Characterization of challenging genes related to genetic diseases: The case of the Filaggrin gene

Victor Grentzinger

Victor Grentzinger
Laboratory of Human Genetics

GIGA-MCB seminar 24/04/25

What is a Challenging Medically Relevant Gene?

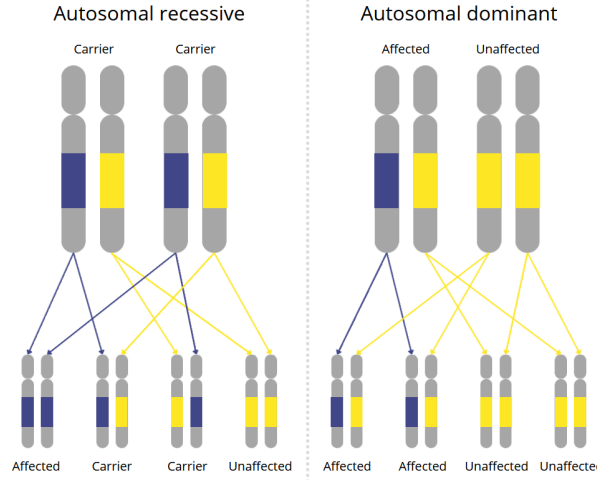
What is a Medically Relevant Gene?

Medically relevant genes

Genes with high-impact variants for genetic diseases.

Important to know if dominant or recessive for diagnosis.

Can also be located on gonosomes (chromosome X or Y).



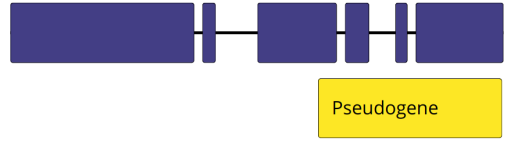
What is a Medically Relevant Gene?

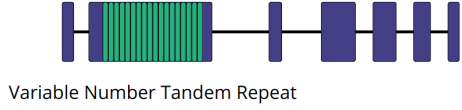
What is a Challenging Gene?

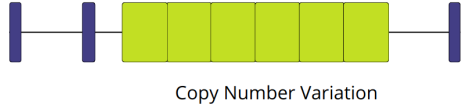
Challenging genes

Highly similar to other DNA sequences within the same genome.

These homologous regions are really complex to characterize.







What is a Challenging Medically Relevant Gene?

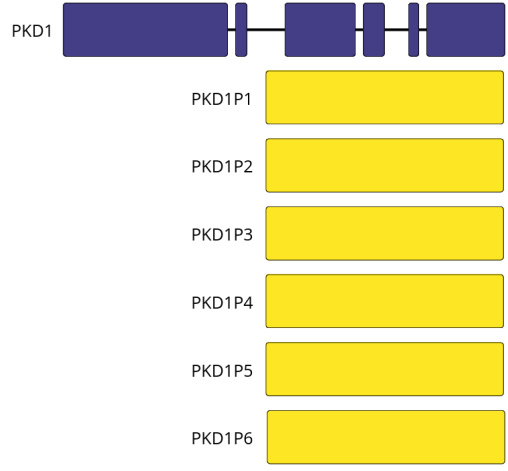
Challenging Medically Relevant Genes (CMRG)

395 genes that are both challenging and with high impact variants.

Two major categories:

- Sanger deadzone.
- NGS deadzone.

Example: PKD1

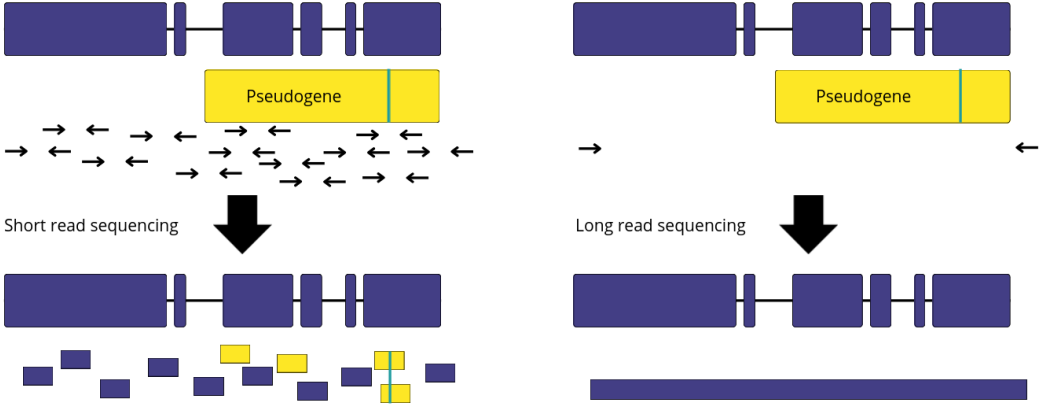


Responsible for 75% of Autosomal Dominant Polycystic Kidney Disease.

6 pseudogenes that are highly similar

Part of the NGS deadzone.

Variant identification for PKD1



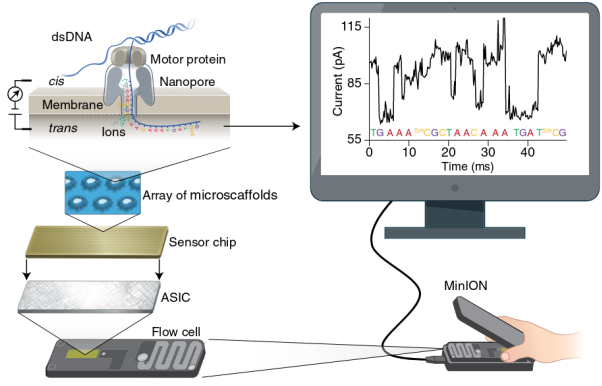
Oxford Nanopore Technologies sequencing

DNA strand passes through a nanopore.

The sequencer records electrical current changes.

Read length: +10k bp.

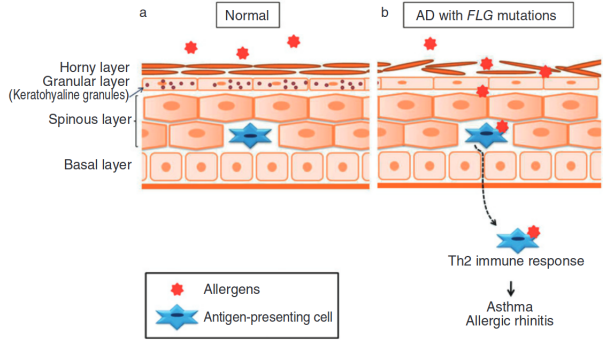
Read Until: Select only reads from given region(s) of interest.



Y. Wang et al. (2021)

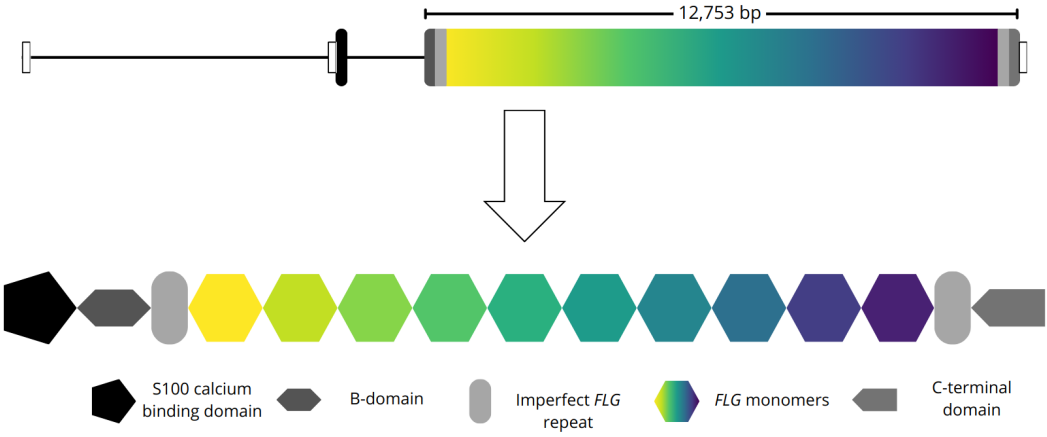
Atopic Dermatitis (AD)

