

**Author: This file is the accepted version of your manuscript, and it shows any changes made by the Editor-in-Chief and the Deputy Editor since you submitted your last revision. This is the version that is being sent to Manuscript Editing for further editing in accordance with NEJM style. You will receive proofs of the edited manuscript, by e-mail. The proofs will contain queries from the manuscript editor, as well as any queries that may be present in this file. The proof stage will be your next opportunity to make changes; in the meantime, please do not make any changes or send any new material to us.**

## The 100,000 Genomes Pilot on Rare Disease Diagnosis in Healthcare – A Preliminary Report

### The 100,000 Genomes Project Pilot Investigators.

Damian Smedley Ph.D.<sup>1,2,\*</sup>, Katherine R Smith Ph.D.<sup>1,2,\*</sup>, Antonio Rueda Martin M.Sc.<sup>1,\*</sup>, Ellen A Thomas M.D.<sup>1,\*</sup>, Ellen M McDonagh Ph.D.<sup>1,3,\*</sup>, Valentina Cipriani Ph.D.<sup>2,4,5,6,\*</sup>, Jamie M Ellingford Ph.D.<sup>7,8,\*</sup>, Gavin Arno Ph.D.<sup>4,5,\*</sup>, Arianna Tucci M.D.<sup>1,2,\*</sup>, Jana Vandrovcova Ph.D.<sup>9,\*</sup>, Georgia Chan Ph.D.<sup>1,\*</sup>, Hywel J Williams Ph.D.<sup>10,11,\*</sup>, Thiloka Ratnaike MBBS, Ph.D.<sup>12,13,14</sup>, Wei Wei Ph.D.<sup>12,13</sup>, Kathleen Stirrups Ph.D.<sup>15,16</sup>, Kristina Ibanez Ph.D.<sup>1</sup>, Loukas Moutsianas Ph.D.<sup>1,2</sup>, Matthias Wielscher Ph.D.<sup>1</sup>, Anna Need Ph.D.<sup>1</sup>, Michael R Barnes Ph.D.<sup>2</sup>, Letizia Vestito M.Sc.<sup>17,18,19</sup>, James Buchanan D.Phil.<sup>20,21</sup>, Sarah Wordsworth Ph.D.<sup>20,21</sup>, Sofie Ashford B.Sc.<sup>15</sup>, Karola Rehmstrom Ph.D.<sup>22</sup>, Emily Li Ph.D.<sup>22</sup>, Gavin Fuller MMedSci<sup>23</sup>, Philip Twiss M.Sc.<sup>23</sup>, Olivera Spasic-Boskovic M.Sc.<sup>23</sup>, Sally Halsall Ph.D.<sup>23</sup>, R. Andres Floto M.D., Ph.D.<sup>22</sup>, Kenneth Poole M.D., Ph.D.<sup>22,23</sup>, Annette Wagner M.D., Ph.D.<sup>23</sup>, Sarju G Mehta M.D.<sup>23</sup>, Mark Gurnell M.D., Ph.D.<sup>24</sup>, Nigel Burrows M.D.<sup>23</sup>, Roger James Ph.D.<sup>15</sup>, Christopher Penkett D.Phil.<sup>15,16</sup>, Eleanor Dewhurst B.A.<sup>15</sup>, Stefan Gräf Ph.D.<sup>15,25,16</sup>, Rutendo Mapeta B.Sc.<sup>15,16</sup>, Mary Kasanicki Ph.D.<sup>15,23</sup>, Andrea Haworth M.Sc. FRCPATH<sup>26</sup>, Helen Savage M.Sc., DipRCPATH<sup>26</sup>, Melanie Babcock Ph.D.<sup>27</sup>, Martin G Reese Ph.D.<sup>27</sup>, Mark Bale<sup>1</sup>, Emma Baple MBBS, Ph.D.<sup>1,28,29</sup>, Christopher Boustred Ph.D.<sup>1</sup>, Helen Brittain M.D.<sup>1</sup>, Anna de Burca MBBS, PhD<sup>30</sup>, Marta Bleda Ph.D.<sup>1</sup>, Andrew Devereau Ph.D.<sup>1</sup>, Dina Halai M.Sc.<sup>1</sup>, Eik Haraldsdottir M.Sc.<sup>1</sup>, Zerin Hyder M.D.<sup>1,8</sup>, Dalia Kasperaviciute Ph.D.<sup>1,2</sup>, Christine Patch Ph.D.<sup>1</sup>, Dimitris Polychronopoulos Ph.D.<sup>1</sup>, Angela Matchan M.Sc.<sup>1</sup>, Razvan Sultana Ph.D.<sup>1</sup>, Mina Ryten M.D., Ph.D.<sup>1,31,18,32</sup>, Ana Lisa Taylor Tavares MBBS<sup>1</sup>, Carolyn Tregidgo Ph.D.<sup>1</sup>, Clare Turnbull M.D., Ph.D.<sup>1,33</sup>, Matthew Welland M.Sc.<sup>1</sup>, Suzanne Wood M.Sc.<sup>1,2</sup>, Catherine Snow Ph.D.<sup>1</sup>, Eleanor Williams Ph.D.<sup>1</sup>, Sarah Leigh Ph.D.<sup>1</sup>, Rebecca E Foulger Ph.D.<sup>1</sup>, Louise C Daugherty M.Sc.<sup>1</sup>, Olivia Niblock M.Sc.<sup>1</sup>,

