data upload**

523 To facilitate faster retrospective patient recruitment, a digital consent form has been  
524 implemented, which allows patients to select conditions for storing their data within the  
525 database and enables the provision of their signature online. To address the specific  
526 requests of patients, this feature was further developed in close collaboration with  
527 patient support groups, e.g., the German Smith-Magenis Syndrome patient  
528 organization Sirius e.V. Patients can access their own cases and provide or withdraw  
529 their consent online. They can also upload images themselves, which greatly simplifies

530 the curation process for longitudinal image data and other prospective data. The fact  
531 that documents such as letters from clinicians or laboratory results can also be  
532 uploaded, while only being visible to the responsible clinician, makes it possible to  
533 obtain molecular and phenotype information on patients recruited retrospectively from  
534 patient support groups. This digital consent is developed in such a way that it could  
535 also, in principle, be used as a dynamic consent model in the future<sup>65</sup>. The consent  
536 form is available in German and English, and other languages will be incorporated in  
537 the near future. Please find them in Supplementary Note (Digital consent, and  
538 Supplementary Figures 5 and 6) for more details.

### 539 **Data curation**

540 The curated data can be broadly categorized as retrospective and prospective.  
541 Retrospective refers primarily to data collected from the literature or from similar  
542 projects with global consent for data sharing (e.g., Minerva&Me<sup>66</sup>). For cases curated  
543 from the literature, the DOI and PubMed ID as well as the contact details of the  
544 corresponding author were collected in order to clarify whether reuse is possible while  
545 respecting intellectual property rights. Following the provision of written informed  
546 consent, our collaboration partners, clinicians from around the world (Figure 1a and  
547 the co-authors), also recruited patients with an established diagnosis from within their  
548 clinical practice or from patient support groups. Prospective curation refers to the  
549 collection of further images or metadata over time. This can be done by the attending  
550 clinician after subsequent consultations, or by the patients themselves.

551 The curation process can be broadly subdivided into three phases. First, medical  
552 students in their final year annotated cases from the literature, mainly searched  
553 PubMed and Google Scholar for publications with images of patients with facial  
554 dysmorphism and monogenic molecular diagnosis.

555 Second, solved patients were recruited from patient support groups. Included patients  
556 were allowed to upload and delete images and findings autonomously and access  
557 their data at any time. To develop a patient-centered, user-friendly platform and  
558 strengthen patient autonomy, feedback was obtained from the recruited patients  
559 during this phase in order to determine whether any adjustments to the process were  
560 required.

561 In the third phase, the database was expanded via international collaborations with  
562 clinicians from different continents. Initially, this focused on patients who had already  
563 been solved but had not yet been published in order to improve the AI's performance.  
564 However, as we progressed, more clinicians shared their unsolved cases with the  
565 scientific community. GMDB then started focusing on facial portraits of patients with  
566 rare monogenic diseases, and is now dominated by, but not limited to, such cases.  
567 Later in the curation process, we also annotated cytogenetic disorders with facial  
568 dysmorphism. In addition to these clinicians, the medical students continued to  
569 annotate data from the literature.

### 570 **Digital Object Identifier assignment**

571 After data submission, the respective case is immediately published on the website.  
572 Subsequently, the author has the option of generating a DOI in order to create a citable  
573 micro-publication<sup>54</sup>. To do this, clinicians must, after uploading the required data and  
574 metadata, enter their own personal identifier (e.g., ORCID), specify all other scientists  
575 or clinicians involved in this case, and provide a title and an abstract. To ensure the  
576 credibility and reliability of the published data, this process will adhere to a rigorous  
577 review similar to that described by Raciti et al.<sup>55</sup>. The DOIs are created and managed  
578 by the University and State Library of Bonn using the DataCite Application  
579 Programming Interface (API) (<https://datacite.org>).

580 Additionally, a dedicated landing page will be created for each case, according to the  
581 specifications of the DataCite metadata schema (Supplementary Figure 2). The  
582 landing page is accessible via the generated DOI, even for individuals without access  
583 to GMDB or those who are not logged in. The landing page contains the full citation  
584 with the DOI as a link, the abstract, and a description of the case data. No phenotypic  
585 information, HPO terms, or images are available. However, the landing page indicates  
586 how many images the micropublication contains.

### 587 **Main components of the GMDB online platform**

588 The GMDB consists of three main components that can in principle be utilized by  
589 registered users (Figure 2c). 1) Search: Clinicians can use the Gallery view to search  
590 the GMDB for disorders or genes of interest and get all patients matching this search  
591 criterion displayed in the database at a glance. 2) Analyze: Clinicians and scientists  
592 can use the GMDB-FAIR data to perform similarity comparisons of cohorts with

593 GestaltMatcher within the research platform of GMDB. 3) Train: The GMDB-FAIR  
594 dataset that can be used by external researchers to train NGP tools. More detailed  
595 information on these features can be found in the Supplement Note.

## 596 **GMDB datasets**

597 All analyses performed in this paper are based on GMDB-FAIR data (v1.1.0). But  
598 actually, the GMDB consists of the GMDB-FAIR dataset and the GMDB-private set  
599 (Supplementary Note and Supplementary Figures 7 and 8). We introduced this  
600 distinction because it is known that patient consent to data sharing is higher when not  
601 shared with a broad mass, but only for a specific study<sup>67</sup>. However, many patients  
602 agree to controlled access for the general scientific community to advance research<sup>67</sup>.  
603 For this reason, patients can decide whether they want to be part of only the GMDB-  
604 private set for AI training or agree to be part of the FAIR data set.

605 The website displays the statistics to the public, showing how many patients are in the  
606 database and how many disorders and disease genes have been curated. When the  
607 user has the link to a specific case in the GMDB (e.g., from a publication in which the  
608 original image may not be branched, but a link to the case is given in the GMDB), if  
609 the user is not logged in, the landing page for the case will show how many images  
610 and metadata are available for the case. Only sex and ancestry, as well as the disease  
611 gene, are given. If it is a case report published with a DOI in the GMDB, the  
612 corresponding title and abstract of the case can also be viewed. The remaining data  
613 can only be viewed after logging in. To visualize the images, the user has to log in to  
614 the platform.

## 615 **GMDB-FAIR data set**

616 The FAIR data set (Supplementary Figure 7b) is accessible to the scientific community.  
617 Data comes from publications and from clinicians or patients themselves. However,  
618 the case is accessible in the Gallery view for all registered users of the GMDB, and  
619 the data sheet with all relevant data and metadata can be viewed. It is also available  
620 to all users of the GMDB to perform similarity comparisons of cohorts in the research  
621 platform (Supplementary Note). The data is used for the GestaltMatcher training and  
622 test set but can also be made available to other scientists to train and test their AI after  
623 they have applied to us with an Institutional Review Board (IRB)-approved study and  
624 proposal.



## 625 **Data Governance and Ethical, Legal and Social Implications of GMDB**

626 Ethical approval for the GMDB was granted by the IRB of the University of Bonn, and  
627 all patients have given informed written consent to participate. During the  
628 GestaltMatcher consent procedure, patients can also indicate whether they agree to  
629 the use of the images in presentations, teaching activities, or in publications in other  
630 journals. This differentiation from other journals is important since patients/parents  
631 show less willingness to consent to publication in open-access journals than to  
632 publication in access-controlled databases that are not publicly accessible<sup>67</sup>. The  
633 patient shown in Figure 3 fully consented to publication of his image data.

