Nanopore sequencing enables allelic phasing of *FLG* loss-of-function variants, intragenic copy number variation and methylation status in atopic dermatitis and ichthyosis vulgaris

Colin WONG, BSc,[1,*] Cheng-Yong THAM, PhD,[2,*] Lin YANG, PhD,[2,*] Miles C. BENTON, PhD,[2] Vipin NARANG, PhD,[3] Simon DENIL, PhD,[1] Kaibo DUAN, PhD,[3] Yik Weng YEW, MD,[4] Bernett LEE, PhD,[3,6,7] Paola Florez de SESSIONS, PhD,[2] John E. A. COMMON, PhD,[1,5,#]

Affiliations

[1]A*STAR Skin Research Laboratory, A*STAR Research Institutes, Singapore

[2]Oxford Nanopore Technologies, Singapore

[3†]Singapore Immunology Network, A*STAR Research Institutes, Singapore

[4]National Skin Centre, National Healthcare Group, Singapore

[5]Skin Research Institute of Singapore

[6 ¶] Centre for Biomedical Informatics, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

[7]A*STAR Infectious Diseases Labs (A*STAR ID Labs), Agency for Science, Technology and Research (A*STAR), Singapore

[*] *Joint first Authors*

[#]*Corresponding Author*

[†]Address and affiliation of Author when work was done

[¶]Current address and affiliation *of Author*

Correspondence: John E. A. Common

Email: john_common@asrl.a-star.edu.sg

Key words: Atopic dermatitis, *FLG*, ichthyosis vulgaris, nanopore, phasing, sequencing

Abbreviations: AD; atopic dermatitis, CNV; copy number variation, EDC; epidermal differentiation complex, IV; ichthyosis vulgaris, IGV; Integrative Genome Viewer, LoF; loss-of-function, ONT; Oxford Nanopore Technologies, RPT; repeat.

LETTER

Loss-of-function (LoF) variants in the *FLG* gene are causative for ichthyosis vulgaris (IV) and the major genetic risk factors for atopic dermatitis (AD) (Barker et al., 2007; Morar et al., 2007; Weidinger et al., 2007, 2006). Due to its extremely repetitive nature and sequence similarity of the repeats, the *FLG* gene is technically challenging to genetically analyse and determine the contribution of LoF variants to disease status.

Compound heterozygous or homozygous carriers of *FLG* LoF variants have been shown to exhibit more severe phenotypes than heterozygous carriers (Sandilands et al., 2007; Smith et al., 2006). Not all clinical presentation adheres to this paradigm and several studies, including ours, have observed a subset of compound heterozygous carriers that present a clinically mild phenotype (Sekiya et al., 2017) which could be caused by LoF variants being in *cis*. To investigate the potential impact of allelic phasing on phenotypic variability in our Singaporean Chinese IV and AD cohorts, we developed two methodologies (amplicon-sequencing and adaptive sampling sequencing) that leverage long read sequencing from Oxford Nanopore Technologies (ONT) to accurately phase LoF variants and ascertain *FLG* copy number variation (CNV) (Brown et al., 2012). These were applied to investigate phasing of compound heterozygous LoF variants in 21 subjects of which a subset of 14 had either mild or severe IV

and mixed AD severities. All LoF variants were found to be in *trans,* indicating that the severity of IV and AD was independent of the allelic distribution of *FLG* LoF variants. Subsequent investigations identified LoF variants that are consistently linked to specific CNV alleles and suggests allelic features that have co-evolved. Adaptive sampling datasets provided methylation profiles across the entire epidermal differentiation complex (EDC), providing proof-of-concept possibilities for performing wide-scale investigations into epigenetic regulation of EDC genes in disease. Thus, these studies highlight the suitability of nanopore long-read sequencing to study genetic features of large repetitive genes.

We amplified *FLG* exon 3 with a single primer pair from six subjects with known genotypes and CNV alleles for amplicon-sequencing (**Table S1**). Pooled amplicon libraries were sequenced (see supplemental materials for details). This provided sufficient coverage (>500X) for high consensus sequence accuracy for all downstream data processing and bioinformatics analyses (**Figure 1a**). *FLG* CNV allele combinations were determined by mapping amplicon reads to the 10-repeat *FLG* sequence in GRCh38, that identified large insertions (>800 bp) as alternate *FLG* alleles, with the proportion of insertions per allele determining whether each *FLG* allele had 10, 11 or 12 total repeats **(Figure S1a)**. Importantly, we detected all known *FLG* LoF variants and these were allelically phased (**Table S2** and **Figure S1b**). Subsequently, in samples 3A and 6A, we performed target sequence enrichment with live sequencing using software-controlled adaptive sampling technology (see supplemental materials for details) (Martin et al., 2022). DNA from both samples were isolated from blood to ensure consistency in quality and tissue type and also with homozygous CNV length (11-repeat and 12-repeat alleles respectively) to apply adaptive sampling within our bioinformatics pipeline with two different *FLG* CNVs lengths. A region of interest corresponding to the EDC (which contains *FLG)* was preferentially selected in real-time to sequence to completion with 21.8x and 15.9x coverage respectively (**Table S3**). Variant calling

and allele phasing concurred with results from amplicon sequencing from the same samples (**Figure S1c and d**). Methylation profiling across the entire EDC was also captured due to methylated bases generating distinct electrical signals upon passing through the nanopore channel (Simpson et al., 2017) (**Figure S2**). Detailed methylation analysis of the entire *FLG* gene for both samples showed hypermethylation with >85% mean CpG methylation (Table S4). Hypermethylation would be expected from DNA originating from whole blood where filaggrin has negligible expression (Figure S3).

To investigate the contribution of *cis* variants towards IV severity, we selected 16 compound heterozygous subjects with known genotypes and CNV allele combinations, of which five had mild IV, nine had severe IV and two subjects with unknown IV and AD status (**Table 1**). We used the amplicon-sequencing workflow as it was more suitable for sample multiplexing. All variants were found to be in *trans* in both mild and severe IV samples (**Table 1 and Figure 1b and c**) and suggests that the phenotypic variability of IV and AD in this cohort is not affected by the presence of any *cis* variants. Additional factors likely contribute to severity of clinical phenotype.

Phasing of variants to specific alleles allowed assignment to haplotype inheritance and evolutionary events. For example, all 11-repeat alleles in our cohort had an extra *FLG*-repeat that mapped to RPT8.2, suggesting this allele arose from a duplication of RPT8.1 rather than RPT10.1 (**Figure 1d**).

Next, we observed a trend of associations between recurring *FLG* LoF variants (with at least two instances) and specific *FLG* CNVs (**Figure 1e**). LoF variants c.3321delA (n=7), c.6950del8 (n=5), c.3222del4 (n=4), p.E2422X (n=3), p.Q2417X (n=3), p.S1515X (n=2) and p.S2706X (n=2) were found only in the 12-repeat *FLG* allele. Another group of variants, c.9040ins19 (n=2), p.G526X (n=2) and p.R826X (n=2) were located exclusively in the 10-repeat *FLG* allele. A final set of variants, p.K4022X (n=2) and p.R501X (n=2) were associated

with the 11-repeat *FLG* allele. Interestingly, only one variant, p.R2447X (n=2), was found in both the 10-repeat allele and the 11-repeat allele, which either indicates an ancestral origin before CNV duplication events or multiple *de novo* mutational events. Additionally, we observed several single nucleotide polymorphisms (SNPs) exclusively associated with 10-repeat or 12-repeat containing alleles (Table S5 and Figure S3). These SNPs could be used to genotype CNV status in our study. We did not identify any specific SNPs that associated with all the 11-repeat alleles to genotype this CNV variant. Taken together, this data reveals further insights of *FLG* gene mutational events and the possibility of using sequence variants as "genetic markers" in population studies.

In conclusion, we describe two long-read nanopore sequencing strategies that improve comprehensive analysis of the *FLG* gene over short-reads. A key advantage of both novel sequencing strategies reported here is that they enable allelic phasing and more accurate detection of CNVs, thus providing deeper layers of genetic information for clinical investigations into disease pathogenesis. Both methods are suitable for adoption across all populations and ethnicities as they do not rely on multiple primer or probe sets that can result in allelic dropout (Navidi and Arnheim, 1991; Shestak et al., 2021). Additionally, nanopore adaptive sampling sequencing of native DNA provides rich methylation information which is extended beyond *FLG* to all the EDC genes and can be revisited for more in-depth analysis in the future. Adaptive sampling could therefore potentially provide new research directions to study gene regulation as well as fully phased sequence variation for diseases associated to the EDC such as psoriasis (*LCE3B* and *LCE3C*) (De Cid et al., 2009) and AD (*FLG2* and *S100A9*) (Berna et al., 2022; Budu-Aggrey et al., 2022; Margolis et al., 2014). Thus, our study provides a framework for future clinical investigations of the EDC region and other complex genomic regions for skin research.

DATA AVAILABILITY

Nanopore raw sequence data of *FLG* amplicon and adaptive sampling runs are deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS50000000166. Further information about the European Genome phenome Archive can be found at https://egaarchive.org and "The European Genome phenome Archive of human data consented for biomedical research."All scripts used in this manuscript are available on our GitHub repository (https://github.com/sgnzapps/flg_variant_phasing).

ORCiDs

Colin Wong: 0009-0002-3510-3328

Cheng Yong Tham: 0000-0002-7594-3025

Lin Yang: 0000-0002-6716-4013

Miles Benton: 0000-0003-3442-965X

Vipin Narang: 0000-0001-8669-9964

Kaibo Duan: 0000-0003-0523-8993

Simon Denil: 0000-0001-7242-9082

Yik Weng Yew: 0000-0001-8915-306X

Bernett Lim: 0000-0002-2709-1972

Paola Florez de Sessions: 0000-0003-0258-9616

John E.A Common: 0000-0002-3280-7365

CONFLICT OF INTEREST

P.F.d.S., C.T., L.Y., and M.C.B., are full-time employees of, and share/phantom shareholders in Oxford Nanopore Technologies. These authors assisted with technical aspects

of nanopore sequencing analysis but were not involved in study design. The other authors do not have any conflicts of interest.

