Genotyping coding-VNTR in the *MUC1* gene using alignment-free method allows genetic diagnosis of *MUC1*-related autosomal dominant tubulointerstitial kidney disease

Hassan **Saei**[1], Vincent **Morinière**[2], Laurence **Heidet**[1,3], Olivier **Gribouval**[1], Said **Lebbah**[2], Frederic **Tores**[4], Bertrand **Knebelmann**[5], Manon **Mautret-Godefroy**[2], Stephane **Burtey**[6], Vincent **Vuiblet**[7], Corinne **Antignac**[1,2], Patrick **Nitschké**[4], Guillaume **Dorval**[1,2]

[1] Labratory of hereditary kidney diseases, Inserm UMR 1163, Institut Imagine, Université Paris Cité, Paris, France
[2] Medical genomics services for rare diseases, Hôpital Necker-Enfants Malades, Assistance publique, Hôpitaux de Paris (AP-HP), Paris, France
[3] Pediatric Nephrology Service, Centre de Référence MARHEA, Hôpital Necker-Enfants Malades, Assistance publique, Hôpitaux de Paris (AP-HP), Paris, France
[4] Bioinformatic Platform, Inserm UMR 1163, Institut Imagine, Université Paris Cité, Paris, France
[5] Nephrology Service, Centre de Référence MARHEA, Hôpital Necker-Enfants Malades, Assistance publique, Hôpitaux de Paris (AP-HP), Paris, France
[6] INSERM, C2VN, INRAE, C2VN, Aix-Marseille Université, Marseille, France
[7] Nephrology Service, CHU de Reims, Reims, France

iScience

CellPress OPEN ACCESS

Article

VNtyper enables accurate alignment-free genotyping of *MUC1* coding VNTR using short-read sequencing data in autosomal dominant tubulointerstitial kidney disease

Scan to read

INTRODUCTION

- Human genome comprises 3% of tandem repeats with variable length, a few of which have been linked to human rare diseases.
- Genotyping VNTRs using short-read sequencing data is challenging due to the poor read mappability.
- Autosomal dominant tubulointerstitial kidney disease-*MUC1* is caused by specific frameshift variants in the coding VNTR of the *MUC1* gene[1].
- *MUC1* encodes mucin-1 protein which is the main component of the mucus expressed in the distal tubules and collecting ducts of the nephrons.