J. Heilman (2010)

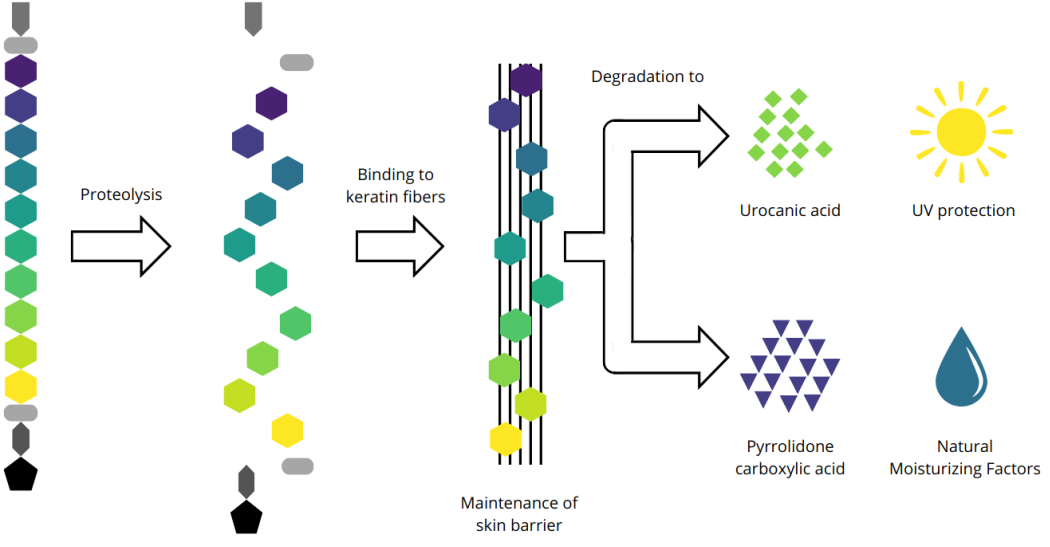


R. Osawa, M. Akiyama and H. Shimizu (2011)

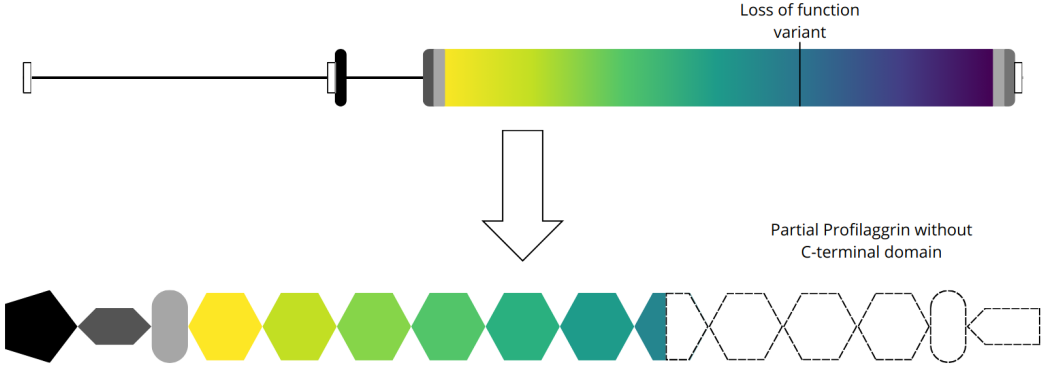
The Filaggrin gene (*FLG*)



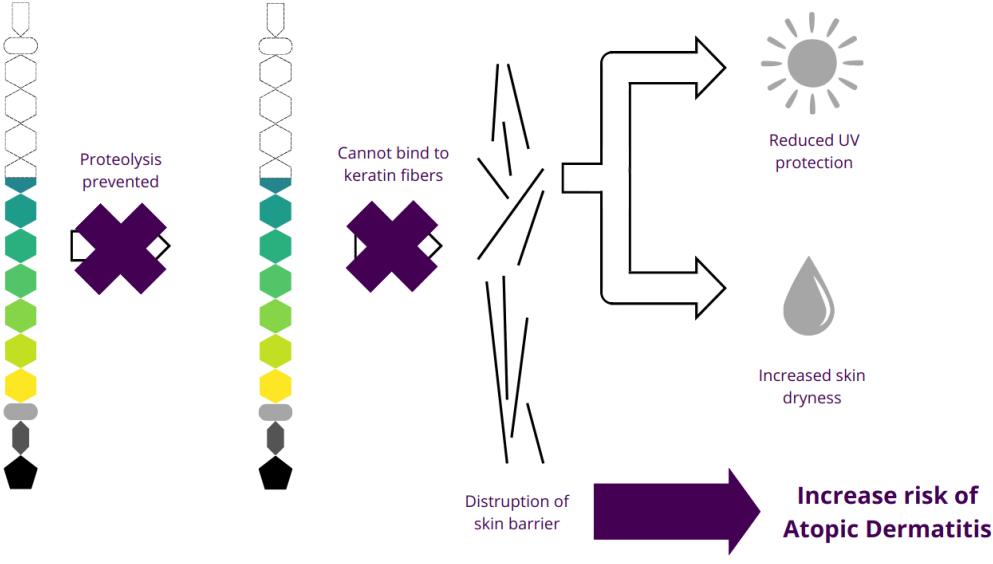
Role of *FLG*


Proteolysis
Binding to keratin fibers
Degradation to
Urocanic acid
UV protection
Pyrrolidone carboxylic acid
Natural Moisturizing Factors
Maintenance of skin barrier

