A novel next-generation sequencing capture-based strategy to report somatic hypermutation status using genomic regions downstream to immunoglobulin rearrangements

Neil McCafferty,[1,2] James Peter Stewart,[1] Nikos Darzentas,[3] Jana Gazdova,[1] Mark Catherwood,[4] Kostas Stamatopoulos,[5] Anton W. Langerak[6] and David Gonzalez[1]

[1]Patrick G Johnston Centre for Cancer Research, Queens University Belfast, Belfast, UK; [2]Barts Cancer Institute, Queen Marys University, London, UK; [3]Department of Hematology, University Hospital Schleswig-Holstein, Kiel, Germany; [4]Belfast City Hospital, Belfast, UK; [5]Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece and [6]Laboratory of Medical Immunology, Department of Immunology, Erasmus University Medical Center, Rotterdam, The Netherlands

Correspondence: D. Gonzalez
D.GonzalezdeCastro@qub.ac.uk

Received: August 15, 2022.
Accepted: December 15, 2022.
Early view: December 29, 2022.

https://doi.org/10.3324/haematol.2022.281928

Supplementary Data

Supplementary Methods

Library preparation and sequencing

An optimised and standardised library preparation and sequencing protocol was followed according to the EuroClonality-NDC protocol[28]. The additional 61 CLL samples matched the EuroClonality-NDC strategy except 500 ng gDNA was used for library preparation and no pre-hybridisation PCR cycles. These samples were pooled and sequenced on the NextSeq 500 (Illumina, Cambridge, UK) using a 75 bp paired-end strategy on a 150-cycle NextSeq 500/550 Mid Output Kit (Illumina, Cambridge, UK).

Distinction of SHM from non-somatic variants

Polymorphic variants in IGHJ-E were filtered out using the population database GnomAD (v3) and several minor allele frequency (MAF) filters for the global population (>2%, >1%, >0.2%, >0.1%, >0.02% and >0.01%, Figure S4)[46]. Recurrent polymorphic and artefactual variants were also identified from the 331 tumour cases if they were present above specified thresholds, using several tumour frequency (TF) filters (>50%, >40%, >30%, >20% and >10%, Figure S5). Minimum variant allele frequency (VAF) filters were also examined to account for variants below the limit of detection (>5%, >4%, >3%, >2% and >1%, Figure S6) since these could represent low-level artefactual variants. All remaining SNVs were visually assessed using IGV to ensure variants were somatic mutations (accounting for remaining alignment artefacts, Figure S7). Variant filters were sequentially optimised, and the best performing filter according to ROC/AUC analysis was selected in each analysis for a combined final filtering strategy containing the following parameters: (i) GnomAD MAF>0.02%, (ii) TF>10%, (iii) VAF>5%, (iv) IGV assessment (Figure S3).

Supplementary Tables

Diagnosed Disease	n (%)	B/T cell
Chronic lymphocytic leukaemia (CLL)[a]	98 (29.6)	B
Mantle cell lymphoma (MCL)	34 (10.3)	B
Follicular lymphoma (FL)	34 (10.3)	B
Plasma cell myeloma (PCM)	24 (7.3)	B
Burkitt lymphoma (BL)	21 (6.3)	B
Diffuse large B-cell lymphoma (DLBCL)	21 (6.3)	B
B Acute Lymphoblastic leukaemia (B-ALL)	16 (4.8)	B
Marginal zone lymphoma (MZL)	7 (2.1)	B
Mucosa-Associated Lymphoid Tissue (MALT)	3 (0.9)	B
T Acute Lymphoblastic leukaemia (T-ALL)	40 (12.1)	T
Anaplastic large-cell lymphoma (ALCL)	12 (3.6)	T
CD30+ anaplastic large cell lymphoma (C-ALCL)	1 (0.3)	T
Angioimmunoblastic T-cell lymphoma (AITL)	9 (2.7)	T
Peripheral T-cell lymphoma, NOS (PTCL)	1 (0.3)	T
Intestinal T-cell lymphoma, NOS (ITCL)	1 (0.3)	T
Mycosis fungoides (MF)	3 (0.9)	T
Sézary syndrome (SS)	3 (0.9)	T
T-lymphoblastic lymphoma (T-LBL)	1 (0.3)	T
Enteropathy-associated T-cell lymphoma (EATL)	2 (0.6)	T
Total	331	258 B 73 T

