bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GRID – Genomics of Rare Immune Disorders: a highly sensitive and specific diagnostic gene panel for patients with primary immunodeficiencies

[1,2]Ilenia Simeoni*, [1,2]Olga Shamardina*, [1,2]Sri VV Deevi*, [3]Moira Thomas, [1,2]Karyn Megy, [4]Emily Staples, [1,2]Deborah Whitehorn, [1,2]Daniel Duarte, [1,2]Rutendo Mapeta, [1,2,5]Willem H Ouwehand, [1,2]Christopher J Penkett, [1,2]Paula Rayner-Matthews, [1,2]Hannah Stark, [1,2]Jonathan C Stephens, [1,2]Kathleen E Stirrups, [1,2,6]Ernest Turro, [2]NIHR BioResource, [7]Adrian Thrasher, [8,9]Taco W Kuijpers, [4]Kenneth GC Smith, [10,11]Sinisa Savic, [12,13]Siobhan O Burns, [14]James E Thaventhiran* and [1,2]Hana Lango Allen*

[1] Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK

[2] NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, UK

[3] Gartnavel General Hospital NHS Greater Glasgow and Clyde, and Queen Elizabeth University Hospital, Glasgow, UK

[4] Department of Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK

[5] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[6] Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK

[7] University College London Great Ormond Street Institute of Child Health, London, UK

[8] Department of Paediatric Haematology, Immunology and Infectious Diseases, Emma Children's Hospital, Academic Medical Center, Amsterdam, The Netherlands

[9] Department of Blood Cell Research, Sanquin, Amsterdam, The Netherlands

[10] The Department of Clinical Immunology and Allergy and the National Institute for Health Research-Leeds Biomedical Research Centre, Leeds, UK

[11] Leeds Institute of Rheumatic and Musculoskeletal Medicine, St James's University Hospital, Leeds, UK

[12] University College London Institute of Immunity and Transplantation, London, UK

[13] Department of Immunology, Royal Free London NHS Foundation Trust, London, UK

[14] Medical Research Council Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK

* equal contribution

Corresponding author: Dr James Thaventhiran, e-mail: jedt2@cam.ac.uk, University of Cambridge, Cambridge, United Kingdom

Key words: Primary immunodeficiencies, gene panel, genetic diagnosis, next generation sequencing, whole genome sequencing, copy number variation

bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Primary Immune disorders affect 15,000 new patients every year in Europe. Genetic tests are usually performed on a single or very limited number of genes leaving the majority of patients without a genetic diagnosis. We designed, optimised and validated a new clinical diagnostic platform called GRID, Genomics of Rare Immune Disorders, to screen in parallel 279 genes, including 2015 IUIS genes, known to be causative of Primary Immune disorders (PID). Validation to clinical standard using more than 58,000 variants in 176 PID patients shows an excellent sensitivity, specificity. The customised and automated bioinformatics pipeline prioritises and reports pertinent Single Nucleotide Variants (SNVs), INsertions and DELetions (INDELs) as well as Copy Number Variants (CNVs). An example of the clinical utility of the GRID panel, is represented by a patient initially diagnosed with X-linked agammaglobulinemia due to a missense variant in the *BTK* gene with severe inflammatory bowel disease. GRID results identified two additional compound heterozygous variants in *IL17RC*, potentially driving the altered phenotype.

Introduction

Primary immunodeficiency disorders (PID) are a heterogeneous group of conditions that lead to infection susceptibility, malignancies and autoimmunity. Although disease-causing variants in over 300 genes have been identified, only about 10-20% of patients receive a molecular diagnosis. Identification of a monogenic cause can inform gene-specific treatment, even for patients with incompletely penetrant variants[1,2]. Targeted high-throughput sequencing (HTS) gene panels offer a cost-effective approach for simultaneous screening of many disease-causing genetic alleles[3].