Ivone U.S. Leong Ph.D.<sup>1</sup>, Caroline F Wright Ph.D.<sup>1,28</sup>, Jim Davies D.Phil<sup>30</sup>, Charles Crichton B.A.<sup>21</sup>, James Welch B.A.<sup>21</sup>, Kerrie Woods B.A.<sup>21</sup>, Lara Abulhoul M.D.<sup>34</sup>, Paul Aurora MRCP, Ph.D.<sup>35</sup>, Detlef Bockenhauer M.D.<sup>17,36</sup>, Alexander Broomfield M.D.<sup>17</sup>, Maureen A Cleary M.D.<sup>17</sup>, Tanya Lam MBBS, MPH<sup>17</sup>, Mehul Dattani FRCP<sup>18,37</sup>, Emma Footitt Ph.D.<sup>17</sup>, Vijeya Ganesan M.D.<sup>17</sup>, Stephanie Grunewald M.D., Ph.D.<sup>34,38</sup>, Sandrine Compeyrot-Lacassagne M.D.<sup>17,38</sup>, Francesco Muntoni M.D.<sup>17,38</sup>, Clarissa Pilkington MBBS<sup>17,38</sup>, Rosaline Quinlivan M.D.<sup>17</sup>, Nikhil Thapar M.D., Ph.D.<sup>39,40</sup>, Colin Wallis M.D.<sup>17</sup>, Lucy R Wedderburn FRCP, Ph.D.<sup>17,35,38</sup>, Austen Worth M.D.<sup>17</sup>, Teofila Bueser M.Sc.<sup>32,41</sup>, Cecilia Compton M.Sc.<sup>32</sup>, Charu Deshpande MRCPCH<sup>32</sup>, Hiva Fassihi FRCP<sup>42</sup>, Eshika Haque M.Sc.<sup>32</sup>, Louise Izatt Ph.D.<sup>32</sup>, Dragana Josifova M.D.<sup>32</sup>, Shehla Mohammed FRCP<sup>32</sup>, Leema Robert MRCPCH<sup>32</sup>, Sarah Rose M.Sc.<sup>32</sup>, Deborah Ruddy Ph.D.<sup>32</sup>, Robert Sarkany FRCP<sup>42</sup>, Genevieve Say M.Sc.<sup>32</sup>, Adam C Shaw M.D.<sup>32</sup>, Agata Wolejko M.Sc.<sup>43</sup>, Bishoy Habib B.Sc.<sup>43</sup>, Gavin Burns Ph.D.<sup>43</sup>, Sarah Hunter M.Sc.<sup>43</sup>, Russell J Grocock Ph.D.<sup>43</sup>, Sean J Humphray B.Sc.<sup>43</sup>, Peter N Robinson M.D.<sup>44</sup>, Melissa Haendel Ph.D.<sup>45</sup>, Michael A Simpson Ph.D.<sup>46</sup>, Siddharth Banka M.D., Ph.D.<sup>7,8</sup>, Jill Clayton-Smith FRCP<sup>7,8</sup>, Sofia Douzgou FRCP, Ph.D.<sup>7,8</sup>, Georgina Hall M.Sc.<sup>7,8</sup>, Huw B Thomas Ph.D.<sup>7</sup>, Raymond T O'Keefe Ph.D.<sup>7</sup>, Michel Michaelides FRCOphth<sup>5,4</sup>, Anthony T Moore FRCOphth<sup>5,4,47</sup>, Sam Malka B.Sc.<sup>5,4</sup>, Nikolas Pontikos Ph.D.<sup>5,4</sup>, Andrew C Browning M.D., Ph.D.<sup>48</sup>, Volker Straub M.D., PhD<sup>49</sup>, Gráinne S Gorman FRCP, Ph.D.<sup>50,51,52</sup>, Rita Horvath M.D., PhD<sup>50,12</sup>, Richard Quinton M.D.<sup>53,54</sup>, Andrew M Schaefer MRCP<sup>50,51</sup>, Patrick Yu-Wai-Man FRCOphth, Ph.D.<sup>55,13,56</sup>, Doug M Turnbull FMedSci, FRS<sup>50,51,52</sup>, Robert McFarland MRCPCH, Ph.D.<sup>50,51</sup>, Robert W Taylor FRCPPath, Ph.D.<sup>50,51</sup>, O'Connor Emer M.D.<sup>9</sup>, Yip Janice MRes<sup>9</sup>, Newland Katrina M.Sc.<sup>9</sup>, Huw R Morris FRCP, Ph.D.<sup>9</sup>, James Polke FRCPPath, Ph.D.<sup>9</sup>, Nicholas W Wood Ph.D., FMedSci<sup>9,6</sup>, Carolyn Campbell FRCPPath<sup>57</sup>, Carme Camps Ph.D.<sup>58,21</sup>, Kate Gibson B.Sc.<sup>57</sup>, Nils Koelling Ph.D.<sup>59</sup>, Tracy Lester Ph.D., FRCPPath<sup>57</sup>, Andrea H Németh FRCP, D.Phil.<sup>60,30</sup>, Claire Palles Ph.D.<sup>61</sup>, Smita Patel FRCP, FRCPPath, Ph.D.<sup>62,21</sup>, Noemi BA Roy FRCPPath, D.Phil.<sup>59,63,21</sup>, Arjune Sen MRCP, Ph.D.<sup>64,21,65</sup>, John Taylor Ph.D.<sup>57,21</sup>, Pilar Cacheiro Ph.D.<sup>2</sup>, Julius O Jacobsen Ph.D.<sup>2</sup>, Eleanor G Seaby M.D.<sup>66</sup>, Val Davison FRCPPath<sup>67</sup>, Lyn Chitty Ph.D. MRCOG<sup>17,18,38</sup>, Angela Douglas Ph.D. FRCPPath<sup>68,67</sup>, Kikkeri Naresh FRCPPath<sup>69</sup>, Dom McMullan Ph.D. FRCPPath<sup>70</sup>, Sian Ellard Ph.D. FRCPPath<sup>71</sup>, I. Karen Temple Ph.D. FRCPPath<sup>72,73</sup>, Andrew D Mumford Ph.D. FRCPPath<sup>74</sup>, Gill Wilson FRCP<sup>75</sup>, Phil Beales FMedSci<sup>18,17,38</sup>, Maria Bitner-Glindzicz MBBS, Ph.D.<sup>18,17,38</sup>, Graeme Black M.D., D.Phil.<sup>7,8</sup>, John R Bradley DM<sup>15</sup>, Paul Brennan FRCP<sup>49</sup>, John Burn MBBS, Ph.D.<sup>76</sup>, Patrick F Chinnery F. MedSci.<sup>12,13,15</sup>, Perry Elliott M.D.<sup>77</sup>, Frances Flinter

M.D.<sup>32</sup>, Henry Houlden M.D.<sup>9</sup>, Melita Irving M.D.<sup>32,78</sup>, William Newman M.D., PhD<sup>7,8</sup>, Shamima Rahman FRCP, FRCPCH, Ph.D.<sup>34,79</sup>, John A Sayer MB ChB, PhD<sup>53,54,80</sup>, Jenny C Taylor Ph.D.<sup>58,21</sup>, Andrew R Webster FRCOphth<sup>5,4</sup>, Andrew OM Wilkie FMedSci, FRS<sup>59</sup>, Willem H Ouwehand FMedSci<sup>15,81,82,16</sup>, F Lucy Raymond M.D., Ph.D.<sup>15,22</sup>, NIHR Bioresource<sup>15</sup>, John Chisholm FEng<sup>1</sup>, Sue Hill Ph.D.<sup>67</sup>, David Bentley D.Phil.<sup>43</sup>, Richard H Scott M.D., Ph.D.<sup>1,17,\*</sup>, Tom Fowler Ph.D.<sup>1,2,\*</sup>, Augusto Rendon Ph.D.<sup>1,16,\*</sup>, Mark Caulfield FRCP, FMedSci<sup>1,2</sup>

<sup>1</sup>Genomics England, Charterhouse Square, London, EC1M 6BQ, United Kingdom, <sup>2</sup>William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, United Kingdom, <sup>3</sup>Open Targets and European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom, <sup>4</sup>UCL Institute of Ophthalmology, University College London, 11-43 Bath St, London, EC1V 9EL, United Kingdom, <sup>5</sup>Moorfields Eye Hospital NHS Foundation Trust, London, EC1V 2PD, United Kingdom, <sup>6</sup>UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, United Kingdom, <sup>7</sup>Division of Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, United Kingdom, <sup>8</sup>Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University Foundation NHS Trust, Manchester, M13 9WL, United Kingdom, <sup>9</sup>The National Hospital for Neurology and Neurosurgery, Queen Square, London, WC1N 3BG, United Kingdom, <sup>10</sup>GOSgene, UCL Great Ormond Street Institute of Child Health, London, WC1N 1EH, United Kingdom, <sup>11</sup>Genetic and Genomic medicine, Institute of Medical Genetics, Cardiff University,, Cardiff, CF14 4AY, United Kingdom, <sup>12</sup>Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, United Kingdom, <sup>13</sup>MRC Mitochondrial Biology Unit, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0XY, United Kingdom, <sup>14</sup>Department of Paediatrics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, United Kingdom, <sup>15</sup>NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, United Kingdom, <sup>16</sup>Department of Haematology, University of Cambridge, Cambridge, CB2 0PT, United Kingdom, <sup>17</sup>Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, United Kingdom, <sup>18</sup>Genetics and Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health, 30 Guilford St, London, WC1N 1EH, United Kingdom, <sup>19</sup>UCL Ear Institute, London, WC1X 8EE, United

