Pioneering an efficient migration of 13,000 whole genomes: Catching up with the latest Human genome assembly

Matthias Haimel[1,2], Johannes Karten[4], Bas Tolhuis[4], Olga Shamardina[1,2], NIHR BioResource[3], Willem H Ouwehand[1,2,3], Chris Penkett[1,2], Kathleen Stirrups[1,2], Ignacio Medina[1], Nicholas W. Morrell[1,2], Augusto Rendon[5], Stefan Gräf[1,2]
[1]University of Cambridge, UK; [2]NIHR BioResource, UK; [3]NHS Blood and Transplant, Cambridge, UK;
[4]GENALICE BV, The Netherlands; [5]Genomics England Ltd, UK

✉ mh719--ASHG@cam.ac.uk @MHaimel

Poster 1425F

Screencast

UNIVERSITY OF CAMBRIDGE

NIHR BioResource Rare Diseases

Genomics england

GENALICE TECHNOLOGY FOR PEOPLE & SCIENCE

Background

The NIHR BioResource – Rare Diseases recruited 13,000 patients and relatives from 15 different rare disease projects over a 4 year period. The 50 participating NHS Trusts and international collaborators collected whole blood samples that were centrally processed following standard protocols. The whole-genome sequence (WGS) data were generated by Illumina to a depth of 30x coverage using PCR free methodology. Sequence and variation results were delivered to the high performance computing service for analysis and amount to 840TB of data.

Genome Variation

Variation data from samples were quality controlled and checked against the recorded gender. 55 billion individual variants were incrementally loaded into a distributed analysis framework. The aggregated 348M single nucleotide variants (SNVs) and insertions / deletions (INDELs) were efficiently annotated and filtered for 170M high quality variants. Rare variants (<1:1,000) take up 88% (150M) of which 106K (0.07%) are protein truncating.

Research findings

Disease cohort specific analysis teams identified 718 disease causing variants in 680 patients. These findings were discussed in multi disciplinary teams (MDT) and evaluated for their pathogenicity. Based on the evaluation, research reports were returned to the NHS Trusts for further clinical testing in accredited laboratories.