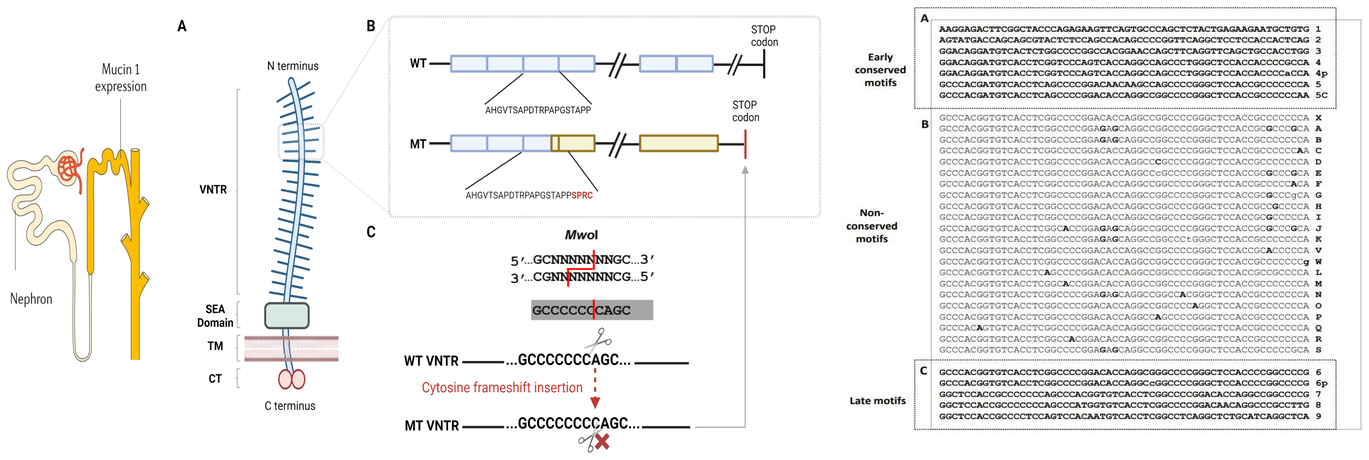

Figure 1. *Illustrates the structural domains of Mucin-1 and highlights the recurrent MUC1 dupC variation hotspot. The MUC1 coding VNTR consists of 34 motifs, each composed of 60-mers, which vary in terms of composition and repetition among individuals.*

AIM

To enhance the genetic diagnosis and detection rate of ADTKD-*MUC1* by implementing standard short-read sequencing technology.

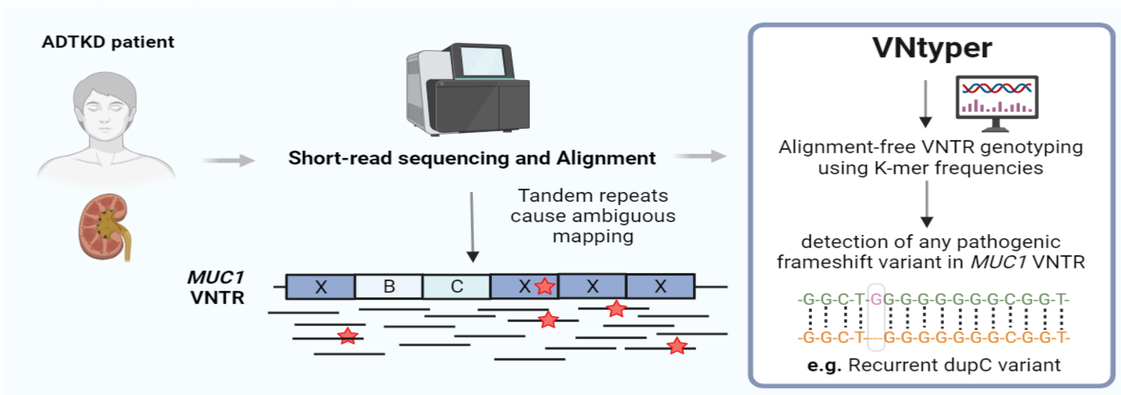

Figure 2. *Short reads originating from the VNTR region may map to multiple motifs and are filtered out during the variant calling step. To address this issue, we have developed a pipeline called VNtyper which uses K-mer frequencies for genotyping.*

METHOD AND STUDY DESIGN

Implementation of a *MUC1*-VNTR-specific motif dictionary

- Thirty-four unique 60-mer motifs exist in the *MUC1* VNTR region. We've created a comprehensive 120-mer motif dictionary that includes all potential motif variations, categorizing them by source and sequence order. This dictionary will be used as a reference for genotyping.

Designing VNtyper pipeline

- We applied the Kestrel mapping-free genotyping algorithm[2], originally designed for genotyping penicillin binding protein (PBP) genes in Streptococcus pneumoniae in 2018, with optimized parameters to accommodate our case-specific reference file.
- We introduced a Python tool called VNtyper, tailored for genotyping, filtering, and prioritizing pathogenic frameshift variations within the coding-VNTR of the *MUC1* gene.
- Furthermore, our analysis incorporates the newly developed code-adVNTR[3] method based on profile-HMMs for comparative assessment.

HISTORICAL COHORT

We have used the cohort of **237 individuals** to test our pipeline[4]. In this cohort we had 118 individuals (94 symptomatic) positive for the *MUC1* VNTR pathogenic variation.

RENOME COHORT

We used our cohort of **2910 patients** with renal symptoms, studied from 2017 to 2022 with NGS each assigned to a group of hereditary renal disease. This cohort was used to study the specificity of our tool.