Kingdom, <sup>20</sup>Health Economics Research Centre, University of Oxford, Oxford, OX3 7LF, United Kingdom, <sup>21</sup>NIHR Oxford Biomedical Research Centre, Oxford, OX3 9DU, United Kingdom, <sup>22</sup>School of Clinical Medicine, University of Cambridge, Cambridge, CB2 0SP, United Kingdom, <sup>23</sup>Addenbrookes Hospital NHS Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, United Kingdom, <sup>24</sup>Wellcome-MRC Institute of Metabolic Science and NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, United Kingdom, <sup>25</sup>Department of Medicine, University of Cambridge, Cambridge Biomedical Campus,, Cambridge, CB2 0QQ, United Kingdom, <sup>26</sup>Congenica Ltd, Wellcome Genome Campus, Cambridge, CB10 1DR, United Kingdom, <sup>27</sup>Fabric Genomics, 1611 Telegraph Ave #500, Oakland, CA 94612, U.S.A, <sup>28</sup>University of Exeter Medical School, Exeter, EX2 5DW, United Kingdom, <sup>29</sup>Peninsula Clinical Genetics Service, Royal Devon & Exeter NHS Foundation Trust, Exeter, United Kingdom, <sup>30</sup>Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, OX3 7LD, United Kingdom, <sup>31</sup>NIHR Great Ormond Street Hospital Biomedical Research Centre,, University College, London, WC1N 3JH, United Kingdom, <sup>32</sup>Clinical Genetics Department, Guy's & St Thomas' NHS Foundation Trust, London, SE1 9RT, United Kingdom, <sup>33</sup>Division of Genetics and Epidemiology, Institute of Cancer Research, London, SM2 5NG, United Kingdom, <sup>34</sup>Metabolic Unit, Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, United Kingdom, <sup>35</sup>Infection, Immunity and Inflammation Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, London, WC1N 1EH, United Kingdom, <sup>36</sup>Department of Renal Medicine, UCL, London, United Kingdom, <sup>37</sup>London Centre for Paediatric Endocrinology and Diabetes, Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, United Kingdom, <sup>38</sup>NIHR GOSH Biomedical Research Centre, London, WC1N 3JH, United Kingdom, <sup>39</sup>Department of Gastroenterology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, United Kingdom, <sup>40</sup>Stem cells and Regenerative Medicine, UCL Great Ormond Street Institute of Child Health, London, WC1N 1EH, United Kingdom, <sup>41</sup>Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, SE1 8WA, United Kingdom, <sup>42</sup>St John's Institute of Dermatology, Guy's & St Thomas' NHS Foundation Trust, London, SE1 9RT, United Kingdom, <sup>43</sup>Illumina Cambridge Ltd, Cambridge, CB21 6DF, United Kingdom, <sup>44</sup>The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT, 06032, USA, <sup>45</sup>Center for Genome Research and Biocomputing, Environmental and Molecular Toxicology, Oregon State University,,

Corvallis, OR, 97331, USA, <sup>46</sup>Division of Genetics and Molecular Medicine, King's College London, London, SE1 9RT, United Kingdom, <sup>47</sup>Ophthalmology Department, UCSF School of Medicine, San Francisco, CA, 94143-0644, USA, <sup>48</sup>Newcastle Eye Centre, Royal Victoria Infirmary, Newcastle upon Tyne, NE1 4LP, United Kingdom, <sup>49</sup>Institute of Genetic Medicine, Newcastle University, International Centre for Life, Newcastle upon Tyne, United Kingdom, <sup>50</sup>Wellcome Centre for Mitochondrial Research, Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom, <sup>51</sup>Highly Specialised Mitochondrial Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, NE1 4LP, United Kingdom, <sup>52</sup>NIHR Newcastle Biomedical Research Centre, Newcastle upon Tyne, NE4 5PL, United Kingdom, <sup>53</sup>Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Central Parkway, Newcastle upon Tyne, NE1 3BZ, United Kingdom, <sup>54</sup>Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, NE7 7DN, United Kingdom, <sup>55</sup>Cambridge Centre for Brain Repair, Department of Clinical Neurosciences, University of Cambridge, Cambridge, CB2 0PY, United Kingdom, <sup>56</sup>NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, EC1V 2PD, United Kingdom, <sup>57</sup>Oxford Genetics Laboratories, Oxford University Hospitals NHS Foundation Trust, The Churchill Hospital, Oxford, OX3 7LE, United Kingdom, <sup>58</sup>Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom, <sup>59</sup>MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DS, United Kingdom, <sup>60</sup>Nuffield Department of Clinical Neurosciences, Level 6, West Wing, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom, <sup>61</sup>Institute of Cancer and Genomic Sciences, Institute of Biomedical Research, University of Birmingham, Edgbaston campus, Birmingham, B15 2TT, United Kingdom, <sup>62</sup>Department of Clinical Immunology, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom, <sup>63</sup>Department of Haematology, Oxford University Hospital Foundation Trust, Oxford, OX3 9DU, United Kingdom, <sup>64</sup>Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom, <sup>65</sup>Department of Neurology, Oxford University Hospitals NHS Foundation Trust, Oxford, OX3 9DU, United Kingdom, <sup>66</sup>Genomic Informatics Group, University Hospital Southampton, Southampton, SO16 6YD, United Kingdom, <sup>67</sup>NHS England and NHS Improvement, Skipton House, 22 London Rd, Elephant and Castle, London, SE1 6JW, United Kingdom, <sup>68</sup>Liverpool Women's NHS Foundation Trust, Liverpool, L8 7SS, United

Kingdom, <sup>69</sup>Imperial College Healthcare NHS Trust, Hammersmith Hospital, London, W12 0HS, United Kingdom, <sup>70</sup>Birmingham Women's Hospital, Birmingham, B15 2TG, United Kingdom, <sup>71</sup>University of Exeter Medical School, Royal Devon and Exeter Hospital, Exeter, EX2 5DW, United Kingdom, <sup>72</sup>University of Southampton, Southampton, SO17 1BJ, United Kingdom, <sup>73</sup>University Hospital Southampton, Southampton, SO16 6YD, United Kingdom, <sup>74</sup>School of Cellular and Molecular Medicine, University of Bristol, Bristol, BS2 8HW, United Kingdom, <sup>75</sup>Yorkshire and Humber, Sheffield Children's Hospital, Sheffield, S10 2TH, United Kingdom, <sup>76</sup>Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, NE1 3BZ, United Kingdom, <sup>77</sup>Institute of Cardiovascular Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom, <sup>78</sup>Division of Medical and Molecular Genetics, 8th Floor Tower Wing, King's College London, London, SE1 9RT, United Kingdom, <sup>79</sup>Mitochondrial Research Group, UCL Great Ormond Street Institute of Child Health, London, WC1N 3JH, United Kingdom, <sup>80</sup>NIHR Newcastle Biomedical Research Centre, Newcastle upon Tyne, NE45PL, United Kingdom, <sup>81</sup>NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom, <sup>82</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

\* contributed equally

Contact info:

Professor Sir Mark Caulfield FMedSci  
Genomics England,  
William Harvey Research Institute  
Queen Mary University of London,  
London EC1M 6BQ  
Email: [m.j.caulfield@qmul.ac.uk](mailto:m.j.caulfield@qmul.ac.uk)  
Telephone: +44(0) 20 7882 3403

## Abstract

### BACKGROUND

The UK 100,000 Genomes Project is in the process of investigating the role of genome sequencing of patients with undiagnosed rare disease following usual care, and the alignment of research with healthcare implementation in the UK's national health service. (Other parts of this Project focus on patients with cancer and infection.)