Supplementary Table 1. Diagnosed disease types of 331 samples. [a]95/98 CLL samples with Sanger sequencing data and in-frame IGH VDJ rearrangements identified by IMGT. CLL cases made up of 37 EuroClonality-NDC study cases and an additional 61 diagnostic CLL cases.

Sanger Sequencing IGHV	n (%)	Detected by IGV (n)	Detected by ARResT/Interrogate (n)
V1-2	1 (1.05%)	1	1
V1-3	2 (2.11%)	1	2
V1-46	1 (1.05%)	1	1
V1-69	13 (13.68%)	12	13
V1-8[a]	3 (3.16%)	0	2
V2-26	1 (1.05%)	0	1
V2-5	3 (3.16%)	1	2
V3-11	2 (2.11%)	0	2
V3-15	2 (2.11%)	0	2
V3-21	9 (9.47%)	6	6
V3-23	7 (7.37%)	4	7
V3-30	7 (7.37%)	3	6
V3-33	3 (3.16%)	2	3
V3-48	6 (6.32%)	5	6
V3-49	1 (1.05%)	1	1
V3-53	1 (1.05%)	1	1
V3-7	5 (5.26%)	4	5
V3-74	4 (4.21%)	2	4
V3-9[b]	1 (1.05%)	0	1

Sanger Sequencing IGHV	n (%)	Detected by IGV (n)	Detected by ARResT/Interrogate (n)
V4-31	1 (1.05%)	0	1
V4-34	9 (9.47%)	2	7
V4-38[a]	1 (1.05%)	0	1
V4-39	5 (5.26%)	4	4
V4-4	1 (1.05%)	0	1
V4-61	2 (2.11%)	0	2
V5-10	1 (1.05%)	1	1
V5-51	3 (3.16%)	2	3
Total	95	53	86

Supplementary Table 2. IGHV genes identified by Sanger Sequencing in 95 CLL samples. IGV concordance: 55.79% (95% CI: 45.23-65.98%). ARResT/Interrogate concordance: 89.47% (95% CI: 81.49-94.84%). Sanger Sequencing SHM data showed 50 M-CLL and 45 U-CLL. [a]IGHV genes not present on GRCh38 or [b]not captured as part of the EuroClonality-NDC panel, and therefore not assessable.

Supplementary Figure Legends

Supplementary Figure 1. Comparison of Sanger sequencing IGHV Identity in 95 CLL cases where the reported clonal IGHV gene could be detected by: A) Visual assessment with IGV (genomic alignment-based), where 53/95 (55.8%) cases were concordant with Sanger sequencing. B) ARResT/Interrogate, where 86/95 (90.5%) cases were concordant. Wilcoxon signed-rank: ns (not significant): $p>0.05$; ****: $p\leq0.0001$.

Supplementary Figure 2. Reporting SNVs in IGHV using an NGS genomic alignment-based strategy. IMGT output of Sanger sequences reported on left (SNVs numbered in blue), IGV track of the rearranged IGHV on right (corresponding SNVs numbered with successful variant calling in green and unsuccessful in red). Example 1. Borderline case where the rearranged IGHV was detected by IGV and ARResT/Interrogate: NGS reads are aligned to the genome, 7/7 SNVs are detected. Example 2. Mutated case where the clonally rearranged IGHV was detected by ARResT/Interrogate but not IGV: NGS reads are aligned to the genome, however, only 1/6 shown SNVs are detected (1/17 in total).