LOF variants of *FLG*

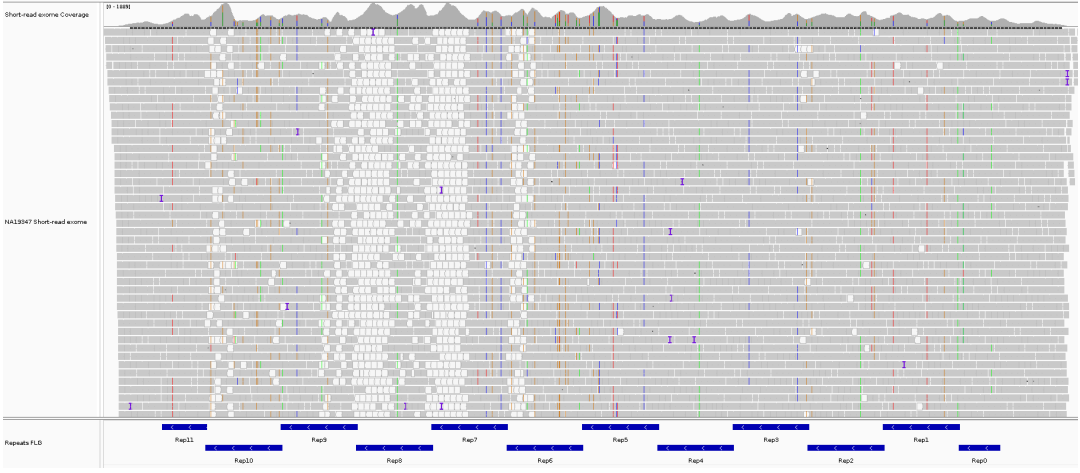


LOF variants prevents proteolysis

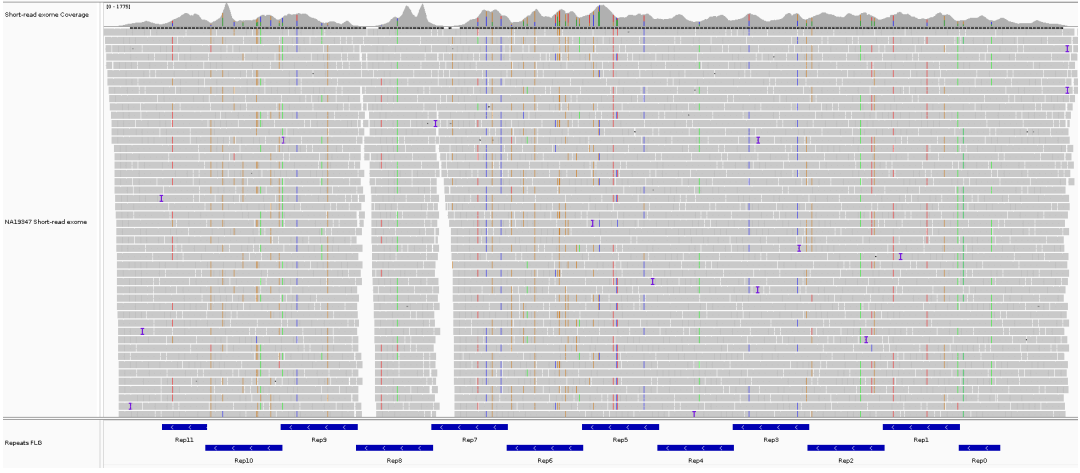


Multiple known alleles of *FLG*

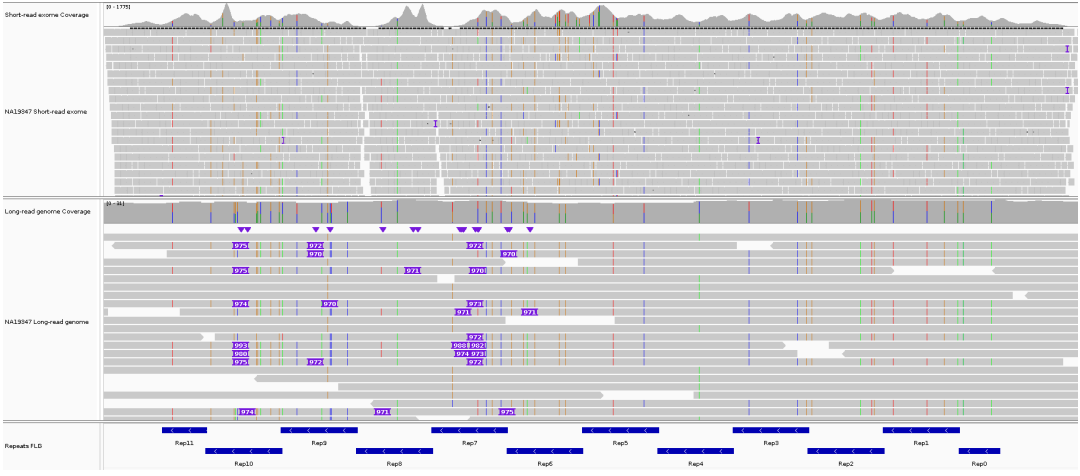
Short read sequencing limitation



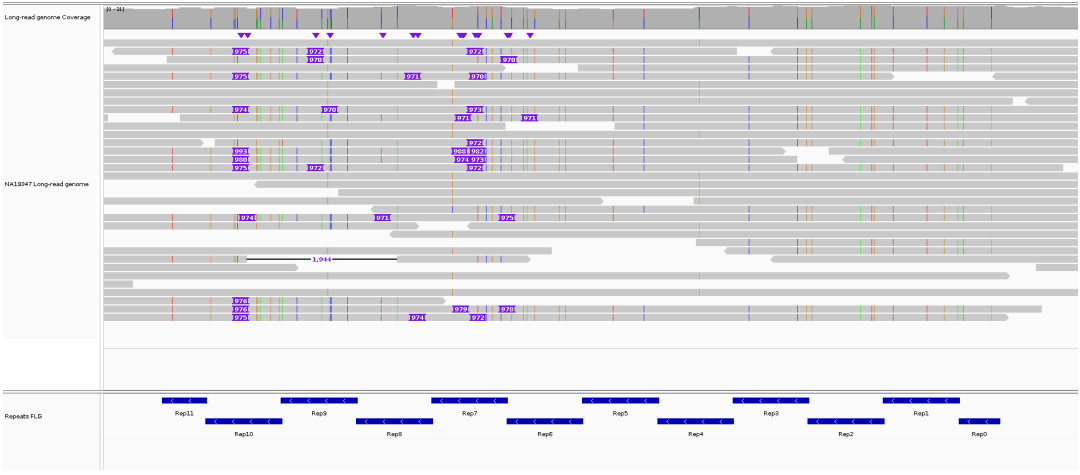
Short read sequencing limitation



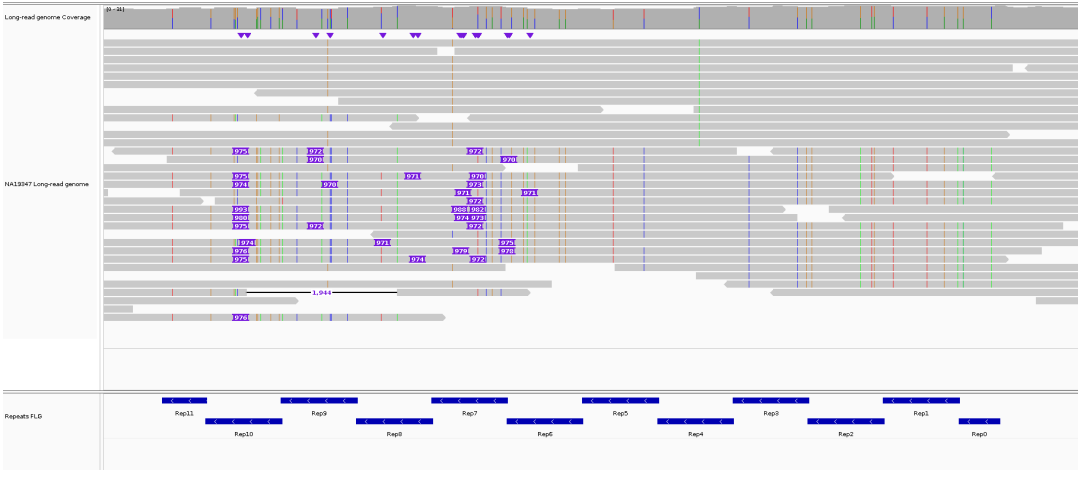
Short read vs Long read



Something is off...

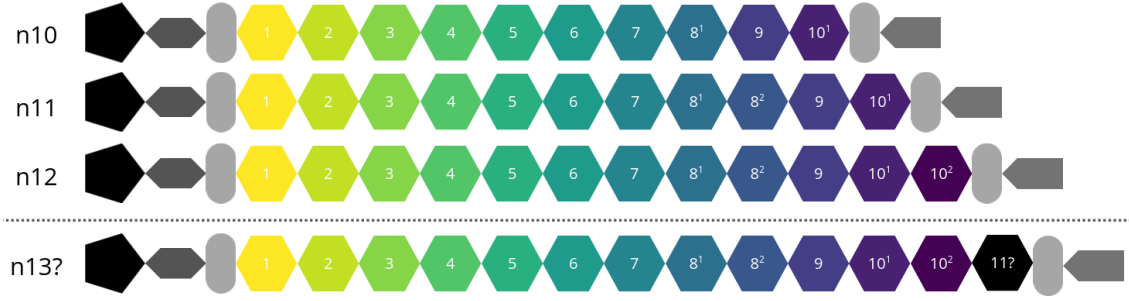


Something is off...



Undescribed *FLG* allele(s)?

Undescribed *FLG* allele(s)?



FLG research project

Capture the genetic variability of the *FLG* gene in a catalog of alleles and infer new pathogenic variants by using long read sequencing technologies.

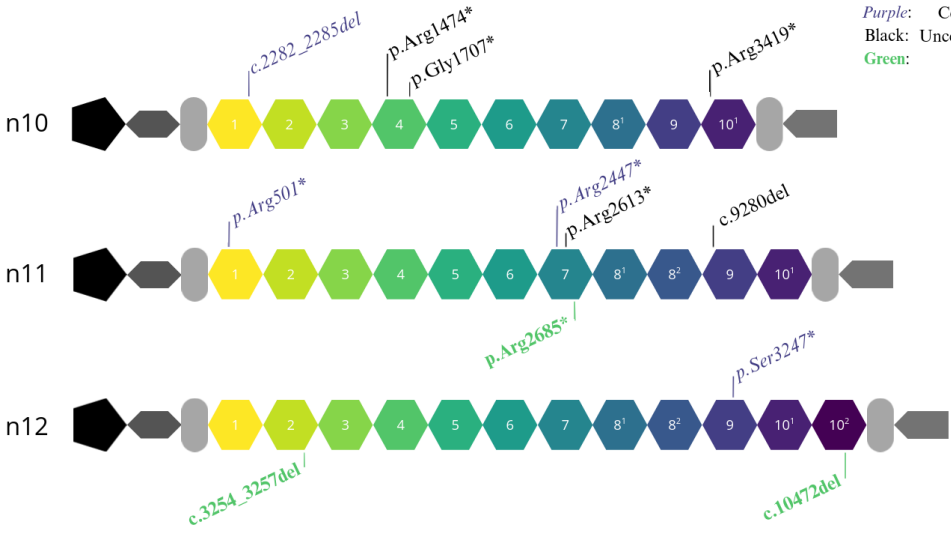
In-house clinical cohort

Origin	Control samples	AD patients	Total
Africa	8	68	76
Europe	83	43	126
Total	91	111	202

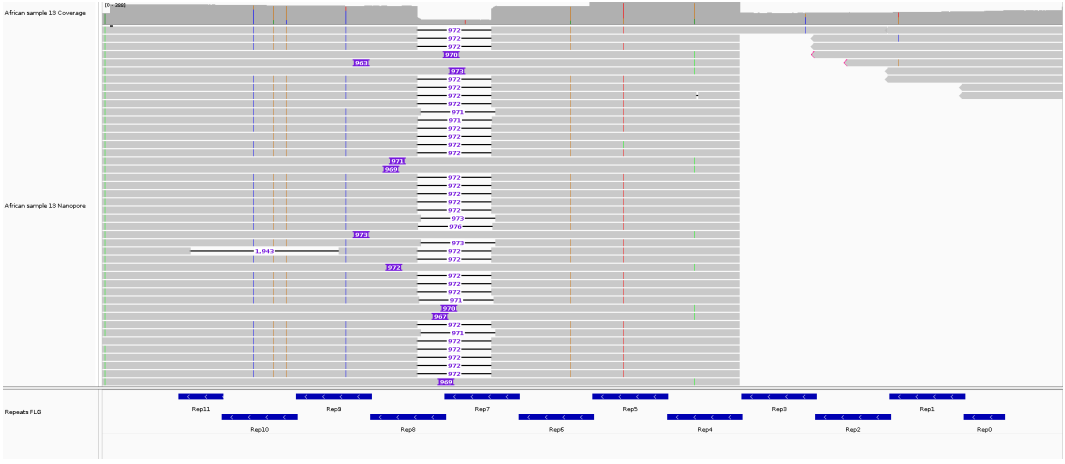
CENTRE HOSPITALIER UNIVERSITAIRE DE KIGALI