ACKNOWLEDGEMENTS

We thank the patients for contributing to this study and the clinical coordinators at National Skin Centre for their diligence in recruitment. We also thank Daniel Turner for his proof-reading and helpful discussions.

FUNDING SOURCES

This work was supported by funding from Agency for Science, Technology and Research (A*STAR) and A*STAR BMRC EDB IAF-PP grants – H17/01/a0/004 "Skin Research Institute of Singapore" and BMRC Central Research Funds (ATR) (C.W, S.D, J.E.A.C). V.N, B.L and K.D are part of the SIgN Immunomonitoring platform (supported by BMRC IAF 311006 grant, BMRC transition funds H16/99/b0/011, BMRC IAF-PP H19/01/a0/024 SIGNAL grant and NRF SIS NRF2017_SISFP09 grant). Oxford Nanopore Technologies provided certain flow cells and services in-kind in support of this research.

AUTHOR CONTRIBUTIONS STATEMENT

Study concept and design: C.W and J.E.A.C. Acquisition of data: C.W, C.T and L.Y. Analysis and interpretation of data: C.W, C.T, L.Y, M.C.B, V.N, D.K and S.D. Writing of the manuscript and preparation of figures: C.W, C.T and L.Y. Critical revision of the manuscript for important intellectual content: J.E.A.C, P.F.d.S, B.L, Y.W.Y. Subject recruitment: Y.W.Y. All authors reviewed the manuscript.

REFERENCES

Barker JNWN, Palmer CNA, Zhao Y, Liao H, Hull PR, Lee SP, et al. Null mutations in the filaggrin gene (FLG) determine major susceptibility to early-onset atopic dermatitis that persists into adulthood. Journal of Investigative Dermatology 2007;127. https://doi.org/10.1038/sj.jid.5700587.

Berna R, Mitra N, Hoffstad O, Wubbenhorst B, Nathanson KL, Margolis DJ. Uncommon variants in FLG2 and TCHHL1 are associated with remission of atopic dermatitis in a large longitudinal US cohort. Arch Dermatol Res 2022;314. https://doi.org/10.1007/s00403-021-02319-7.

Brown SJ, Kroboth K, Sandilands A, Campbell LE, Pohler E, Kezic S, et al. Intragenic copy number variation within filaggrin contributes to the risk of atopic dermatitis with a dose-dependent effect. Journal of Investigative Dermatology 2012;132. https://doi.org/10.1038/jid.2011.342.

Budu-Aggrey A, Kilanowski A, Sobczyk MK, Shringarpure SS, Mitchell R, Reis K, et al. European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation. MedRxiv 2022.

De Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, Dannhauser EN, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 2009;41. https://doi.org/10.1038/ng.313.

Margolis DJ, Gupta J, Apter AJ, Ganguly T, Hoffstad O, Papadopoulos M, et al. Filaggrin-2 variation is associated with more persistent atopic dermatitis in African American subjects. Journal of Allergy and Clinical Immunology 2014;133. https://doi.org/10.1016/j.jaci.2013.09.015.

Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome Biol 2022;23. https://doi.org/10.1186/s13059-021-02582-x.

Morar N, Cookson WOCM, Harper JI, Moffatt MF. Filaggrin mutations in children with severe atopic dermatitis. Journal of Investigative Dermatology 2007;127. https://doi.org/10.1038/sj.jid.5700739.

Navidi W, Arnheim N. Using PCR in preimplantation genetic disease diagnosis. Human Reproduction 1991;6. https://doi.org/10.1093/oxfordjournals.humrep.a137438.

Sandilands A, Terron-Kwiatkowski A, Hull PR, O'Regan GM, Clayton TH, Watson RM, et al. Comprehensive analysis of the gene encoding filaggrin uncovers prevalent and rare mutations in ichthyosis vulgaris and atopic eczema. Nat Genet 2007;39:650–4. https://doi.org/10.1038/ng2020.

Sekiya A, Kono M, Tsujiuchi H, Kobayashi T, Nomura T, Kitakawa M, et al. Compound heterozygotes for filaggrin gene mutations do not always show severe atopic dermatitis. Journal of the European Academy of Dermatology and Venereology 2017;31:158–62. https://doi.org/10.1111/jdv.13871.

Shestak AG, Bukaeva AA, Saber S, Zaklyazminskaya E V. Allelic Dropout Is a Common Phenomenon That Reduces the Diagnostic Yield of PCR-Based Sequencing of Targeted Gene Panels. Front Genet 2021;12. https://doi.org/10.3389/fgene.2021.620337.

211 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine
212 methylation using nanopore sequencing. Nat Methods 2017;14.
213 https://doi.org/10.1038/nmeth.4184.

214 Smith FJD, Irvine AD, Terron-Kwiatkowski A, Sandilands A, Campbell LE, Zhao Y, et al. Loss-of-function
215 mutations in the gene encoding filaggrin cause ichthyosis vulgaris. Nat Genet 2006;38:337–42.
216 https://doi.org/10.1038/ng1743.

217 Weidinger S, Illig T, Baurecht H, Irvine AD, Rodriguez E, Diaz-Lacava A, et al. Loss-of-function
218 variations within the filaggrin gene predispose for atopic dermatitis with allergic sensitizations.
219 Journal of Allergy and Clinical Immunology 2006;118. https://doi.org/10.1016/j.jaci.2006.05.004.

220 Weidinger S, Rodríguez E, Stahl C, Wagenpfeil S, Klopp N, Illig T, et al. Filaggrin mutations strongly
221 predispose to early-onset and extrinsic atopic dermatitis [2]. Journal of Investigative Dermatology
222 2007;127. https://doi.org/10.1038/sj.jid.5700630.

223

224

225

226

227

228

229

			***FLG* CNV status**	**Called LoF variants**	**Nanopore**			
Sample ID	**IV severity**	**AD severity**	**(Gel/ nanopore)**	**(Illumina/ nanopore)**	**Haplotype**	**CNV status**	**Repeat location of LoF variant**	**Allelic distribution**
1B	Mild	Moderate	12_12	p.E2422X	HP1	12	7	*Trans*
				p.S1515X	HP2	12	4	
2B	Mild	Moderate	12_12	c.3321delA	HP1	12	2	*Trans*
				p.E2422X	HP2	12	7	
3B	Mild	Severe	12_12	p.Q2417X	HP1	12	6	*Trans*

				c.3321delA	HP2	12	2	
4B	Mild	Severe	12_12	p.S2706X p.Q2417X	HP1 HP2	12 12	7 6	*Trans*
5B	NR	NR	10_12	c.6950_6957del8 p.R2447X	HP1 HP2	12 10	6 7	*Trans*
6B	Mild	Moderate	10_12	c.7945delA c.6950_6957del8	HP1 HP2	10 12	7 6	*Trans*
7B	NR	NR	12_12	c.6950_6957del8 c.678delA	HP1 HP2	12 12	6 0	*Trans*
8B	Severe	Severe	12_12	c.3321delA p.S2706X	HP1 HP2	12 12	2 7	*Trans*
9B	Severe	Severe	10_12	p.G526X c.3321delA	HP1 HP2	10 12	1 2	*Trans*
10B	Severe	Moderate	10_12	p.R826X c.6950_6957del8	HP1 HP2	10 12	2 6	*Trans*
11B†	Severe	Moderate	11_12	c.3222del4 p.K4022X	HP1 HP2	12 11	2 11	*Trans*
12B	Severe	NR	12_12	c.3222del4 c.3321delA	HP1 HP2	12 12	2 2	*Trans*
13B	Severe	No AD	11_12	p.R501X p.E2422X	HP1 HP2	11 12	1 7	*Trans*
14B	Severe	Severe	10_11	c.9040_9058dup19 p.Q1790X	HP1 HP2	10 11	8.1 5	*Trans*
15B	Severe	Moderate	12_12	p.Q368X c.3321delA	HP1 HP2	12 12	0 2	*Trans*

16B	Severe	Severe	11_12	c.8393delA p.S1515X	HP1 HP2	11 12	8.1 4	*Trans*

Table 1: Nanopore sequencing enables investigation of *FLG* genotype-phenotype associations in mild and severe IV samples. Nanopore long reads detected and phased *FLG* CNV and LoF variants accurately in all 16 samples, which are in concordance to gel electrophoresis and Illumina short-read sequencing. LoF variants in all samples were detected as t*rans* compound heterozygous. LoF, loss-of-function; CNV, copy number variation; HP, haplotype. † Sample 11B is the same subject as sample 5A from Tables S1 and S2 and serves as a bridge between the two batches of sequencing. NR = Not Recorded. AD severity determined by SCORAD values.

FIGURE LEGEND

Figure 1: Nanopore sequencing facilitates phasing of LoF variants and association with CNV alleles in the *FLG* gene: a) Bioinformatics analysis pipeline used in both amplicon sequencing and adaptive sampling protocols. **b**) IGV plot of read alignments for sample 1B (mild IV) showing in *trans* phasing of LoF variants: two single-nucleotide substitutions (p.S1515X; Chr1: 152310342 and p.E2422X; Chr1: 152307622). **c**) IGV plot of read alignments for sample 14B (severe IV) showing in *trans* phasing of LoF variants: a 19 bp duplication (c.9040dup19; Chr1: 152305825) and a single-nucleotide substitution (p.Q1790X; Chr1:152309518). IGV colour schemes are set to default with bases that match the reference

displayed in gray, purple indicating sequence insertions, black horizontal lines representing gaps within read alignments that correspond to sequence deletions and coloured line markings representing single base substitutions (Red=T; Green=A; Blue=C; Orange=G). HP, haplotype. Reads are grouped according to haplotype and separate by a gray, dotted horizontal line. **d)** The 11-repeat *FLG* allele contains RPT8.2 and not RPT10.2. The vast majority of 11-repeat containing *FLG* sequences map to a reference sequence containing RPT8.2 rather than containing RPT10.2. **e)** Examples of LoF variants in this study associating with specific *FLG* CNVs.

