LIÈGE université
GIGA

Lifting the veil on Challenging Medically Relevant Genes

Victor Grentzinger

Laboratory of Human Genetics
GIGA Institute - Liège, Belgium

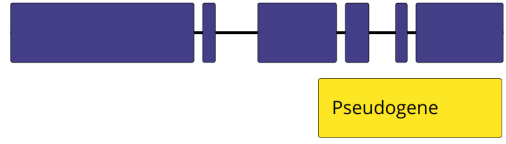
ISHG 07/05/25

What is a Challenging Medically Relevant Gene?

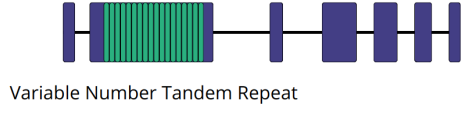
What is a Challenging Gene?

Challenging genes

Highly similar to other DNA sequences within the same genome.



These homologous regions are really complex to characterize.

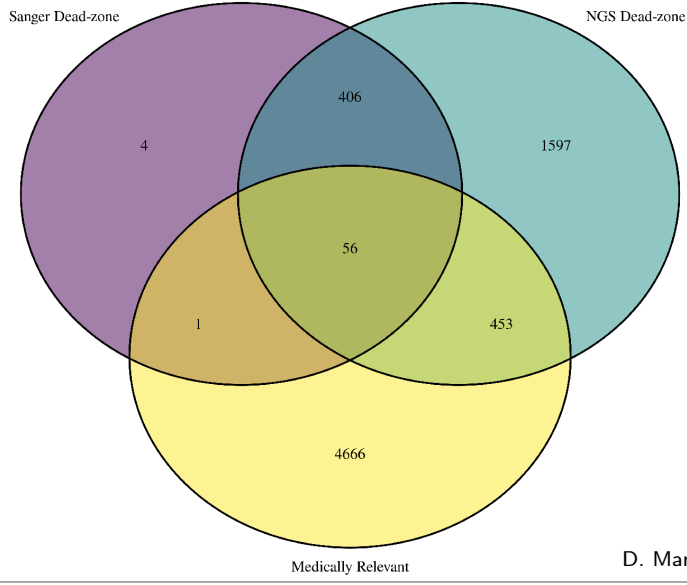


Two major categories:

- Sanger deadzone.
- NGS deadzone.

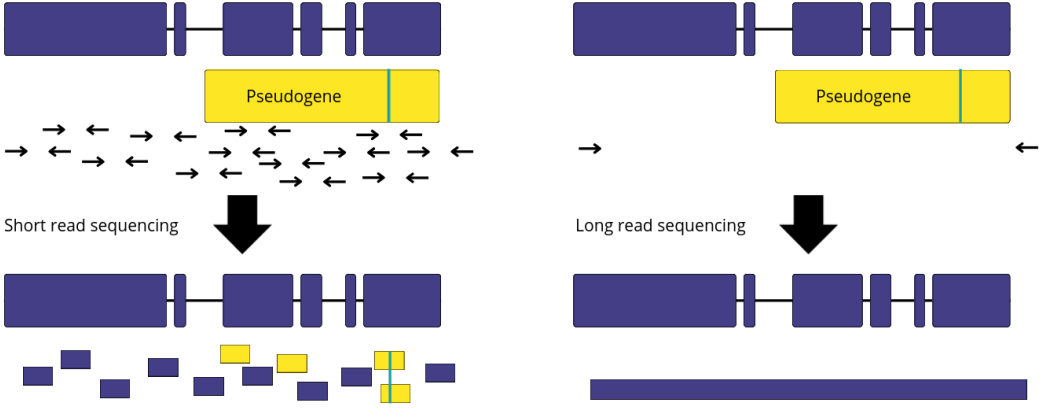
What is a Challenging Medically Relevant Gene?

Challenging Medically Relevant Genes (CMRG)



D. Mandelker et al. (2016)

Example: Gene with pseudogene



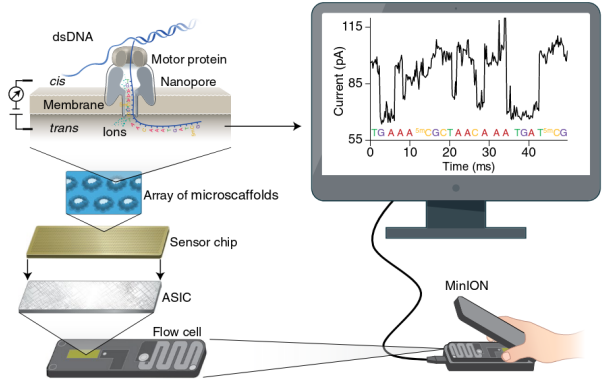
Oxford Nanopore Technologies sequencing

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)

My research goals

- Improve our knowledge about the sequence structure of these challenging genes.
- Understand how can we overcome the NGS limitation, mostly using long-read sequencing.
- Fast and cost-effective methods of genetic diagnosis.

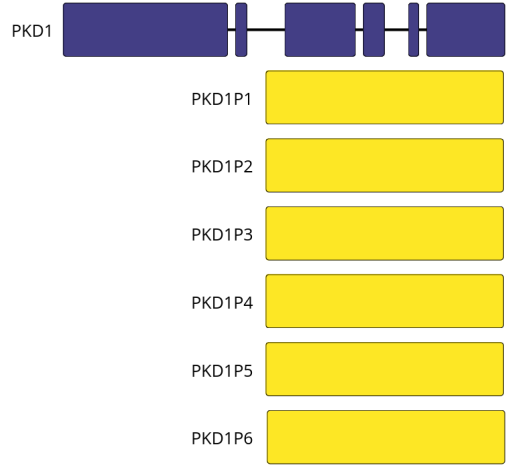
Genes of interest

1. *PKD1*: Autosomal Dominant Polycystic Kidney Disease (ADPKD).
2. *FLG*: Atopic Dermatitis (AD).

Genes of interest

1. *PKD1*: Autosomal Dominant Polycystic Kidney Disease (ADPKD).
2. *FLG*: Atopic Dermatitis (AD).
3. *MUC1*: Autosomal Dominant Tubulointerstitial Kidney Disease (ADTKD).
4. *SMN1*: Spinal Muscular Atrophy (SMA).

The *PKD1* case



Responsible for 75% of Autosomal Dominant Polycystic Kidney Disease.

6 pseudogenes that are highly similar

Part of the NGS deadzone.

PKD1 strategy

Gold standard: Sanger sequencing.

Dataset of 34 patients with 56 pathogenic variants, sequenced both with:

- Whole Exome Sequencing with Illumina.
- Amplicon sequencing with Oxford Nanopore Technologies.

PKD1 strategy

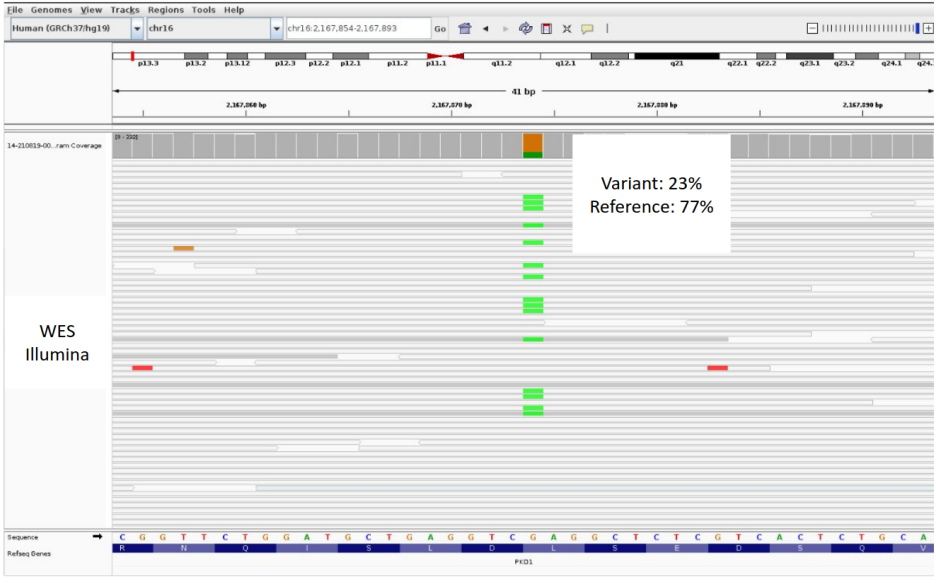
Gold standard: Sanger sequencing.