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Identification of new LOF variants



Identification of novel alleles



Multi-ethnic cohort

Origin	Number of samples
Africa	392
Europe	326
East Asia	213
South Asia	223
Latin America	182
Total	1336

CENTRE HOSPITALIER UNIVERSITAIRE DE KIGALI

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Quality Control of Public Data

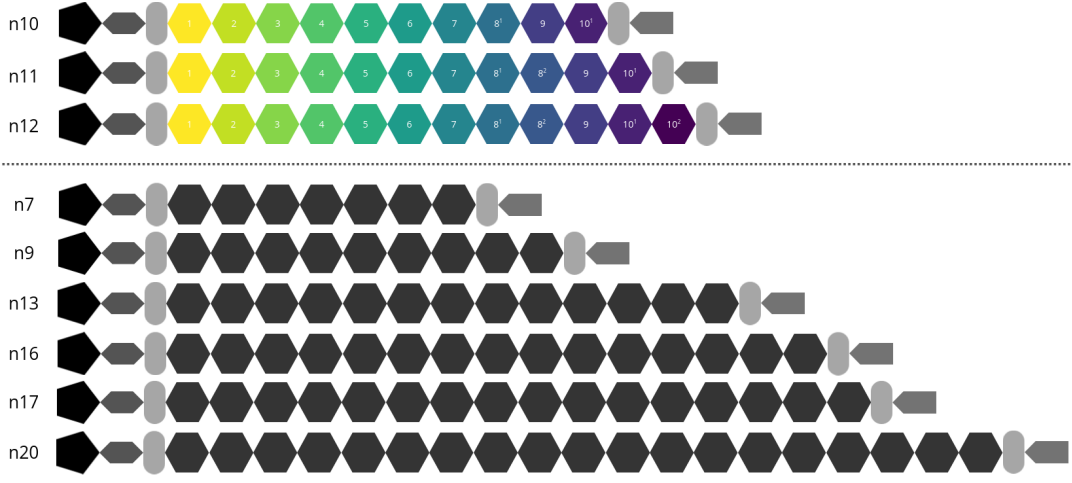
Quality Control of Public Data

Multi-ethnic cohort after Quality Control

Origin	Number of samples	After QC
Africa	392	355
Europe	326	311
East Asia	213	193
South Asia	223	191
Latin America	182	163
Total	1336	1193

Catalog of novel alleles

Catalog of novel alleles

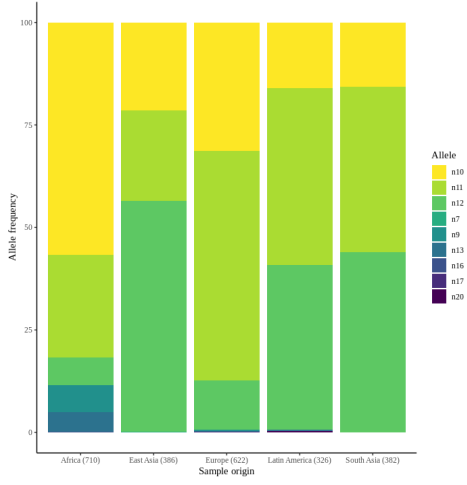


Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	**403**	195	83	60	52
n11	177	**348**	85	154	**141**
n12	48	75	***217***	**168**	131
n7	0	0	1	0	0
n9	47	2	0	0	0
n13	34	1	0	0	1
n16	1	0	0	0	0
n17	0	1	0	0	0
n20	0	0	0	0	1

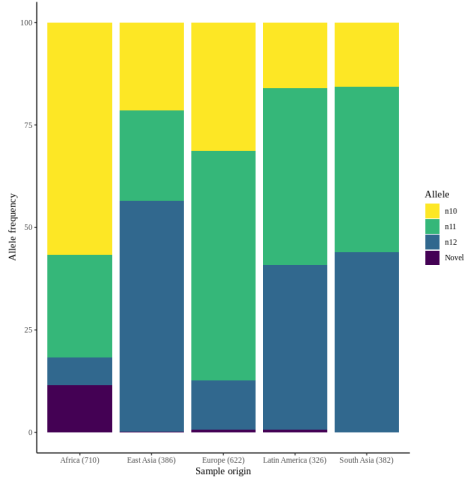
Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	**403**	195	83	60	52
n11	177	**348**	85	154	**141**
n12	48	75	***217***	**168**	131
n7	0	0	1	0	0
n9	47	2	0	0	0
n13	34	1	0	0	1
n16	1	0	0	0	0
n17	0	1	0	0	0
n20	0	0	0	0	1



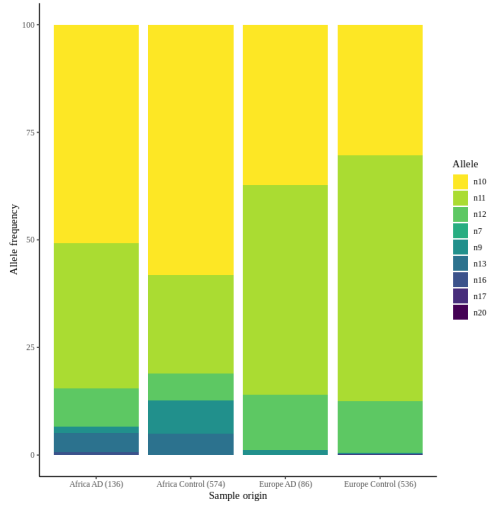
Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	**403**	195	83	60	52
n11	177	**348**	85	154	**141**
n12	48	75	***217***	**168**	131
n7	0	0	1	0	0
n9	47	2	0	0	0
n13	34	1	0	0	1
n16	1	0	0	0	0
n17	0	1	0	0	0
n20	0	0	0	0	1



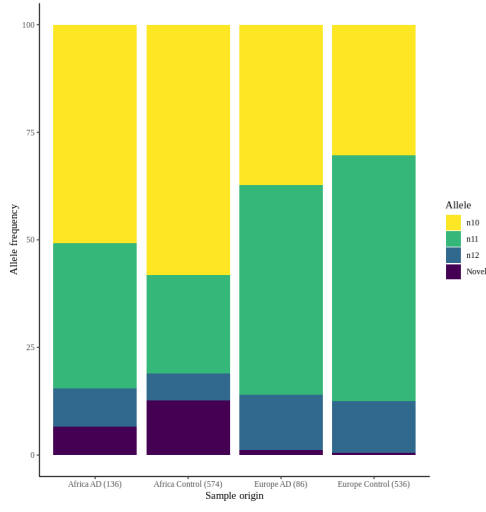
Frequencies comparison AD/controls

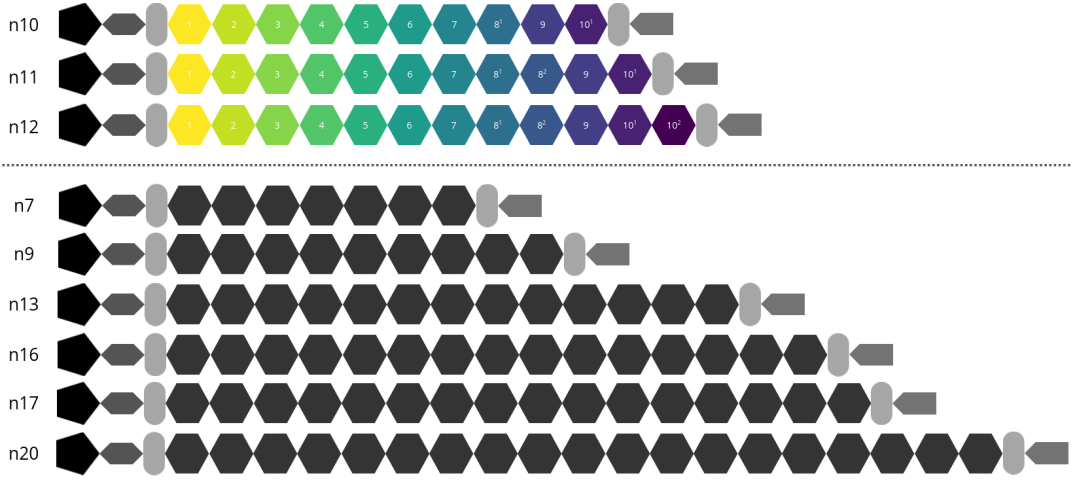
Allele	Africa		Europe	
	AD	Control	AD	Control
n10	**69**	**334**	32	163
n11	46	131	**42**	**306**
n12	12	36	11	64
n7	0	0	0	0
n9	2	45	1	1
n13	6	28	0	1
n16	1	0	0	0
n17	0	0	0	1
n20	0	0	0	0