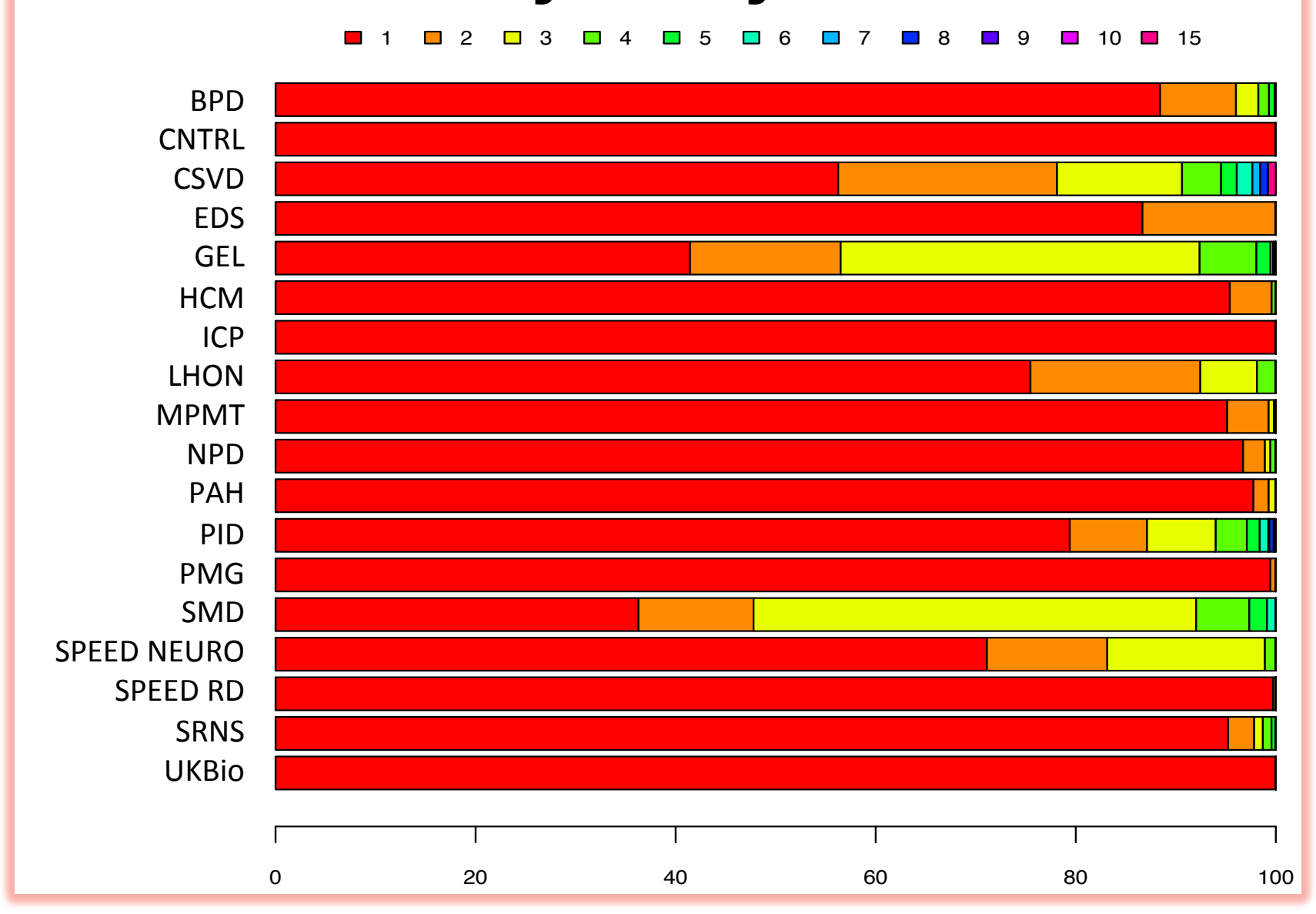
Transition to GRCh38

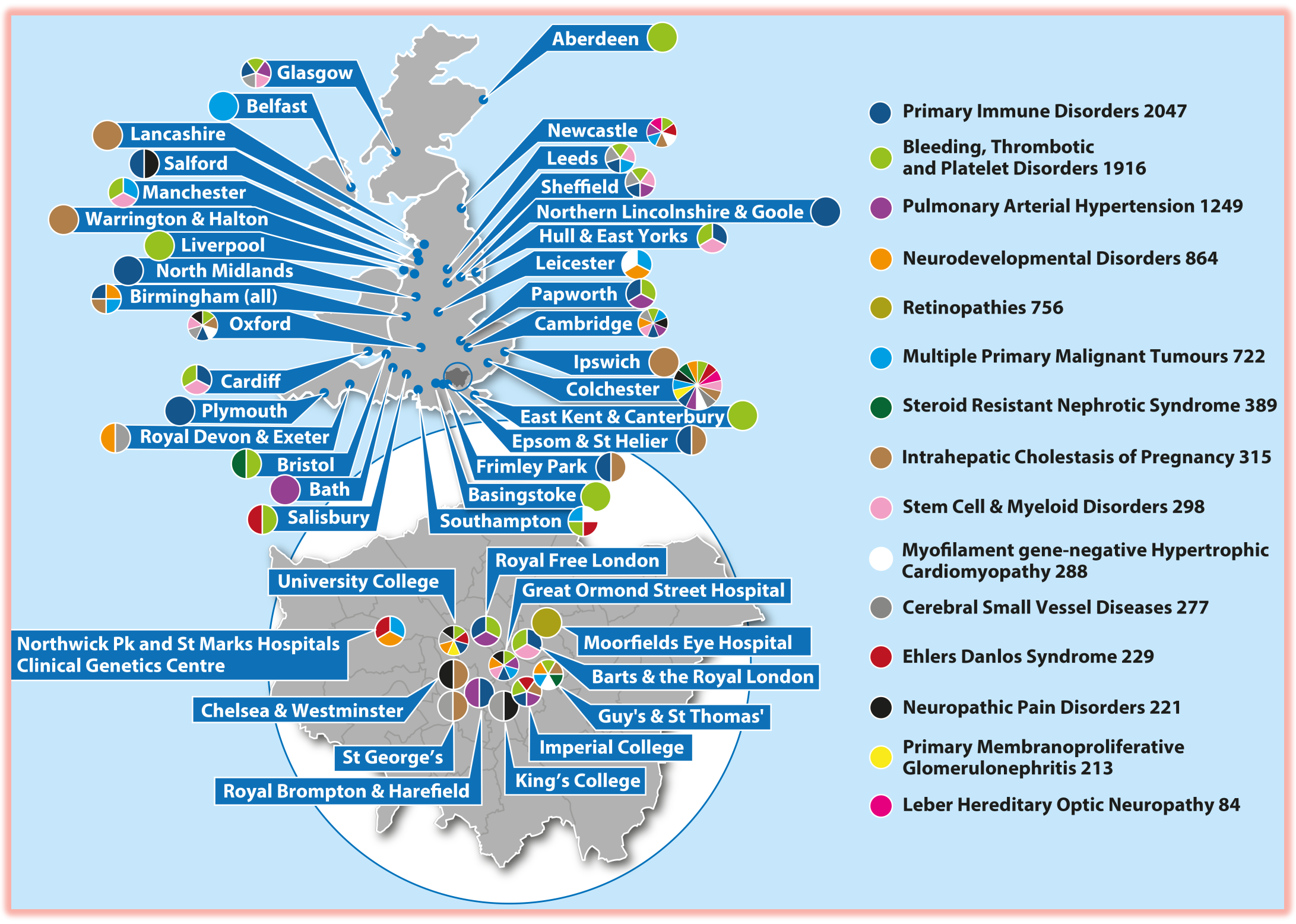
13K samples were aligned to GRCh38 and 1K samples aligned to GRCh37 for comparison by GENALICE using the same methodology. We quantified the increase in covered bases of the genome and the increased yield of variants between matching samples in GRCh37 and GRCh38. A common variant comparison in GRCh37 with the Non-Finnish European (NFE) gnomAD allele frequencies found a high correlation. Alignment and variant calling for GRCh38 was completed in 20 days using 10 compute nodes.