Dataset of 34 patients with 56 pathogenic variants, sequenced both with:

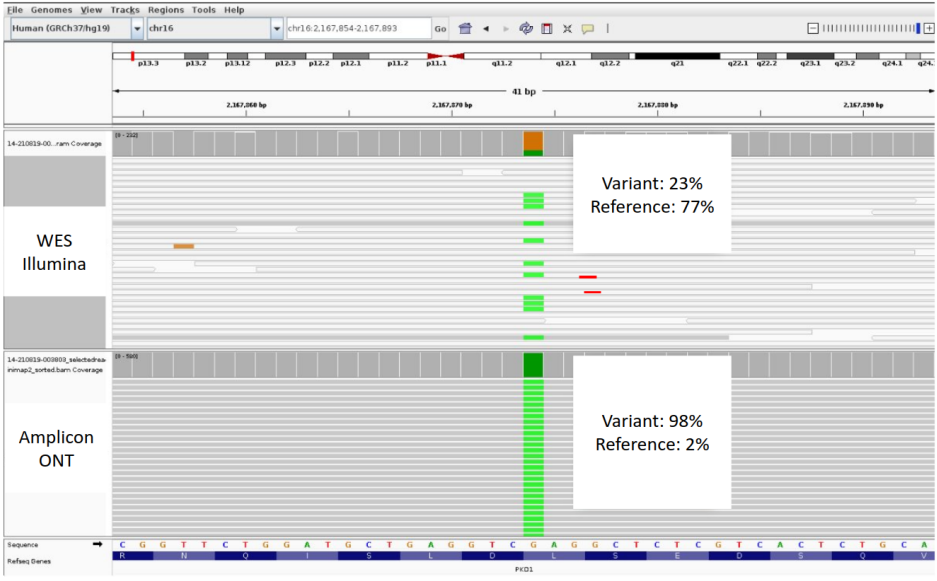
- Whole Exome Sequencing with Illumina.
- Amplicon sequencing with Oxford Nanopore Technologies.

All variants have been found with *both* techniques.

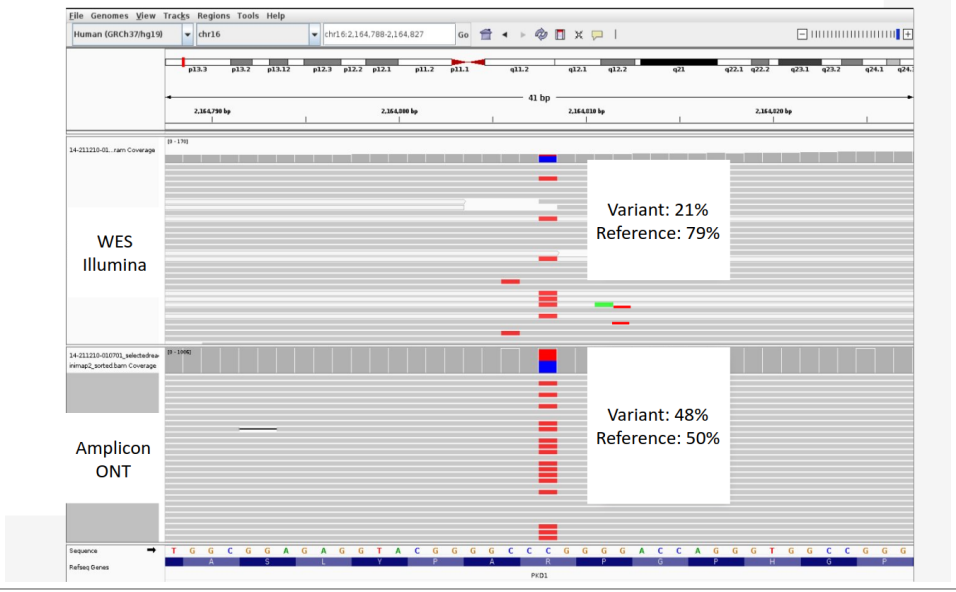
Unbalanced allele frequency in WES



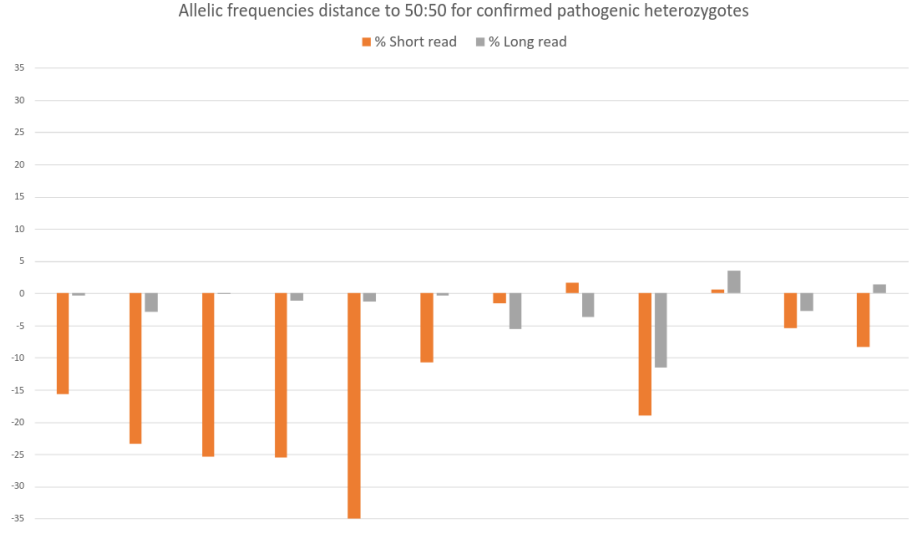
Allele frequency resolved with long-read