### METHODS

We enrolled participants, collected clinical features with human phenotype ontology terms, undertook genome sequencing and applied automated variant prioritization based on virtual gene panels (PanelApp) and phenotypes (Exomiser), alongside identification of novel pathogenic variants through research analysis. We report results on a pilot study of 4660 participants from 2183 families with 161 disorders covering a broad spectrum of rare disease.

### RESULTS

Diagnostic yields varied by family structure and were highest in trios and larger pedigrees. Likely monogenic disorders had much higher diagnostic yields (35%) with intellectual disability, hearing and vision disorders, achieving yields between 40 and 55%. Those with more complex etiologies had an overall 25% yield. Combining research and automated approaches was critical to 14% of diagnoses in which we found etiologic non-coding, structural and mitochondrial genome variants and coding variants poorly covered by exome sequencing. Cohort-wide burden testing across 57,000 genomes enabled discovery of 3 new disease genes and 19 novel associations. Of the genetic diagnoses that we made, 24% had immediate ramifications for the clinical decision-making for the patient or their relatives.

### CONCLUSION

Our pilot study of genome sequencing in a national health care system demonstrates diagnostic uplift across a range of rare diseases.

(Funded by National Institute for Health Research and others)

Rare disease is a worldwide healthcare challenge with approximately 10,000 disorders affecting 6% of the population in Western societies.<sup>1,2</sup> Over 80% of rare diseases have a genetic component and these conditions are disabling and expensive to manage. One-third of children with a rare disease die before their fifth birthday.<sup>1</sup> The adoption of next generation sequencing has improved rare disease diagnostic rates over the past decade.<sup>3-5</sup> However, the majority of rare disease patients remain without a molecular diagnosis following standard diagnostic testing.<sup>3-5</sup> To address this, the UK Government launched the 100,000 Genomes Project (100KGP) in 2013 to apply whole genome sequencing (WGS) to rare disease, cancer and infection in national healthcare.<sup>6</sup>

To assess impact of this WGS approach on the genetic diagnosis of rare disease in the UK's National Health Service, we carried out a pilot study in which we enrolled families and undertook detailed clinical phenotyping of the proband.<sup>4</sup> We collected electronic health records from all participants in a multi-petabyte research environment.<sup>5</sup> When necessary, we carried out wet bench orthogonal tests and *in-silico* approaches.

## Methods

### Patients

Following ethical approval, consenting participants (identified by healthcare professionals and researchers) with a broad range of rare diseases without diagnoses after undergoing usual care in the NHS (which ranged from no available test through approved tests which did not include genome sequencing) were recruited by nine English hospitals and consented through the National Institute for Health Research (NIHR) BioResource for Rare Diseases. To test the broad applicability of genome sequencing, participants were eligible if they had a rare disease (as defined in the UK as a disorder affecting 1 in 2000 or less), were likely to have a single gene or oligogenic aetiology, and no genomic diagnosis. Data on prior proband testing was collected where possible including single-gene tests, karyotyping, single nucleotide polymorphism (SNP) arrays, next generation sequencing panels, and exomes. Probands and, where feasible, parents and/or other family members were enrolled by multiple clinical specialties in the NHS. Standardized baseline clinical data were recorded using the Human Phenotype Ontology (HPO)<sup>7</sup> against disease specific data models<sup>8</sup> and



whole blood was drawn for DNA extraction. The participants are followed over their life course using electronic health records (all hospital episodes, registries and cause of death).

### Genome Sequencing

Genome sequencing<sup>9</sup> was performed using the Illumina TruSeq DNA PCR-Free sample preparation kit by Illumina Laboratory Sciences, Cambridge UK on an HiSeq 2500 sequencer, generating a mean depth of 32× (range from 27× to 54×) and greater than 15× for at least 95% of the reference human genome. WGS reads were aligned to the Genome Reference Consortium human genome build 37 (GRCh37) using Isaac Genome Alignment Software. Family-based variant calling of single variant nucleotides and insertion deletions (indels) for chromosomes 1 to 22, X, and the mitochondrial genome (mean 2814x coverage, range 142-16581) was performed using the Platypus variant caller.<sup>10</sup>

### The Diagnostic Pipeline

We constructed an automated analytical pipeline to filter the genome down to rare, segregating and predicted damaging candidate variants in coding regions. To limit the possibility of overlooking, or inefficiently prioritizing diagnoses we focussed initially on virtual gene panels based on both the recruited clinical indication/disease and submitted HPO terms (applied virtual panels). To address the issue of which genes have sufficient evidence to attribute causation and include in these virtual gene panels, we used our PanelApp software to enable expert, crowd-sourced review and curation of genes with diagnostic-grade evidence for each of our disease categories e.g. evidence in at least three, unrelated families.<sup>11</sup> Loss of function (LoF) or *de novo*, protein altering variants affecting genes in the applied virtual panels were classified as tier 1, other variant types such as missense variants affecting these genes were classified as tier 2, and all other filtered variants were classified as tier 3 (Figure S1 in the Supplementary Appendix). To further reduce the possibility of missing, or inefficient prioritization of diagnoses, we ran Exomiser<sup>12</sup>, a phenotype-based approach to look across all genes in the genome for a diagnosis. Exomiser prioritizes rare, segregating, predicted pathogenic variants in genes where the patient phenotypes match previous reference knowledge from human disease or model organism databases. The ontology-driven phenotype matching can detect patients possessing atypical profile for a disease.

Decision support systems and clinical genetics teams provided by Congenica Ltd and Fabric Genomics<sup>13,14</sup> assisted us in variant prioritization and return of candidate variants to the 13 NHS Genomic Medicine Centres (GMC). These variants were reviewed by NHS clinical scientists and clinicians using the American College of Medical Genetics and Genomics guidelines and a diagnostic report was issued for each proband.<sup>15</sup> Final clinical outcomes included whether a genetic diagnosis was obtained, the variant(s) involved, whether they explained all, or some of the phenotypes and whether an intervention was deployed.

The pilot participants were recruited and sequenced throughout 2014-2016, while the infrastructure to collect, QC, process and return data was being established. Results were returned to the GMCs from May 2016 to April 2019. In our post-pilot phase with an established pipeline, we now return results to the GMCs within 6 weeks of sample collection.