Figure 1A

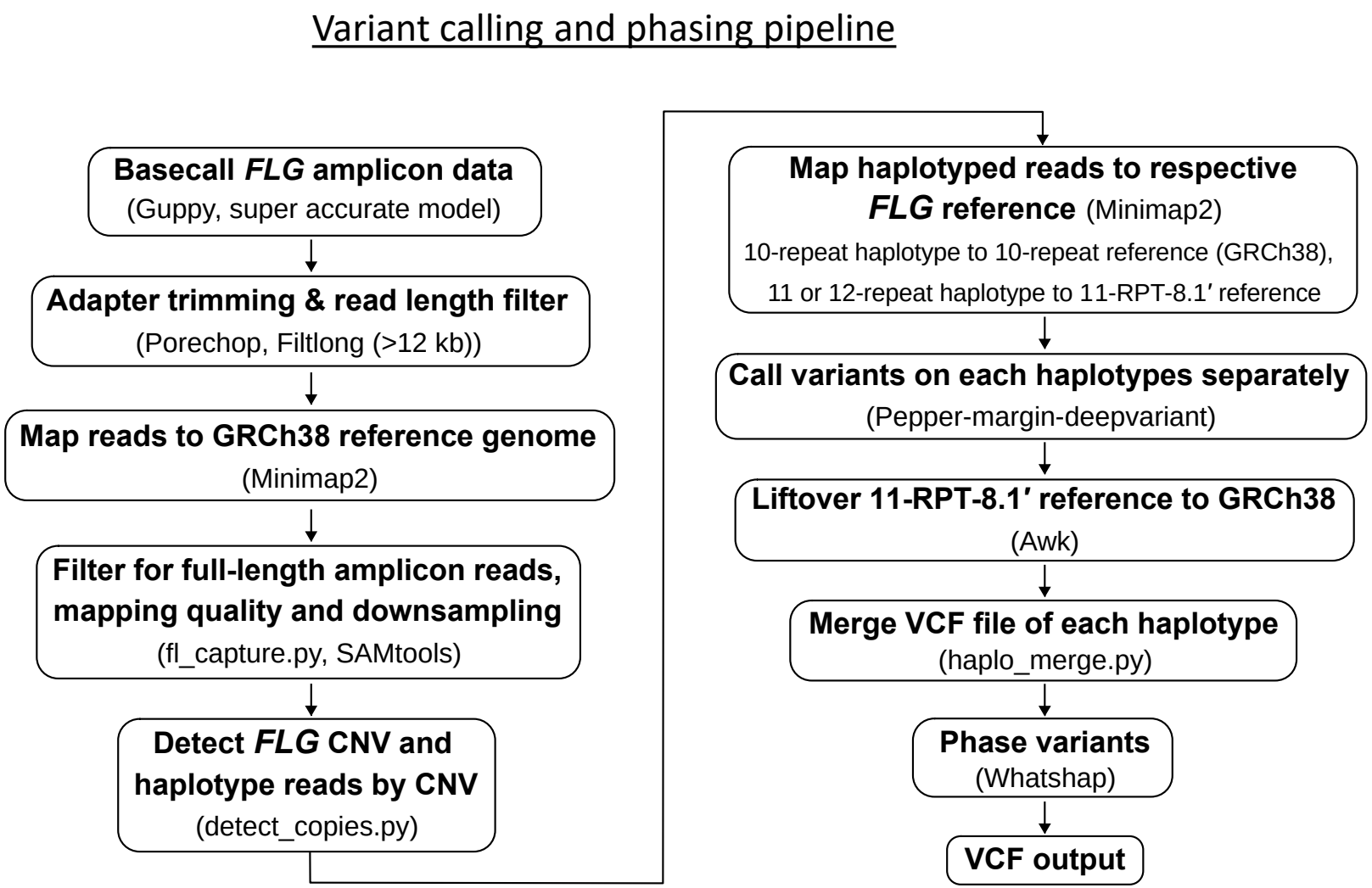

Variant calling and phasing pipeline
Basecall *FLG* amplicon data
(Guppy, super accurate model)
Adapter trimming & read length filter
(Porechop, Filtlong (>12 kb))
Map reads to GRCh38 reference genome
(Minimap2)
Filter for full-length amplicon reads, mapping quality and downsampling
(fl_capture.py, SAMtools)
Detect *FLG* CNV and haplotype reads by CNV
(detect_copies.py)
Map haplotyped reads to respective *FLG* reference (Minimap2)
10-repeat haplotype to 10-repeat reference (GRCh38),
11 or 12-repeat haplotype to 11-RPT-8.1′ reference
Call variants on each haplotypes separately
(Pepper-margin-deepvariant)
Liftover 11-RPT-8.1′ reference to GRCh38
(Awk)
Merge VCF file of each haplotype
(haplo_merge.py)
Phase variants
(Whatshap)
VCF output