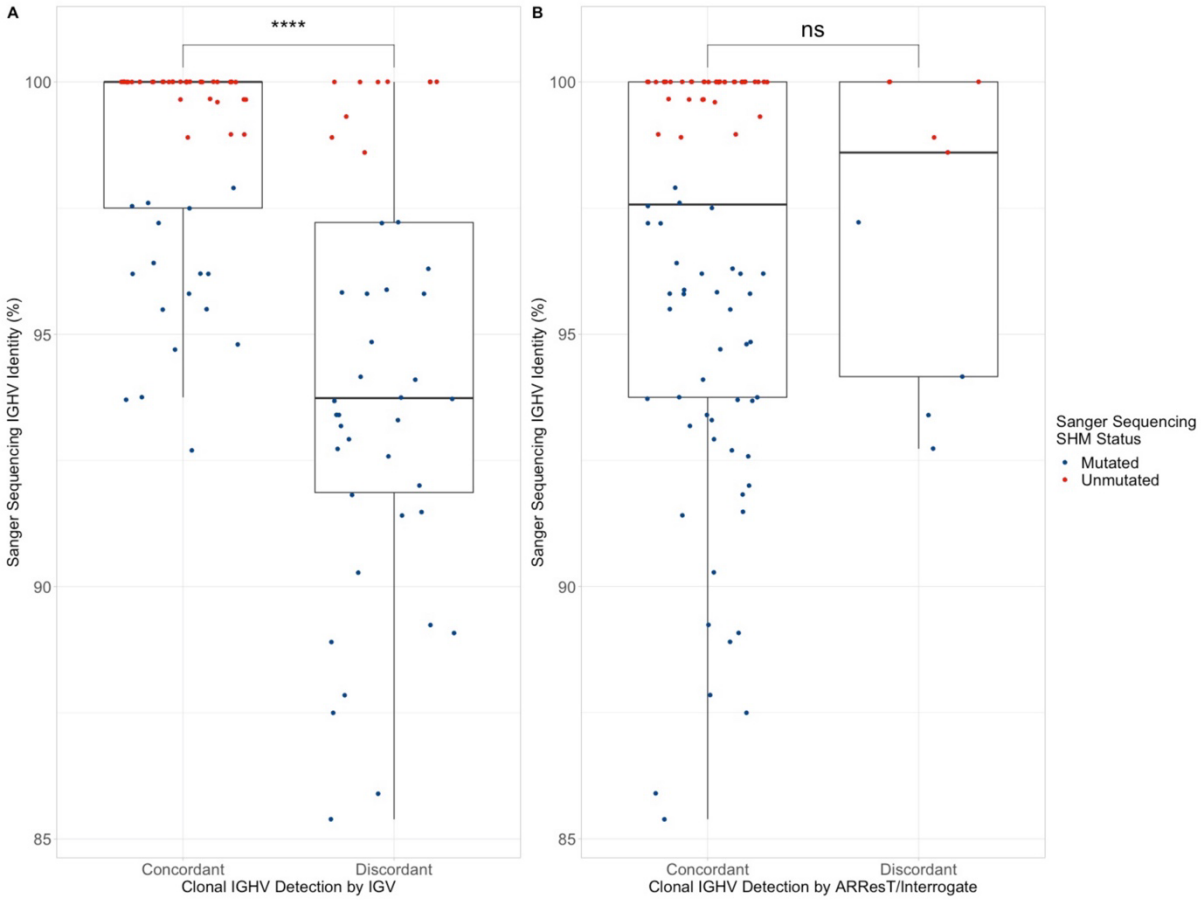
Supplementary Figure 3. ROC curves of IGHJ-E with Sanger Sequencing SHM status, selected variant filters and final IGHJ-E following SNV assessment for alignment artefacts. Somatic variant filters were selected sequentially, e.g., the best performing GnomAD filter (MAF>0.02) was selected and carried through to the following analyses.

Supplementary Figure 4. ROC curves of IGHJ-E with Sanger sequencing SHM status. Population variants were selected based on minor allele frequency (MAF) 0.01-2% applied to filter SNPs. Population SNPs were sourced from GnomAD using all available GRCh38 sequences.