### Novel Pathogenic Variants

Researchers investigated coding and non-coding regions for novel diagnoses in genes matching the patients' phenotypes, including the presence of *de novo* variants in highly constrained coding regions<sup>16</sup> with 95% confidence. We used a novel methodology for mitochondrial DNA that accounts for heteroplasmy,<sup>17</sup> Genomiser,<sup>18</sup> and ExpansionHunter for simple tandem repeat expansions.<sup>19</sup> Finally we employed a novel random forest method to analyse Canvas<sup>20</sup> and Manta<sup>21</sup> calls and identify potentially pathogenic copy number and structural variants.

Gene-based burden testing to detect enrichment of rare, predicted pathogenic, segregating variants in novel genes in specific disease cohorts relative to controls was performed on the pilot genomes as well as additional genomes from the rest of the 100KGP to increase power (57,002 genomes; see Supplementary Methods).

Access to the pilot genomic and clinical data is freely accessible by becoming a member of a Genomics England Clinical Interpretation Partnership (GeCIP) domain (<https://www.genomicsengland.co.uk/about-gecip/>).

### Statistical Analysis

Testing was performed using the R (version 3.6.0) and Stata (version 16) statistical packages. Further detail on individual methods is given in the Supplementary Appendix.

## Results

### Patients

We enrolled 4660 participants (2183 probands and 2477 family members) from 161 broad categories across rare disease (Table 1), with neurologic, ophthalmologic and tumor syndromes commonly represented. Participants were recruited with varying numbers of affected and unaffected family members. We aimed, with varying degrees of success, to recruit trios or larger family structures to facilitate more effective variant prioritization. Of the probands with multiple bowel polyps whom we recruited, 93% were singletons. In contrast, 12% of probands with intellectual disability were singletons. Adult probands were more commonly enrolled than pediatric probands (age at recruitment 18 years or younger) (74% vs. 26%), in line with the general population (79% vs. 21%; 2011 census of England and Wales). The preponderance of adults is unusual compared to previous sequencing projects and reflects an eligibility criterion: probands had already undergone usual care: in many cases, usual care involved standard genetic testing (mostly single-gene or panel-based). A lower percentage of female probands were recruited, especially for pediatric cases, where the difference was significant (232 female vs. 339 male;  $P < 0.001$ ) based on the expected female proportion of 51% from 2011 census of England and Wales) across most disease categories. The increased susceptibility of males to recessive X-linked conditions may account for this sex bias: over 6% of total diagnoses involved variants on the X chromosome (which represents approximately 5% of the genome). The inferred ancestry of the probands (see Supplementary Appendix) was in line with that expected from the population (86% white, 7.5% Asian, 3.3% black, 2.2% mixed, 1% other: 2011 census of England and Wales). However, significantly more pediatric probands were of South Asian ancestry compared to adult probands (16% vs. 4%,  $P < 0.001$ ); our results indicated potential consanguinity in 43% of pediatric South Asian probands and 1% for the other pediatric probands (Table 1).

### Clinical Data and Sequencing

We collected HPO terms for each participant (median of 4 present terms, range 1-61 and median of 4 absent terms (phenotypes not exhibited by the proband), range 0-144). We

then carried out genome sequencing followed by quality assurance to check coverage, sequence quality, presence of repeat sample submissions or sample swaps, and consistency with reported family structures (see Supplementary Appendix).

### The Diagnostic Yield

We obtained genetic diagnoses for 25% of probands and deposited the genotypes into the ClinVar repository (accession numbers XXXX to YYYY). Of these diagnoses, 60% were made on the basis of coding SNV/indels in the applied virtual panels, 26% from coding SNV/indels affecting well-established disease genes outside the virtual panels using phenotype-based prioritization and/or expert review by the clinicians, Congenica Ltd, or Fabric Genomics, and 14% from genome-wide, phenotype-agnostic research analysis looking beyond SNV/indels, coding regions, and disease genes in the virtual panels (Figure 1). Following international guidelines<sup>15</sup> a further 10% of probands were classified with variants of unknown significance in genes consistent with the phenotype by clinical review at the site, but with further functional validation required. Fewer candidate variants were returned after filtering in larger family structures (Table 3), making it easier to identify causative variants, in turn leading to higher diagnostic rates for trios, quads and more complex family structures (Figure 2a), even within a disorder e.g. for hereditary ataxia the diagnostic rate increased from 21% for singletons to 32% for trios (Table S4 in the Supplementary Materials).

Unsurprisingly, we obtained a higher diagnostic yield for diseases that were considered more likely to have a monogenic cause (Table S4 in the Supplementary Appendix) than those we considered more likely to have complex etiology (35% vs 11%) (Figure 2a). Likely monogenic diseases equate to those with a presence in OMIM and where genetic testing is part of the standard diagnostic workup, based on the consensus blinded review of three clinical geneticists. Diagnostic yield was highly variable by disease (Figure 2b, Table S3 in the Supplementary Appendix), varying from 40-55% for intellectual disability and various vision and hearing disorders to 6% for tumor syndromes.

We obtained data on the presence or absence of prior genetic testing for a subset (1177) of the participants. The number of tests per proband ranged from 0-16 with a median of 1 (IQR 0-2), and approximately half of the probands in this subset had been tested at least once. The overall diagnostic uplift from genome sequencing in this subset

was 32% with only a slight difference depending on whether prior testing had been performed (33%), or not (31%). However, many of these prior tests were not recent. The diagnostic yield provided by genome sequencing varied between 28 to 45% depending on the type of prior testing (Figure 2c, Table S5 in the Supplementary Appendix) which, for the most part, involved targeted single gene and panel testing (Table S6 in the Supplementary Appendix).

### Diagnostic Pipeline

The aim of the automated, diagnostic pipeline is to identify a few, potentially causative candidate variants, from the millions in a whole genome, through removal of extremely unlikely candidates (filtering) and identification of the most likely in the remainder (prioritization). This allows the GMCs to efficiently perform manual, clinical interpretation and issue a diagnostic report. The virtual panel-based pipeline identified 322 (66%) of the 490 SNV/indel-based diagnoses from the genomes, with a high positive predictive value given the millions of variants in the whole genomes: of 1041 of returned candidate variants, 291 (28%) proved to be diagnostic. We re-ran this analysis in December 2019 to assess the impact of using updated versions of the virtual panels containing the latest disease gene discoveries, improved virtual panel selection based on the patient's phenotype and advances in variant filtering strategies, e.g. allowing for incomplete penetrance where suspected. This increased the number of genetic diagnoses detected from 322 to 377 (77%) with a positive predictive value of 15% (Figure 2d), demonstrating effective filtering and prioritization of the variants with only a median of 1 (IQR 0-2) candidate variant in panels returned to the clinicians at the GMCs per case (Table 3). Ongoing evolution of the virtual panels with new disease genes is expected to continue increasing the yield from this approach.

Phenotype-based prioritization using Exomiser detected 77%, 86%, and 88% of these diagnoses in the top, top 3 and top 5 ranked candidates respectively (Figure 2d). Exomiser and use of virtual panels were complementary, with 92% of these diagnoses re-called when used combined (last blue bar in Figure 2d). Precision phenotyping of our patients was essential both for Exomiser and for the selection of additional virtual panels, without which only 54% of these diagnoses would have been prioritized in the recruited disease virtual panel and presented to the GMCs as a likely candidate (first blue bar in Figure 2d).