Another example

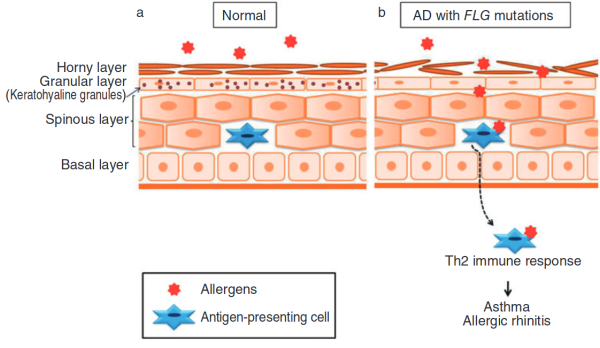


Long-read improves allele frequency



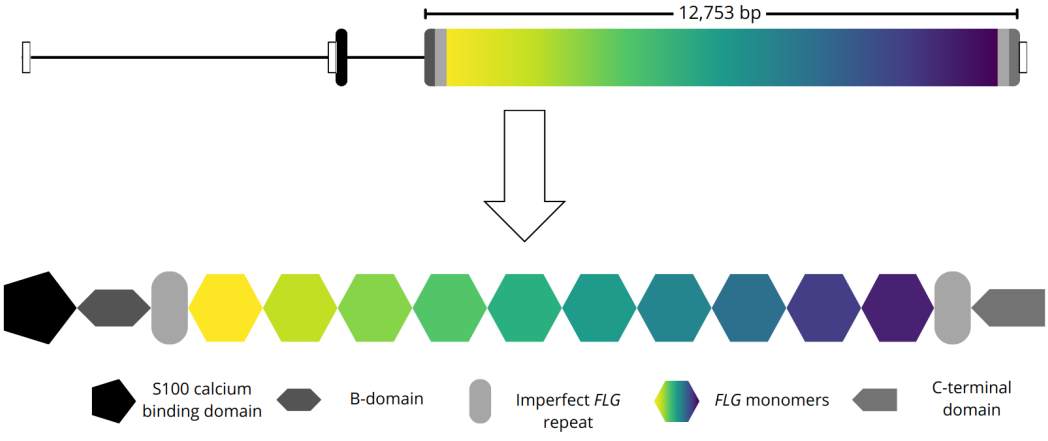
Atopic Dermatitis (AD)

J. Heilman (2010)

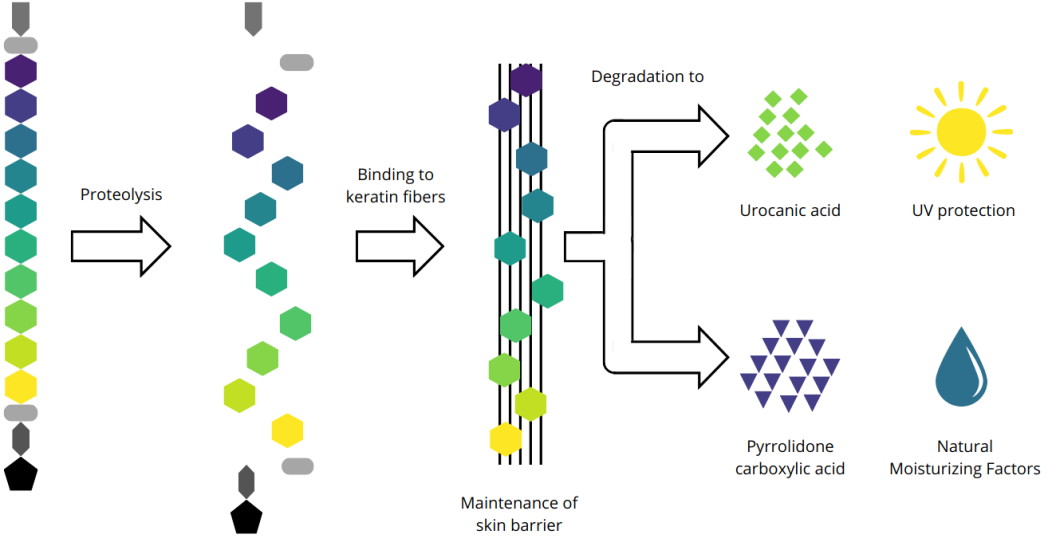


R. Osawa, M. Akiyama and H. Shimizu (2011)

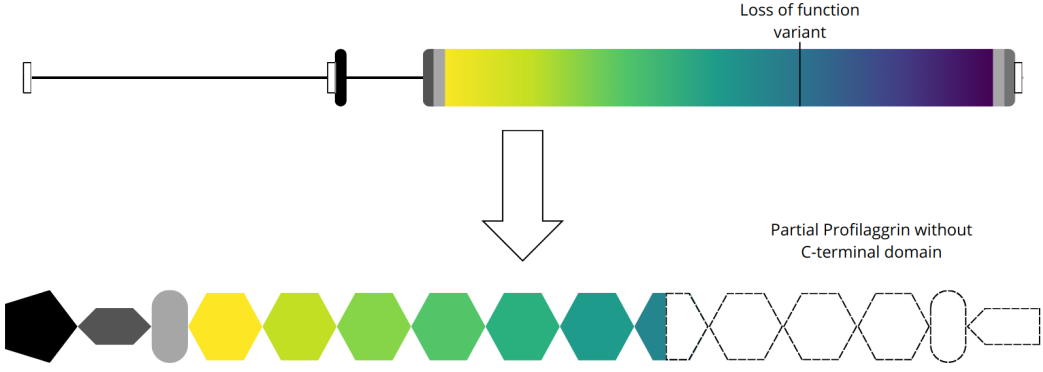
The Filaggrin gene (*FLG*)