634 The GMDB has four different levels of data access (Supplementary Figure 8): 1) The  
635 public data, which includes a summary of the GMDB statistics on the website and a  
636 landing page for case reports with DOI (Supplementary Figure 2), requires no login  
637 and is openly accessible. 2) The FAIR data, which can be viewed with a GMDB user  
638 account, and in principle, downloaded by external AI researchers. 3) The restricted  
639 data, which is not accessible to GMDB users and external AI researchers and can only  
640 be used to train the GestaltMatcher AI. 4) Patient-shared data: Patients can only view  
641 their own case and upload data if they are invited to do so by the attending clinician.

642 External scientist in the field of AI can apply to download of GMDB-FAIR data for the  
643 development of NGP approaches. Prerequisites for this are IRB approval and  
644 submission of a proposal to [info@gestaltmatcher.org](mailto:info@gestaltmatcher.org). In addition, external scientists  
645 must sign and adhere to the GDPR. The Advisory Board will conduct a thorough review  
646 of all applications. If the majority of the members of the Board approve the application,  
647 access (under the extent permissible by law) will be granted to applicants within two  
648 to three weeks.

## 649 **Advisory Board**

650 Advisory Board comprises the following co-authors: Benjamin D. Solomon, Koen  
651 Devriendt, Shahida Moosa, Christian Netzer, Martin Mücke, Christian Schaaf, Alain  
652 Verloes, Christoffer Nellåker, Markus M. Nöthen, Gholson J. Lyon, Aleksandra Jezela-  
653 Stanek, and Karen W. Gripp.

## 654 **HPO-defined symptom groups**

655 In one of our previous works<sup>68</sup>, twelve distinct and non-overlapping categories of HPO  
656 terms were defined by clinical experts (“HPO defined symptom groups”). All GMDB  
657 cases for which HPO terms were annotated were then assigned to each of those  
658 groups, if at least one of the HPO terms in this group was annotated; i.e., each GMDB  
659 case can be assigned to several HPO-defined symptom groups. For each case, the  
660 most pronounced HPO-defined symptom group was defined as the single group  
661 comprising the largest number of the case’s annotated HPO terms. The HPO-defined  
662 symptom group “Others” was only assigned as the leading HPO-defined symptom  
663 group if no other HPO-defined symptom group was present for the case.

664 Phenotypic similarity between cases was calculated using the R-package  
665 ontologySimilarity (version 2.5). Pairwise similarities were calculated for the combined  
666 data set of GMDB cases with HPO terms (n=4,474), the TRANSLATE-NAMSE exome  
667 sequencing data set (n=1,577), and data on known diseases and their clinical features  
668 downloaded from the HPO website (n=7,765,  
669 <https://hpo.jax.org/app/download/annotation>, file: genes\_to\_phenotype.txt,  
670 downloaded on 10 April 2021). The resulting distance matrix was projected in a four-  
671 dimensional space via Uniform Manifold Approximation and Projection (UMAP). The  
672 first two dimensions were plotted using ggplot2 (version 3.4.4). To analyze which  
673 HPO-defined symptom groups occur jointly, the proportion of patients assigned to the  
674 first group that were also assigned to the second group was assessed. All analyses  
675 were conducted in R (version 4.3.2).

## 676 **GestaltMatcher Algorithm**

677 GestaltMatcher<sup>15</sup> is the extension of the DeepGestalt approach<sup>17</sup>. DeepGestalt is a  
678 deep learning-based NGP tool using frontal face photos to classify up to 216  
679 syndromes it has seen during training. However, it needed a lot of training data to  
680 achieve a reasonable performance on these syndromes. That also meant it could not  
681 classify unseen syndromes during training (ultra-rare syndromes). This led to the  
682 development of GestaltMatcher, which uses a clustering approach. As such, if at least  
683 one image of the sought-after syndrome is in the gallery set, a test image can be  
684 matched to/clustered with that image using some similarity metric. Later, this approach  
685 was further enhanced by Hustinx et al.<sup>19</sup>, using a more recent architecture (iResNet)

686 and training loss (ArcFace Loss), as well as test-time augmentation and a model  
687 ensemble to improve robustness. That is also the approach we used for our  
688 experiments. Thus, for fine-tuning, we utilized the Adam optimizer, cross-entropy loss,  
689 and class weighting to deal with the imbalance in data availability between disorders.  
690 In this study, we used 7,787 images representing 275 disorders as the training set and  
691 a validation set of 1,007 images during the model training. We then tested the model  
692 on a test set consisting of 882 images.

693 The overall idea behind the methodology is to train a classifier on a more frequent  
694 subset of the syndromes, achieving a model that generalizes well on those seen  
695 syndromes. In practice, the authors of both papers decided to use syndromes with at  
696 least seven patients as the training set for this classifier. Thereafter, everything up to  
697 the penultimate layer of the classifier is used as an encoder, obtaining feature  
698 embeddings of images of interest. These could be images for the gallery set or images  
699 for the test set.

700 The aforementioned gallery set is the set of images (and their feature embeddings)  
701 with known syndromes. This can include the syndromes used for training (seen) and  
702 syndromes with too few images to train on (unseen). The theory is that similar facial  
703 phenotypes form clusters in the feature space, which is spanned by the feature  
704 embeddings in 512 dimensions and which we refer to as clinical face phenotype space.  
705 The similarity between images and clusters is computed using the cosine distance,  
706 where a lower distance implies a higher similarity. Contrary to the approach by  
707 Gurovich et al.<sup>17</sup>, this approach can easily increase support for ultra-rare syndromes.  
708 The quality and diversity of the gallery set is crucial for this approach to match test  
709 images to clusters in the gallery set.

#### 710 **Performance metric (top-k accuracy)**

711 The applied performance metric was top-k accuracy. Top-1 indicates that the disorder  
712 was correctly classified as the first guess, while top-5 indicates the correct class was  
713 in the first five guesses. We reported top-k accuracies (k=1, 5, 10, and 30) as the  
714 performance readout.

## 715 **Ancestry analysis**

716 The genetic ancestry of each individual was documented as precisely as possible  
717 using self-reported data. For instance, if an individual was born in Germany and all of  
718 the respective grandparents also originated from there, this individual was assigned  
719 to Germany (country) and Europe (continent). The same approach was used for all  
720 individuals with no self-reported migration history in previous generations. For  
721 individuals with mixed ancestry, the respective ancestries were combined. For  
722 example, an individual with a father from Gambia and a mother from Eastern Europe  
723 was assigned European-African mixed ancestry.

724 The performance of GestaltMatcher is highly dependent on the training set and the  
725 gallery set. To investigate the impact of incorporating diverse ancestry on the  
726 performance, we have therefore conducted two sets of experiments for those two  
727 components, respectively. First, we analyzed the influence on the models'  
728 performance when including only European versus both European and non-European  
729 data into the training set. And second, we analyzed the same performance when  
730 iteratively increasing the amount of non-European data into the gallery set.

731 In the first experiment, a subset of images of European patients (EU) was extended  
732 by either the inclusion of a different subset of images of European patients (EU\*), or a  
733 subset of patients with non-European ancestries (non-EU) (Supplementary Figure 9).  
734 Random sampling of these subsets was performed five times. EU consisted of on  
735 average 3,139.2 images, and EU\* comprised on average 1,567.6 images. First, the  
736 model was trained on the EU + EU\* set containing on average 4,706.8 images of  
737 patients of solely European ancestry. For EU + non-EU, a subset containing on  
738 average 1,567.6 images of patients with any non-European ancestry was used,  
739 totaling to 4,706.8 images. The experiment design ensured the maintenance of the  
740 same distribution of disorders as that found in the training data.