HOW VNtyper WORKS?

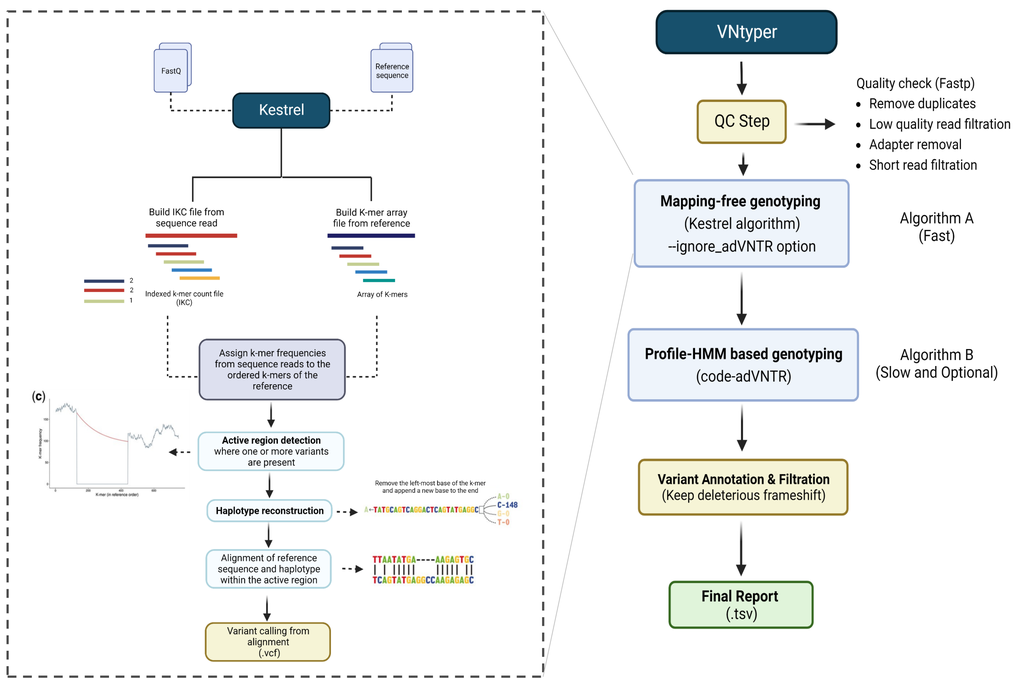

Figure 3. *Schematic representation of the VNtyper pipeline, featuring two integrated genotyping algorithms.*

UTILITY ANALYSIS – HISTORICAL COHORT

- A cohort comprising **237** *MUC1* positive and negative individuals, was employed to evaluate the pipeline.
- We have computed a depth-dependent score and threshold to distinguish between true positives and false positives.