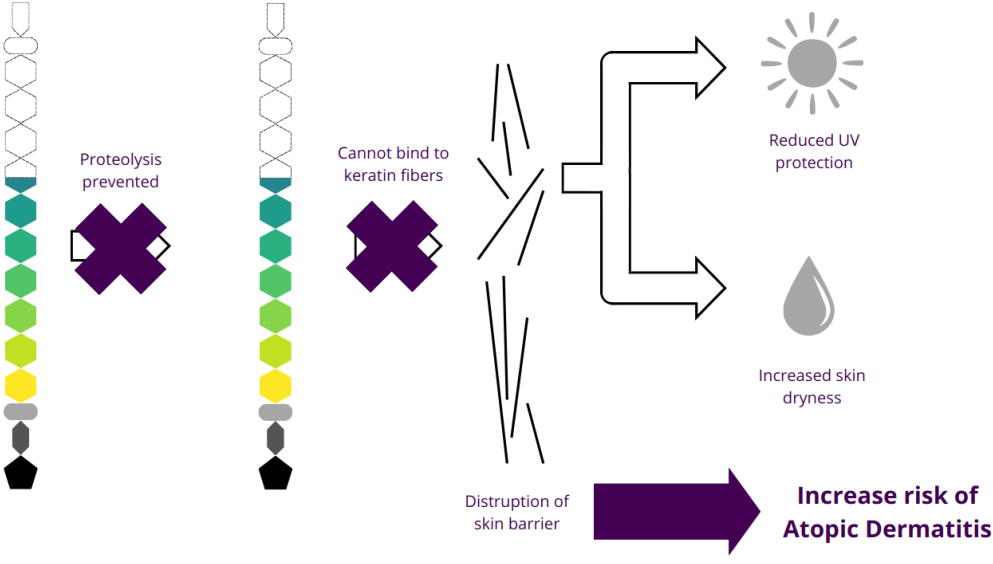
Role of *FLG*


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*

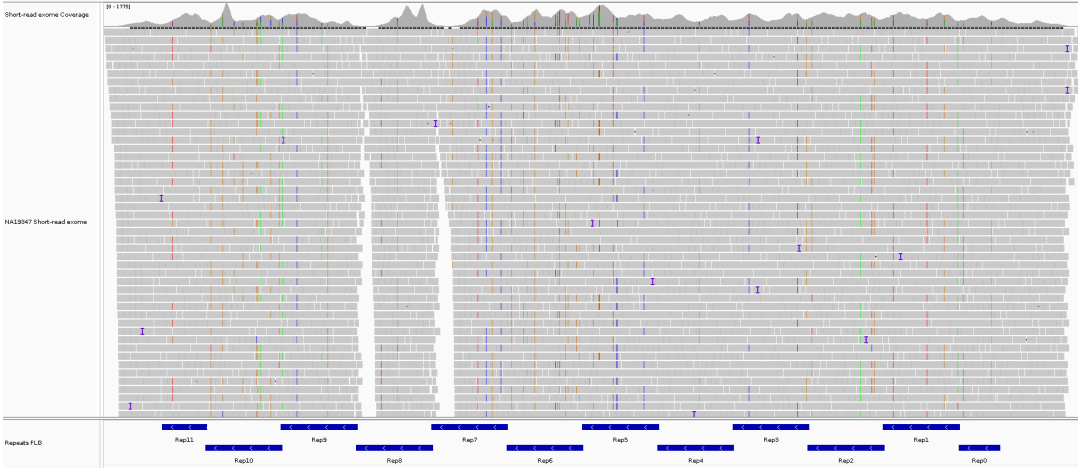


LOF variants prevents proteolysis

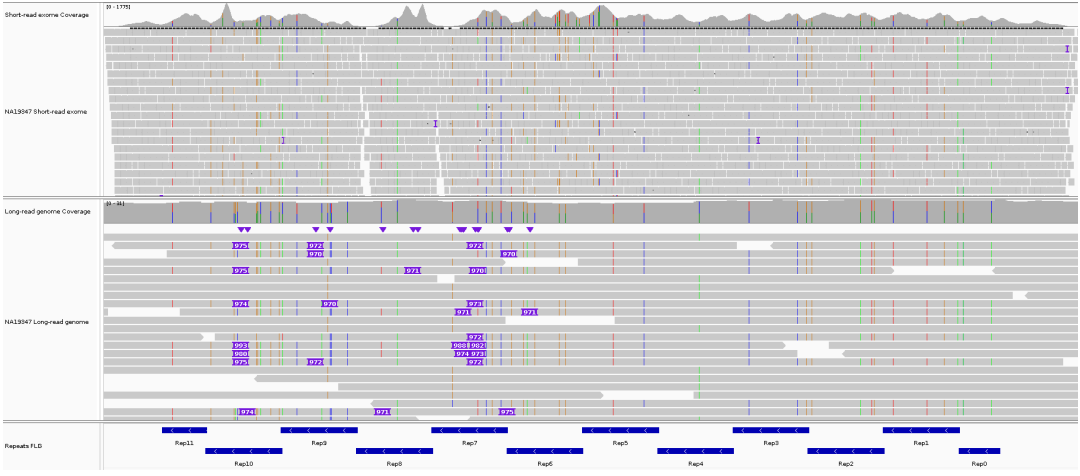


Multiple known alleles of *FLG*

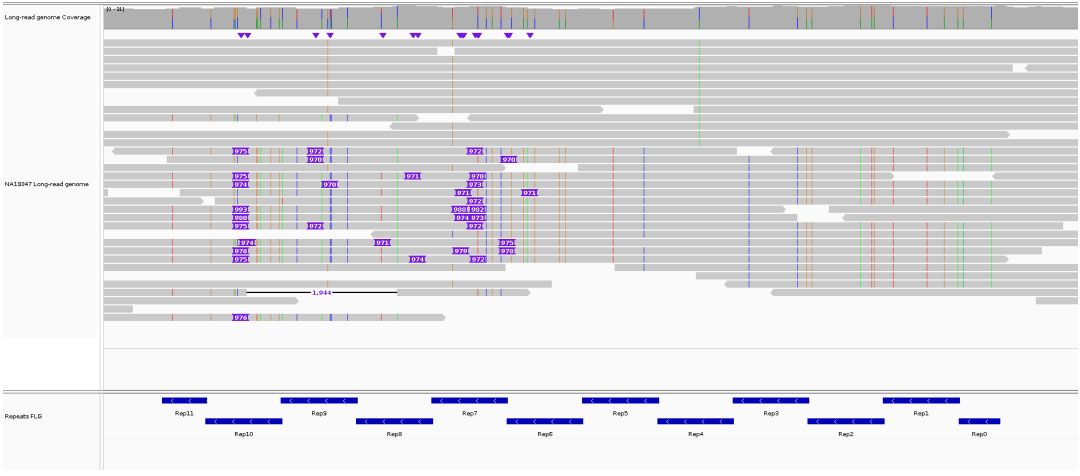
Short read sequencing limitation



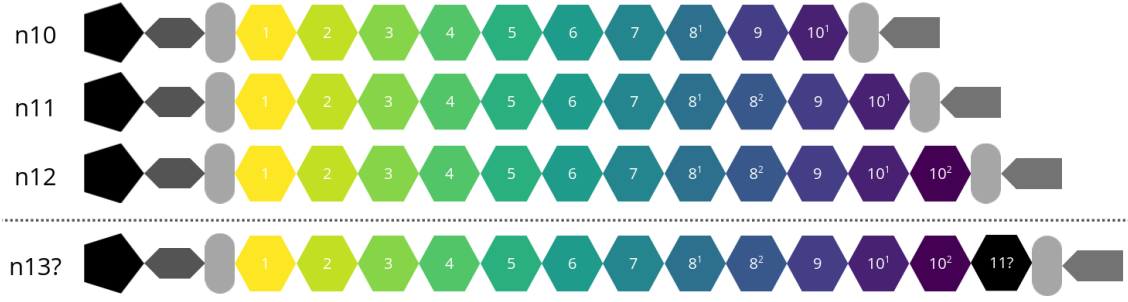
Short read vs Long read



Unusual number of *FLG* repeats



Undescribed *FLG* allele(s)?



Multi-ethnic public data cohort

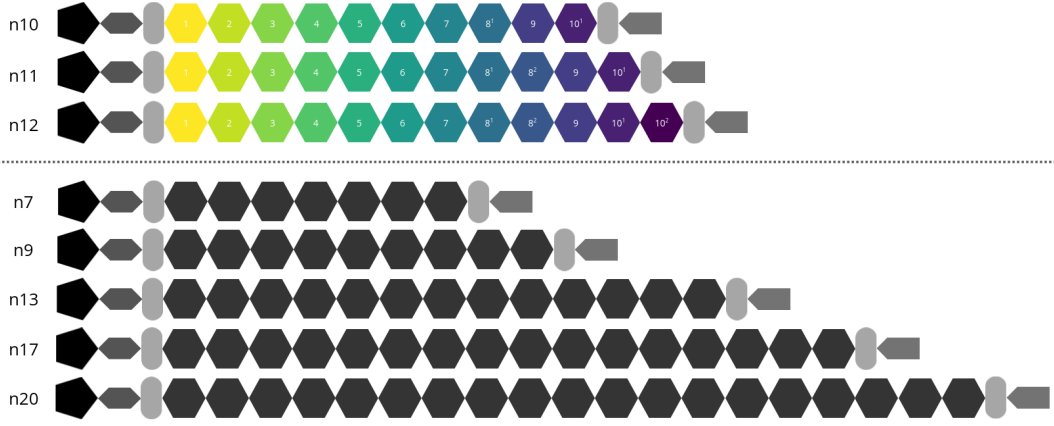
Origin	Number of samples
Africa	316
Europe	200
East Asia	213
South Asia	223
Latin America	182
Total	1134

Quality Control of Public Data

Multi-ethnic cohort after Quality Control

Origin	Number of samples	After QC
Africa	316	279
Europe	200	185
East Asia	213	193
South Asia	223	191
Latin America	182	163
Total	1134	1011