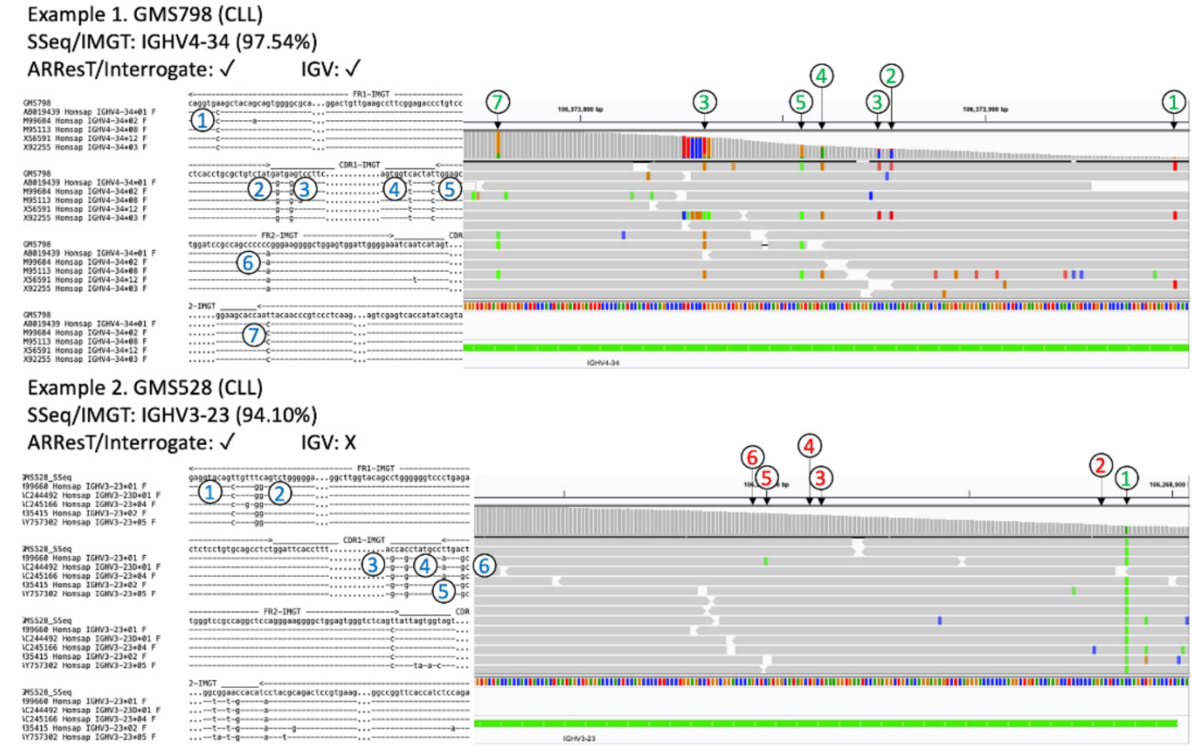
Supplementary Figure 5. ROC curves of IGHJ-E with Sanger sequencing SHM status. Recurrent SNPs with a tumor frequency (TF) 10-50% were used. A list of tumor SNPs were generated using all available EuroClonality-NDC study and diagnostic CLL cases (n=331)

Supplementary Figure 6. ROC curves of IGHJ-E with Sanger sequencing SHM status. Minimum variant allele frequencies (VAF) 1-5% were applied to filter SNVs**.**