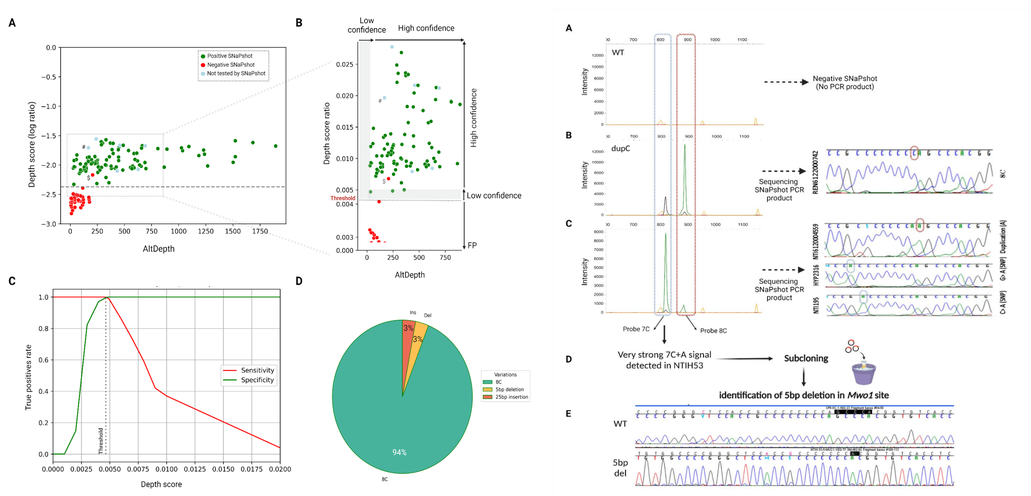

Figure 4. *Historical cohort characterization. Left panel explains the depth-score adapted threshold which separates true positives from false positives. Right panel, illustrates the SNaPshot results for validated variations.*

VALIDATION – RENOME COHORT

- The second cohort, consisting of **2,910 patients** with renal symptoms, was utilized to assess the pipeline's specificity.
- Our method identified **30** previously overlooked patients with renal symptoms, leading to their diagnosis.

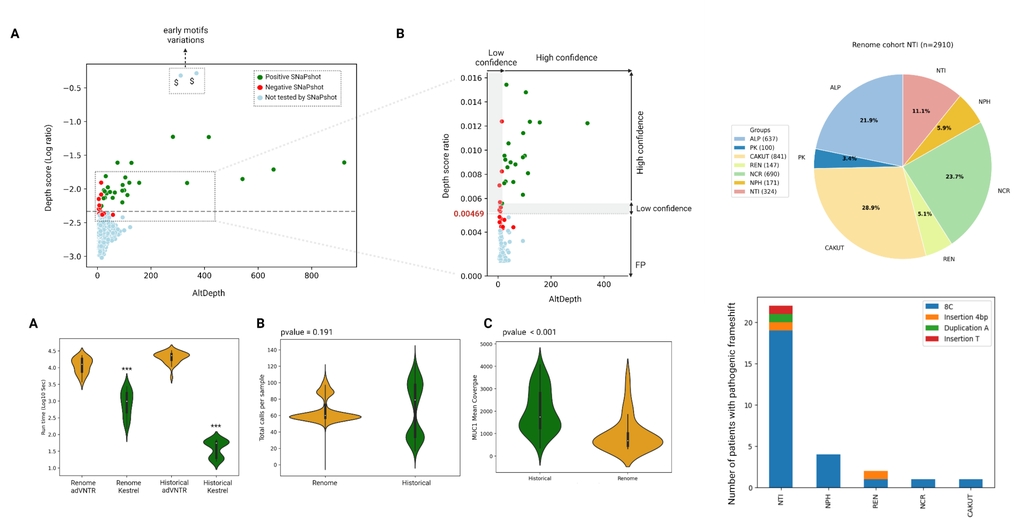

Figure 5. *Renome Cohort characterization: Upper left validates the depth-score threshold for TP and FP differentiation. Upper right, shows patient groups. Lower left, confirms our method's faster genotyping compared to code-adVNTR. Stacked plot highlights newly diagnosed patients in the renome cohort.*

REFERENCES

[1] **Kirby et al**. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. **Nature Genetics**, 2013.

[2] **Audano et al**. Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. **Bioinformatics**, 2018.

[3] **Park et al**. Detecting tandem repeat variants in coding regions using code-adVNTR. **iScience**, 2022.

[4] **Saei 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.

ABOUT THE AUTHOR

Hassan Saei

- PhD Candidate (PPU-Imagine International Doctoral Program scholar)
- I am passionate about leveraging computational methods to enhance genetic diagnosis. My profound interest lies in the development of disease models, such as organoids, and the application of genome editing techniques to delve into disease pathobiology and advance therapeutic solutions.

LET'S CONNECT

@HassanSaei

Hassan.saei@inserm.fr
Guillaume.dorval@aphp.fr

Paris, France

Inserm

Université Paris Cité