Catalog of novel alleles

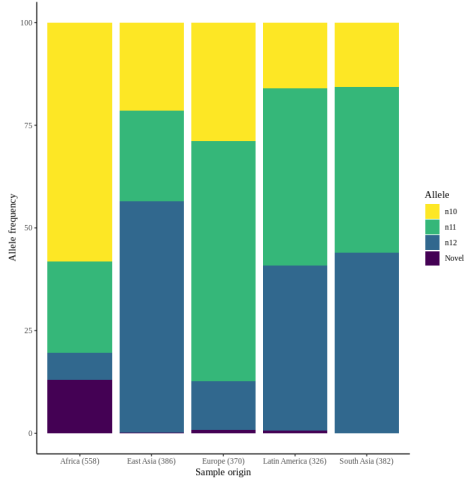


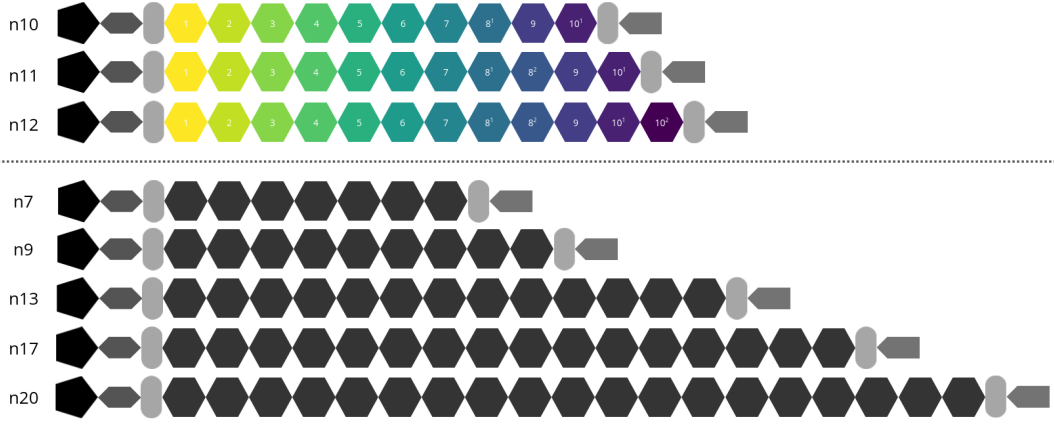
Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	**325**	107	83	60	52
n11	124	**217**	85	154	**141**
n12	36	44	***217***	**168**	131
n7	0	0	1	0	0
n9	45	1	0	0	0
n13	28	1	0	0	1
n17	0	1	0	0	0
n20	0	0	0	0	1

Frequencies of novel alleles

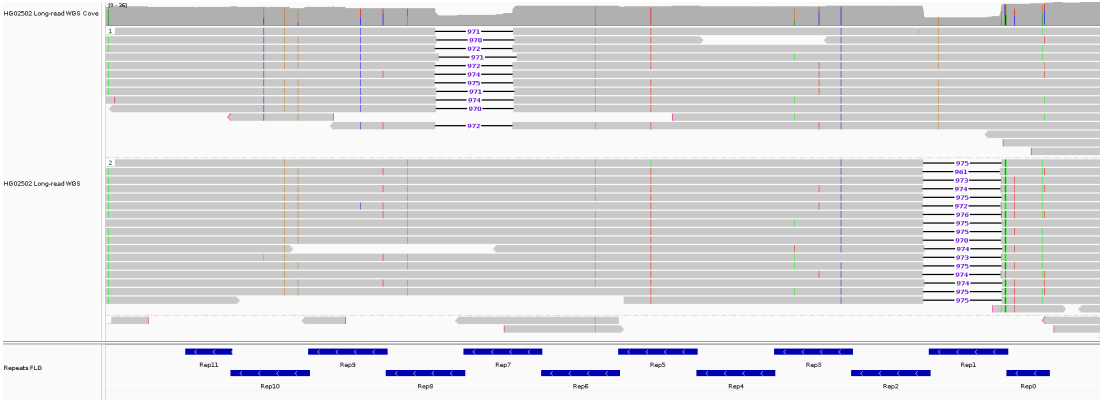
Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	**325**	107	83	60	52
n11	124	**217**	85	154	**141**
n12	36	44	***217***	**168**	131
n7	0	0	1	0	0
n9	45	1	0	0	0
n13	28	1	0	0	1
n17	0	1	0	0	0
n20	0	0	0	0	1