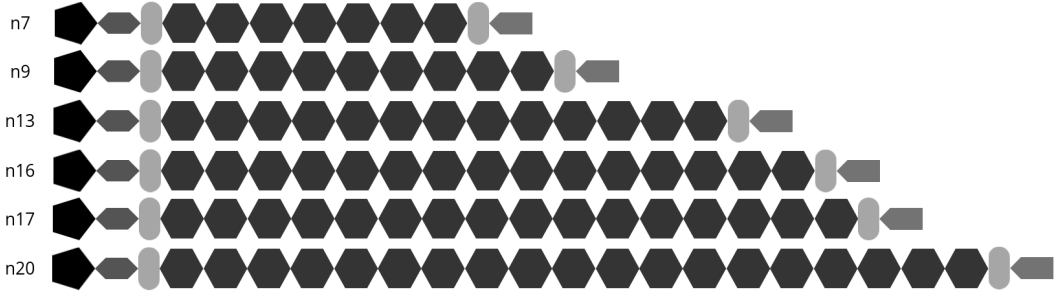


Frequencies comparison AD/controls

Allele	Africa		Europe	
	AD	Control	AD	Control
n10	**69**	**334**	32	163
n11	46	131	**42**	**306**
n12	12	36	11	64
n7	0	0	0	0
n9	2	45	1	1
n13	6	28	0	1
n16	1	0	0	0
n17	0	0	0	1
n20	0	0	0	0




n10
n11
n12
n7
n9
n13
n16
n17
n20


n7
n9
n13
n16
n17
n20

Characterization of novel alleles

CNV events have been called for each sample and stored in VCF files.

The position, the length and the sequence of each event is known.

Just need to remap these sequence to the human genome!

Characterization of novel alleles: the smarter way

CNV events have been called for each sample and stored in VCF files.

The position, the length and the sequence of each event is known.

Just need to remap these sequence to the human genome!

Characterization of novel alleles: the other way

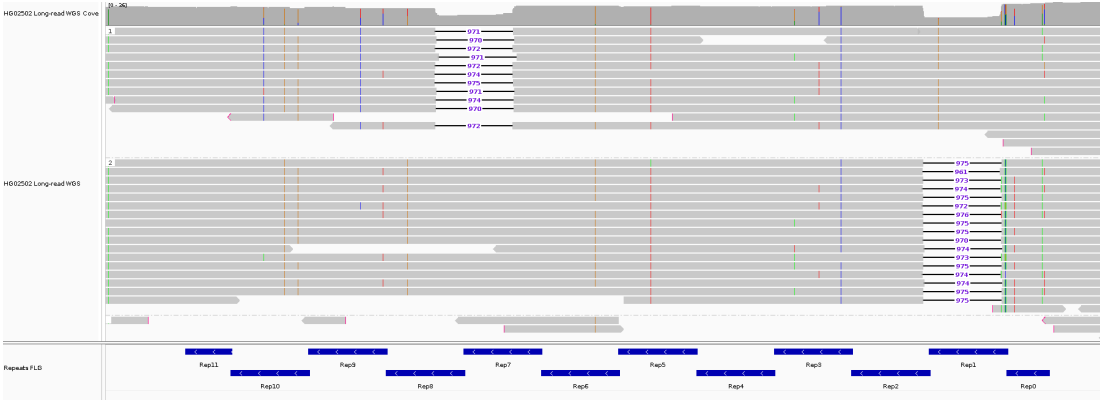
Take all 1193 samples and visualize them on IGV.

Note by hand which sample have which CNV event.

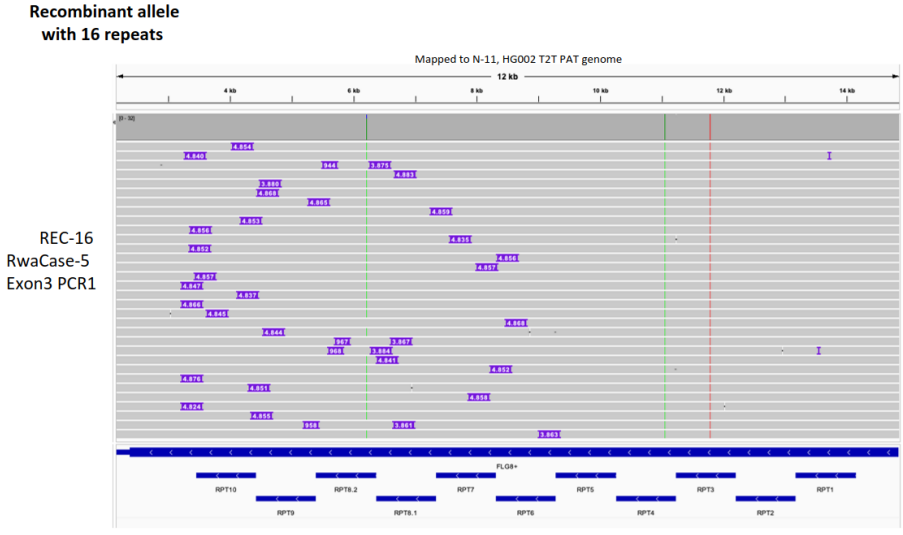
Map novel alleles to n10, n11 and n12 alleles.

Extract manually the CNV sequence and remap it.