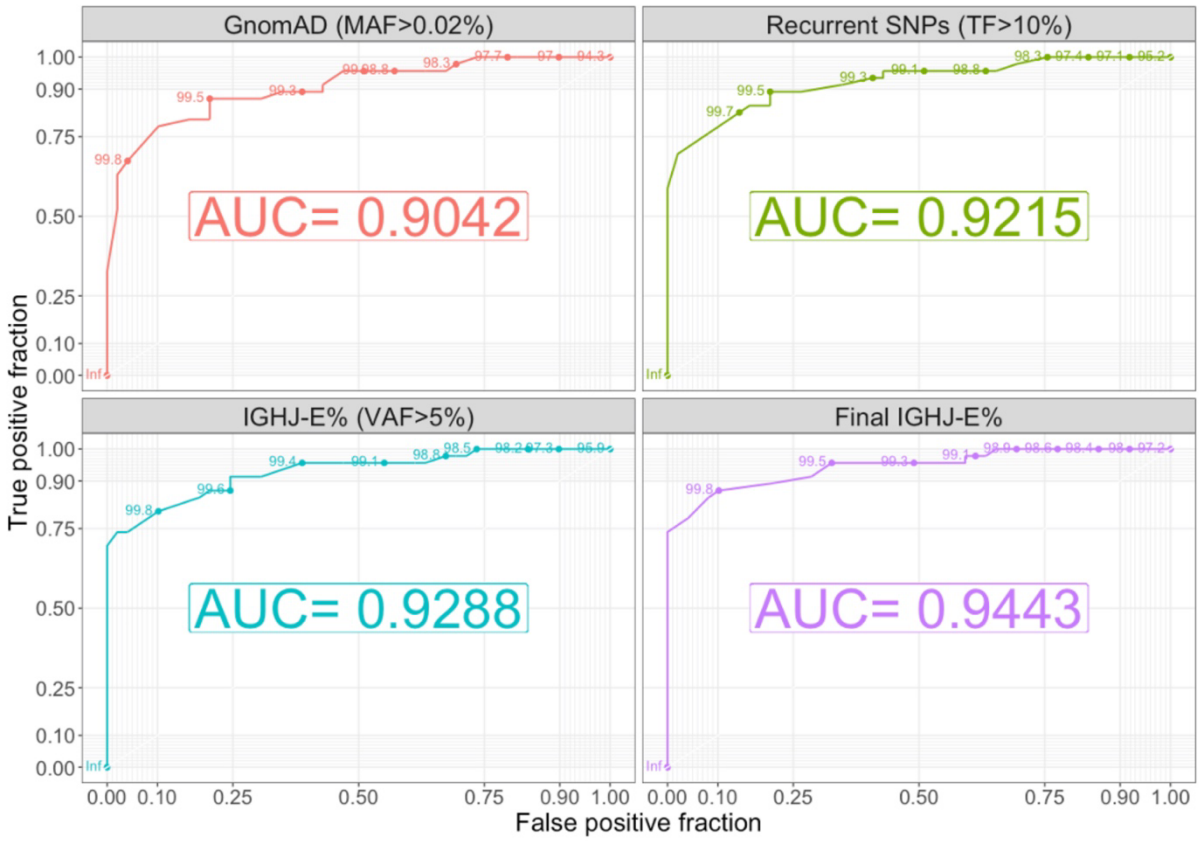
Supplementary Figure 7. A) IGV screenshot of IGHJ-E analysed region. This sample was reported as unmutated by Sanger sequencing. B) Zoomed region highlights sequencing/alignment artefact resulting in several false somatic variants. Variants were removed from the analysis and IGHJ-E identity was updated.