Figure 1B

Sample 1B: mild IV

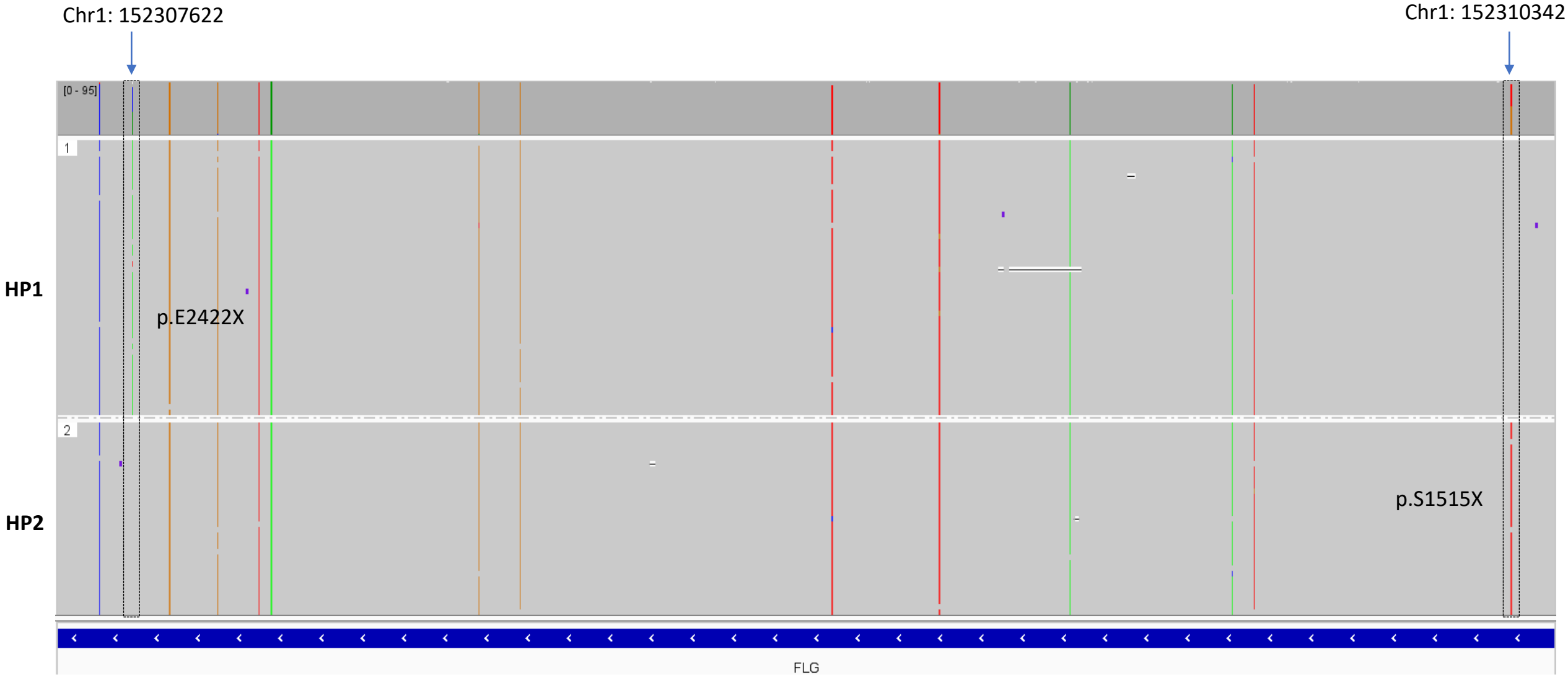


Figure 1C

Sample 14B: severe IV

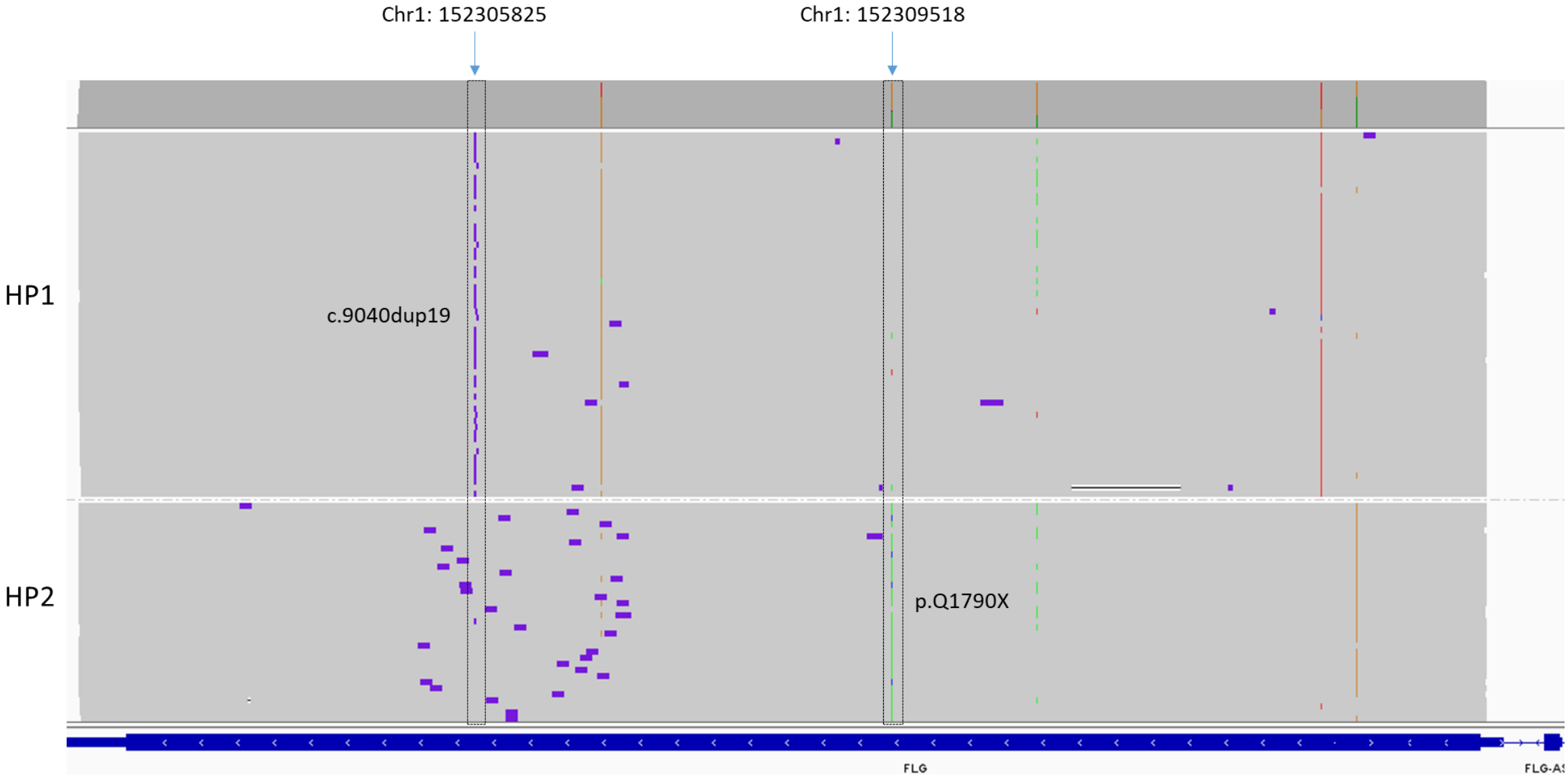


Figure 1D

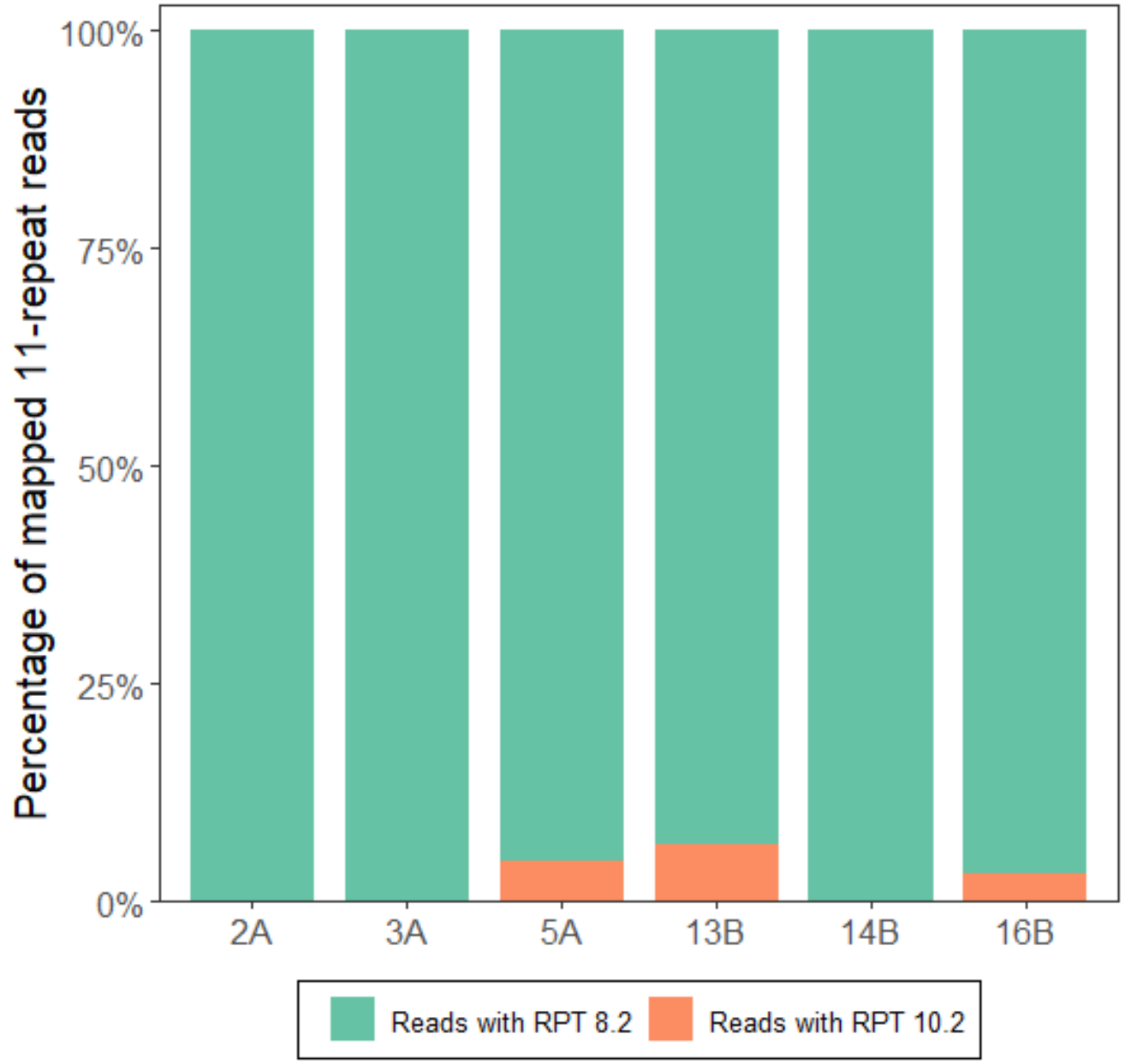
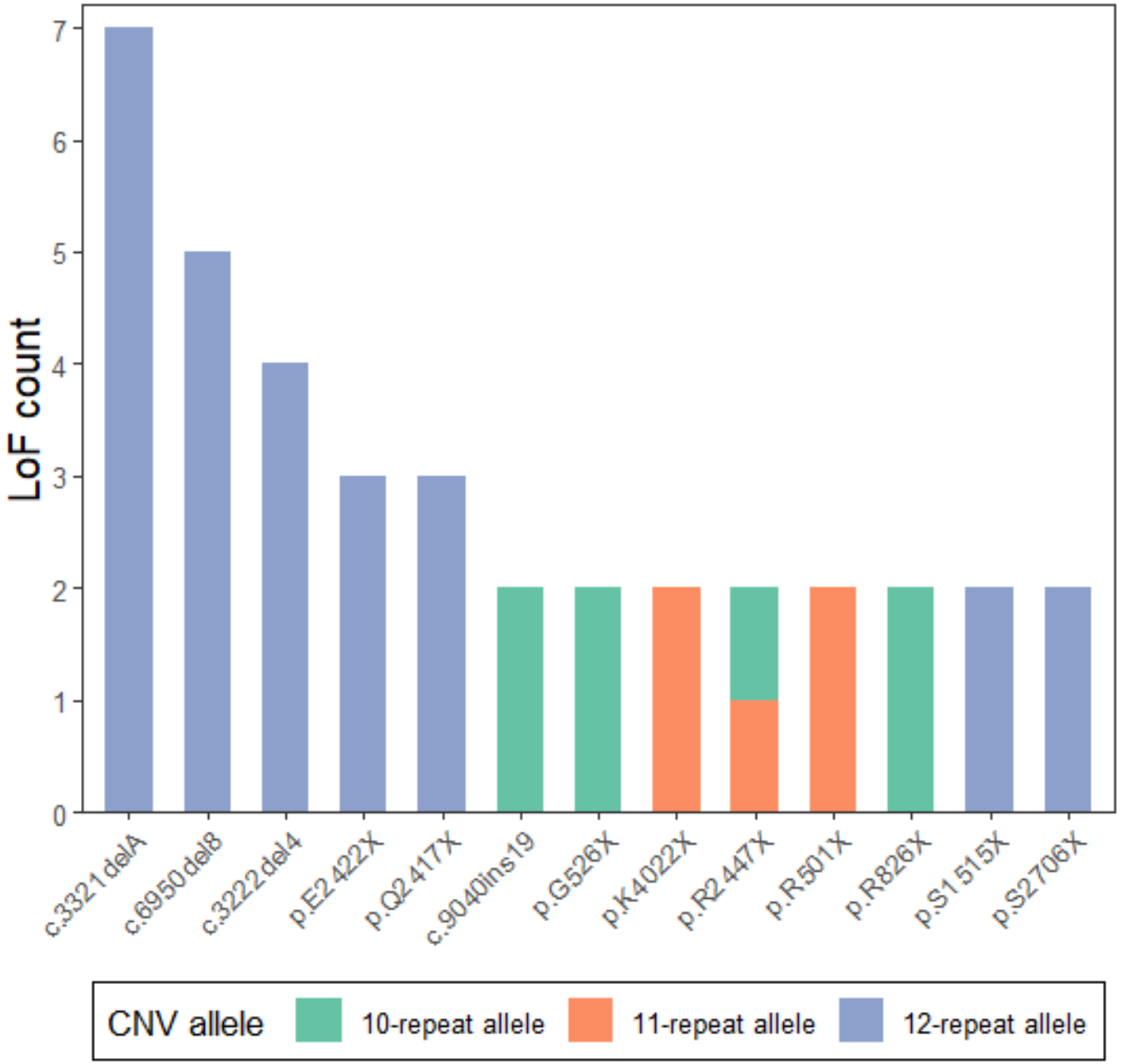

Percentage of mapped 11-repeat reads
100%
75%
50%
25%
0%
2A
3A
5A
13B
14B
16B
Reads with RPT 8.2
Reads with RPT 10.2

Figure 1E


LoF count
0
1
2
3
4
5
6
7
c.3321delA
c.6950del8
c.3222del4
p.E2422X
p.Q2417X
c.9040ins19
p.G526X
p.K4022X
p.R2447X
p.R501X
p.R826X
p.S1515X
p.S2706X
CNV allele
10-repeat allele
11-repeat allele
12-repeat allele

SUPPLEMENTARY MATERIALS

Materials and methods

Recruitment of study subjects

Six unrelated subjects were selected from a pre-existing cohort at the National Skin Centre (NSC) with known *FLG* genotypes and CNV, for initial analysis with nanopore sequencing. DNA was obtained either from processed whole blood (GE Healthcare, USA) or saliva (DNA Genotek Inc., Kanata, ON, Canada) according to the manufacturer's instructions.