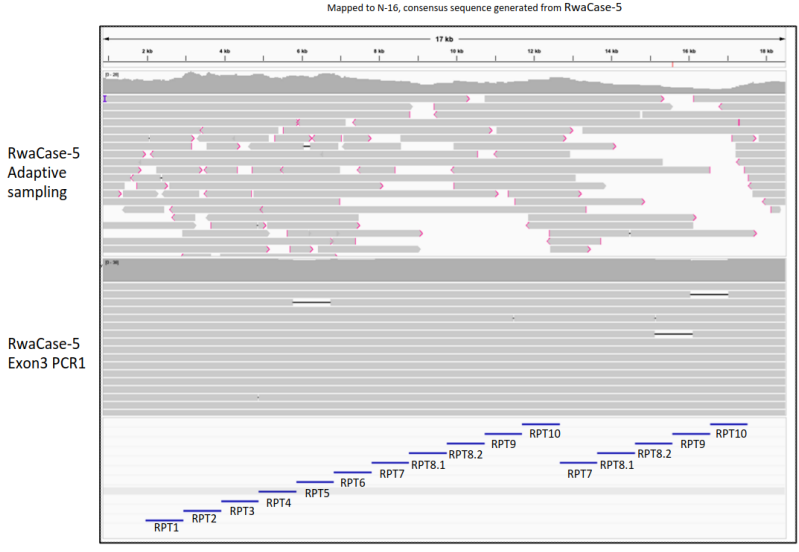
Alleles with deletions

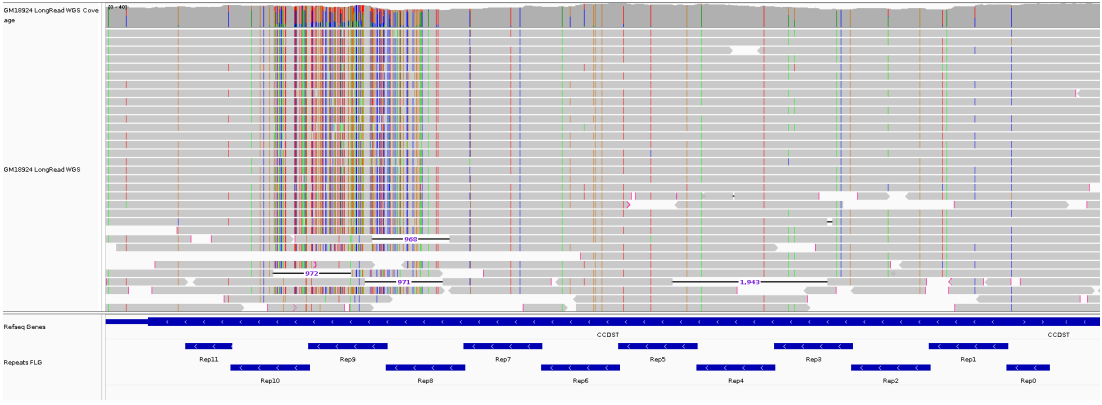


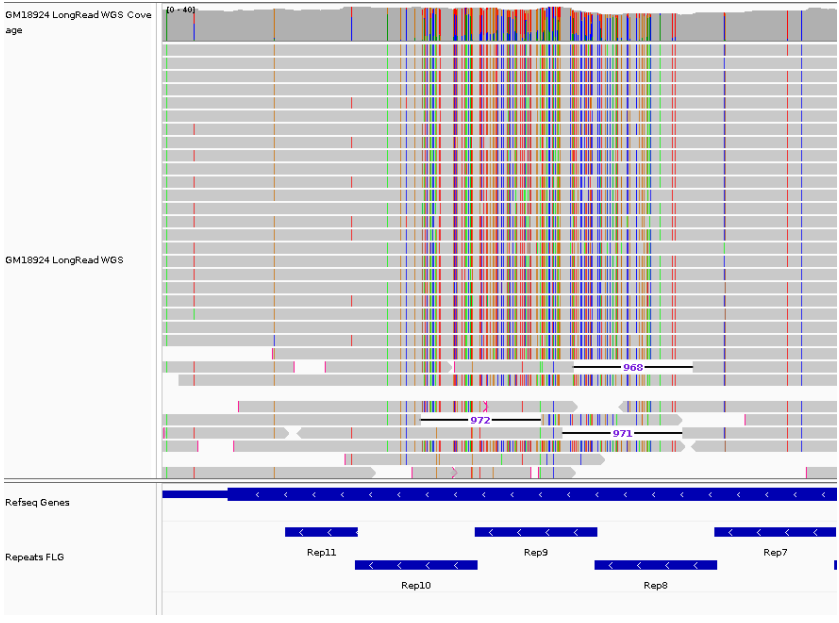
Alleles with insertions

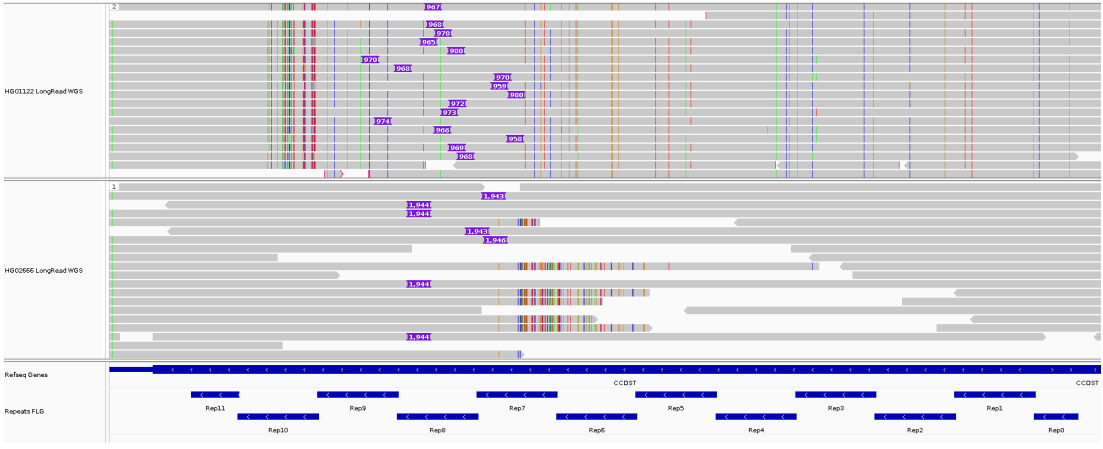


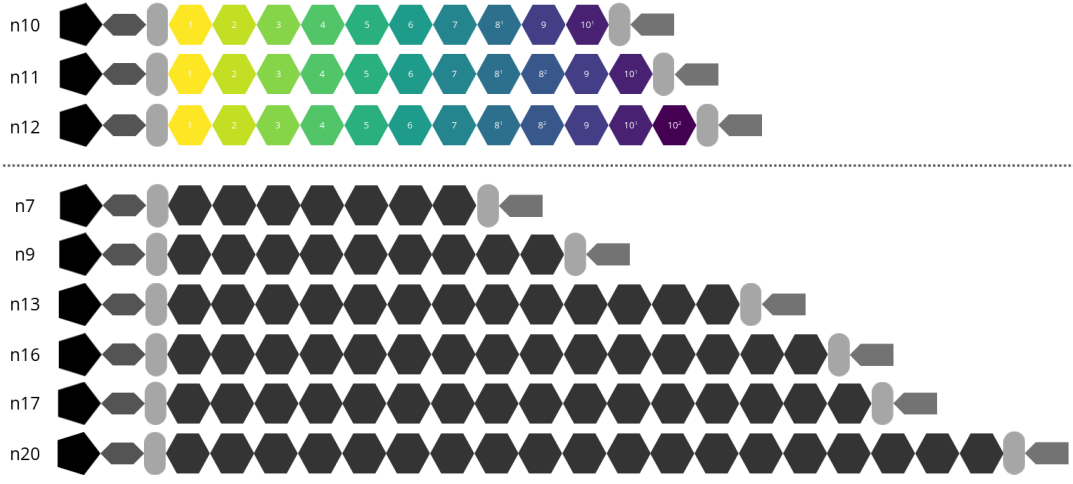
Insertions: Consensus and motif searching

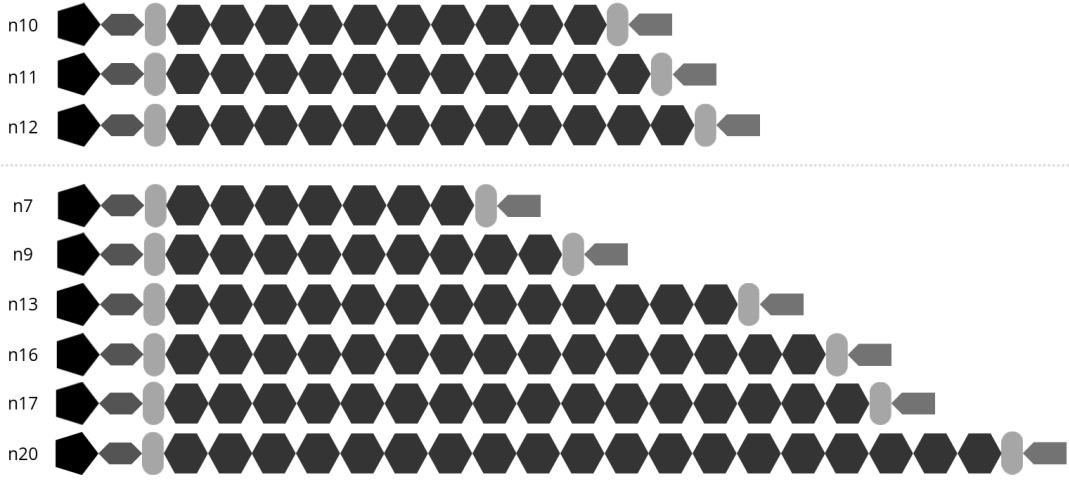



GM18924 LongRead WGS Coverage
GM18924 LongRead WGS
Refseq Genes
Repeats FLG
CCDST
CCDST
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0


GM18924 LongRead WGS Cove
age
[0 - 40]
GM18924 LongRead WGS
968
972
971
Refseq Genes
Repeats FLG
Rep11
Rep9
Rep7
Rep10
Rep8

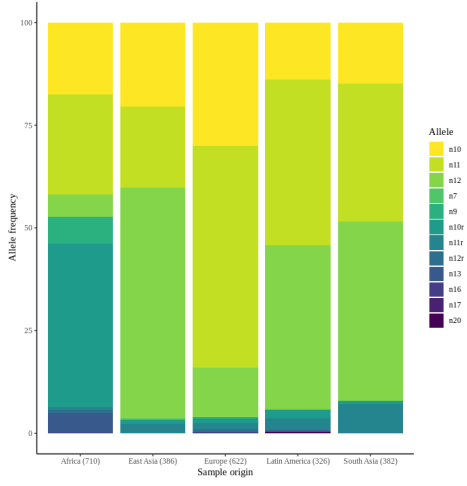

HG01122 LongRead WGS
HG02666 LongRead WGS
Refseq Genes
CCDST
CCDST
Repeats FLG
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0


n10
n11
n12
n7
n9
n13
n16
n17
n20


n10
n11
n12
n7
n9
n13
n16
n17
n20

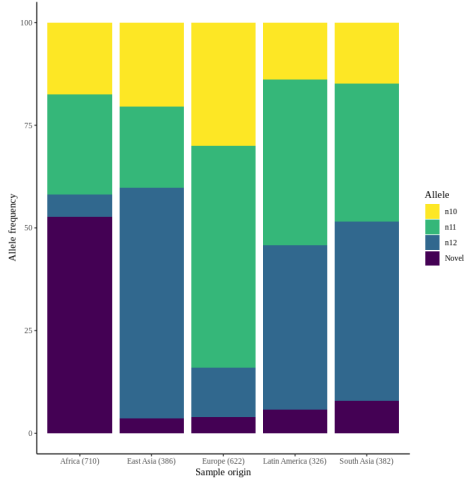
Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	124	188	79	57	45
n11	172	**338**	76	128	**132**
n12	38	75	***217***	**167**	130
n7	0	0	1	0	0
n9	47	2	0	0	0
n10	***279***	7	4	3	7
n11	5	10	9	26	9
n12	6	4	0	1	1
n13	34	1	0	0	1
n16	1	0	0	0	0
n17	0	1	0	0	0
n20	0	0	0	0	1