741 The model was fine-tuned for 50 epochs on subsets EU + EU\* and EU + non-EU of  
742 GMDB (v1.1.0). All other hyperparameters were left unchanged. It is important to note  
743 that the model was not tasked with learning to classify the ancestry, only with learning  
744 to classify the disorder.

745 Post-training, the models' performances were measured on the same evaluation set,  
746 containing images of patients with diverse ancestral backgrounds. This evaluation set  
747 consisted of 649 images and was sampled in such a manner that there was no overlap  
748 between patients or images in any subset. Top-k accuracy was averaged over each  
749 ancestry rather than each image in order to address the imbalance in ancestry  
750 frequency. As such, the performance of any infrequent group weighed equally with  
751 those of the more frequent groups.

752 In the second set of experiments, we trained the models of Hustinx et al.<sup>19</sup> using the  
753 GMDB-FAIR training set, including different proportions of non-EU data for the gallery  
754 set. We compared the performance of the syndromes our models have seen during  
755 training. For completeness, Table 1 shows the top-k accuracy (over all images) for  
756 different categories (sex, ancestry, and age range) using the entire gallery set  
757 consisting of 8,794 images (100% EU [4911] + 100% non-EU [3883]). For the  
758 experiments, we computed the performance when including different proportions of  
759 non-EU data, extending the gallery set by +10% per iteration. This experiment was  
760 repeated tenfold, randomly sampling patients with different ancestries and all their  
761 photos for the gallery set. As such, at 0%, we include only data from EU patients in  
762 the gallery set, and at 100%, we include all patient data for the relevant syndromes.

763 We further computed the performance on syndromes that occur in both the European-  
764 group and each non-European group to more accurately reflect the performance  
765 differences, avoiding the imbalance between offered support for each ancestral group.

## 766 Data and code availability

767 GMDB-FAIR can be downloaded in GMDB after the application is approved by the  
768 advisory board. Please find more details in the Data Governance and ELSI section.  
769 Code is available in the GitHub repository ([github.com/igsb/GestaltMatcher-  
770 Arc/tree/gmdb](https://github.com/igsb/GestaltMatcher-Arc/tree/gmdb)).

## 771 Acknowledgments

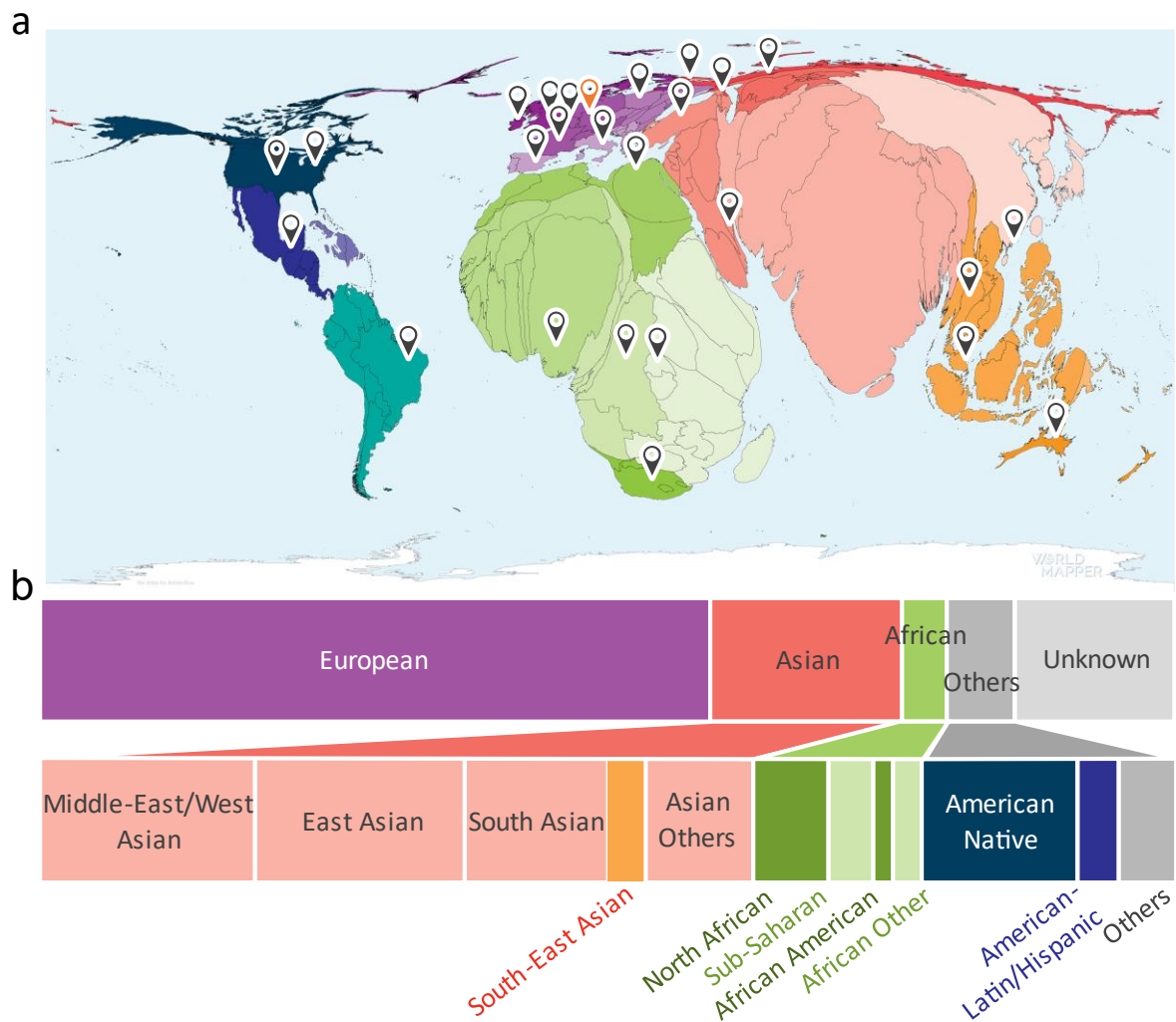
772 This research was supported in part by the Intramural Research Program of the  
773 National Human Genome Research Institute, National Institutes of Health, the United  
774 States of America. Tzung-Chien Hsieh, Peter M. Krawitz and Annabelle Arlt are