NIHR BioResource – Rare Diseases

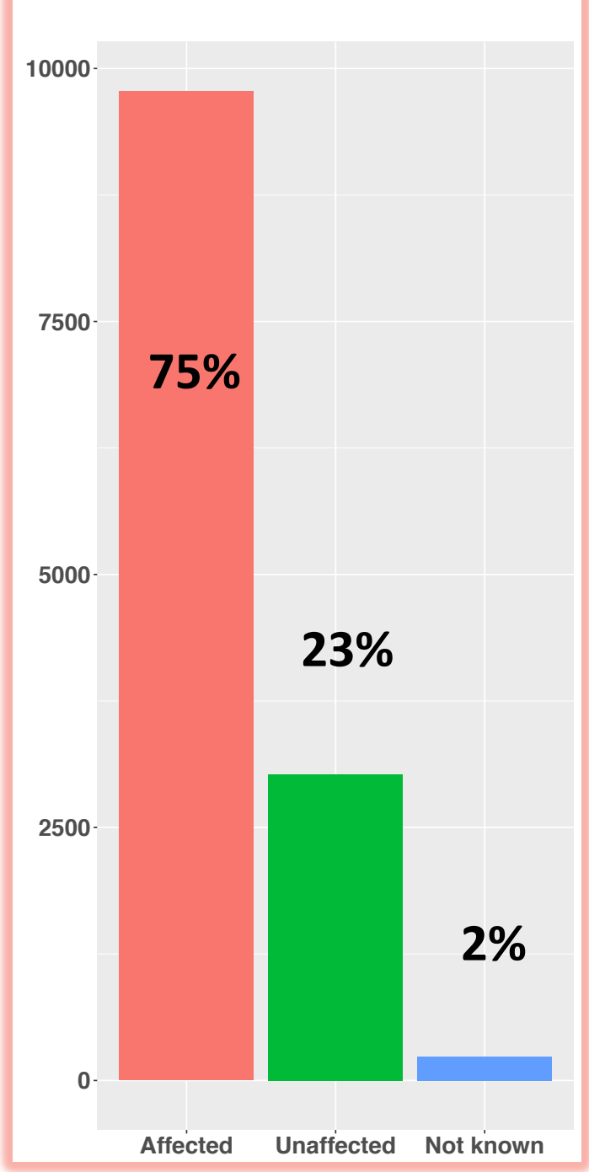
Individuals were recruited by 50 NHS Trusts and international collaborators, of which 75% were patients affected by a rare disease. The majority of the cohort are primary index cases with some larger families and trios for segregation studies. We identified 80.2% European, 9.2% Other, 7.2% South-Asian, 2.3% African, 0.08% East-Asian and 0.02% Finnish-European as part of the cohort.