## Research-based Diagnoses

14% of the genetic diagnoses required research outside the diagnostic pipeline (Figure 1). This research involved comparisons with the genome sequences and clinical data in our research environment, with validation using wet bench orthogonal tests and *in-silico* approaches (Table S7 in the Supplementary Appendix). Additional diagnoses were made by screening for the presence of *de novo* variants in highly constrained coding regions<sup>16</sup>. These diagnoses included a *de novo* *EBF3* missense variant in a patient with hereditary ataxia. Mitochondrial genome analysis, taking into account heteroplasmy, detected 4 new diagnoses as well as the 9 that had already been detected by the main pipeline). Twelve probands had intronic splicing variants prioritized by Exomiser due to the known pathogenic status of these variants in ClinVar.<sup>23</sup> Nine novel non-coding diagnoses involving previously undescribed variants required exploration of the whole genome and *in vitro* functional validation via reverse transcription polymerase chain reaction, mini-gene, or luciferase assays.<sup>24,25,26</sup> Here, unsolved probands were queried for non-coding variants affecting genes in the applied virtual panels, either alone, or in compound heterozygosity with loss-of-function variants. These were identified using either Genomiser or, for retinal disorder probands, systematic analysis of the untranslated regions, promoter or introns. A further 43 probands were fully or partially explained by structural variants or simple tandem repeat expansions in the genes *HTT* or *FXN* in probands with hereditary spastic paraplegia.

## Novel Disease Gene Associations

We performed burden testing to discover novel Mendelian disease gene associations and potential genetic diagnoses for unsolved probands; 828 significant disease-gene associations (q value < 0.1) were identified, including 249 known and 579 novel genes (novel with respect to their association with disease), with only  $0.03 \pm 0.2$  (range 0-3) associations from 10,000 permutations where cases and controls were assigned randomly. Twenty two candidates represent the most likely new, fully penetrant, Mendelian disease genes (Table S8 in the Supplementary Appendix and ClinVar accession numbers SCV001759972 - SCV001760540) with three recently independently confirmed diagnoses: *UBAP1* in hereditary spastic paraplegia,<sup>27</sup> *FOXJ1* in non-CF bronchiectasis,<sup>28</sup> and *SORD* in Charcot-Marie Tooth disease.<sup>29</sup> Diagnostic reports were issued for three probands with these genes

(Figure 1) and we are investigating others in GeneMatcher and by functional validation studies in model organisms.

### Diagnostic Sequelae

These findings ended long diagnostic odysseys for some patients and their families (the median duration of odyssey was 75 months and number of hospital visits was 68); Table S1 in Supplementary Appendix); we speculate that they will mitigate NHS resource costs (183,273 episodes of hospital care costing £87 million for affected participants; Table S3 in Supplementary Appendix). In addition, 134 (25%) of the 533 genetic diagnoses were reported by clinicians to be of immediate clinical actionability with only 11 (0.2%) described as having no benefit. As of now, the remainder of the diagnoses are of unknown utility. Healthcare benefits included 4 diagnoses leading to a suggested change in medication, 26 suggesting additional surveillance for the proband or relatives, 13 allowing clinical trial eligibility, 59 informing future reproductive choices, and 32 with other benefits (Table S9 in the Supplementary Appendix).

In several specific probands, diagnoses have had important clinical actionability. In a 36-yr-old male with suspected choroideraemia, we detected a novel, *CHM* promoter variant causing loss of gene expression<sup>26</sup> and offering eligibility for a gene-replacement trial. A male neonate proband presented with severe infection and transient neurologic symptoms immediately after birth and died at 4 months with no diagnosis but healthcare costs of approximately £80,000 (Table S10 in Supplementary Appendix). A diagnosis of transcobalamin 2 deficiency due to a homozygous frameshift in *TCN2* was made from this study which enabled predictive testing to be offered to the younger brother within one week of birth. The younger child, who received a positive result, received weekly hydroxocobalamin injections to prevent metabolic decompensation. A 10-year-old girl was admitted to intensive care with life-threatening chicken pox. She had endured a diagnostic odyssey over seven years at a total cost of £356,571 across 307 secondary care episodes (Table S11 in Supplementary Appendix). We were able to diagnose *CTPS1* deficiency due to a homozygous, known pathogenic splice acceptor variant. A diagnosis enabled a curative bone marrow transplant (cost £70,000) and predictive testing of her siblings showed no further family members to be at risk. One proband had waited till his sixth decade for a genomic diagnosis of an *INF2* mutation causing focal segmental glomerulosclerosis. His

father, brother and uncle had all died of renal failure. He had received two kidney transplants, had transmitted the condition to his daughter and was concerned about whether his 15-year-old grand daughter, who was under surveillance, was at risk. After he received his genetic diagnosis, the grand-daughter was tested, found to be negative, and discharged from regular medical surveillance.

## Discussion

Our findings demonstrate a substantial uplift in genomic diagnoses achieved for patients by genome sequencing across a broad spectrum of rare disease. The enhanced diagnostic benefit was observed regardless of whether participants had undergone prior genetic testing (31% in those who had received testing and 33% in those who had not). For 25% of those who received a genetic diagnosis, there was immediate clinical actionability. Standardizing procedures, from enrolment of patients to the return of NHS-validated results to clinicians, was critical to our success. For example, clinical data collection using disease-specific data models and HPO terms enabled diagnoses confirming the value of standardization through ontologies and clinical annotation in precision medicine.<sup>30</sup> . These additional diagnoses, beyond the 264 (49% of total diagnoses) observed in the single disease virtual panel, came from Exomiser and additional, applied virtual panels. The diagnostic discoveries derived by combining research, decision support and clinical validation and assessment leveraged an additional 72 diagnoses.

Diagnostic yield was influenced by family structure, and for disorders with a likely Mendelian inheritance and a single gene etiology our yield increased to 35%: ophthalmological, metabolic and neurologic disorders yielded the greatest percentage of diagnoses. The scale of our dataset enabled cohort-wide burden testing which identified numerous novel disease–gene associations including three that have now been confirmed and 19 with compelling evidence that are likely to be confirmed in independent datasets.

Of the diseases we diagnosed through genome-sequencing, 13% were caused by mutations in non-coding sequence or mitochondrial genomes, tandem repeat expansions in Huntington disease, and a wide range of structural variants with nucleotide resolution of breakpoints using a novel random forest method. An additional 2% of diagnoses involved coding variants in regions of low coverage on exome sequencing. Our results provide new



evidence of the value of genome sequencing and mirror previous studies where 53% of participants who received new diagnoses from genome sequencing had previously received testing by exome sequencing.<sup>5</sup>

Previous studies have demonstrated how next-generation sequencing can reveal diagnoses with yields of between 25% and 29% from exome sequencing in persons who had received no prior genetic testing.<sup>32-34</sup> The Undiagnosed Disease Network reported a 26% yield from a mixture of exome and genome sequence analysis of 382 patients<sup>5</sup> and another genome sequencing study gave a 42% yield in 50 families with intellectual disability in whom prior testing had previously been carried out.<sup>35</sup> We obtained similar results with a broad range of disorders (160) with unmet diagnostic need. Our approach is limited to diagnoses that are readily made through short-read genome sequencing. Fully phased, long-read sequencing better detects structural variation and delivers sequence from parts of the genome that are poorly captured by short read sequencing.<sup>31</sup>