16 subjects of Singaporean Chinese ancestry with known compound heterozygous LoF variants and CNV, were selected from our previous studies to investigate the contribution of in *cis* LoF variants on IV severity. Five had mild IV, nine had severe IV and the remaining two had unknown IV and AD status. IV severity was graded with a global assessment scale (UK Working Party's Diagnostic Criteria for Atopic Dermatitis) that takes into account the effect of scaling and number of body sites involved (Williams et al., 1994). The grades were: mild IV, mild scaling and dryness on one body site (e.g. legs); moderate IV, moderate scaling and dryness and/or two affected body sites; severe IV, severe scaling and dryness with a general bodily distribution. This brought the total number of subjects to 21. All subjects provided written informed consent approved by the Domain-specific Institutional Review Board (NHG DSRB references: 2006/00221 and 2013/01212).

Long-range amplification of FLG exon 3 and nanopore sequencing

A primer pair consisting of the following: forward primer FilF3 (5' - GCT GAT AAT GTG ATT CTG TCT G - 3') that was previously published (Smith et al., 2006) and reverse primer PCR2R1 (5' – AAG ATG TGC TAG CCC TGA TGT TGA – 3') was used to amplify the entire exon 3/CDS of *FLG* in a 50 µL reaction with the LA Taq Hotstart kit (Takara) reaction

mixture which consisted of the following: 5 µL of buffer, 8 µL of dNTPs, 2.35 µL forward primer (3.2 µM), 2.35 µL reverse primer (3.2µM), 0.5 µL LA Taq hotstart polymerase, 200 ng DNA and an appropriate volume of nuclease-free water (Promega) to make up the final reaction volume of 50 µL and cycled under the following conditions: 1 minute initial denaturation at 94°C, 28 cycles of denaturation at 94°C for 30 seconds and annealing and extension at 68°C for 15 minutes followed by a final extension at 72°C for 10 minutes. 5 µL of each product was run on a 0.7% agarose gel to visualise success of the PCR and also to assess the size of the *FLG* allelic variants. Where possible, multiple PCR reactions for each sample were run (5-8 reactions) to ensure adequate PCR product starter material for downstream nanopore sequencing.

Clean-up of PCR products was performed using HighPrep™ beads (Magbio) according to the manufacturer's instructions. PCR reactions from the same sample were pooled together into a 1.5 mL tube and mixed with DNA-binding magnetic beads at a volume 0.8X that of the sample volume. Samples were eluted in an appropriate volume of nuclease-free water (Promega) and quantified using the Qubit dsDNA broad-range assay (ThermoFisher).

A minimum of 500 ng (although 1 µg is recommended) of PCR products from each sample was taken as input into the Ligation Sequencing Kit (SQK-LSK109, ONT) and library preparation protocol (EXP-NBD104/114, ONT), where the libraries were barcoded and pooled. The pooled library was then sequenced on a R9.4.1 MinION flow cell on the ONT GridION sequencer.

Amplicon read mapping and filtering

Sequencing adapters were trimmed from reads using Porechop (v0.2.4) (Wick et al., 2017), and read filtering for reads with a minimum length of 12 kb was performed using Filtlong (v0.2.1) (https://github.com/rrwick/Filtlong). Reads were mapped to the GRCh38 reference

genome (Schneider et al., 2017) using Minimap2 (v2.24-r1122) (Li, 2018) with the preset (-x map-ont). SAMtools (v1.14) (Danecek et al., 2021) was used for SAM to BAM conversion, sorting, filtering by MAPQ (-q 60) and removal of secondary alignments, supplementary alignments, and unmapped reads (-F 2308). The "fl_capture.py" custom script was used to capture full-length *FLG* amplicon reads that spans at least 98% lengthwise of the expected PCR product and have soft clippings of no more than 300 bp, while also performing read strand balancing. The script also carried out down sampling of reads to a depth of 100X for downstream variant calling analysis.

Generation of 11-RPT-8.1 ' FLG reference sequence and mapping

The insertion position of RPT8.2 was observed to vary greatly amongst reads when aligned to GRCh38. This mapping inaccuracy is a result of the repetitive structure of *FLG*, which creates ambiguity in RPT insertion placement by the sequence aligner tool. If left uncorrected, these misalignments can cause unreliable downstream variant calling. To address this, we constructed a synthetic 11-repeat *FLG* reference sequence, named 11-RPT-8.1′. This was performed by tandemly duplicating RPT8.1 within the *FLG* exon 3 sequence (chr1:152300001-152317000) in GRCh38. RPT coordinates were identified according to their respective published sequences (Smith et al., 2006). Through Minimap2, this 11-RPT-8.1′ *FLG* sequence was used to map haplotyped 11- or 12-repeat reads for improved alignment, downstream variant calling, and phasing accuracy. For haplotyped 10-repeat reads, the default GRCh38 10-repeat *FLG* reference was used.

FLG CNV detection and haplotyping

Basecalling of nanopore data was performed using the super-accurate model through Guppy (v6.0.7) (Wick et al., 2019) (https://community.nanoporetech.com). Reads were mapped to the GRCh38 reference genome using Minimap2 (v2.24-r1122).

FLG monomer repeat (RPT) copy number detection was performed using our custom script "detect_copies.py". The script uses the alignment information of reads that mapped to the 10-repeat *FLG* reference in GRCh38. Using the CIGAR information in BAM files, the genomic region, chr1:152303959-152307308, was queried for insertions larger than 800 bp within each read. We took each detected large insertion as an additional single RPT copy. Hence, to calculate the number of RPTs per read, we added the number of additional RPT copies to the default 10 RPT copies in GRCh38. For instance, a read with one additional RPT copy indicates a total of 11 RPT copies, while a read with two additional RPT copies indicates a total of 12 RPT copies. Next, we calculated the proportion of reads having either the 10, 11, or 12 total RPT alleles. The RPT alleles with a minimum allele frequency of 0.2 were designated the major RPT alleles. Reads were then haplotyped according to the major RPT alleles and subsequently analysed as separate haplotypes. In samples with only one major RPT allele, no read haplotyping was performed and all reads belonging to the major RPT allele were analysed together in the next stage.

***FLG* variant calling and phasing**

Variant calling on *FLG* exon 3 was carried out by PEPPER-Margin-DeepVariant (r0.8) (Shafin et al., 2021) on each RPT haplotype separately against their respective *FLG* reference using default parameters. Variants labeled as "refCall" or sub-haplotype variants (Presenting '0/1' genotype in a haplotype) were omitted from the VCF files. Next, the coordinates of variants called using the 11-RPT-8.1′ reference were lift-over to those of GRCh38 using mawk (v1.3.4).

This was done by adding 152300000 to the coordinates of variants occurring before RPT8.1′ and adding 152299028 to the coordinates of variants after RPT8.1′. Subsequently, the VCF files arising from each haplotype were merged into one using our custom script "haplo_merge.py", resulting in one VCF file per sample. The script resolves the genotype (GT) information of variants during the merging, where variants present only in one haplotype would have their genotype converted and phased from "1/1" to "0|1" or "1|0", and variants present in both haplotypes would be represented as one entry with their variant quality (QUAL) and allele frequency (AF) averaged. For both cases, read depth (DP) would be summed. Finally, Whatshap (v1.6)(Martin et al., 2016) was used to revise the variant phasing for the final VCF output, and haplotag the final BAM file for display on IGV (Thorvaldsdóttir et al., 2013) with GRCh38 as the reference and the "—ignore-read-groups" and "--indels" options enabled.

Adaptive sampling for EDC analysis and generation of methylation profiles

3 µg of genomic DNA was used for library preparation (SQK-LSK110, ONT). 28 fmol of library per sample was loaded onto the R9.4.1 flow cell and sequenced on the GridION running MinKNOW 22.03.4. A BED file specifying a 2.9 Mb region spanning the entire EDC was supplied, along with the human reference genome GRCh38. A flow cell wash and sample reload was performed after 18 hours, followed by a further 18 hours of sequencing.

Basecalling with modified 5mC for methylation profiling was performed using the super-accurate model through Guppy (v6.4.2) Read alignment to the GRCh38 reference assembly was also performed during basecalling. SAMtools (v1.14) was used to concatenate, sort and index all BAM files from each run, and subsequently filter reads by MAPQ (-q 60) and remove secondary alignments, supplementary alignments, and unmapped reads (-F 2308). Furthermore, SAMtools was used to extract reads mapping to the *FLG* region which were used for downstream CNV and variant calling. RPT copy number was detected in a similar way to

the amplicon using a modified “detect_copies_AS.py” script which added a read alignment filter for the region chr1:152303642-152307641 since reads do not all start and end at the same position. As both samples have the 11/12-repeat *FLG* structure, their reads were mapped against our 11-RPT-8.1′ FLG reference sequence through Minimap2 (v2.24-r1122). PEPPER-Margin-DeepVariant (r0.8) was used to call variants on the *FLG* gene against the 11-RPT-8.1′ *FLG* reference, and variant coordinate lift-over to GRCh38 was done according to the instructions for amplicon analysis. Lastly, Whatshap (v1.6) was used for variant phasing and haplotagging the final BAM file, with GRCh38 as the reference and the ”—ignore-read-groups” and “--indels” options enabled. Visualization of variants and CpG methylation information was carried out on IGV.