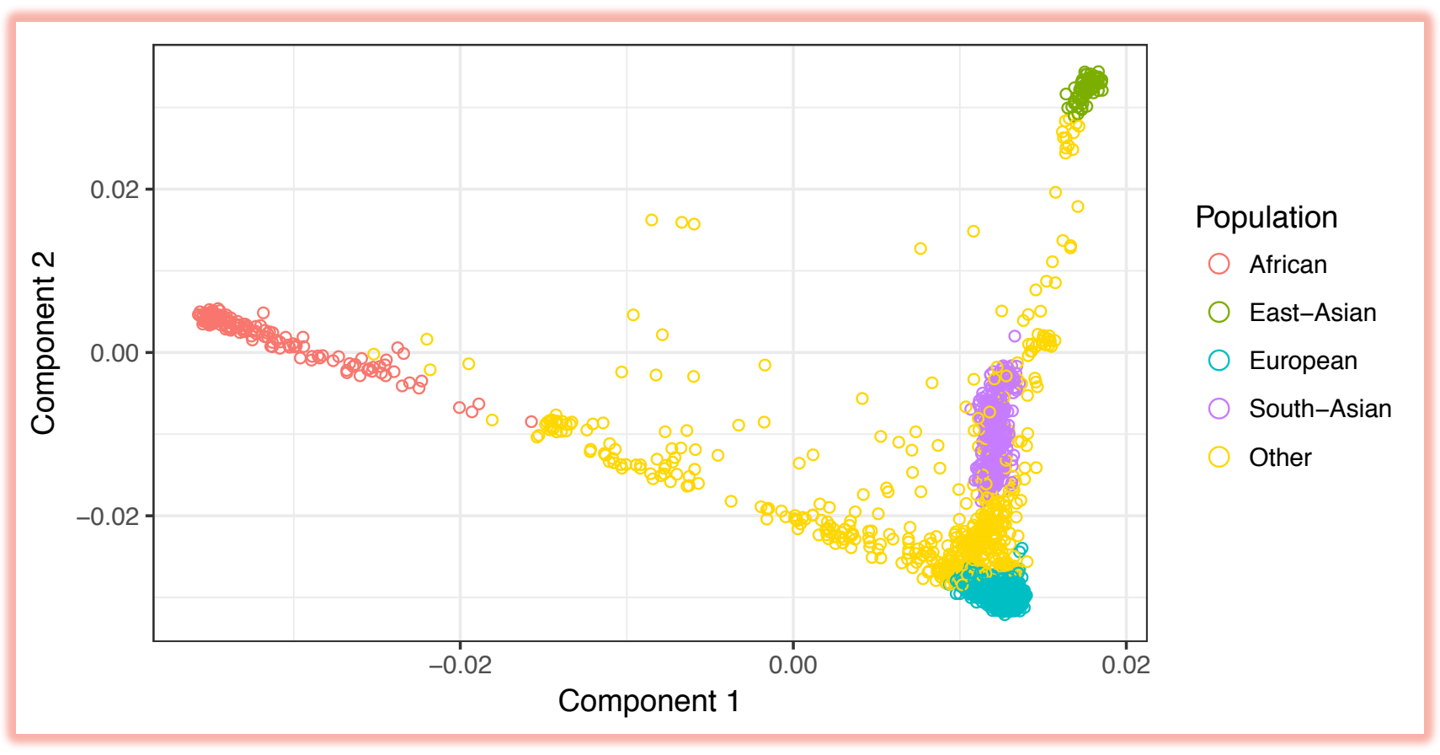
% of Cohort by family size recruited




Aberdeen
Glasgow
Belfast
Lancashire
Salford
Manchester
Warrington & Halton
Liverpool
North Midlands
Birmingham (all)
Oxford
Newcastle
Leeds
Sheffield
Northern Lincolnshire & Goole
Hull & East Yorks
Leicester
Papworth
Cambridge
Ipswich
Colchester
East Kent & Canterbury
Epsom & St Helier
Frimley Park
Basingstoke
Southampton
Cardiff
Plymouth
Royal Devon & Exeter
Bristol
Bath
Salisbury
Royal Free London
University College
Great Ormond Street Hospital
Northwick Pk and St Marks Hospitals Clinical Genetics Centre
Moorfields Eye Hospital
Barts & the Royal London
Chelsea & Westminster
Guy's & St Thomas'
Imperial College
St George's
King's College
Royal Brompton & Harefield
Primary Immune Disorders 2047
Bleeding, Thrombotic and Platelet Disorders 1916
Pulmonary Arterial Hypertension 1249
Neurodevelopmental Disorders 864
Retinopathies 756
Multiple Primary Malignant Tumours 722
Steroid Resistant Nephrotic Syndrome 389
Intrahepatic Cholestasis of Pregnancy 315
Stem Cell & Myeloid Disorders 298
Myofilament gene-negative Hypertrophic Cardiomyopathy 288
Cerebral Small Vessel Diseases 277
Ehlers Danlos Syndrome 229
Neuropathic Pain Disorders 221
Primary Membranoproliferative Glomerulonephritis 213
Leber Hereditary Optic Neuropathy 84