Here we describe the design and validation to clinical standard of GRID (Genomics of Rare Immune Disorders, www.gridgenomics.org.uk), a targeted HTS panel for screening all known genetic causes of PID. We show that this panel covers 99% of both the full IUIS 2015 gene list[4] and the Human Gene Mutation Database (HGMD) variants associated with PID. In addition to having a considerably wider gene coverage compared to other diagnostic PID panels[5,6], our platform provides good coverage of GC-rich regions previously reported as difficult to capture. We provide sensitivity and specificity estimates by direct comparison of over 58,000 variants in 176 PID patients with both whole genome sequencing (WGS) and GRID panel data. We describe a customised and automated bioinformatics analysis pipeline to prioritise and report pertinent Single Nucleotide Variants (SNVs), INsertions and DELetions (INDELs) as well as Copy Number Variants (CNVs). Finally, we demonstrate the clinical utility of an unbiased gene screen on 86 PID patients who had not been genetically tested previously.

GRID panel design and coverage optimisation

We designed the GRID panel to capture all genes known to contain PID-causing variants, including the complete 2015 IUIS classification[4], currently 279 genes (Table S1). Regions targeted in the

bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

panel design are defined in Figure 1A, with an illustration of targeted and captured regions over *TBX1* gene. For prioritisation of variants for diagnostic reporting we defined regions of interest (ROI) as: coding or lincRNA exons plus 15bp either side; and non-coding PID-associated HGMD DM variants. The validation of the panel described henceforth is based on the ROI.

To optimise panel coverage, we initially sequenced 59 samples and identified regions of sub-optimal (<20 reads) coverage in at least 5% of samples (problematic regions). We increased the number of probes in those regions through a process known as bait rebalancing (Figure 1B and Supplementary Methods). We then sequenced 176 samples from PID patients, and use these to show the coverage improvement by bait rebalancing and to report our final coverage metrics. We capture 643,088 of the 655,348 bases representing the ROI by at least 20 reads, corresponding to the average per gene fraction of ROI coverage of 98.96% (median 100%) (Figures 1B and 1C). The number of HGMD variants covered at ≥20X increased from 99.0% to 99.4% (Figure 1B). Overall, the mean coverage of the ROI across the 176 samples was 215X (min 148, max 314).

An example of improved coverage in one gene is shown in Figure 1B. Furthermore, in a number of genes previously reported to be difficult to capture, such as *JAK3*[7,8], we obtained 100% coverage of at least 20X for all coding bases (Figure S1). Overall, all but 13 genes have >99.9% ROI coverage of ≥20X (Figure S2). Regions of systematically low coverage are primarily restricted to 4 genes (*C4A*, *C4B*, *IKBKG* and *NCF1*, see Figures 1D and S2), which have a high proportion of non-uniquely mapped reads owing to regions of high sequence homology with these genes elsewhere in the genome. This is a general limitation of short-read sequencing technology that cannot be overcome by bait rebalancing.

GRID bioinformatics analysis and validation

We further developed the automated bioinformatics pipeline previously reported by Simeoni et al.[9] Improvements were made in: pertinent variant prioritisation within the ROI (Figure 2A); gene coverage quality control for each sample; and the CNV calling algorithm (Figure 2B and Supplementary Methods).

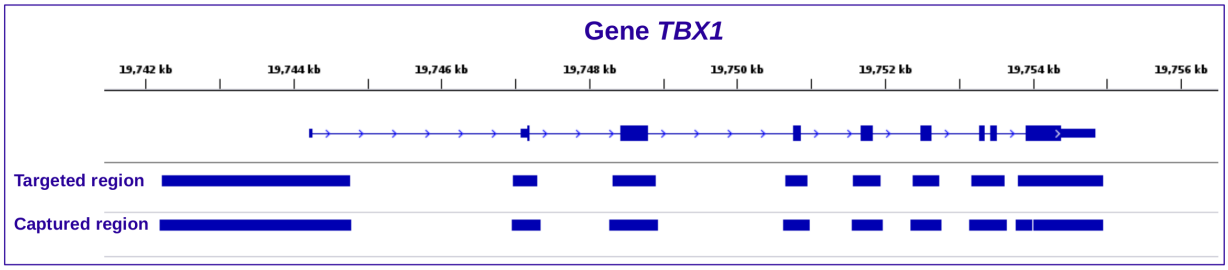
We assessed the sensitivity and specificity of the GRID panel using the 176 PID patient samples previously analysed by WGS, of which 53 patients had at least one pathogenic variant identified. We based these analyses on the ROI of the 279 PID genes. Before computing sensitivity and specificity metrics, we attempted to resolve all discrepant calls by Sanger sequencing to create a gold standard set of variant calls. Based on 52,731 SNV and 1,178 INDEL gold standard calls across all 176 samples, we obtained sensitivities of 99.87% (95% CI 98.87%-99.89%) and 99.83% (95% CI 99.47%-99.97%), respectively. The specificity values obtained for the 56,819 SNVs and 1,315 INDELs called by the GRID platform where the WGS data had ≥20X coverage and pass

bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A

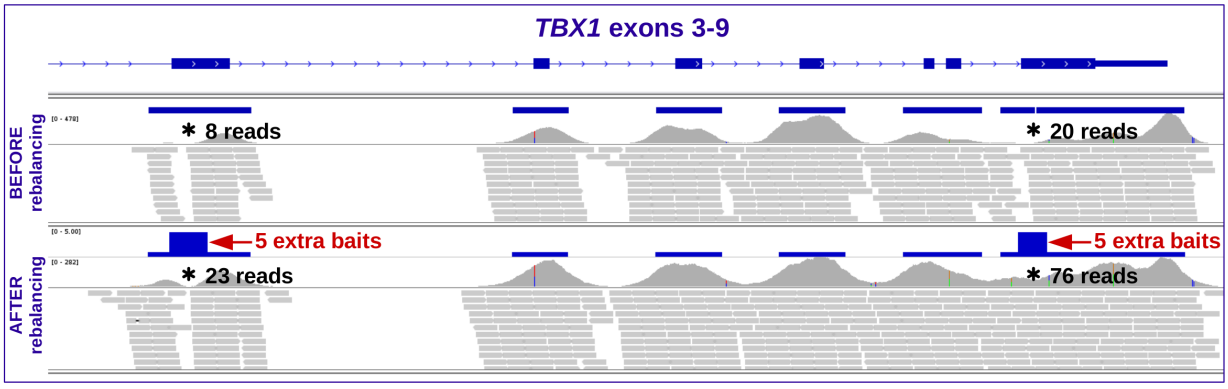
GENE PANEL CONTENT: 279 genes including all the genes in the IUIS 2015 Classification (total 2454646 bp)

Targeted regions for all selected transcripts of each gene:
All exons ± 100bp; 2000bp at the 5' and 500bp at the 3' of the first exon; all HGMD variants (± 25bp) including non coding

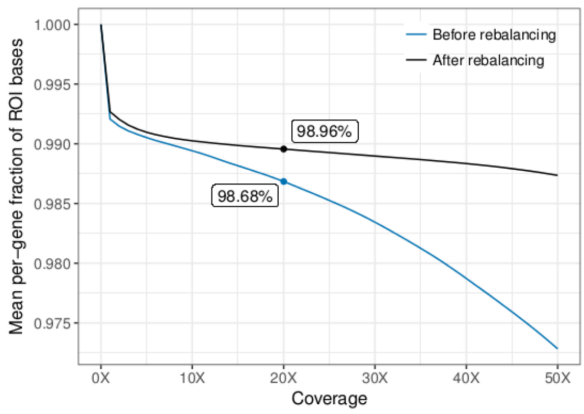


B

Metrics (coverage of at least 20X)	BEFORE rebalancing	AFTER rebalancing
Mean per gene fraction of ROI bases covered	98.68%	98.96%
Problematic HGMD variants (of total 9966 variants)	98	62
Number of genes covered at:		
100%	219	245
99.9-100%	29	20
90-99.9%	24	9
<90%	6	4