Supplementary tables and figures

Sample ID	LoF variant 1	LoF variant 2	*FLG* CNV status by gel	Total reads	Yield (Mb)	N50 length (Kb)	Mean qscore
1A	p.G526X	NIL	10_10	139,976	883	12.5	14.1
2A	c.9040dup19	p.G323X	10_11	89,159	572	11.2	13.9
3A	p.R2447X	p.R501X	11_11	125,709	769	10.5	13.9
4A	c.3321delA	p.R826X	10_12	112,829	786	12.0	14.1
5A	p.K4022X	c.3222del4	11_12	288,480	2,017	11.4	14.2
6A	c.6950del8	p.Q2417X	12_12	126,864	770	9.9	13.8

Table S1: Details of pilot samples and nanopore sequencing output. Samples with known LoF variants and *FLG* CNV status were sequenced with long reads. LoF, loss-of-function; CNV, copy number variation; Mb, megabase; Kb, kilobase.

Sample ID	*FLG* CNV status (Gel/nanopore)	Called LoF variants (Illumina/nanopore)	Nanopore		
			Haplotype	CNV status	Allelic distribution
1A	10_10	p.G526X	HP2	10	N/A
2A	10_11	p.G323X	HP1	11	*Trans*
		c.9040dup19	HP2	10	
3A	11_11	p.R2447X	HP1	11	*Trans*
		p.R501X	HP2	11	
4A	10_12	c.3321delA	HP1	12	*Trans*
		p.R826X	HP2	10	
5A	11_12	c.3222del4	HP1	12	*Trans*
		p.K4022X	HP2	11	
6A	12_12	c.6950del8	HP1	12	*Trans*
		p.Q2417X	HP2	12	

Table S2: *FLG* variant calling and phasing with method development cohort samples. Nanopore long reads accurately detected *FLG* CNV and LoF variants, which are in concordance to gel electrophoresis and Illumina short-read sequencing. Furthermore, nanopore long reads enable phasing of the variants. LoF variants in samples 2A-6A were all detected as *trans* compound heterozygous. LoF, loss-of-function; CNV, copy number variation; HP, haplotype. N/A, not applicable, due to the presence of only one LoF variant.

		Adaptive sampling accepted reads						
Sample ID	Total passed reads	Total passed	N50 length (kb)	Mean qscore	Primary alignments	Alignments on target	% alignments on target	Average coverage of target region
3A	6,754,854	6,094k	22.9	14.87	6,073k	5,937k	97.8%	21.8X
6A	5,946,857	4,804k	21.6	14.84	4,786k	4,693k	98.1%	15.9X

149 **Table S3: Nanopore adaptive sampling sequencing for the EDC region at Chr1q21.**
150 Sample DNA was used to construct sequencing libraries with SQK-LSK110, and sequencing
151 was carried out on the GridION with a BED file specifying the 2.9 Mb EDC region at Chr1q21.
152 Read coverage of 21.8X and 15.9X was obtained for the EDC region for sample 3A and 6A
153 respectively.

Sample 3A (11_11)

Region	Promoter (1kb from TSS)		Exon1		Intron1		Exon2		Intron2		Exon3	
Strand	+	-	+	-	+	-	+	-	+	-	+	-
N_{CpGs}	5	5	0	0	26	26	1	1	0	0	274	284
$\%_{methylation}$	81.8%	73.7%	na	na	85.8%	74.7%	83.3%	100.0%	na	na	91.0%	92.0%
$N_{valid\ cov}$	33	57	na	na	190	233	6	4	na	na	1813	2842
N_{5mC}	26	41	na	na	161	171	5	4	na	na	1519	2403
N_{5hmC}	1	1	na	na	2	3	0	0	na	na	130	211
$N_{canonical}$	6	15	na	na	27	59	1	0	na	na	164	228

Region	S-100 domain		RPT0		RPT1		RPT2		RPT3		RPT4		RPT5		RPT6	
Strand	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
N_{CpGs}	2	1	15	16	22	23	25	25	22	22	25	25	26	27	17	18
$\%_{methylation}$	90.9%	100.0%	94.1%	94.6%	90.1%	94.0%	94.9%	94.2%	85.2%	94.2%	83.2%	95.0%	93.2%	89.2%	89.9%	92.6%
$N_{valid\ cov}$	11	4	101	130	141	201	175	223	162	208	191	241	206	286	109	163
N_{5mC}	10	4	88	110	118	171	153	188	129	177	141	204	183	232	89	133
N_{5hmC}	0	0	7	13	9	18	13	22	9	19	18	25	9	23	9	18
$N_{canonical}$	1	0	6	7	14	12	9	13	24	12	32	12	14	31	11	12

Region	RPT7		RPT8.1		RPT8.2		RPT9		RPT10.1		RPT10.2		RPT11		3'UTR	
Strand	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
N_{CpGs}	20	22	21	21	20	21	23	28	24	26	na	na	28	29	2	2
$\%_{methylation}$	88.6%	85.4%	89.5%	91.0%	91.7%	93.3%	94.3%	90.3%	94.7%	93.4%	na	na	94.4%	88.7%	76.9%	84.0%
$N_{valid\ cov}$	114	185	114	199	121	225	123	299	132	286	na	na	178	337	13	25
N_{5mC}	90	145	89	168	98	198	109	261	118	253	na	na	159	281	9	21
N_{5hmC}	11	13	13	13	13	12	7	9	7	14	na	na	9	18	1	0
$N_{canonical}$	13	27	12	18	10	15	7	29	7	19	na	na	10	38	3	4

154

Sample 6A (12_12)

Region	Promoter (1kb from TSS)		Exon1		Intron1		Exon2		Intron2		Exon3	
Strand	+	-	+	-	+	-	+	-	+	-	+	-
N_{CpGs}	3	3	0	0	26	24	1	1	0	0	251	247
$\%_{methylation}$	80.0%	85.0%	na	na	80.4%	78.7%	100.0%	100.0%	na	na	90.9%	91.3%
$N_{valid\ cov}$	15	20	na	na	168	150	8	5	na	na	1465	1520
N_{5mC}	12	17	na	na	128	112	8	3	na	na	1247	1250
N_{5hmC}	0	0	na	na	7	6	0	2	na	na	85	137
$N_{canonical}$	3	3	na	na	33	32	0	0	na	na	133	133

Region	S-100 domain		RPT0		RPT1		RPT2		RPT3		RPT4		RPT5		RPT6	
Strand	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
N_{CpGs}	2	2	14	15	22	22	25	23	21	13	20	21	24	18	14	16
$\%_{methylation}$	93.8%	100.0%	89.2%	88.7%	88.6%	86.3%	92.1%	91.8%	87.3%	91.4%	89.9%	94.1%	94.9%	96.6%	88.5%	93.1%
$N_{valid\ cov}$	16	9	83	97	132	146	165	146	142	70	129	101	157	88	78	87
N_{5mC}	15	7	71	77	111	117	143	125	115	61	108	86	146	80	69	69
N_{5hmC}	0	2	3	9	6	9	9	9	9	3	8	9	3	5	0	12
$N_{canonical}$	1	0	9	11	15	20	13	12	18	6	13	6	8	3	9	6

Region	RPT7		RPT8.1		RPT8.2		RPT9		RPT10.1		RPT10.2		RPT11		3'UTR	
Strand	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
N_{CpGs}	20	21	19	17	17	21	23	26	18	20	11	15	26	27	1	2
$\%_{methylation}$	92.0%	86.1%	95.2%	92.5%	100.0%	91.6%	91.5%	93.5%	90.5%	94.3%	90.4%	89.8%	86.4%	88.8%	87.5%	95.7%
$N_{valid\ cov}$	113	122	104	107	80	131	117	153	95	123	52	108	132	197	8	23
N_{5mC}	98	98	87	92	70	105	97	116	79	103	45	92	109	157	7	22
N_{5hmC}	6	7	12	7	10	15	10	27	7	13	2	5	5	18	0	0
$N_{canonical}$	9	17	5	8	0	11	10	10	9	7	5	11	18	22	1	1

Table S4: CpG methylation profile across *FLG* subregions of sample 3A and 6A as detected by nanopore sequencing. N_{CpGs} refers to the number of called CpGs per strand of a subregion that passed filterings. $N_{valid\ cov}$ refers to the collective coverage of reads at each CpG in a subregion that contains passed methylation information (equivalent to the sum of N_{5mC}, $N_{5hmC,}$ and $N_{canonical}$). N_{5mC}, $N_{5hmC,}$ and $N_{canonical}$ refer to the collective number of 5-methylcytosine, 5-hydroxymethylcytosine, and canonical cytosine calls, respectively, from reads covering CpGs in a subregion. $\%_{methylation}$ refers to the collective percentage of methylation (5mC and 5hmC) across CpGs per strand in a subregion.

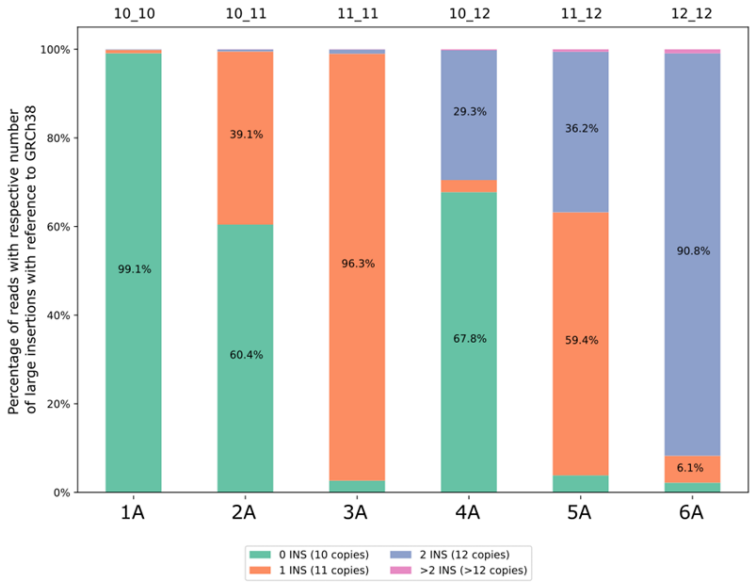
	1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	7B	8B	9B	10B	11B	12B	13B	14B	15B	16B
CNV status	10/10	10/11	11/11	10/12	11/12	12/12	12/12	12/12	10/12	12/12	12/12	10/12	12/12	10/12	11/12	12/12	10/12	12/12	11/12	10/11	12/12	11/12
rs192116923 (T > G)	1/1	0/1	0/0	0/1	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/1	0/0	0/1	0/0	0/0	0/1	0/0	0/0	0/1	0/0	0/0
rs11588170 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs2011331 (T>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs11584340 (G>A)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs74129461 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3120653 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs66831674 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs58001094 (G>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs11586631 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs11581433 (T>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs11204978 (G>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs12407807 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs12405278 (G>A)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs12405241 (G>A)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs12407748 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1

rs3126079 (G>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs2184954 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs2184953 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs6664985 (G>A)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs139476473 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs71625202 (C>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs71625201 (C>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs71625200 (T>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs55650366 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3126074 (G>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3126072 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs57672167 (C>A)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs2065958 (A>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs2065957 (A>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3126069 (T>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs6681433 (T>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1

rs2065956 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3091276 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs2065955 (C>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3126067 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3126066 (A>G)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs72697000 (C>A)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs3126075 (G>C)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1
rs77422831 (C>T)	0/0	0/0	0/0	0/1	0/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/0	1/1	0/1