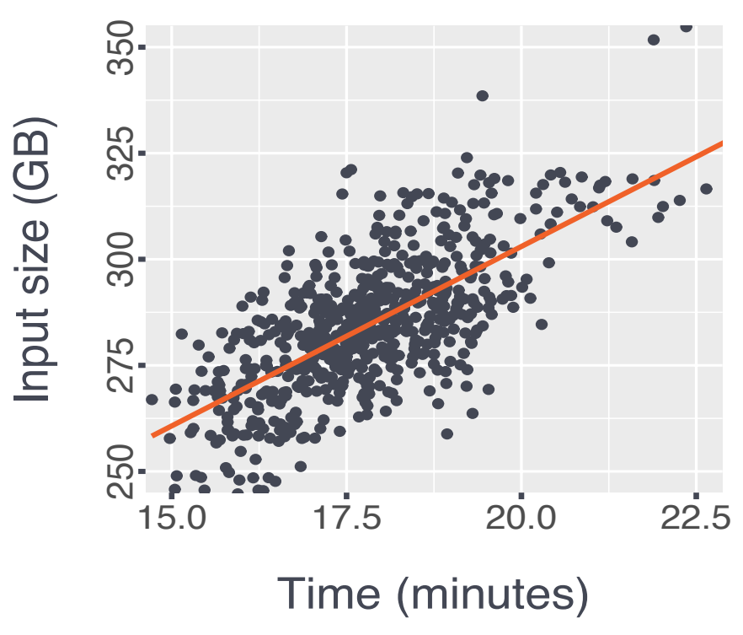
Phenotype





GRCh37 vs. GRCh38: Gain or pain?