C



D

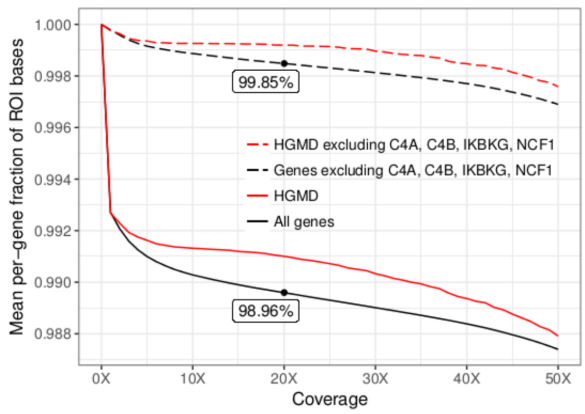


Figure 1. GRID panel design and coverage optimisation. (A) GRID panel content and example of targeted and captured regions in *TBX1* gene. (B) Improved coverage metrics and example of improved read depth in *TBX1*. (C) Mean per-gene fraction of ROI bases before and after rebalancing. (D) Mean per-gene fraction of ROI bases in all genes (solid black line) and HGMD variants (solid red line) in 176 samples; the same metrics excluding problematic genes (dashed lines). ROI = regions of interest.

bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

filter call over those positions (even if called as homozygous reference), were 99.88% (95% CI 98.88%-99.90%) and 99.47% (95% CI 99.01%-99.75%), respectively. All 69 pertinent variants reported across the 53 patients based on WGS data were correctly called and prioritised by the GRID pipeline, including 6 large deletions.

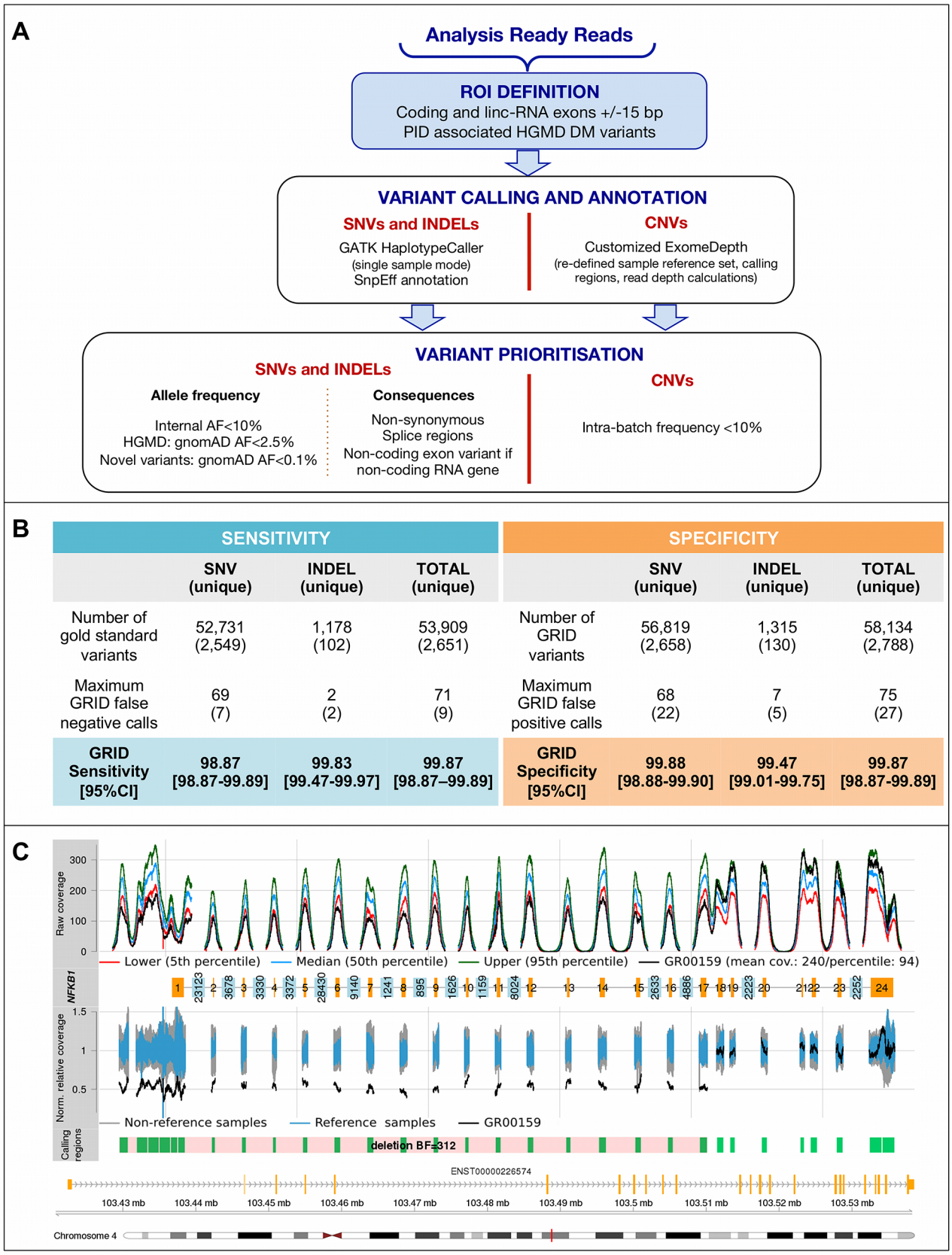
Intra-run reproducibility of the platform was assessed by sequencing three library preparations of the same sample on a single plate. The average pairwise genotype concordance of within ROI was 98.9%. Another set of three samples with three library preparations each was used to assess inter-run reproducibility, and the average pairwise ROI genotype concordance rates across three plates was 97.3% (Supplementary Methods and Table S2). Within the exonic and HGMD regions the intra- and inter-run concordances were even higher, at 99.2% and 98.9%, respectively. Together, these data show that the GRID panel generates highly reproducible variant calls, especially within the regions known to contain PID-causing variants.