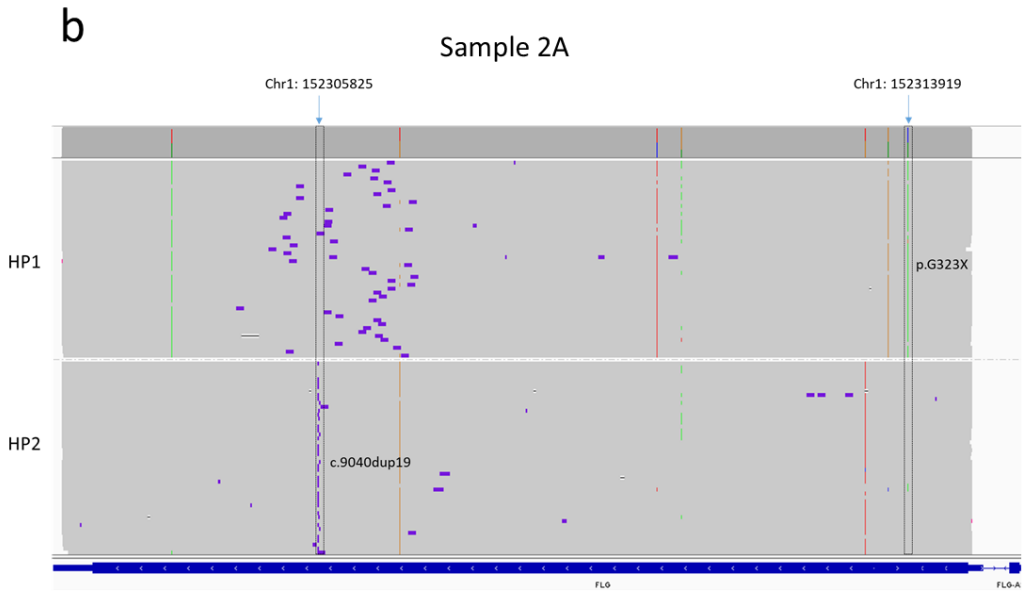
Table S5: Inheritance patterns of SNPs specific to *FLG* CNV alleles inform CNV status. Observation of homozygous or heterozygous SNP in the 10-repeat *FLG* allele (highlighted in Blue) and 37 SNPs located in the 12-repeat *FLG* allele (highlighted in Red) are informative to genotype 10- and 12-CNV alleles in all samples in this cohort (sample IDs are in bold). The 11-repeat allele can be inferred from the inheritance patterns of the 10-repeat and 12-repeat SNPs, however, we did not identify any SNPs that were present in all 11 repeat alleles. Homozygous was defined by a minor allele frequency of 80-100% and heterozygous by 20-50%. 0/0=wildtype; 0/1=heterozygous; 1/1=homozygous. Nucleotide substitutions are indicated in brackets below each SNP ID (dbSNP database, NCBI).

a

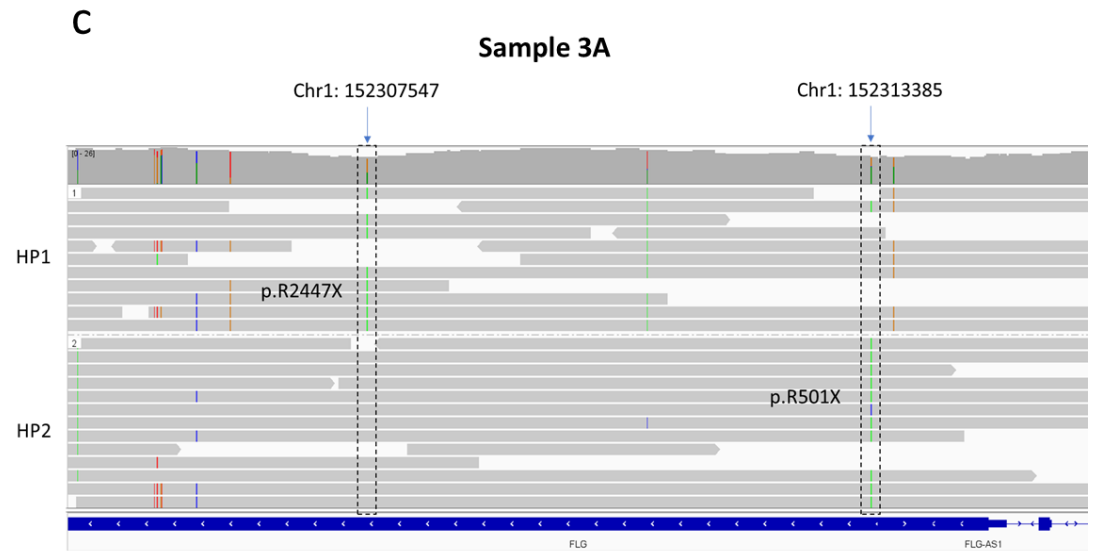

10_10
10_11
11_11
10_12
11_12
12_12
100%
80%
60%
40%
20%
0%
Percentage of reads with respective number of large insertions with reference to GRCh38
99.1%
39.1%
60.4%
96.3%
29.3%
67.8%
36.2%
59.4%
90.8%
6.1%
1A
2A
3A
4A
5A
6A
0 INS (10 copies)
1 INS (11 copies)
2 INS (12 copies)
>2 INS (>12 copies)

169

b


Sample 2A
Chr1: 152305825
Chr1: 152313919
HP1
HP2
p.G323X
c.9040dup19
FLG
FLG-A

170



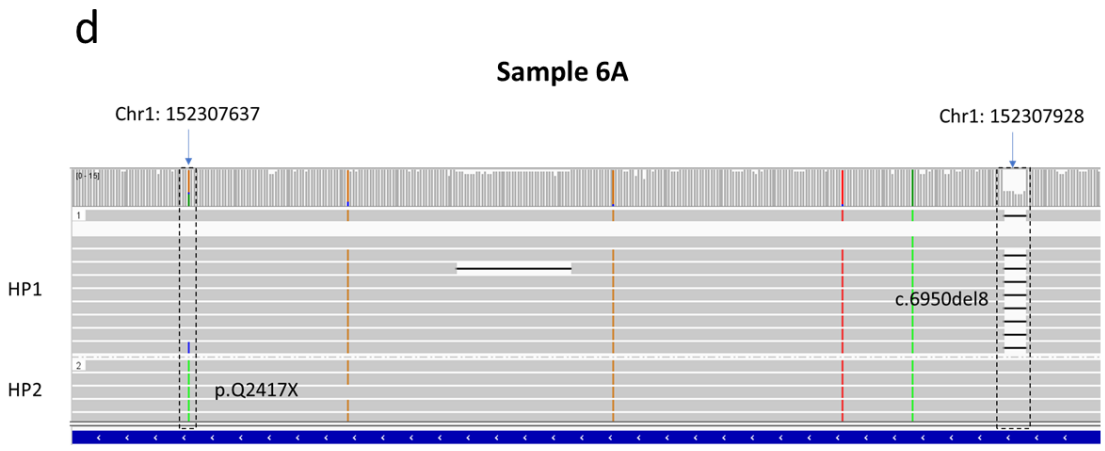
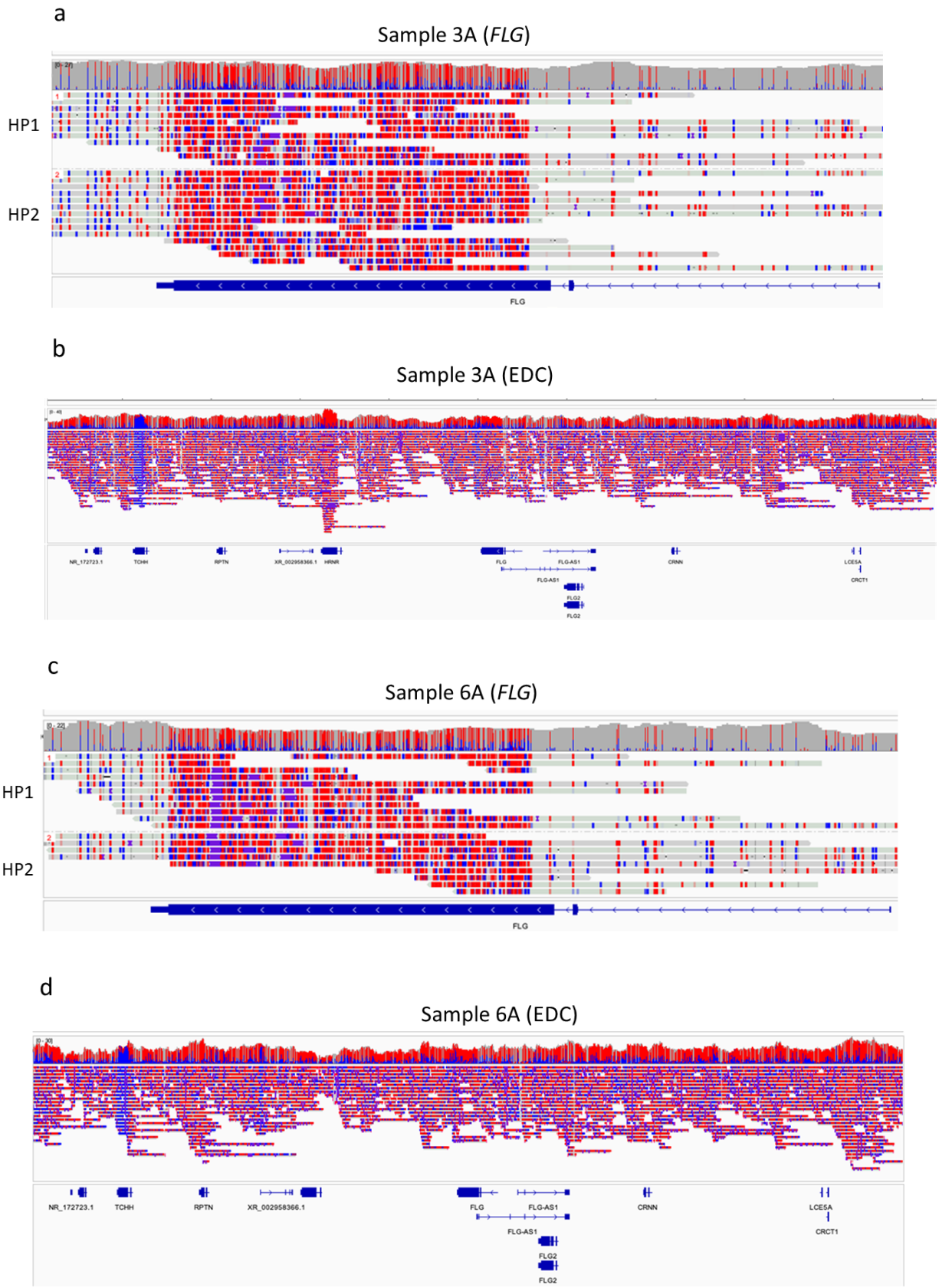