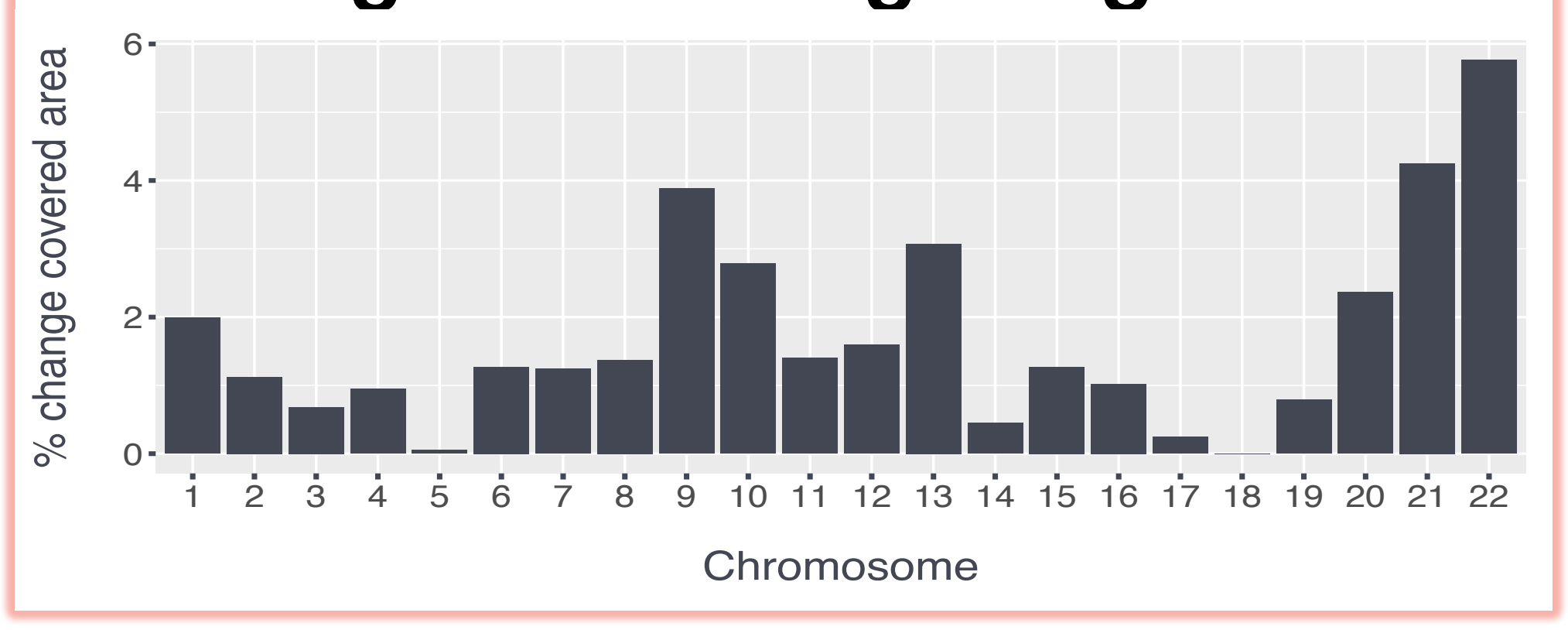
Read Alignment



Alignment of reads

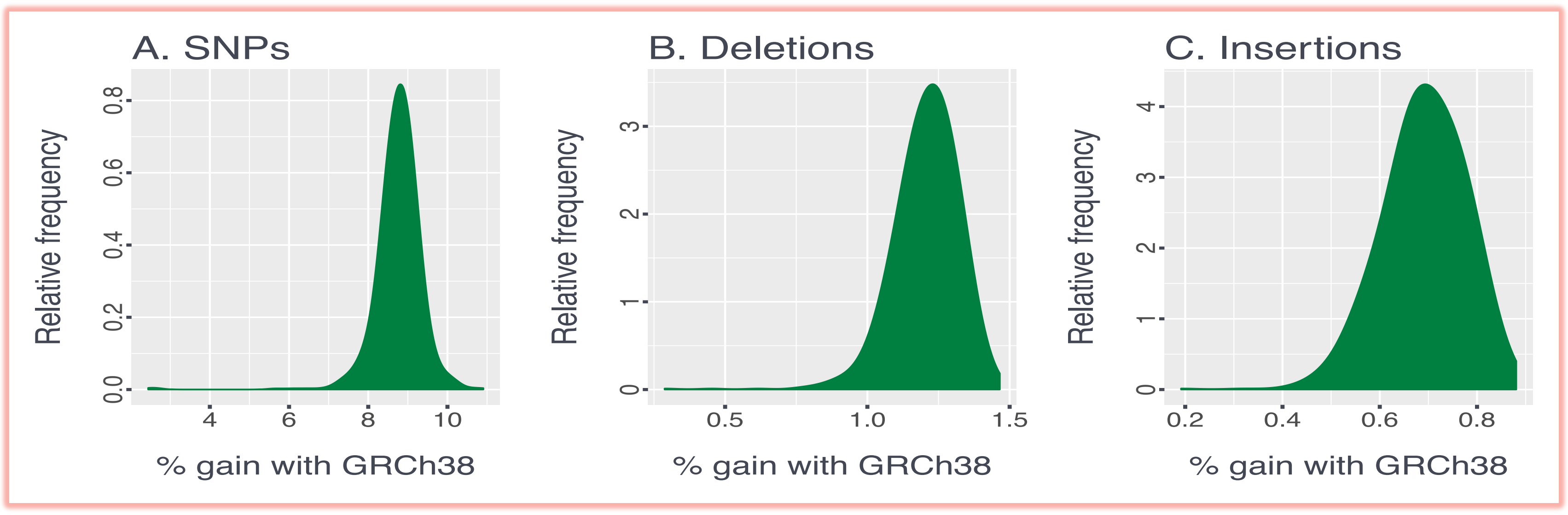
Read alignment time increases linearly with the amount of data independent of the reference genome. Changing to GRCh38 showed an increase in the number of covered bases for each chromosome while reducing the number of unmapped reads.