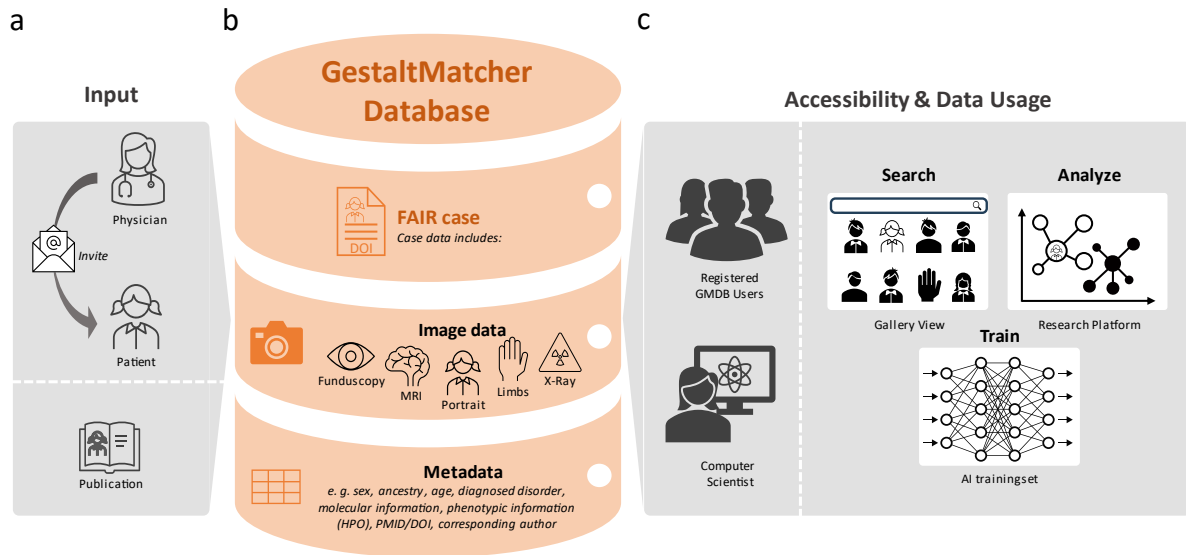
775 partner of the European Joint Programme on Rare Diseases (EJP RD) for the project  
776 ANR-22-RAR4-0001-01 (UPS36NDDiag). Sofia Douzgou Houge was supported by  
777 the Norwegian National Advisory Unit on Rare Disorders (grant number #43066).  
778 Tahsin Stefan Barakat was supported by the Netherlands Organisation for Scientific  
779 Research (ZonMw Vidi, grant 09150172110002). Heidi Beate Bentzen is supported by  
780 EU grant 101071203 and Research Council of Norway grants 322672 and 324278.  
781 Nina-Maria Wilpert was supported by the DFG Research Unit 2841 “Beyond the  
782 Exome” and is a participant in the BIH Charité Junior Clinician Scientist Program  
783 funded by the Charité - Universitätsmedizin Berlin, the Berlin Institute of Health at  
784 Charité (BIH), the Alliance4Rare, and the Berliner Sparkassenstiftung Medizin.  
785 Cristina Dias was supported by the Wellcome Trust [grant number 209568/Z/17/Z].  
786 The authors thank the Asia Pacific Society of Human Genetics, the Wirtgen  
787 Foundation, the Eva Luise und Horst Köhler Foundation, Kabuki Syndrome  
788 Foundation, Kleefstra support group, German Smith-Magenis Syndrome patient  
789 organization Sirius e.V. and the Focus Foundation for their support.

790

791 **Figures**



792  
793 **Figure 1: a)** Birth rate distribution worldwide. The size of country is scaled in  
794 accordance with the respective birth rate. The map indicates countries from which  
795 unpublished images were obtained (source: <https://worldmapper.org/faq/>, modified).  
796 **b)** Distribution of ancestry groups in GestaltMatcher Database. 16% of the patients  
797 without ancestral information were categorized as Unknown. The breakdown of  
798 ancestries in the dataset with known ancestry is as follows: European 67%, Asian 19%,  
799 African 7%, and Others 7%.



800

801 **Figure 2: GestaltMatcher Database (GMDB) Architecture and Dataflow. a)**

802 Retrospective data are collected from the literature and annotated by data curators or  
803 are uploaded by collaborating attending clinician. Patients can also upload images of

804 their own cases, incorporate prospective data, and view their own data at any time. **b)**

805 The data (multimodal image data, including portrait images as well as magnetic  
806 resonance imaging, X-ray, funduscopy and extremity images) are stored in the GMDB

807 (MySQL database) together with the relevant meta information (such as sex, age,

808 ancestry, molecular, and phenotypic information). **c)** Registered users can view and

809 search the FAIR data in the GMDB Gallery. The patient image can also be analyzed

810 using the Next-Generation Phenotyping tool GestaltMatcher within the Research

811 Platform. In addition, once their application has been approved by the Advisory Board,

812 external computer scientists can use the GMDB-FAIR data set for training purposes

813 for their projects.

814

815 **Figure 3: An example case presentation of a FAIR case with a Digital Object**

816 **Identifier (DOI). a)** FAIR cases in the GestaltMatcher Database (GMDB) are displayed

817 to GMDB users via the data sheet. Each FAIR case can also be assigned a DOI in

818 order to render it a citable micro-publication. This micro-publication contains the image

819 data and metadata, including demographic, molecular, and phenotype information.

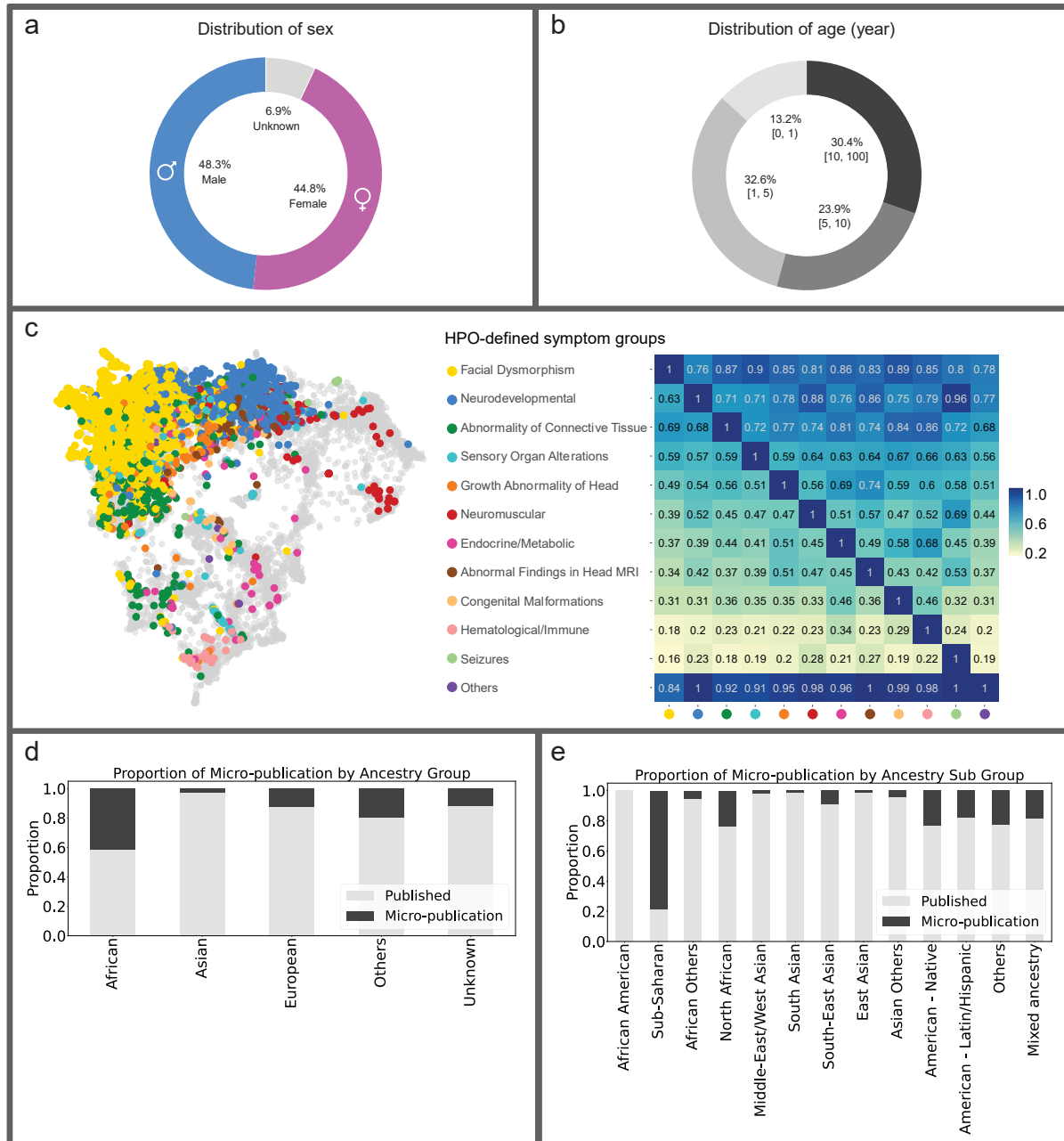
820 The dynamic nature of the GMDB case report enables longitudinal image data storage