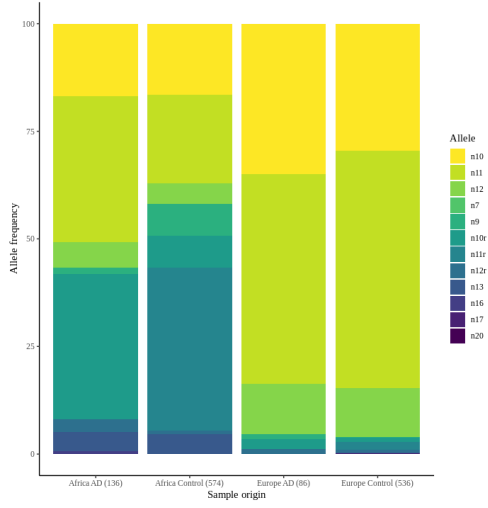
Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	124	188	79	57	45
n11	172	**338**	76	128	**132**
n12	38	75	***217***	**167**	130
n7	0	0	1	0	0
n9	47	2	0	0	0
n10	***279***	7	4	3	7
n11	5	10	9	26	9
n12	6	4	0	1	1
n13	34	1	0	0	1
n16	1	0	0	0	0
n17	0	1	0	0	0
n20	0	0	0	0	1



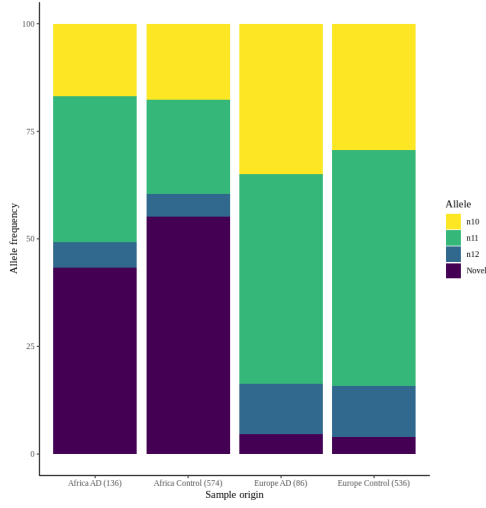
Frequencies comparison AD/controls

Allele	Africa		Europe	
	AD	Control	AD	Control
n10	23	101	30	158
n11	**46**	126	**42**	**296**
n12	8	30	10	61
n7	0	0	0	0
n9	2	45	1	1
n10	***46***	***233***	2	5
n11	0	5	0	10
n12	4	6	1	3
n13	6	28	0	1
n16	1	0	0	0
n17	0	0	0	1
n20	0	0	0	0

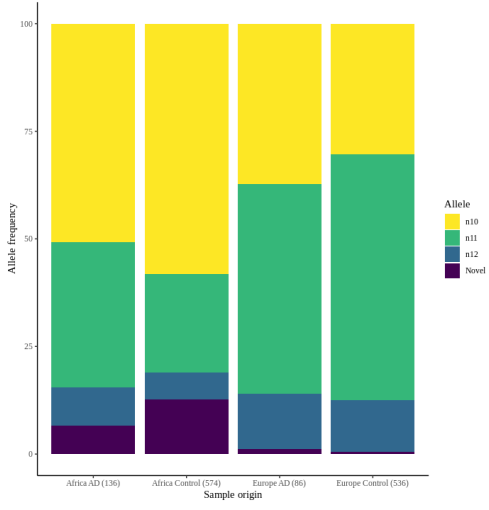


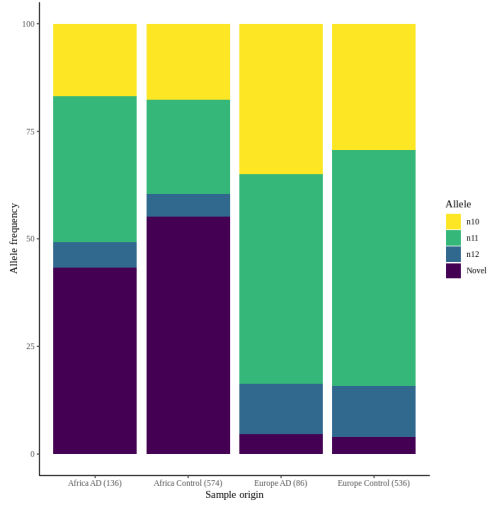
Frequencies comparison AD/controls

Allele	Africa		Europe	
	AD	Control	AD	Control
n10	23	101	30	158
n11	**46**	126	**42**	**296**
n12	8	30	10	61
n7	0	0	0	0
n9	2	45	1	1
n10	***46***	***233***	2	5
n11	0	5	0	10
n12	4	6	1	3
n13	6	28	0	1
n16	1	0	0	0
n17	0	0	0	1
n20	0	0	0	0