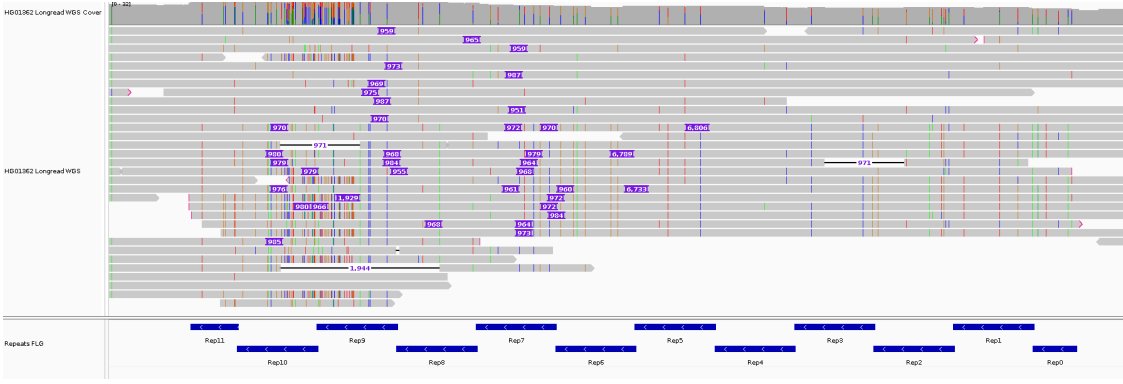

n10
n11
n12
n7
n9
n13
n17
n20

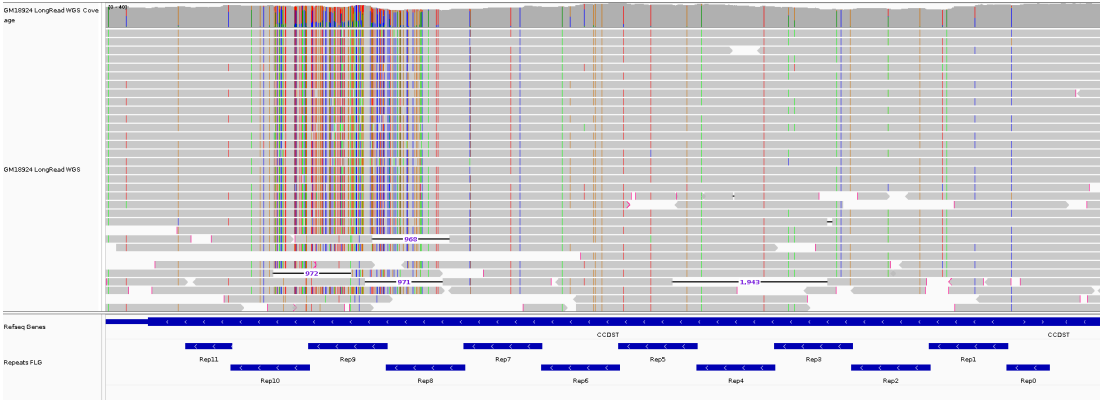

n7
n9
n13
n17
n20

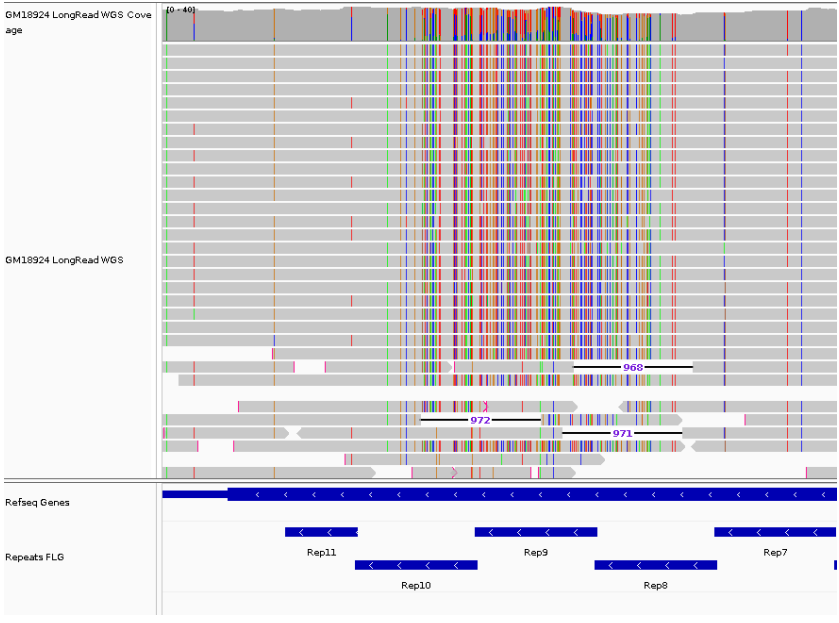
Alleles with deletions

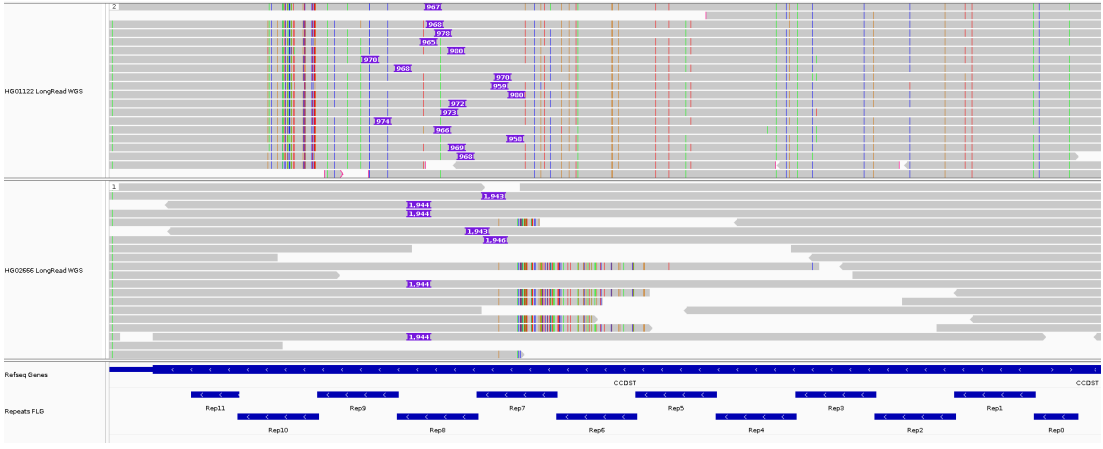


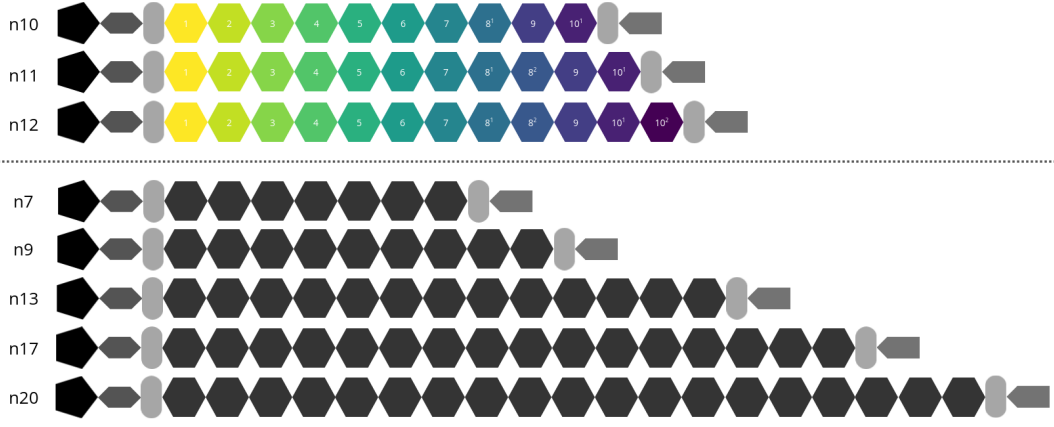
Alleles with insertions

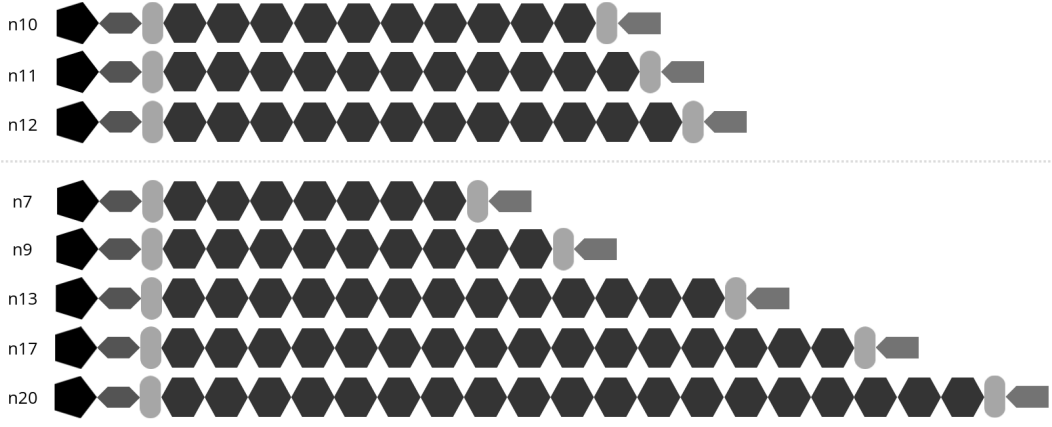



GM18924 LongRead WGS Coverage
(0 - 40)
GM18924 LongRead WGS
960
972
971
1,943
Refseq Genes
CCDST
CCDST
Repeats FLG
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0


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

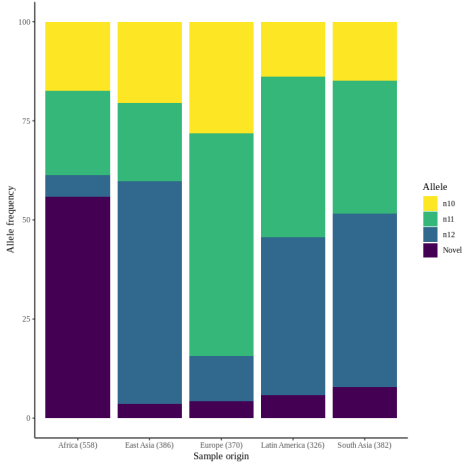

HG01122 LongRead WGS
HG02666 LongRead WGS
Refseq Genes
Repeats FLG
CCDST
CCDST
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0


n10
n11
n12
n7
n9
n13
n17
n20


n10
n11
n12
n7
n9
n13
n17
n20

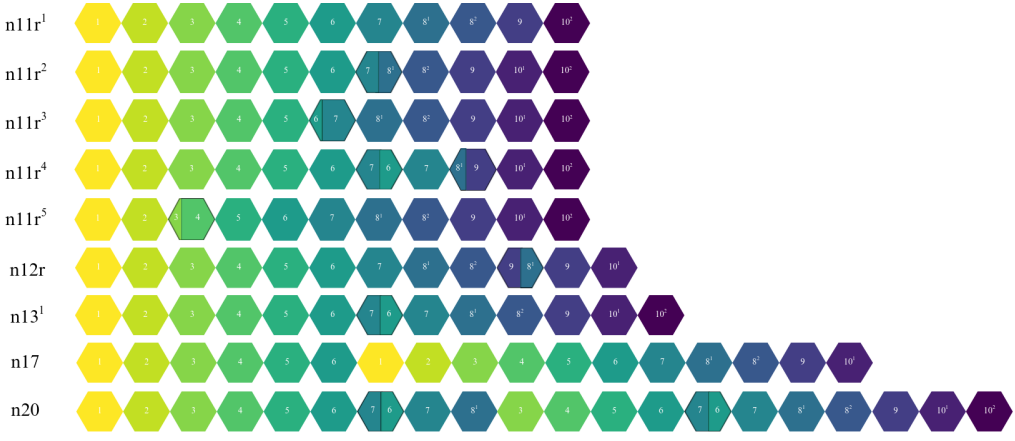
Frequencies of novel alleles

Allele	Africa	Europe	East Asia	South Asia	Latin America
n10	97	104	79	57	45
n11	119	**208**	76	128	**132**
n12	30	42	***217***	**167**	130
n7	0	0	1	0	0
n9	45	1	0	0	0
n10	***228***	3	4	3	7
n11	5	9	9	26	9
n12	6	2	0	1	1
n13	28	0	0	0	1
n17	0	1	0	0	0
n20	0	0	0	0	1



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.

