A Needle in a Haystack: Improving Genetic Analysis of Challenging Medically Relevant *MUC1* Gene

Victor Grentzinger[1,2], Leonor Palmeira[2], Keith Durkin[1,2], Maria Artesi[1,2], Benoit Charloteaux[2], Vinciane Dideberg[2], Vincent Bours[1,2]

[1] Laboratory of Human Genetics, GIGA Research Institute, Liège, Belgium

[2] Center of Genetics, University Hospital, Liège, Belgium

Background

Autosomal Dominant Tubulointerstitial Kidney Disease (ADTKD) is a genetic disease with a prevalence of 0.7 to 4 per million. Multiple genes can induce this disease, one of them, *MUC1*, contains a Variable Number Tandem Repeat (VNTR) region of a 60-mer repeat. A known frameshift variant in this region leads to a truncated, misfolded Mucin-1 protein, whose accumulation leads to ADTKD[1]. This VNTR region is complex to characterize by short-read sequencing, and the usual method used for genetic diagnosis is to screen exclusively for this known variant. However, other potential pathogenic variants may be overlooked.

Methods

100 publicly available samples[2] analyzed with our in-house pipeline:

- Aligned to human genome reference hg19 and hg38.
- Haplotypes phazed based on small variants called on aligned data.
- Consensus sequences assembled for each haplotype.

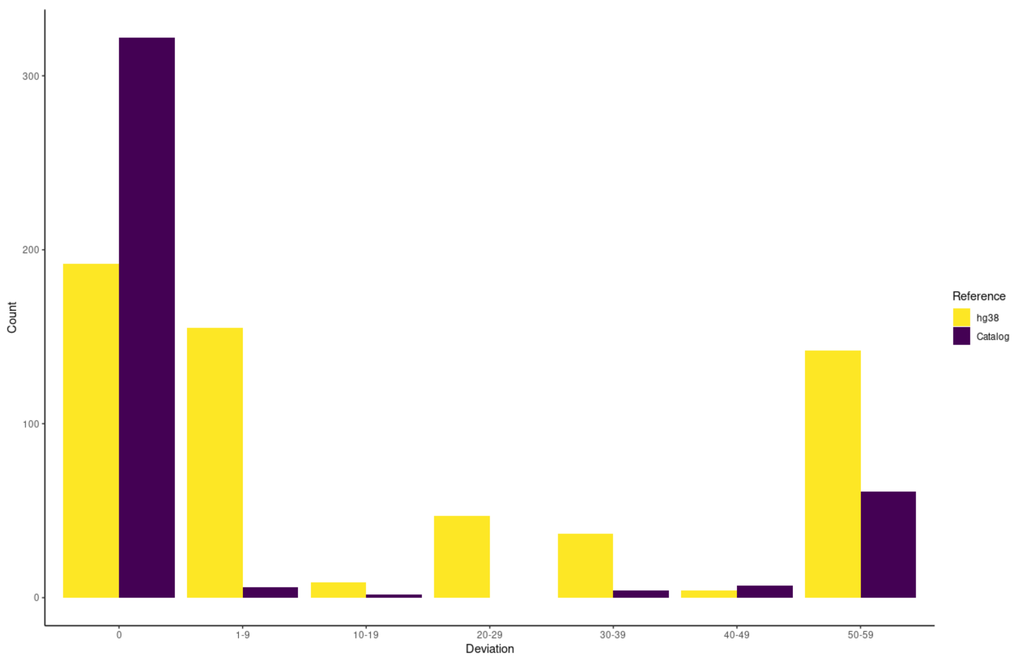
We re-aligned our data to 10 consensus sequences with different numbers of repetition obtained to see if we could better characterize the alleles for each sample.

Results

68% of haplotypes have been assembled into a consensus sequence covering the whole *MUC1* gene.

Samples	100
Haplotypes	200
Consensus sequences	199
Consensus sequences after QC	136
Samples with both consensus sequences after QC	**45**

Calling CNVs on reads mapped to our *MUC1* consensus catalog **improves accuracy** of repeat length.



Relevance

- Mapping long-reads to a *MUC1* consensus catalog improves the characterization of the number of repeats.
- Another bioinformatics tool uses an mapping-free genotyping of *MUC1* short-read data to detect frameshifts due to insertion or deletion within the VNTR region[3].
- However, this method cannot detect LOF variants originating from SNPs.
- A continuous improvement of our *MUC1* consensus catalog would help to pinpoint small variants within the homologous region of this gene, and may lead to new pathogenic variants discovery.

References

1: Kirby A, Gnirke A, Jaffe DB, Barešová V, Pochet N, Blumenstiel B, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat Genet 2013;45:299–303

2: Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, et al. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. Genome Res 2024;34:2061–73

3: Saei H, Morinière V, Heidet L, Gribouval O, Lebbah S, Tores F, et al. VNtyper enables accurate alignment-free genotyping of MUC1 coding VNTR using short-read sequencing data in autosomal dominant tubulointerstitial kidney disease. iScience 2023;26:107171

Contact me!

LIÈGE université
GIGA institute

CHU de Liège