821 even after initial publication, which is not possible in conventional journals. **b)** After

822 uploading, case reports can be viewed and searched by other users in the Gallery

823 view. **c)** The image data can also be used for inter-cohort comparisons of the gestalt  
 824 scores within the research platform.

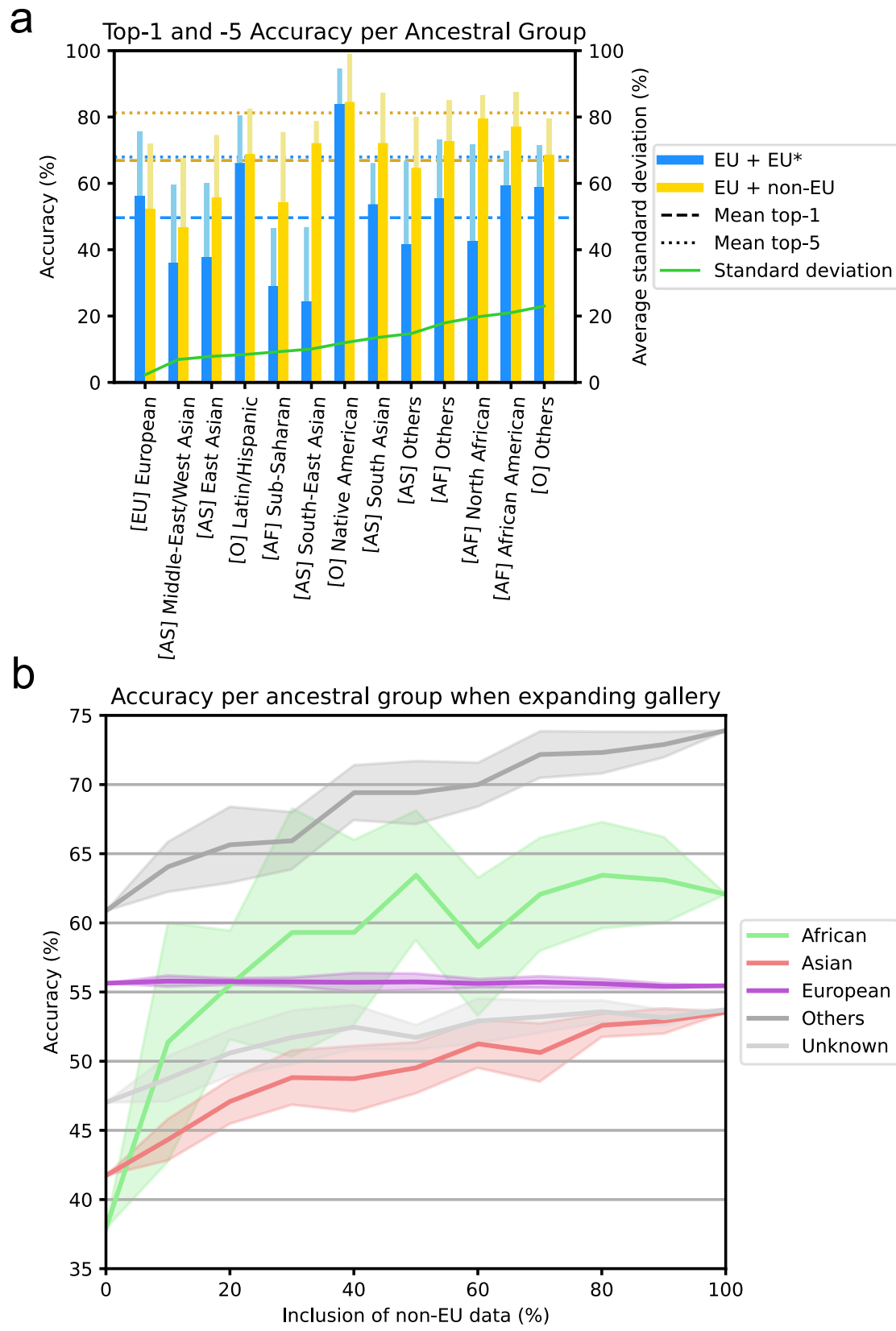
825



826  
 827 **Figure 4: Overview of the GestaltMatcher Database (GMDB)-FAIR dataset. a)** Sex  
 828 distribution. Number of images shown in brackets. **b)** Distribution of patient age in  
 829 years. **c)** Left: Two-dimensional representation of phenotypic similarities between  
 830 patients, as calculated on the basis of Human Phenotype Ontology (HPO) terms via  
 831 Uniform Manifold Approximation and Projection (UMAP). HPO terms were annotated  
 832 for 4,474 individuals in the GMDB, and expert clinicians defined twelve distinct HPO-

833 defined symptom groups. Based on the annotated HPO terms, each case was  
834 assigned to one or more HPO-defined symptom groups. All OMIM diseases were  
835 included, using their HPO annotations (gray background dots) as a reference. GMDB  
836 cases are color-coded according to their most pronounced HPO-defined symptom  
837 group, i.e., the group that includes the majority of their HPO terms. The dataset is  
838 dominated by two major clusters (facial dysmorphism in yellow and  
839 neurodevelopmental in blue) but shows cases from across the complete disease  
840 landscape. Right: Heatmap of the proportion of GMDB individuals within the HPO-  
841 defined symptom group on the X-axis who are also assigned to the HPO-defined  
842 symptom group on the Y-axis. Notably, facial dysmorphism is present in at least 70%  
843 of the cases of each HPO-defined symptom group. **d)** Proportion of the unpublished  
844 and published images in each ancestry group. **e)** Proportion of the unpublished and  
845 published images in each sub-ancestry group.





846  
847  
848  
849

**Figure 5: Performance of ancestry analysis.** a) Top-1 and top-5 accuracy of GestaltMatchers' disorder classification accuracy per ancestral group. Top-1 and top-5 accuracy of the models' disorder classification accuracy per ancestral group, where

850 (blue) belongs to the EU only subset, and (yellow) belongs to the diverse subset. Each  
851 wide, darker bar and each light, thinner bar indicate the top-1 and top-5 accuracy per  
852 ancestral group, respectively. The horizontal dashed lines and dotted lines indicate  
853 the top-1 and top-5 overall accuracy averaged over all ancestral groups, respectively.  
854 The order of the ancestry group in the x-axis is ranked according to standard deviation  
855 between top-1 accuracies of the 5-fold experiment. **b)** Top-1 accuracy of  
856 GestaltMatcher when including different proportion of non-European patients in the  
857 gallery. The x-axis is the proportion of non-European data included in the gallery. The  
858 y-axis is the top-1 accuracy. The colored region along the line indicates the standard  
859 deviation.