We made further use of the WGS data to optimise the specificity of automated CNV calls, which has been rarely performed in previously described diagnostic gene panels. The GRID CNV pipeline identified 11 deletions and 21 duplications in 176 PID samples. WGS data confirmed 9 deletions and 10 duplications, ranging from 1 exon to multi-gene CNVs. All false calls had a lower confidence as measured by the Bayes factor statistic, allowing us to set a threshold above which to report CNVs (Figure S3). The 9 reliable deletion calls spanned 18 genes (single gene calls: *ARPC1B*, *CFHR4*, *DOCK8*, *IGKC*, *LRBA*, *NFKB1*, *TBX1* and one multi-gene call: *ATM*, *CD3D*, *CD3E*, *CD3G*, *CTSC*, *IL10RA*, *IL18*, *MRE11*, *SLC37A4*, *TIRAP*, *ZBTB16*), and 10 duplication calls spanned 9 genes (*CD8A*, *DOCK8*, *IKBKB*, *LYST*, *MASP2*, *MCM4*, *PRKDC*, *STAT2*, *UNG*). Each CNV can be visually inspected in an automatically generated plot (Figure 2B).

Clinical application

We assessed the clinical utility of the GRID panel using new referrals of 86 PID patients without a genetic diagnosis. The automated prioritisation of variants generated an average of 5 variants (4.3 SNVs, 0.5 INDELs and 0.2 CNVs) per sample. All prioritised variants were reviewed in a Multi-Disciplinary Team meeting and categorised as clearly pathogenic, likely pathogenic or of uncertain significance, and for their contribution to the patient's phenotype. This resulted in a genetic diagnosis for 10 of patients (11.6%). One patient with previously diagnosed Bruton's agammaglobulinaemia and inflammatory bowel disease had disease-causing variants in both *IL17RC (p.Arg378* and p.Arg751Serfs*14)* and *BTK (p.Glu479Lys)*. Conventionally, this patient would have had a Sanger sequencing test of the *BTK* gene as a confirmation of the clinical diagnosis, albeit with an unusual presentation. The unbiased GRID approach identified an additional variant in a second gene that explains what was initially thought to be an expanded *BTK* phenotype.

bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



	SENSITIVITY				SPECIFICITY		
	SNV (unique)	INDEL (unique)	TOTAL (unique)		SNV (unique)	INDEL (unique)	TOTAL (unique)
Number of gold standard variants	52,731 (2,549)	1,178 (102)	53,909 (2,651)	Number of GRID variants	56,819 (2,658)	1,315 (130)	58,134 (2,788)
Maximum GRID false negative calls	69 (7)	2 (2)	71 (9)	Maximum GRID false positive calls	68 (22)	7 (5)	75 (27)
GRID Sensitivity [95%CI]	**98.87 [98.87-99.89]**	**99.83 [99.47-99.97]**	**99.87 [98.87–99.89]**	**GRID Specificity [95%CI]**	**99.88 [98.88-99.90]**	**99.47 [99.01-99.75]**	**99.87 [98.87-99.89]**

Figure 2. (A) GRID data processing, variant calling and prioritisation. (B) Sensitivity and Specificity of SNV and INDEL detection. (C) CNV plot example showing *NFKB1* heterozygous deletion call (pink block in the "Calling regions" track) made by the CNV pipeline automatically combining individual calls over exons 1-17 (green bars). "Norm. relative coverage" track visually confirms ~50% relative coverage in this sample (black lines) compared to reference samples (blue lines). Full explanation of the plot is provided in Supplementary Methods.

bioRxiv preprint doi: https://doi.org/10.1101/431544; this version posted January 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In conclusion, we report the design, optimisation and validation of a new HTS panel for PID patients. We show that the GRID panel has excellent sensitivity and specificity for SNVs, INDELs and CNVs, and overcomes previously encountered challenges with coverage of GC-rich regions harbouring clinically relevant variants. We demonstrate its potential as a first-line, unbiased genetic test to provide a conclusive molecular diagnosis for patients suspected of having a highly penetrant genetic basis of their immune disorder.

Acknowledgements

Funding for the project was provided by the National Institute for Health Research (NIHR, grant number RG65966). J.E.D.T. is supported by the MRC (RG95376 and MR/L006197/1). A.J.T. is supported by the Wellcome Trust (104807/Z/14/Z) and the NIHR Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London.

Conflict of Interest statement

S.O.B. has received personal fees from CSL Behring, Immunodeficiency Canada/IAACI, CSL Behring, Baxalta, and Octagam. A.J.T. has received personal fees from Orchard Therapeutics, Rocket Pharmaceuticals, and Torus Therapeutics. S.S. has received personal fees from Grifols, Octapharma, and Shire. The rest of the authors declare that they have no relevant conflicts of interest.

References

1. Chinen J et al., “Advances in basic and clinical immunology in 2016” J Allergy Clin Immunol 2017;140:959-973.
2. Hye Sun Kuehn et al, “Immune dysregulation in human subjects with heterozygous germline mutations in *CTLA4*” Science 2014; 345:1623–1627.
3. Seleman et al., “Uses of Next-Generation Sequencing Technologies for the Diagnosis of Primary Immunodeficiencies” Front Immunol 2017.
4. Bousfiha A. et al., “The 2015 IUIS Phenotypic Classification for Primary Immunodeficiencies” J. Clin Immunol 2015;35:727-38.
5. Stoddard et al., “Targeted NGS: A Cost-Effective Approach to Molecular Diagnosis of PIDs” Front Immunol 2014;5:531.
6. Al-Mousa et al., “Unbiased targeted next-generation sequencing molecular approach for primary immunodeficiency diseases” J Allergy Clin Immunol 2016;137:1780-1787.
7. Yu et al., “Rapid molecular diagnostics of severe primary immunodeficiency determined by using targeted next-generation sequencing” J Allergy Clin Immunol 2016;1142-1151.
8. Moens et al., “Diagnostics of Primary Immunodeficiency Diseases: A Sequencing Capture Approach” PLoS ONE 2014;9(12): e114901.
9. Simeoni et al., “A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders” Blood 2016;127:2791-803.