Whole genome sequencing for the investigation of rare anaemias: Challenges and real-world outcomes

Brierley CK[1,2], Babbs C[1], McGowan S[1], Downes D[1], Aydinok Y[3], Thakur I[4], Connor P[4], Tedgar U[5], Bechensteen AG[6], Harvet G[7], Josu de la Fuente J[8], Mettananda S[9], Hendricks M[10], Carey P[11], Lausen B[12], Ancliff P[13], Qureshi A[14], Mapeta R[15,16], Megy K[15,16], Ouwehand W[15-17], Roberts I[1,15], Layton M[8]*, Roy N[1,2]*

*these authors contributed equally to this workMRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK

1. Dept of Haematology, Oxford University Hospitals NHS Foundation trust, Oxford, UK
2. Department of Pediatric Hematology, Ege University Hospital, Izmir, Turkey
3. Department of Paediatric Haematology/Oncology, Children's Hospital for Wales, Cardiff, Wales, UK
4. Department of Paediatrics, Skåne University Hospital, Lund, Sweden
5. Department of Pediatrics, Oslo University Hospital, Norway
6. University Hospital Southampton NHS FT, Southampton, UK
7. Department of Haematology, Imperial College London, Hammersmith Hospital Campus, London, UK
8. Department of Pediatrics, Faculty of Medicine, University of Kelaniya, Sri Lanka
9. Department of Paediatric Medicine and Child Health, Red Cross War Memorial Children's Hospital, Cape Town, South Africa
10. Department of Clinical Haematology, Royal Victoria Infirmary, Newcastle-upon-Tyne, UK
11. Department of Paediatrics and Adolescent Medicine, The University Hospital Rigshospitalet, Copenhagen, Denmark
12. Department of Haematology, Great Ormond Street Hospital, London, UK
13. Dept of Paediatrics, Oxford University Hospitals NHS Foundation trust, Oxford, UK

Background

Next generation sequencing (NGS) is a powerful tool to identify novel genetic mutations underlying haematological diseases. Targeted NGS panels evaluating 30-100 genes are in routine clinical use to evaluate patients presenting with unexplained anaemia. Typically, pathogenic variants are identified in <50% of cases.[1] For the remainder, whole genome sequencing (WGS) may help identify novel candidate genetic variants. The NIHR Bioresource for Stem & Myeloid Disorders study recruited patients with a clinical phenotype of suspected inherited unexplained anaemia and negative targeted NGS panel analyses.

Methods:

Deep phenotyping was completed for all cases. A "pertinent findings" MDT comprising clinicians, clinical geneticists and bioinformaticians was performed to assess variants identified on a virtual panel of 150 known stem and myeloid disease-associated genes. Patients with no pathogenic mutations in known disease genes were referred for research analysis (see **Figure 1** for study outline). Trio analysis was performed where possible to prioritise coding variants as *de novo*, compound heterozygous, and homozygous in proband/heterozygous in parents. Variants were annotated using ANNOVAR, filtered based on rarity (allele frequency <0.02) in the 1000G Project, and a pathogenicity score assigned based on 8 different *in silico* scoring systems (PolyPhen 2, SIFT, LRT, MutationTaster, MutationAssessor, FATHMM, GERP, CADD). A systematic approach was developed to assess the likely clinical significance of each genetic variant on a case-by-case basis and generate a list of prioritised variants awaiting functional work-up.

Results

42 patients with unexplained anaemia and 102 first degree relatives underwent WGS (Table 1). 4 proband (10%) and 5 relative samples failed WGS quality control. In 6 (14%) cases, the anaemia was felt to be acquired as it had resolved by the time pertinent findings analysis was completed. Causative mutations, deletions or copy number variations in known genes were identified in 12 cases (29%) (PIEZO1, HBB, PKLR, SEC23B, NPHP1, CECR1, GPI, RUNX1, TMPRSS6, SPTA1) and a clinical grade research report was issued. 14 cases (33%) with a severe phenotype and complete trio were selected for the first phase of research analysis.

A median of 59 (range 19-78) *in silico* predicted pathogenic variants were identified per patient. These were then filtered manually. Violin plots of erythroblast RNA sequencing data were generated to prioritise genes expressed in erythroblasts. Public databases were used to review the predicted effect of the variant on the transcript expressed in erythroblasts (Ensembl, Exac, ClinVar) and explore known phenotype associations (OMIM, Genecards). The impact of gene knock-out in animal models was reviewed using Mouse Genome Informatics.

Using this approach, a median of 4 (range 3-6) plausibly pathogenic, prioritised variants were identified per patient. These were listed on GeneMatcher, an online platform to connect investigators researching the same gene[2]. A median of 5 (range 2-12) researchers were contacted per case. One case led to a phenotype match with another research team.

Discussion

WGS provided initial clinical answers to ~29% of cases and a strategy for the successful interpretation of genetic variation in negative cases was developed. Functional validation of prioritised variants will now follow in the 14 trios with severe phenotypes. High-throughput methods for experimental validation of large numbers of rare variants, alongside machine learning algorithms for predicting the pathogenicity of rare coding variants should allow WGS to realise its full contribution to personalised medicine and clinical diagnosis of patients with rare anaemias.

1. Roy, N. B., et al., *Br J Haematol* **2016,** *175* (2), 318-330.
2. Sobreira, N., et al., *Hum Mutat* **2015,** *36* (10), 928-30.

Table 1: Patient characteristics (n=46)

Age /years (median, range)	7.2 (0.9 - 45.6)
Haemoglobin / g/L (median, range)	80 (19-158)
Clinical classification of anaemia	
Diamond Blackfan anaemia	9 (21%)
Congenital dyserythropoietic anaemia	8 (19%)
Haemolytic anaemia	9 (21%)
Iron refractory iron deficiency anaemia	2 (5%)
Schwachman Diamond syndrome	1 (2%)
Megaloblastic anaemia	1 (2%)
Unexplained anaemia NOS	13 (31%)
Transfusion dependency	16 (38%)
Consanguinity	19 (45%)

Figure 1: Study outline