860 Tables

861 **Table 1: Performance of GestaltMatcher on different categories of sex, ancestry,**  
 862 **and age.** The top-1, top-5, top-10, and top-30 accuracy are reported. For the top-1 to  
 863 top-30 columns, the best performance in each category is boldfaced. In the ancestry  
 864 category, the sampling influences European and other ancestry groups' performance  
 865 due to the significant difference in the test image size. They may evaluate the different  
 866 sets of disorders. We, therefore, presented the performance of the overlapped  
 867 disorders in Table 2. In the age category, the notation [x, y) represents a half-open  
 868 interval, which includes the starting point x but excludes the endpoint y. For example,  
 869 [0, 1) years range from birth but do not include one year old.

Category	Test images	Top-1	Top-5	Top-10	Top-30	
<b>Overall</b>	882	56.58%	76.08%	82.61%	90.36%	
<b>Ancestry</b>	African	29	62.07%	<b>82.76%</b>	82.76%	86.21%
	Asian	127	53.54%	78.74%	<b>85.04%</b>	89.76%
	European	523	55.45%	75.14%	82.60%	90.25%
	Others	69	<b>73.91%</b>	81.16%	81.16%	<b>92.75%</b>
	Unknown	134	53.73%	73.13%	81.34%	91.04%
<b>Sex</b>	Male	419	55.37%	74.22%	80.67%	88.78%
	Female	393	55.98%	75.83%	83.21%	91.09%
	Unknown	70	<b>67.14%</b>	<b>88.57%</b>	<b>91.43%</b>	<b>95.71%</b>
<b>Age</b>	[0, 1) years	53	52.83%	71.70%	79.25%	90.57%
	[1, 5) years	137	56.20%	75.91%	81.02%	90.51%
	[5, 10) years	115	57.39%	<b>83.48%</b>	<b>86.09%</b>	90.43%
	[10, ∞) years	165	<b>58.18%</b>	71.51%	77.58%	85.45%
	Unknown	412	56.31%	76.46%	84.71%	<b>92.23%</b>

870  
871  
872

873 **Table 2: Performance comparison between European and other ancestry groups**  
 874 **on the overlapping disorders.** This table is an extension of the ancestry section in  
 875 Table 1, taking the overlapped disorders between European and other ancestry

876 groups. Each category compares European and non-European ancestry groups'  
 877 performance on the same set of disorders. The number of overlapped disorders is  
 878 reported in the 'Disorders' column. In comparing African and European groups, six  
 879 disorders exist in the test sets of both ancestry groups. The top-1, top-5, top-10, and  
 880 top-30 accuracy are reported. For the top-1 to top-30 columns, the best performance  
 881 in each category is boldfaced.

Category	Disorders	Test images	Top-1	Top-5	Top-10	Top-30	
[African, European]	6	African	14	64.29%	85.71%	85.71%	92.86%
		European	32	<b>81.25%</b>	<b>96.88%</b>	<b>96.88%</b>	<b>100.00%</b>
[Asian, European]	36	Asian	83	57.83%	79.52%	84.34%	87.95%
		European	139	<b>64.75%</b>	<b>82.73%</b>	<b>88.49%</b>	<b>91.37%</b>
[Others, European]	20	Others	53	<b>81.13%</b>	<b>90.57%</b>	90.57%	<b>100.00%</b>
		European	115	69.56%	84.35%	<b>93.91%</b>	96.52%
[Unknown, European]	32	Unknown	77	59.74%	<b>81.81%</b>	88.31%	<b>96.10%</b>
		European	170	<b>62.35%</b>	81.18%	<b>89.41%</b>	94.12%

882  
 883  
 884

885 **Table 3: Training accuracy with EU + non-EU and EU + EU\* datasets.** Within the  
 886 European training row, numbers annotated with \* in brackets indicate the training  
 887 images from EU + EU. Higher top-1 and top-5 accuracies between EU + EU\* and EU  
 888 + non-EU training are denoted in bold.

	Number of images		Performance EU + non-EU		Performance EU + EU*	
	Training	Testing	Top-1	Top-5	Top-1	Top-5
European	(4706.2 ± 24.4)* 3139.2 ± 15.1	444.6 ± 22.2	52.35 ± 2.30%	72.05 ± 2.66%	<b>56.17 ± 2.27%</b>	<b>75.66 ± 2.70%</b>
East Asian	283.2 ± 5.0	31 ± 6.2	<b>55.78 ± 10.25%</b>	<b>74.56 ± 5.90%</b>	37.77 ± 5.45%	60.13 ± 5.77%
Latin/Hispanic	257.8 ± 7.0	28.4 ± 4.7	<b>68.86 ± 8.92%</b>	<b>82.56 ± 7.77%</b>	66.16 ± 7.89%	80.51 ± 6.58%
Middle-East/ West Asian	211.2 ± 6.8	30 ± 5.8	<b>46.76 ± 7.01%</b>	<b>67.59 ± 7.52%</b>	36.10 ± 6.81%	59.67 ± 3.68%
South Asian	200.2 ± 5.4	18.8 ± 2.7	<b>72.15 ± 12.24%</b>	<b>87.32 ± 10.04%</b>	53.70 ± 14.86%	66.13 ± 13.40%
Asian Others	170.6 ± 2.4	16.4 ± 4.1	<b>64.66 ± 10.93%</b>	<b>80.06 ± 13.87%</b>	41.64 ± 18.51%	66.84 ± 11.87%
Sub-Saharan	119 ± 2.7	18.2 ± 3.4	<b>54.23 ± 7.32%</b>	<b>75.49 ± 15.27%</b>	28.91 ± 11.18%	46.55 ± 11.71%

North African	64.8 ± 3.2	7.4 ± 1.6	<b>79.64 ± 20.19%</b>	<b>86.64 ± 14.62%</b>	42.71 ± 19.34%	71.64 ± 18.38%
Native American	63.2 ± 6.8	14.8 ± 1.6	<b>84.55 ± 12.77%</b>	<b>99.09 ± 1.82%</b>	83.94 ± 11.18%	94.65 ± 8.62%
African Others	53.2 ± 2.0	6.2 ± 2.2	<b>72.78 ± 18.29%</b>	<b>85.00 ± 13.33%</b>	55.56 ± 17.57%	73.33 ± 22.61%
South-East Asian	51.4 ± 2.0	5.4 ± 1.3	<b>72.12 ± 13.36%</b>	<b>78.81 ± 24.61%</b>	24.40 ± 6.76%	46.83 ± 17.97%
Others	54.6 ± 2.3	6 ± 2.9	<b>68.71 ± 23.67%</b>	<b>79.43 ± 22.38%</b>	59.00 ± 22.35%	71.57 ± 21.09%
African American	38.4 ± 4.3	3.8 ± 2.6	<b>77.08 ± 18.04%</b>	<b>87.50 ± 21.65%</b>	59.38 ± 24.00%	69.79 ± 18.49%
Overall	4706.8 ± 26.7	631 ± 23.8	<b>66.90%</b>	<b>81.24%</b>	49.65%	67.95%