This pilot has underpinned the case for genome-sequencing in the diagnosis of certain specific rare diseases in the new NHS National Genomic Test Directory<sup>36</sup>. For patients in the National Health Service for specific disorders, such as intellectual disability, genome-sequencing will now be the first-line test (Table S12 in the Supplementary Appendix) and the NHS in England, through a new National Genomic Medicine Service, is in the process of sequencing 500,000 whole genomes in rare disease and cancer in healthcare. We hope our findings will assist other health systems in considering the role of genome sequencing in the care of patients with rare diseases.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

|

## Acknowledgements

We thank all the participants and healthcare teams in Addenbrooke's Hospital in Cambridge, Great Ormond St Hospital NHS Foundation Trust, University College London NHS Foundation Trust, Guys and St Thomas's Hospital, Barts Health, Oxford University Hospitals NHS Foundation Trust, Manchester University NHS Foundation Trust and The Newcastle Hospitals NHS Foundation Trust. Mark Caulfield and Willem H Ouwehand are NIHR Senior Investigators. This work is part of the portfolio of translational research at the NIHR Biomedical Research Centres at Barts, Cambridge University Hospitals NHS Foundation Trust, Great Ormond Street Foundation NHS Trust, Manchester University NHS Foundation Trust, Moorfield's NHS Foundation Trust, The Newcastle Hospitals NHS Foundation Trust, Oxford University Hospitals NHS Foundation Trust, and University College London NHS Foundation Trust. This work was made possible through the generosity of NHS patients and their families and uses clinical data from the NHS and NHS Digital.

We thank all those across the world who have contributed to the PanelApp knowledgebase and to the validation and reporting working group (Dom McMullan, Helen Firth, Steve Abbs, Sian Ellard) for their role in supporting the development of the bioinformatics pipeline and reporting process. We received extremely valuable feedback on our work from Dr. David Bick and Prof. Gil McVean. We are grateful for the support from Professor Dame Sue Hill and the team in NHS England for the work to fund and establish the 13 Genomic Medicine Centres and that enabled the NHS contribution including the clinical return of results within the NHS in a standardized and validated format which both led to the confirmation of the diagnoses, provided additional information and led to the patient benefit reported. Maria Bitner-Glindzicz from Great Ormond Street Hospital and Institute of Child Health was a key contributor to the 100,000 Genomes Project Pilot but died during the preparation of this manuscript.

## FUNDING

Genomics England and the 100,000 Genomes Project was funded by the National Institute for Health Research, the Wellcome Trust, the Medical Research Council, Cancer Research UK, the Department of Health and Social Care and NHS England. The NIHR BioResource is funded by the NIHR.

PFC is a Wellcome Trust Principal Research Fellow (212219/Z/18/Z), and an NIHR Senior Investigator, who receives support from the Medical Research Council Mitochondrial Biology Unit (MC\_UU\_00015/9), the Medical Research Council (MRC) International Centre for Genomic Medicine in Neuromuscular Disease (MR/S005021/1), and the NIHR Biomedical Research Centre based at Cambridge University Hospitals NHS Foundation Trust and the University of Cambridge. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. LRW is supported by Grants from Versus Arthritis (21593) , The NIHR BRC at GOSH and Medical Research Council ( MR/R013926/1).

We are grateful for the support of the Monarch Initiative on HPO and Exomiser, funded by the National Institutes of Health (NIH) Office of the Director (OD) [1R24OD011883]. DS, PCM and VC were funded by National Institutes of Health Grant 5-UM1-HG006370. GA is supported by a Fight for Sight (UK) Early Career Investigator Award (5045/46), NIHR-BRC at Great Ormond Street Hospital Institute for Child Health and Moorfields Eye Charity (Stephen and Elizabeth Archer in memory of Marion Woods). The Moorfields/UCL Institute of Ophthalmology team are additionally funded by NIHR-BRC at Moorfields Eye Hospital and UCL Institute of Ophthalmology. We gratefully acknowledge the Illumina Laboratory Services team at Hinxton for genome sequencing and secondary analysis.

## References

1. Annual Report of the Chief Medical Officer 2016. Generation Genome. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/631043/CMO\\_annual\\_report\\_generation\\_genome.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/631043/CMO_annual_report_generation_genome.pdf)
2. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A.* 2019. 179, 885–892.
3. Boycott KM, Rath A, Chong JX, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet* 2017;100(5).
4. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* 2015 Jul;47(7):717-726.
5. Splinter K, Adams DR, Bacino CA, et al. Undiagnosed Diseases Network. Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N Engl J Med* 2018; 29;379(22):2131-2139.
6. The Genomics England Protocol 2017. [https://www.genomicsengland.co.uk/wp-content/uploads/2017/03/GenomicEnglandProtocol\\_151117-v4-Wales.pdf](https://www.genomicsengland.co.uk/wp-content/uploads/2017/03/GenomicEnglandProtocol_151117-v4-Wales.pdf)
7. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 2019;47(D1):D1018-D1027.
8. Genomics England data models 2018. <https://www.genomicsengland.co.uk/?wpdmdl=5500>
9. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; 456(7218):53-9.
10. Rimmer A, Phan H, Mathieson I, Iqbal Z, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 2014 Aug;46(8):912-918.
11. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019 Nov;51(11):1560-1565.
12. Smedley D, Jacobsen JO, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc.* 2015 Dec;10(12):2004-15.
13. Congenica platform. <https://www.congenica.com/platform>.
14. Fabric Genomics platform. <https://fabricgenomics.com/>.
15. Richards S, Aziz N, Bale S, Bick D et al; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015; 17(5):405-24. doi: 10.1038/gim.2015.30.
16. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. *Nat Genet.* 2019 Jan;51(1):88-95.xs
17. Wei W, Tuna S, Keogh MJ, Smith KR, et al. Germline selection shapes human mitochondrial DNA diversity. *Science.* 2019 May 24;364(6442). pii: eaau6520. doi: 10.1126/science.aau6520.
18. Smedley D, Schubach M, Jacobsen JOB, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet.* 2016 Sep 1;99(3):595-606.