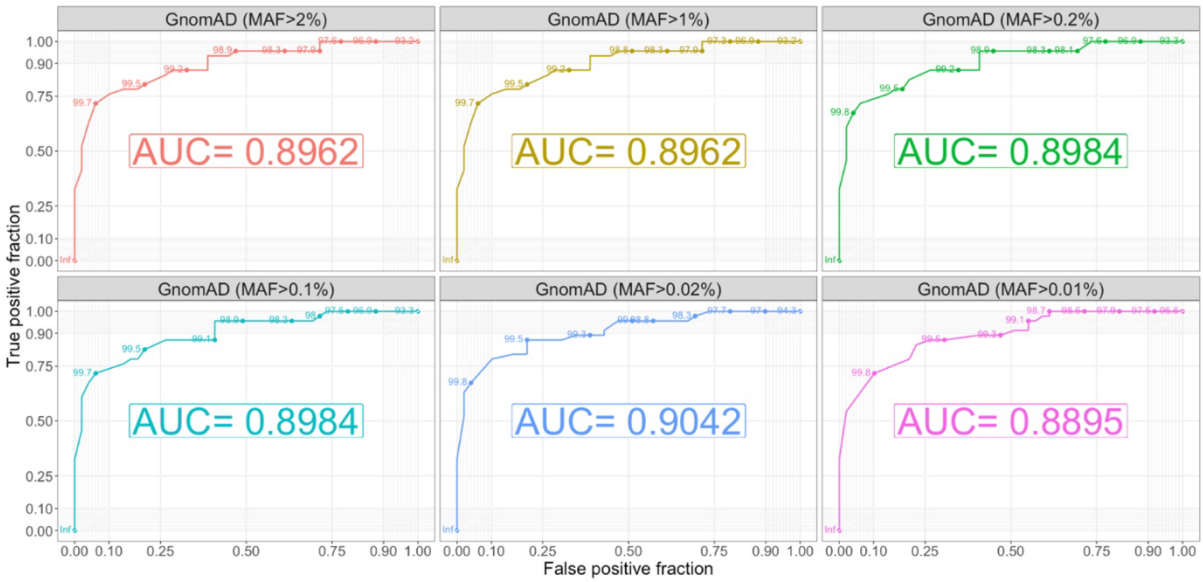
Supplementary Figure 1.



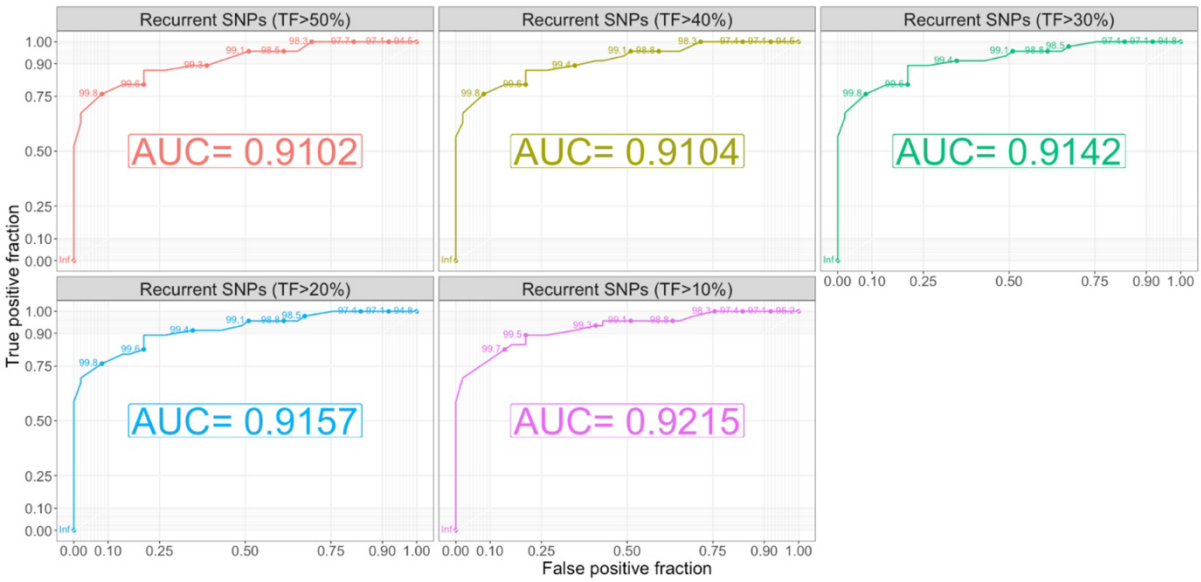
Supplementary Figure 2.