889  
890  
891

## 892 References

- 893 1. Hart, T. C. & Hart, P. S. Genetic studies of craniofacial anomalies: clinical implications  
894 and applications. *Orthod. Craniofac. Res.* **12**, 212–220 (2009).
- 895 2. Lesmann, H., Klinkhammer, H. & Dr. med. Dipl. Phys. Peter M. Krawitz. The future role  
896 of facial image analysis in ACMG classification guidelines. *Med. Genet.* **35**, 115–121  
897 (2023).
- 898 3. Tekendo-Ngongang, C. *et al.* Rubinstein-Taybi syndrome in diverse populations. *Am. J.*  
899 *Med. Genet. A* **182**, 2939–2950 (2020).
- 900 4. Kruszka, P., Tekendo-Ngongang, C. & Muenke, M. Diversity and dysmorphology. *Curr.*  
901 *Opin. Pediatr.* **31**, 702–707 (2019).
- 902 5. Hadj-Rabia, S. *et al.* Automatic recognition of the XLHED phenotype from facial images.  
903 *Am. J. Med. Genet. A* **173**, 2408–2414 (2017).
- 904 6. Martínez-Abadías, N. *et al.* Facial biomarkers detect gender-specific traits for bipolar  
905 disorder. *FASEB J.* **35**, (2021).
- 906 7. Fang, F., Clapham, P. J. & Chung, K. C. A systematic review of interethnic variability in  
907 facial dimensions. *Plast. Reconstr. Surg.* **127**, 874–881 (2011).
- 908 8. Vorravanpreecha, N., Lertboonnum, T., Rodjanadit, R., Sriplienchan, P. & Rojnueangnit,  
909 K. Studying Down syndrome recognition probabilities in Thai children with de-identified  
910 computer-aided facial analysis. *Am. J. Med. Genet. A* **176**, 1935–1940 (2018).
- 911 9. Kruszka, P. *et al.* Down syndrome in diverse populations. *Am. J. Med. Genet. A* **173**,



- 912 42–53 (2017).
- 913 10. Porras, A. R., Summar, M. & Linguraru, M. G. Objective differential diagnosis of Noonan  
914 and Williams-Beuren syndromes in diverse populations using quantitative facial  
915 phenotyping. *Mol Genet Genomic Med* **9**, e1636 (2021).
- 916 11. Lumaka, A. *et al.* Facial dysmorphism is influenced by ethnic background of the patient  
917 and of the evaluator. *Clin. Genet.* **92**, 166–171 (2017).
- 918 12. Burchard, E. G. *et al.* The importance of race and ethnic background in biomedical  
919 research and clinical practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
- 920 13. Martínez-Abadías, N. *et al.* Phenotypic evolution of human craniofacial morphology after  
921 admixture: a geometric morphometrics approach. *Am. J. Phys. Anthropol.* **129**, 387–398  
922 (2006).
- 923 14. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**,  
924 243–250 (2022).
- 925 15. Hsieh, T.-C. *et al.* GestaltMatcher facilitates rare disease matching using facial  
926 phenotype descriptors. *Nat. Genet.* **54**, 349–357 (2022).
- 927 16. Dudding-Byth, T. *et al.* Computer face-matching technology using two-dimensional  
928 photographs accurately matches the facial gestalt of unrelated individuals with the same  
929 syndromic form of intellectual disability. *BMC Biotechnol.* **17**, 90 (2017).
- 930 17. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep  
931 learning. *Nat. Med.* **25**, 60–64 (2019).
- 932 18. Porras, A. R., Rosenbaum, K., Tor-Diez, C., Summar, M. & Linguraru, M. G.  
933 Development and evaluation of a machine learning-based point-of-care screening tool  
934 for genetic syndromes in children: a multinational retrospective study. *Lancet Digit*  
935 *Health* (2021) doi:10.1016/S2589-7500(21)00137-0.
- 936 19. Hustinx, A. *et al.* Improving Deep Facial Phenotyping for Ultra-rare Disorder Verification  
937 Using Model Ensembles. in *2023 IEEE/CVF Winter Conference on Applications of*  
938 *Computer Vision (WACV)* 5007–5017 (IEEE, 2023).
- 939 20. Muenke, M., Adeyemo, A. & Kruszka, P. An electronic atlas of human malformation

- 940 syndromes in diverse populations. *Genet. Med.* **18**, 1085–1087 (2016).
- 941 21. Mishima, H. *et al.* Evaluation of Face2Gene using facial images of patients with  
942 congenital dysmorphic syndromes recruited in Japan. *J. Hum. Genet.* **64**, 789–794  
943 (2019).
- 944 22. Narayanan, D. L. *et al.* Computer-aided Facial Analysis in Diagnosing Dysmorphic  
945 Syndromes in Indian Children. *Indian Pediatr.* **56**, 1017–1019 (2019).
- 946 23. Elmas, M. & Gogus, B. Success of Face Analysis Technology in Rare Genetic Diseases  
947 Diagnosed by Whole-Exome Sequencing: A Single-Center Experience. *Mol. Syndromol.*  
948 **11**, 4–14 (2020).
- 949 24. Hennocq, Q. *et al.* Next generation phenotyping for diagnosis and phenotype-genotype  
950 correlations in Kabuki syndrome. *Sci. Rep.* **14**, 2330 (2024).
- 951 25. 1000 Genomes Project Consortium *et al.* A global reference for human genetic  
952 variation. *Nature* **526**, 68–74 (2015).
- 953 26. Winter, R. M. & Baraitser, M. The London Dysmorphology Database. *J. Med. Genet.* **24**,  
954 509–510 (1987).
- 955 27. Murdoch Children’s Research Institute. POSSUMweb. *POSSUMweb*  
956 <https://www.possum.net.au/>.
- 957 28. Patrinos, G. P. Chapter 6 - Incentives for Human Genome Variation Data Sharing. in  
958 *Human Genome Informatics* (eds. Lambert, C. G., Baker, D. J. & Patrinos, G. P.) 109–  
959 129 (Academic Press, 2018).
- 960 29. Mons, B. *et al.* The value of data. *Nat. Genet.* **43**, 281–283 (2011).
- 961 30. Patrinos, G. P. *et al.* Microattribution and nanopublication as means to incentivize the  
962 placement of human genome variation data into the public domain. *Hum. Mutat.* **33**,  
963 1503–1512 (2012).
- 964 31. Giardine, B. *et al.* Systematic documentation and analysis of human genetic variation in  
965 hemoglobinopathies using the microattribution approach. *Nat. Genet.* **43**, 295–301  
966 (2011).
- 967 32. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and