19. Dolzhenko E, van Vugt JJFA, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* 2017; 27(11):1895-1903. doi: 10.1101/gr.225672.117.
20. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput Biol.* 2019 May 28;15(5):e1007069. doi: 10.1371/journal.pcbi.1007069. eCollection 2019 May. Erratum in: *PLoS Comput Biol.* 2019 Sep 20;15(9):e1007367.
21. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 2019 Jun 3;20(1):117. doi: 10.1186/s13059-019-1720-5.
22. Genomics England Eligibility Criteria 2018. [https://www.genomicsengland.co.uk/wp-content/uploads/2018/06/Rare-Disease-Eligibility-Criteria-v1.9.0-PAR-GUI-058\\_approved-version-1.pdf](https://www.genomicsengland.co.uk/wp-content/uploads/2018/06/Rare-Disease-Eligibility-Criteria-v1.9.0-PAR-GUI-058_approved-version-1.pdf)
23. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D1062-D1067.
24. Carss KJ, Arno G, Erwood M, et al. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am J Hum Genet.* 2017 Jan 5;100(1):75-90.
25. <https://www.biorxiv.org/content/10.1101/781088v1>
26. Radziwon A, Arno G, K Wheaton D, et al. Single-base substitutions in the CHM promoter as a cause of choroideremia. *Hum Mutat.* 2017 Jun;38(6):704-715.
27. Farazi Fard MA, Rebelo AP, Buglo E, et al. Truncating Mutations in UBAP1 Cause Hereditary Spastic Paraplegia. *Am J Hum Genet.* 2019 Apr 4;104(4):767-773.
28. Wallmeier J, Frank D, Shoemark A, et al. De Novo Mutations in FOXJ1 Result in a Motile Ciliopathy with Hydrocephalus and Randomization of Left/Right Body Asymmetry. *Am J Hum Genet.* 2019 Nov 7;105(5):1030-1039.
29. Cortese A, et al. Biallelic mutations in *SORD* cause a common and potentially treatable hereditary neuropathy with implications for diabetes. *Nat Genet.* 2020 52, 473–481.
30. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. *N Engl J Med* 2018; 379:1452-1462
31. Eichler EE. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N Engl J Med.* 2019 Jul 4;381(1):64-74.
32. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA.* 2014 Nov 12;312(18):1870-9.
33. Hu X, Li N, Xu Y, et al. Proband-only medical exome sequencing as a cost-effective first-tier genetic diagnostic test for patients without prior molecular tests and clinical diagnosis in a developing country: the China experience. *Genet Med.* 2018 Sep;20(9):1045-1053.
34. Vissers LELM, van Nimwegen KJM, Schieving JH, et al. A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genet Med.* 2017 Sep;19(9):1055-1063.
35. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature.* 2014 Jul 17;511(7509):344-7.
36. NHS national genomic test registry 2020. <https://www.england.nhs.uk/publication/national-genomic-test-directories/>

## FIGURE

## LEGENDS

Figure 1. Overview of the diagnostic and research pipeline and source of diagnoses. Results were returned to the Genomic Medicine Centres (GMCs) of the recruiting hospitals on an 2183 pilot probands. 25% received a positive diagnosis, 10% had variant(s) of unknown significance (VUS) in genes consistent with the phenotype according to clinical geneticists at the recruiting site, but with further functional validation required. The remaining 65% received a negative report at the time but will be reanalysed. Numbers and source of these positive diagnoses is shown at each stage of the automated diagnostic pipeline and additional research where a clear diagnosis was not immediately obvious.

Figure 2 Diagnoses in the rare disease pilot. (a) Percentage diagnostic yield for all samples and sub-divided by family structure or whether likely monogenic (35% yield) vs more complex aetiologies (11% yield) with the numbers of probands shown on bars, (b) Percentage diagnostic yield by disease area (numbers of closed probands shown on bars), (c) Percentage diagnostic yield for probands with/without prior genetics testing and broken down by most extensive testing type: chromosomal (karyotyping , arrayCGH, SNP arrays), targeted single gene tests, NGS panels or WES (numbers of closed probands shown on bars) (d) Performance of virtual panel-based and Exomiser prioritization for identifying the diagnoses. Virtual disease panel only: a single panel for the recruited disease category. Applied panels - all applied virtual panels used in the pipeline including the recruited disease associated panel as well as 0 or more additionally selected panels based on the patient phenotypes (HPO terms). Proportion of diagnoses detected are in blue (sensitivity) along with proportion of prioritized variants leading to a positive diagnosis in

orange (positive predictive value). Proportions are also shown on bars. Here, diagnosed variant(s) are true positives and other returned candidate variants are false positives.





Table 1. Demographics (including inferred ancestry) of the 100,000 Genomes Project pilot.

Variable	All probands (N=2183)	Paediatric (age at recruitment <= 18) probands (N=571)	Adult (age at recruitment > 18) probands (N=1612)
<b>Sex — no. (%)</b>			
Male	1138 (52)	339 (16)	799 (37)
Female	1045 (48)	232 (11)	813 (37)
	2183 (100)	571 (26)	1612 (74)
<b>Median (IQR) age in years at recruitment</b>	35 (18-54)	9 (5-14)	45 (31-60)
<b>Race or ethnic group — no. (%), %consanguinity suggested in record</b>			
African	50 (2), 0	25 (4), 0	25 (2), 0
Ad Mixed American	26 (1), 23	12 (2), 25	14 (1), 21
East Asian	8 (<1), 0	2 (<1), 0	6 (<1), 0
European	1931 (88), <1	438 (77), <1	1493 (93), <1
South Asian	163 (7), 36	93 (16), 43	70 (4), 25
Not determined	5 (<1), 0	1 (<1), 0	4 (<1), 0
	2183 (100), 3	571 (26), 8	1612 (74), 2

Table 2. Clinical features of the 100,000 Genomes Project pilot

Primary symptoms — no. (%)	All Families	Singletons	Duos	Trios	Larger families
Cardiovascular	147 (7)	56 (3)	24 (1)	49 (2)	18 (1)
Ciliopathies	69 (3)	34 (2)	14 (1)	16 (1)	5 (<1)
Dermatological	38 (2)	9 (<1)	5 (<1)	22 (1)	2 (<1)
Dysmorphic and congenital abnormalities	20 (1)	10 (<1)	2 (<1)	7 (<1)	1 (<1)
Endocrine	87 (4)	57 (3)	14 (1)	12 (1)	4 (<1)
Gastroenterological	32 (1)			18 (1)	14 (1)
Growth	3 (<1)	3 (<1)			
Haematological and immunological	5 (<1)	2 (<1)	3 (<1)		
Haematological	7 (<1)		3 (<1)	2 (<1)	2 (<1)
Hearing and ear	35 (2)	6 (<1)	5 (<1)	17 (1)	7 (<1)
Metabolic	93 (4)	24 (1)	12 (1)	48 (2)	9 (<1)
Intellectual disability (ID)	130 (6)	10 (<1)	24 (1)	78 (4)	18 (1)
Neurology and neurodevelopmental (excl. ID)	521 (24)	193 (9)	93 (4)	194 (9)	41 (2)
Ophthalmological	348 (16)	74 (3)	62 (3)	199 (9)	13 (1)
Renal and urinary tract	176 (8)	125 (6)	21 (1)	26 (1)	4 (<1)
Respiratory	2 (<1)	1 (<1)		1 (<1)	
Rheumatological	48 (2)	14 (1)	6 (<1)	25 (1)	3 (<1)
Skeletal	62 (3)	15 (1)	11 (1)	23 (1)	13 (1)
Tumour syndromes	293 (13)	231 (11)	31 (1)	27 (1)	4 (<1)
Other	67 (3)	17 (1)	12 (1)	34 (2)	4 (<1)
	2183 (100)	881 (40)	343 (16)	797 (37)	162 (7)



Table 3. Number of candidate variants returned to the NHS per case by automated virtual panel-based analysis pipeline. Duos refer strictly to parent-child pairs and trios to both parents and a child in a family. Values shown are median (IQR).

	All family structures	Singletons	Duos	Trios	Other family structures
Variants after filtering	221 (49-288)	292 (258-327)	149 (117-213)	29 (17-136)	22.5 (9-71)
In virtual panels	1 (0-2)	1 (0-2)	1 (0-3)	1 (0-2)	0 (0-1)