MUC1 and *SMN1/SMN2*

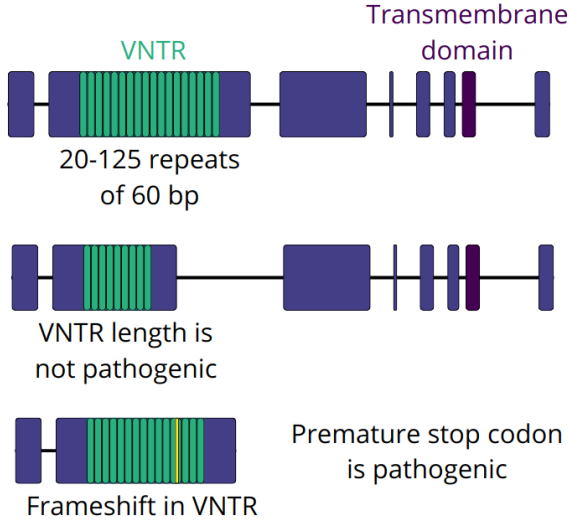
MUC1:

- Autosomal Dominant Tubulointerstitial Kidney Disease (ADTKD)
- Challenge: Variable Number Tandem Repeat (VNTR)

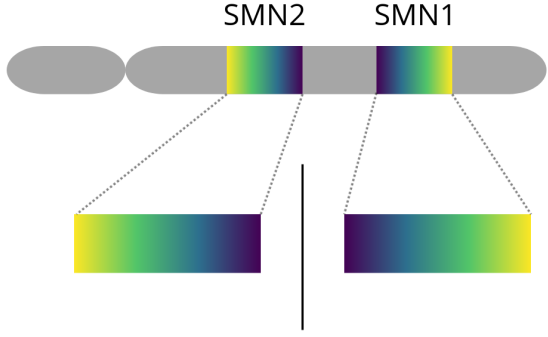
SMN1:

- Spinal Muscular Atrophy (SMA)
- Challenge: Gene duplicated nearby (*SMN2*)

MUC1: Variable Number Tandem Repeat



SMN1 and *SMN2*



100 WGS ONT samples data analysis

Gene	Samples with X>2	Consensus
MUC1	100	136
SMN1	42	10
SMN2	32	8

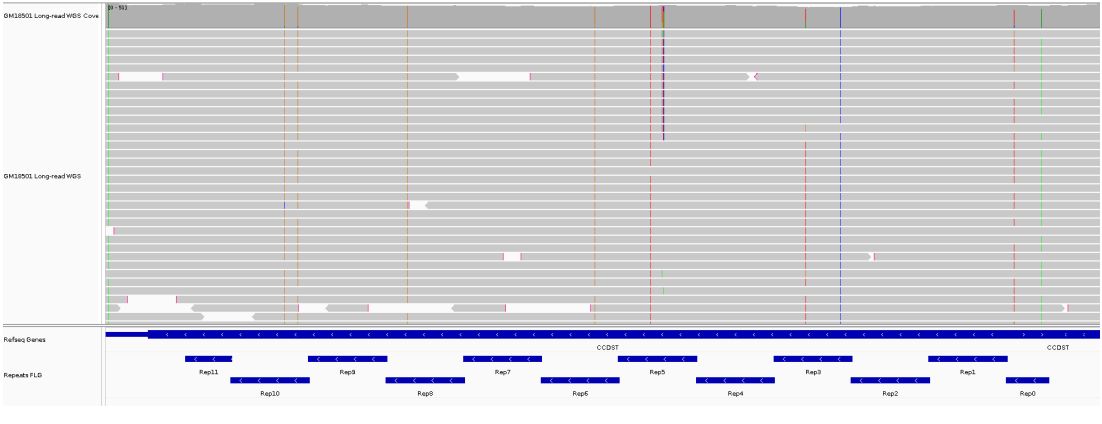
Conclusion

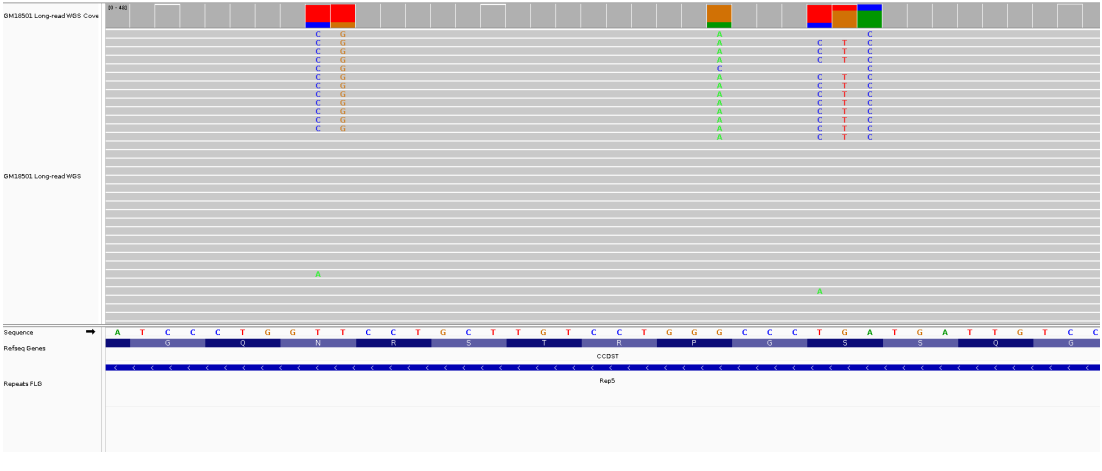
- Long-read sequencing can improve the characterization of CMRG, and even be used in a clinical setting.
- Novel *FLG* alleles have been identified, mostly in African and African-descent populations.
- However, some regions are still difficult to characterize.
  - Adaptive sampling or amplicon sequencing could help.