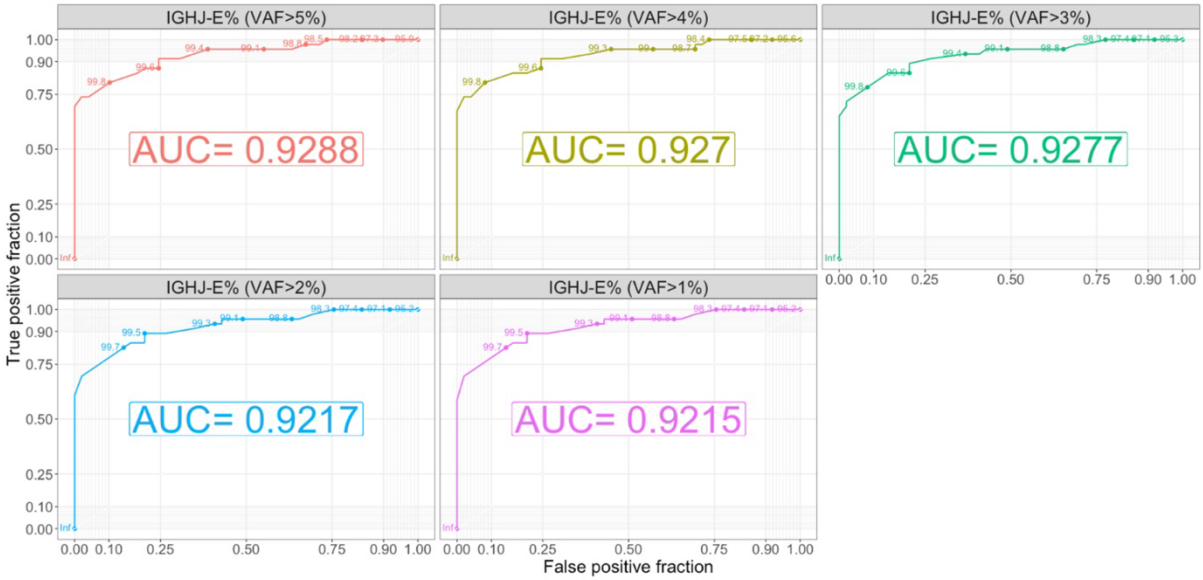
Supplementary Figure 3.



Supplementary Figure 4.

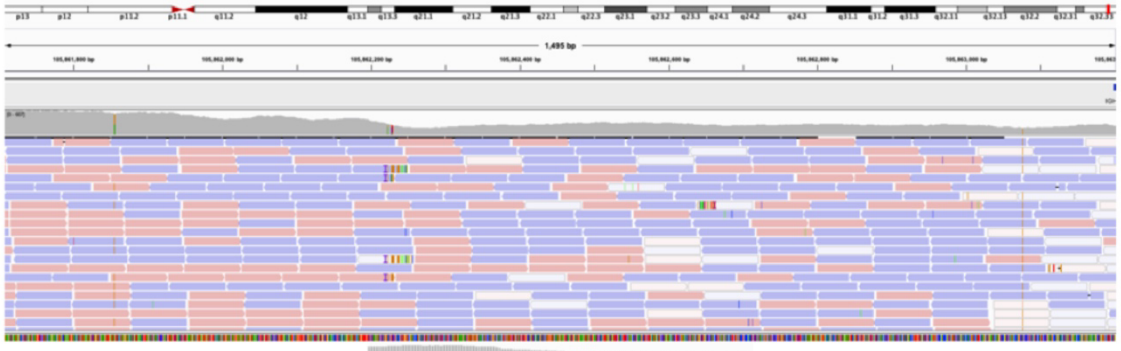


Supplementary Figure 5.

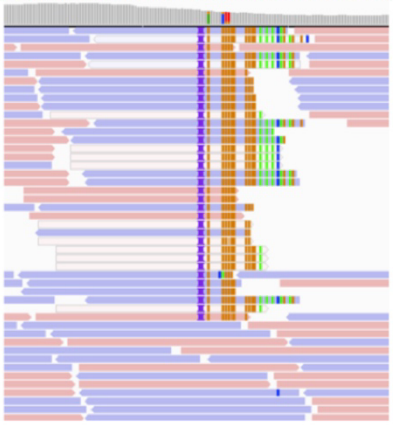


Supplementary Figure 6.

A.



B.



Supplementary Figure 7.