Figure S1: *FLG* CNV detection, variant calling and phasing using nanopore long reads. **a**) Bar chart showing the percentage of amplicon-sequencing reads mapping to *FLG* alleles containing specific CNVs. *FLG* extra repeats (RPT) were identified by detecting large insertions (>800 bp) when reads were aligned to the GRCh38. RPT alleles with a minimum allele frequency of 0.2 were designated the major RPT alleles. The predefined CNV status of each sample by gel electrophoresis was accurately determined by nanopore sequencing and is labelled at the top of each bar (for example, denoted as 10_10 to represent that both alleles had 10 *FLG* repeats). **b**) Representative IGV plot of read alignments for sample 2A showing in *trans* phasing of two LoF variants 8 kb apart with amplicon sequencing: a 19 bp duplication (c.9040dup19; Chr1: 152305825) and a single-nucleotide substitution (p.G323X; Chr1:

152313919). **c**) IGV plot of read alignments for sample 3A showing in *trans* phasing of LoF variants with adaptive sampling: two single-nucleotide substitutions (p.R501X; Chr1:152313385 and p.R2447X; Chr1:152307547). **d**) IGV plot of read alignments for sample 6A showing in *trans* phasing of LoF variants with adaptive sampling: an 8 bp deletion (c.6950del8; Chr1:152307928) and a single-nucleotide substitution (p.Q2417X; Chr1:152307637). IGV colour schemes are set to default with bases that match the reference displayed in gray, purple indicating sequence insertions, black horizontal lines corresponding to sequence deletions and coloured line markings representing single base substitutions (Red=T; Green=A; Blue=C; Orange=G). HP, haplotype. Reads are grouped according to haplotype and separate by a gray dotted horizontal line.



193 **Figure S2: Nanopore adaptive sampling sequencing enables generation of base**

194 **methylation profiles. a)** IGV plot of the *FLG* locus methylation pattern for sample 3A. **b)** IGV

plot of the EDC methylation pattern for sample 3A. **c)** IGV plot of the *FLG* locus methylation pattern for sample 6A. **d**) IGV plot of the EDC methylation pattern for sample 6A. IGV colour schemes are set to default with hypermethylated bases denoted in red, hypomethylated bases denoted in blue and unmethylated bases denoted in gray. Reads are grouped according to haplotype (HP1=haplotype 1; HP2=haplotype 2) and separate by a gray, dotted horizontal line.

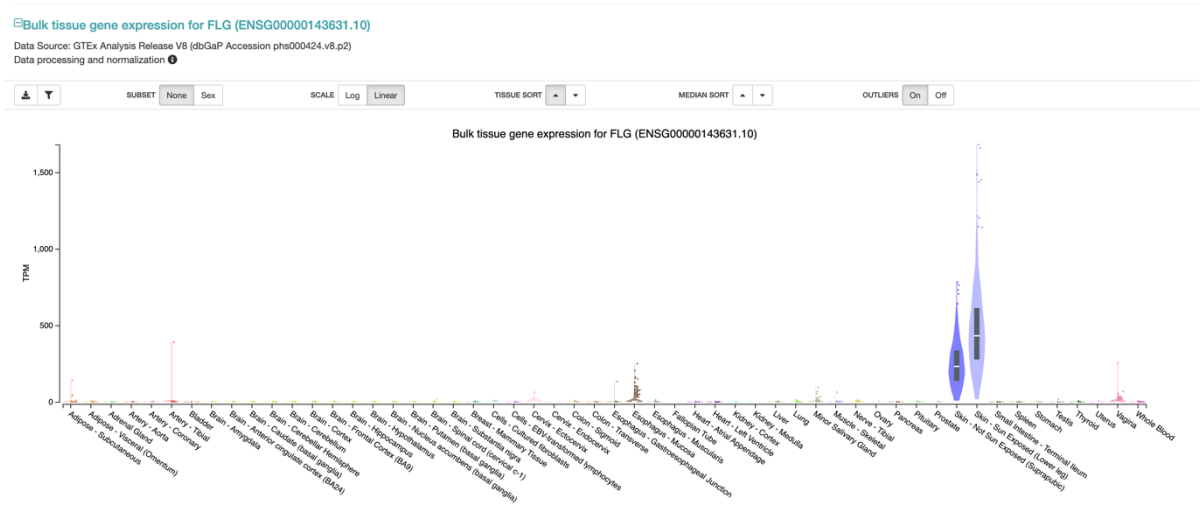


Figure S3: *FLG* expression level across different bulk tissues. Violin plot obtained from the Genotype-Tissue Expression (GTEx) portal showing high expression of *FLG* almost exclusively in skin tissue. The data used for this analyses were obtained from dbGaP accession number phs000424.v8.p2 on 12/20/2023. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. TPM, transcript per million.

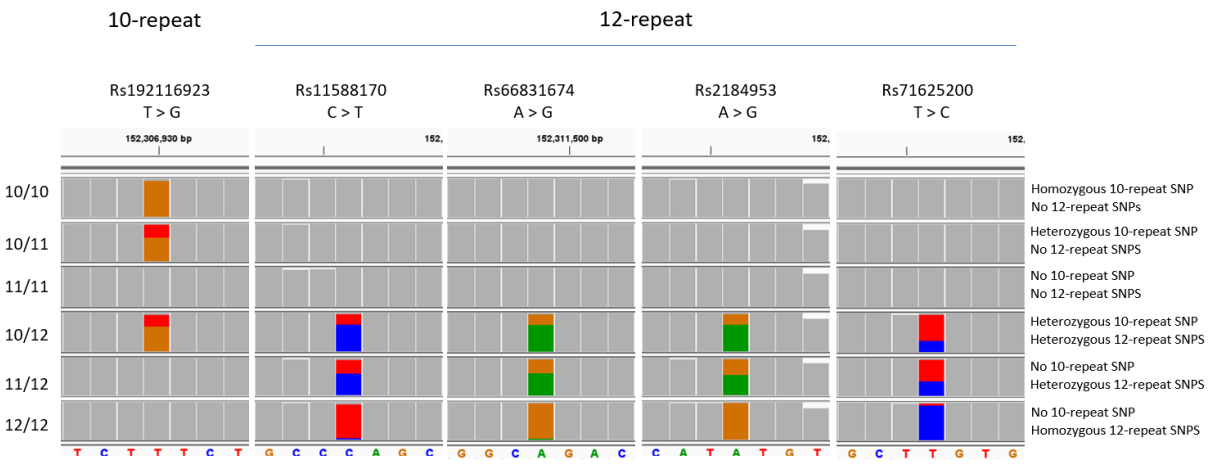


Figure S3. IGV boxplot to demonstrate the analysis with five SNPs associated with the 10- and 12-repeat alleles. Genotypes of the 10-repeat SNP (rs192116923) and four representative 12-repeat SNPs (rs11588170, rs66831674, rs2184953 and rs71625200) of the n=37 we have identified in the six samples from our method development cohort samples.

Supplementary References

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10. https://doi.org/10.1093/gigascience/giab008.

Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 2018;34. https://doi.org/10.1093/bioinformatics/bty191.

Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, et al. WhatsHap: fast and accurate read-based phasing. BioRxiv 2016.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 2017;27. https://doi.org/10.1101/gr.213611.116.

Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods 2021;18. https://doi.org/10.1038/s41592-021-01299-w.

Smith FJD, Irvine AD, Terron-Kwiatkowski A, Sandilands A, Campbell LE, Zhao Y, et al. Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. Nat Genet 2006;38. https://doi.org/10.1038/ng1743.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 2013;14. https://doi.org/10.1093/bib/bbs017.

Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom 2017;3. https://doi.org/10.1099/mgen.0.000132.

Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol 2019;20. https://doi.org/10.1186/s13059-019-1727-y.

Williams HC, Jburney PG, Hay RJ, Archer CB, Shipley MJ, Ahunter JJ, et al. The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. I. Derivation of a minimum set of discriminators for atopic dermatitis. British Journal of Dermatology 1994;131. https://doi.org/10.1111/j.1365-2133.1994.tb08530.x.