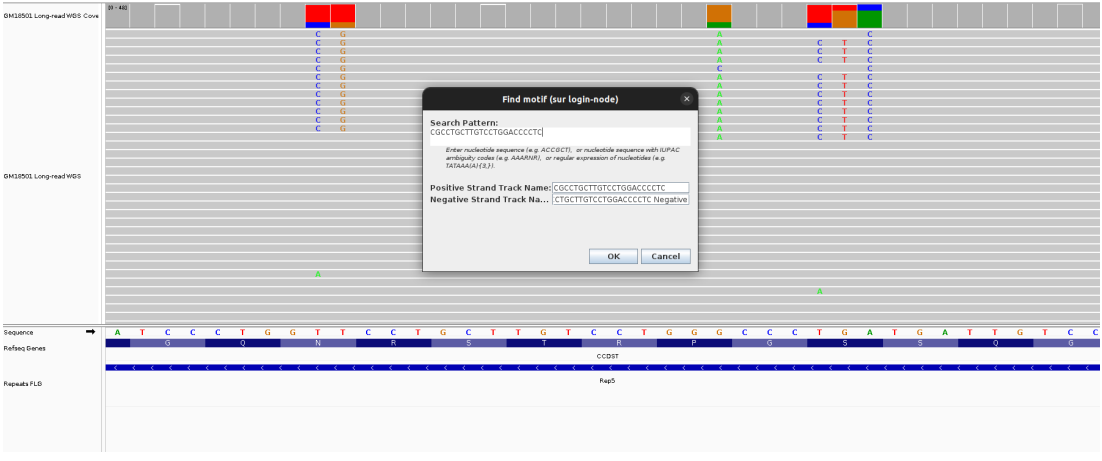
Acknowledgement

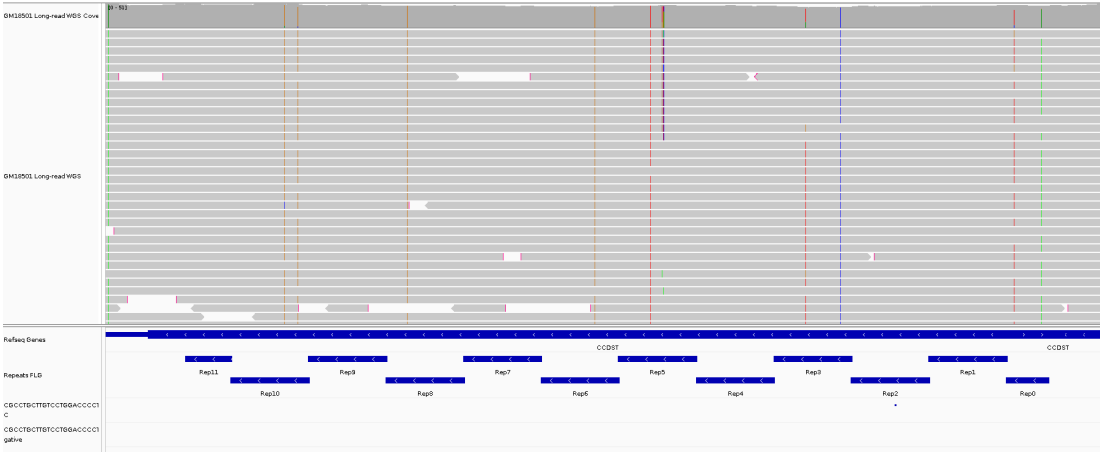
Pr. Vincent Bours
Dr. Keith Durkin
Dr. Leonor Palmeira
Dr. Maria Artesi
Dr. Benoit Charloteaux
Dr. Laura Helou
Pr. Vinciane Dideberg
Nadine Cambisano
Nathalie Renotte
Dr. Sabine Olivier
Dr. Annette Uwineza
Dr. Alice Amani Uwajeni
Pr. Alan Irvine

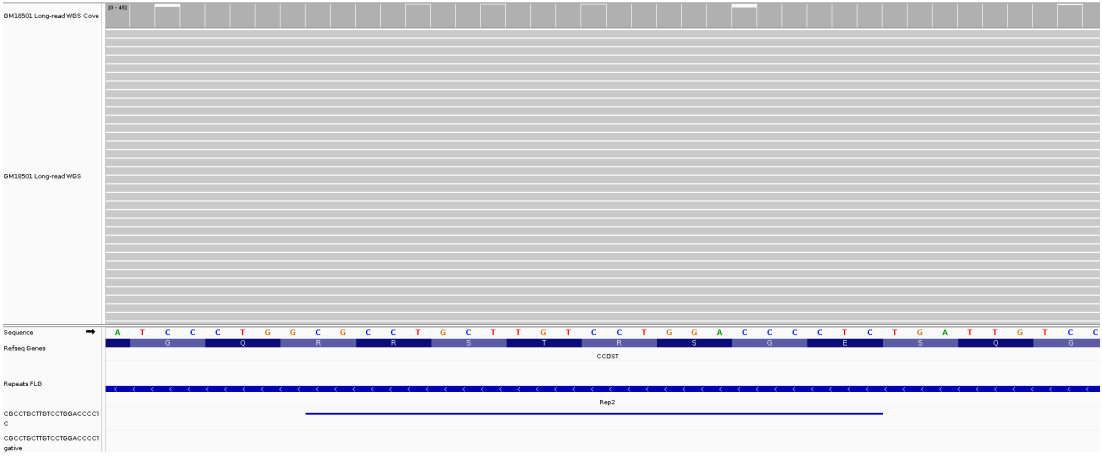
Contact me!

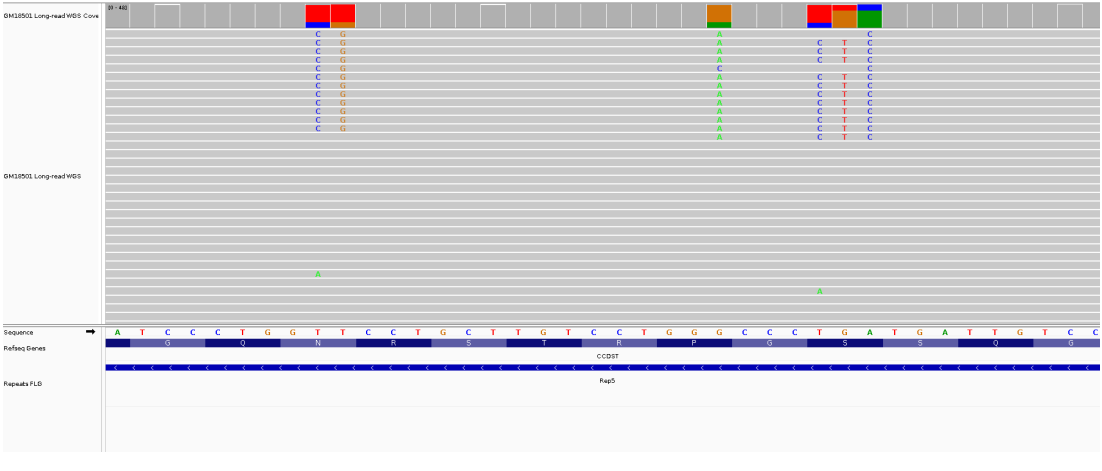

GM18501 Long-read WGS Cove
0 - 51
GM18501 Long-read WGS
Refseq Genes
CCDST
CCDST
Repeats FLG
Rep11
Rep9
Rep7
Rep5
Rep3
Rep1
Rep10
Rep8
Rep6
Rep4
Rep2
Rep0


GM18501 Long-read WGS Cove
[0 - 44]
GM18501 Long-read WGS
Sequence
A T C C C T G G T T C C T G C T T G T C C T G G G C C C T G A T G A T T G T C C
Refseq Genes
G Q N R S T R P G S S Q G
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){8,}).
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
0 - 513
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
Refseq Genes
G Q R R S T R S G E S Q G
CCDST
Repeats FLB
Rep2
CGCCTGCTTGTCCTGGACCCC1
C
CGCCTGCTTGTCCTGGACCCC1
gative


GM18501 Long-read WGS Cove
[0 - 44]
GM18501 Long-read WGS
Sequence
A T C C C T G G T T C C T G C T T G T C C T G G G C C C T G A T G A T T G T C C
Refseq Genes
G Q N R S T R P G S S Q G
CCDST
Repeats FLD
Rep5