- 968           stewardship. *Sci Data* **3**, 160018 (2016).
- 969   33. Lesmann, H. & Weiland, H. Atypical presentation of a case with Noonan syndrome with  
970           multiple lentiginos (Version 1). (2024) doi:10.60723/10693.
- 971   34. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and  
972           analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
- 973   35. Sümer, Ö., Hellmann, F., Hustinx, A., Hsieh, T.-C. & Krawitz, P. Few-Shot Meta-  
974           Learning for Recognizing Facial Phenotypes of Genetic Disorders. in *Caring is Sharing*  
975           – *Exploiting the Value in Data for Health and Innovation* 932–936 (IOS Press, 2023).
- 976   36. Campbell, J., Dawson, M., Zisserman, A., Xie, W. & Nellåker, C. Deep Facial  
977           Phenotyping with Mixup Augmentation. in *Medical Image Understanding and Analysis*  
978           133–144 (Springer Nature Switzerland, 2024).
- 979   37. Wu, D. *et al.* Multimodal Machine Learning Combining Facial Images and Clinical Texts  
980           Improves Diagnosis of Rare Genetic Diseases. *arXiv [q-bio.QM]* (2023).
- 981   38. Hsieh, T.-C., Lesmann, H. & Krawitz, P. M. Facilitating the Molecular Diagnosis of Rare  
982           Genetic Disorders Through Facial Phenotypic Scores. *Curr Protoc* **3**, e906 (2023).
- 983   39. Ebstein, F. *et al.* PSMC3 proteasome subunit variants are associated with  
984           neurodevelopmental delay and type I interferon production. *Sci. Transl. Med.* **15**,  
985           eabo3189 (2023).
- 986   40. Asif, M. *et al.* De novo variants of CSNK2B cause a new intellectual disability-  
987           craniodigital syndrome by disrupting the canonical Wnt signaling pathway. *HGG Adv* **3**,  
988           100111 (2022).
- 989   41. Kampmeier, A. *et al.* PHIP-associated Chung-Jansen syndrome: Report of 23 new  
990           individuals. *Front Cell Dev Biol* **10**, 1020609 (2022).
- 991   42. Lyon, G. J. *et al.* Expanding the phenotypic spectrum of NAA10-related  
992           neurodevelopmental syndrome and NAA15-related neurodevelopmental syndrome. *Eur.*  
993           *J. Hum. Genet.* **31**, 824–833 (2023).
- 994   43. Aerden, M. *et al.* The neurodevelopmental and facial phenotype in individuals with a  
995           TRIP12 variant. *Eur. J. Hum. Genet.* **31**, 461–468 (2023).

- 996 44. Blackburn, P. R. *et al.* Loss-of-function variants in CUL3 cause a syndromic  
997 neurodevelopmental disorder. *medRxiv* (2023) doi:10.1101/2023.06.13.23290941.
- 998 45. Oppermann, H. *et al.* CUX1-related neurodevelopmental disorder: deep insights into  
999 phenotype-genotype spectrum and underlying pathology. *Eur. J. Hum. Genet.* **31**,  
1000 1251–1260 (2023).
- 1001 46. Blackburn, P. R. *et al.* Loss-of-function variants in *CUL3* cause a syndromic  
1002 neurodevelopmental disorder. *medRxiv* (2023) doi:10.1101/2023.06.13.23290941.
- 1003 47. Averdunk, L. *et al.* Biallelic variants in CRIPT cause a Rothmund-Thomson-like  
1004 syndrome with increased cellular senescence. *Genet. Med.* **25**, 100836 (2023).
- 1005 48. Oppermann, H. *et al.* CUX1-related neurodevelopmental disorder: deep insights into  
1006 phenotype-genotype spectrum and underlying pathology. *Eur. J. Hum. Genet.* (2023)  
1007 doi:10.1038/s41431-023-01445-2.
- 1008 49. Schmetz, A. *et al.* Delineation of the adult phenotype of Coffin-Siris syndrome in 35  
1009 individuals. *Hum. Genet.* **143**, 71–84 (2024).
- 1010 50. Küry, S. *et al.* Unveiling the crucial neuronal role of the proteasomal ATPase subunit  
1011 gene PSMC5 in neurodevelopmental proteasomopathies. *medRxiv* (2024)  
1012 doi:10.1101/2024.01.13.24301174.
- 1013 51. Li, D. *et al.* Spliceosome malfunction causes neurodevelopmental disorders with  
1014 overlapping features. *J. Clin. Invest.* **134**, (2024).
- 1015 52. Rigter, P. M. F. *et al.* Role of CAMK2D in neurodevelopment and associated conditions.  
1016 *Am. J. Hum. Genet.* **111**, 364–382 (2024).
- 1017 53. Laugwitz, L. *et al.* ZSCAN10 deficiency causes a neurodevelopmental disorder with  
1018 characteristic oto-facial malformations. *Brain* (2024) doi:10.1093/brain/awae058.
- 1019 54. Clark, T., Ciccarese, P. N. & Goble, C. A. Micropublications: a semantic model for  
1020 claims, evidence, arguments and annotations in biomedical communications. *J. Biomed.*  
1021 *Semantics* **5**, 28 (2014).
- 1022 55. Raciti, D., Yook, K., Harris, T. W., Schedl, T. & Sternberg, P. W. Micropublication:  
1023 incentivizing community curation and placing unpublished data into the public domain.

- 1024        *Database* **2018**, (2018).
- 1025    56. Liu, J. *et al.* Natural History and Real-World Data in Rare Diseases: Applications,  
1026        Limitations, and Future Perspectives. *J. Clin. Pharmacol.* **62 Suppl 2**, S38–S55 (2022).
- 1027    57. European Union. Charter of Fundamental Rights of the European Union, 2016. EUR-  
1028        Lex. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12016P%2FTXT>  
1029        (2016).
- 1030    58. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April  
1031        2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No  
1032        178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives  
1033        90/385/EEC and 93/42/EEC. [https://eur-lex.europa.eu/legal-](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745)  
1034        [content/EN/TXT/?uri=CELEX%3A32017R0745](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745).
- 1035    59. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April  
1036        2016 on the protection of natural persons with regard to the processing of personal data  
1037        and on the free movement of such data, and repealing Directive 95/46/EC (General  
1038        Data Protection Regulation). [https://eur-lex.europa.eu/legal-](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679)  
1039        [content/EN/TXT/?uri=CELEX:32016R0679](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679).
- 1040    60. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE*  
1041        *Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
- 1042    61. Huang, G. B., Ramesh, M., Berg, T. & Learned-Miller, E. *Labeled Faces in the Wild: A*  
1043        *Database for Studying Face Recognition in Unconstrained Environments*. [http://vis-](http://www.cs.umass.edu/lfw/)  
1044        [www.cs.umass.edu/lfw/](http://www.cs.umass.edu/lfw/). (2007).
- 1045    62. Boyadjiev, S. A. & Jabs, E. W. Online Mendelian Inheritance in Man (OMIM) as a  
1046        knowledgebase for human developmental disorders. *Clin. Genet.* **57**, 253–266 (2000).
- 1047    63. den Dunnen, J. T. *et al.* HGVS Recommendations for the Description of Sequence  
1048        Variants: 2016 Update. *Hum. Mutat.* **37**, 564–569 (2016).
- 1049    64. Stevens-Kroef, M., Simons, A., Rack, K. & Hastings, R. J. Cytogenetic Nomenclature  
1050        and Reporting. in *Cancer Cytogenetics: Methods and Protocols* (ed. Wan, T. S. K.) 303–  
1051        309 (Springer New York, New York, NY, 2017).



- 1052 65. Kaye, J. *et al.* Dynamic consent: a patient interface for twenty-first century research  
1053 networks. *Eur. J. Hum. Genet.* **23**, 141–146 (2015).
- 1054 66. Nellåker, C. *et al.* Enabling Global Clinical Collaborations on Identifiable Patient Data:  
1055 The Minerva Initiative. *Front. Genet.* **10**, 611 (2019).
- 1056 67. Schoeman, L., Honey, E. M., Malherbe, H. & Coetzee, V. Parents' perspectives on the  
1057 use of children's facial images for research and diagnosis: a survey. *J. Community*  
1058 *Genet.* **13**, 641–654 (2022).
- 1059 68. Schmidt, A. *et al.* Next-generation phenotyping integrated in a national framework for  
1060 patients with ultra-rare disorders improves genetic diagnostics and yields new molecular  
1061 findings. *medRxiv* 2023.04.19.23288824 (2023) doi:10.1101/2023.04.19.23288824.  
1062