Change in coverage of genome

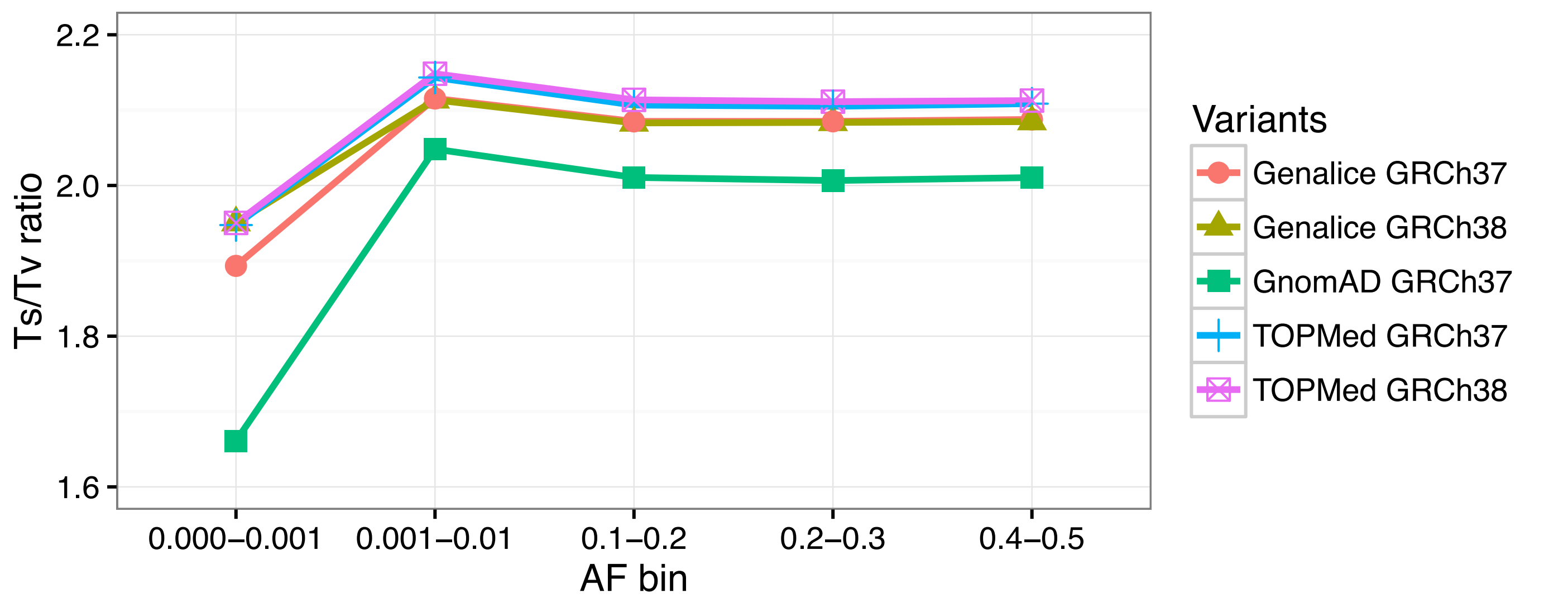


Variant call increase

In addition to the increase in covered bases, we found a gain of variants in GRCh38. Autosomal variants showed a change of 8.8%, 1.5% for SNVs and INDELs respectively.



SNPs: Ts/Tv ratio for AF ranges in different variant data sets



Quality metrics

We compared the variant calls from 1K selected GRCh37 and GRCh38 samples with available public datasets to assess the quality. The transition / transversion (Ts/Tv) ratio was calculated and compared for different allele frequency bins. TOPMed was lifted back from GRCh38 to GRCh37, yet use was limited due to lack of available ethnic specific frequencies.

HGMD

	GRCh37	GRCh38
Variants	13,471	13,450

Pathogenic variants

To measure the ability to recall variants in GRCh38 we used 1K selected samples and 2 sources: the Human Gene Mutation Database (HGMD) and pathogenic variants reported by MDTs. The concordance between the releases was 99.6% for HGMD entries and 100% for MDT reported variants.

Disease variants

	GRCh37	GRCh38
Cases	426	426

Conclusion

Analysis of 13,000 whole genomes shows that GRCh38 delivers better coverage and significantly more variants without detriment to quality. Rapid realignment and calling at scale to match changing genome builds is feasible and beneficial. The NIHR RD Sequence Variation browser will become publically available for both GRCh37 and GRCh38 providing variant summary information through a fast interactive browser (IVA).

References

- **NIHR BioResource – Rare Diseases**
https://bioresource.nihr.ac.uk/rare-diseases/welcome
- **GENALICE**
http://www.genalice.com
- **NIHR BR-RD Sequence Variation Browser**
https://goo.gl/ZQtmJF