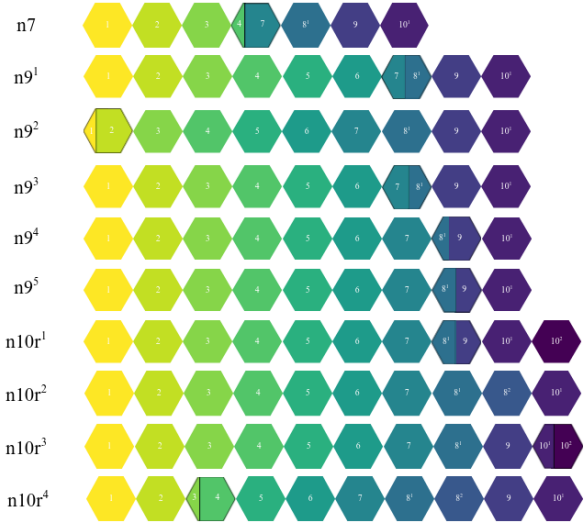


Sheding new light on data





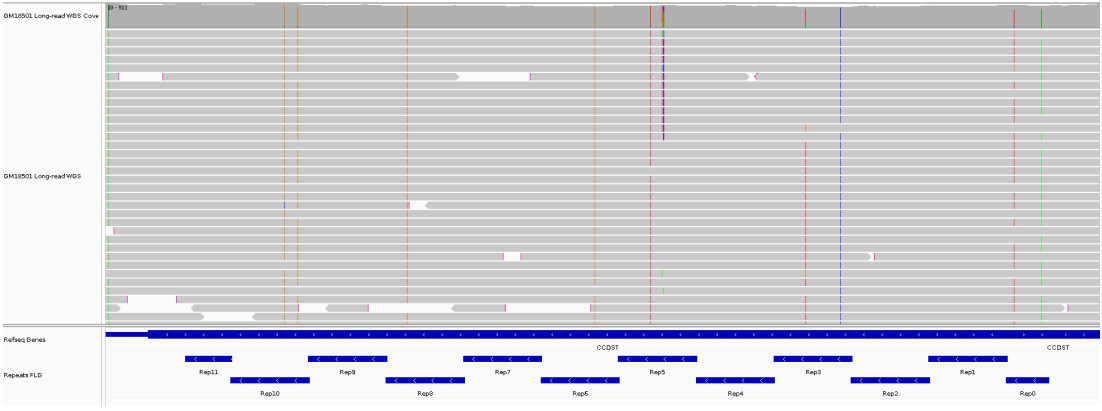
Current state of the catalog - 1

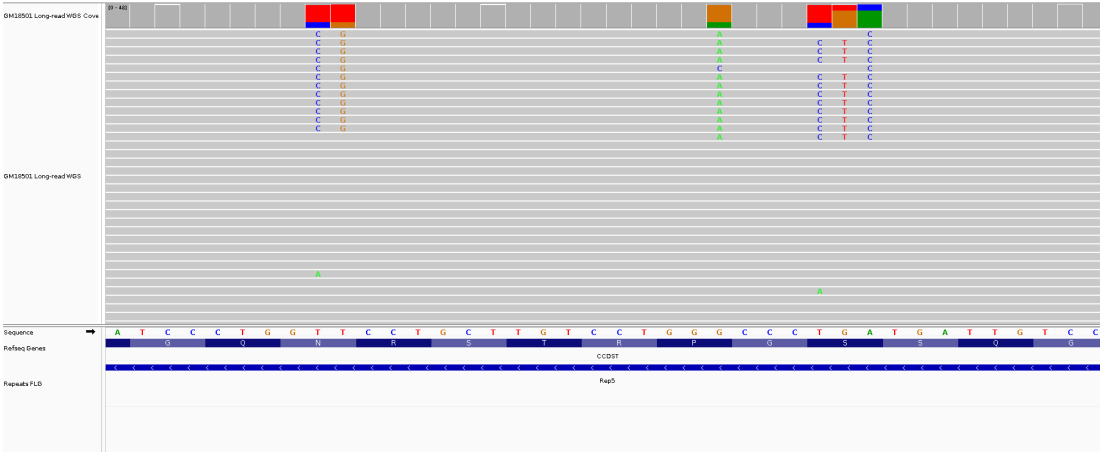


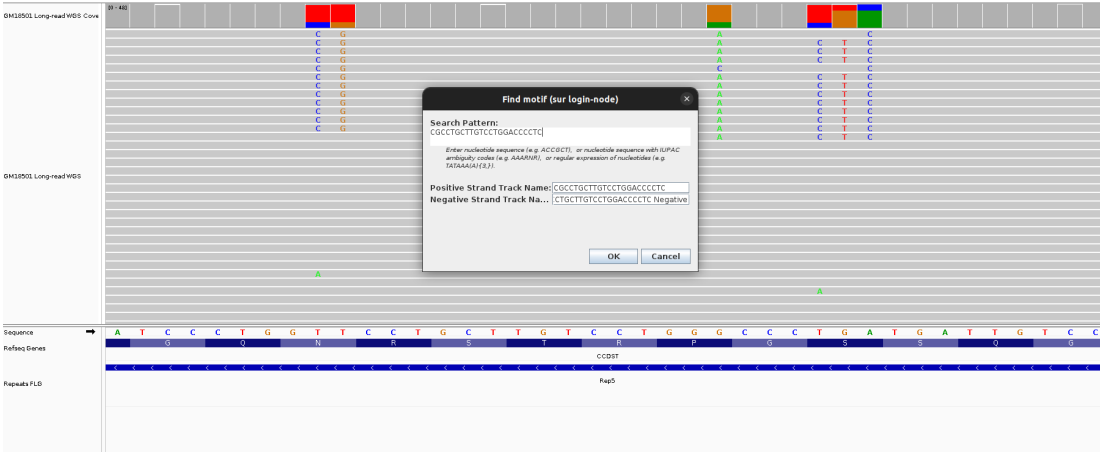
Current state of the catalog - 2

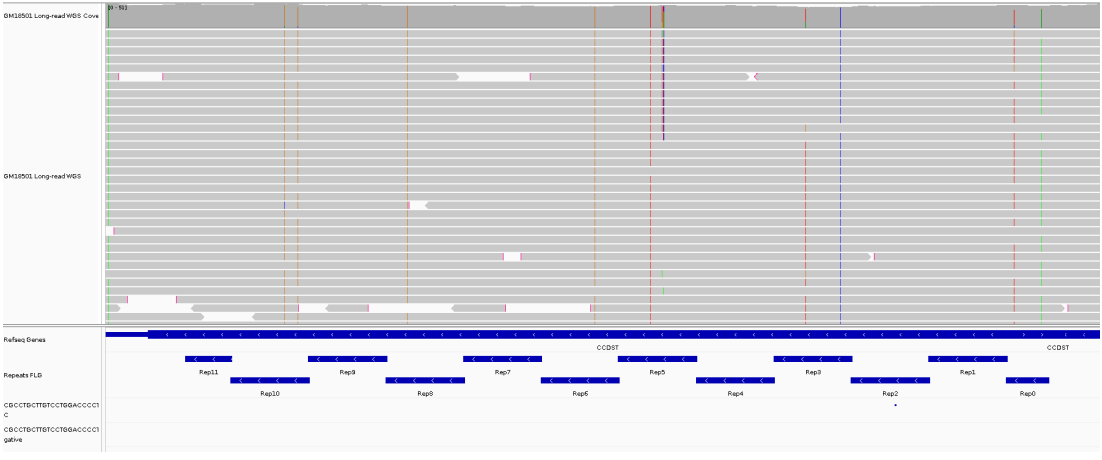
Potential origin of these novel alleles

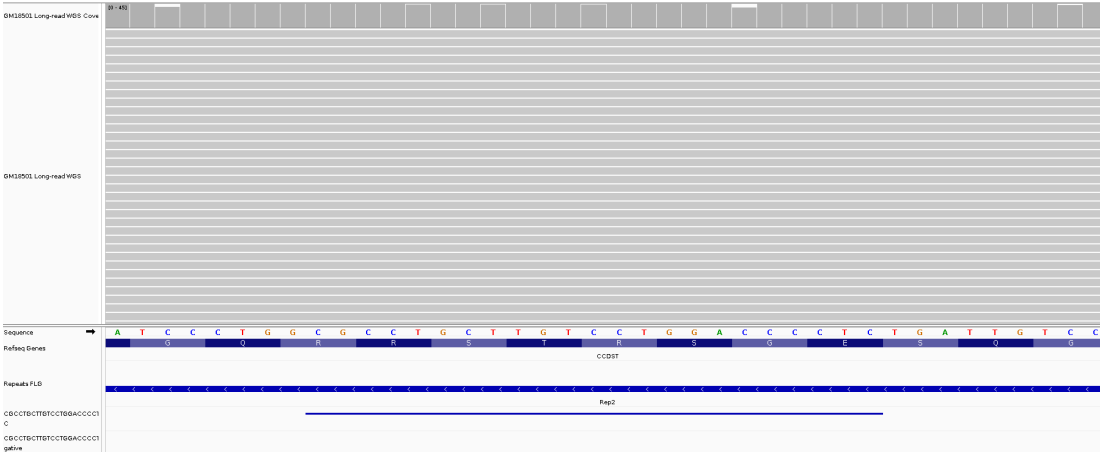
- Seems to originate from misalignment during homologous recombination.
- Due to the high similarity between repeats, an misalignment could happen during DNA repair.

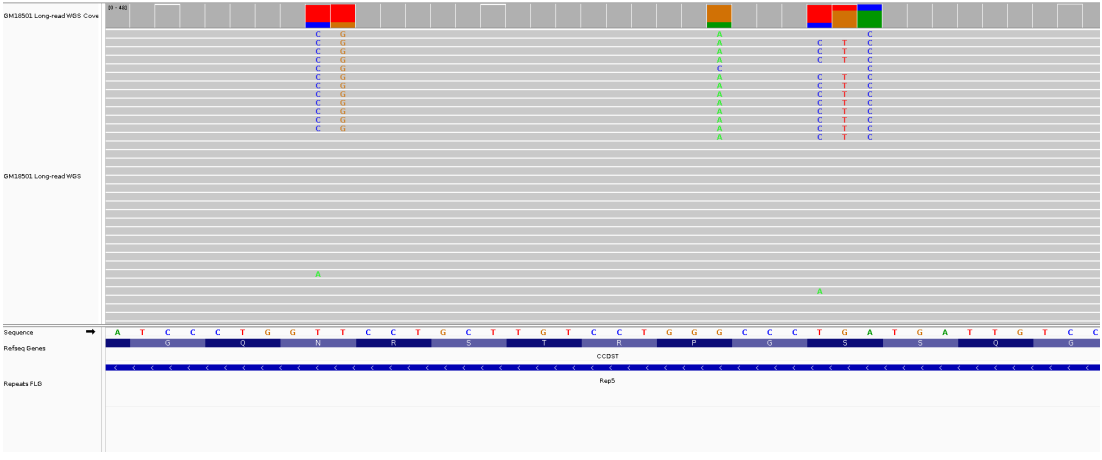

GM18501 Long-read WGS Cove
0 - 51
GM18501 Long-read WGS
Refseq Genes
CCDST
CCDST
Repeats FLG
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0


GM18501 Long-read WGS Cove
GM18501 Long-read WGS
Sequence
Refseq Genes
CCDST
Repeats FLD
Rep5


GM18501 Long-read WGS Cove
GM18501 Long-read WGS
Find motif (sur login-node)
Search Pattern:
CGCCTGCTTGTCCTGGACCCCTC
Enter nucleotide sequence (e.g. ACCGCT), or nucleotide sequence with IUPAC ambiguity codes (e.g. AAARNR), or regular expression of nucleotides (e.g. TATAAA(A|T)).
Positive Strand Track Name:
CGCCTGCTTGTCCTGGACCCCTC
Negative Strand Track Na...
CTGCTTGTCCTGGACCCCTC Negative
OK
Cancel
Sequence
Refseq Genes
CCDST
Repeats FLD
Rep5


GM18501 Long-read WGS Cove
GM18501 Long-read WGS
Refseq Genes
CCDST
CCDST
Repeats FLG
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0
CGCCTGCTTGTCCTGGACCCC1
C
CGCCTGCTTGTCCTGGACCCC1
gative


GM18501 Long-read WGS Cove
0 - 45
GM18501 Long-read WGS
Sequence
A T C C C T G G C G C C T G C T T G T C C T G G A C C C C T C T G A T T G T C C
G Q R R S T R S G E S Q G
Refseq Genes
CCDST
Repeats FLB
Rep2
CGCCTGCTTGTCCTGGACCCC1
C
CGCCTGCTTGTCCTGGACCCC1
gative


GM18501 Long-read WGS Cove
GM18501 Long-read WGS
Sequence
Refseq Genes
CCDST
Repeats FLD
Rep5

Acknowledgement

Pr. Vincent Bours

Dr. Keith Durkin

Romain Thissen

Dr. Sabine Olivier

Dr. Leonor Palmeira

Dr. Maria Artesi

Dr. Benoit Charloteaux

Dr. Laura Helou

Pr. Vinciane Dideberg

Nadine Cambisano

Nathalie Renotte

Dr. Annette Uwineza

Dr. Alice Amani Uwajeni

Pr. Alan Irvine