iScience, Volume *26*

Supplemental information

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

Hassan Saei, Vincent Morinière, Laurence Heidet, Olivier Gribouval, Said Lebbah, Frederic Tores, Manon Mautret-Godefroy, Bertrand Knebelmann, Stéphane Burtey, Vincent Vuiblet, Corinne Antignac, Patrick Nitschké, and Guillaume Dorval

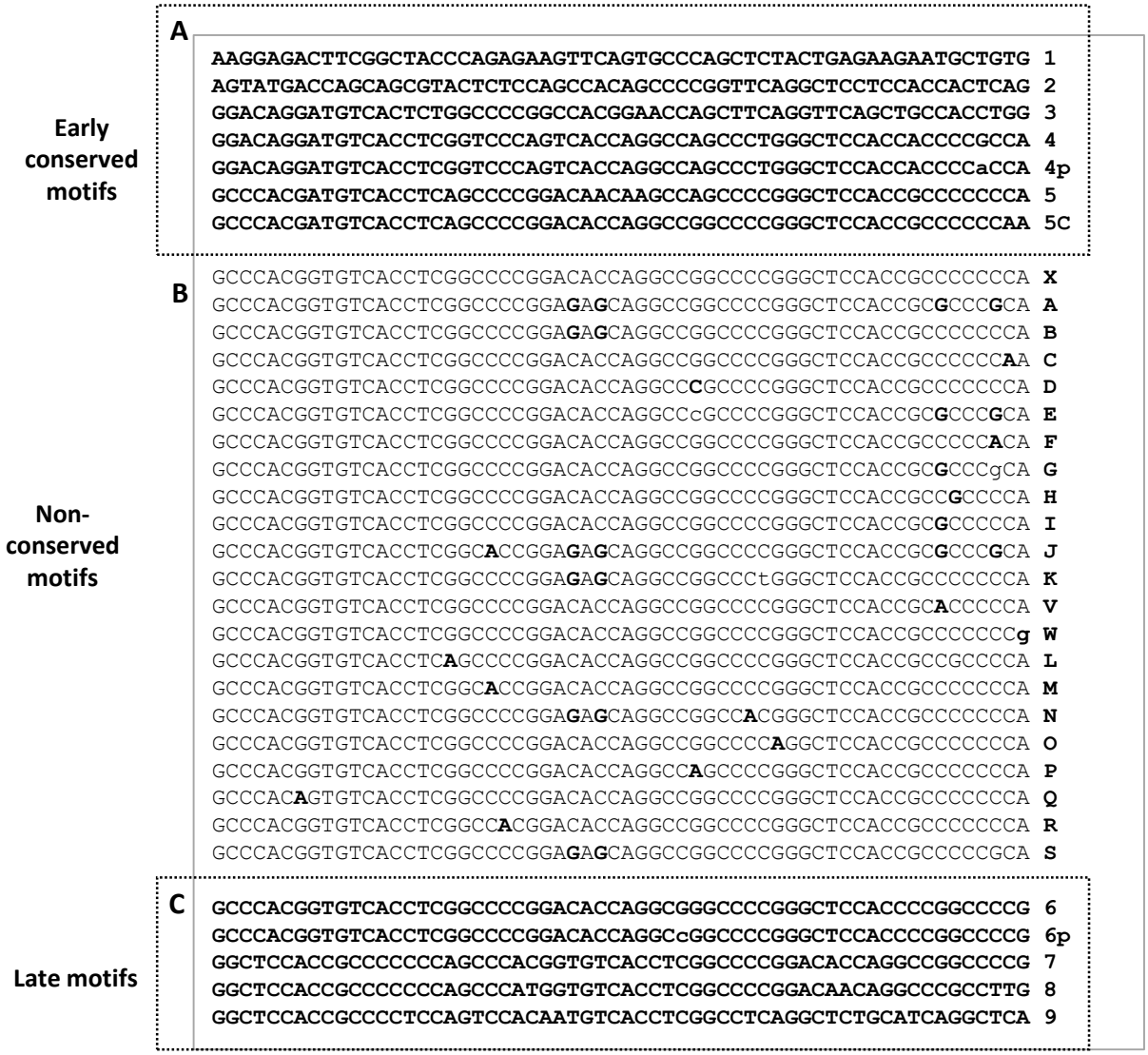

Figure S1. VNTR motifs and their sequences in *MUC1*, related to STAR methods. Thirty-four motifs of 60-mer units within exon 2 of the *MUC1* gene have been identified. **A-C)** The hypervariable non-conserved region of the VNTR is flanked by distinct conserved 60-mer units: motifs 1-5 including 4' upstream and motifs 6-9 including 6' downstream. To construct a 120-mer motif dictionary that will serve as the *MUC1* reference sequence, all conceivable motif re-ordering of the hypervariable zone and the neighbouring early and late conserved motifs were considered (n=558). The coding strand is on the reverse strand of the hg19 assembly of the human genome.

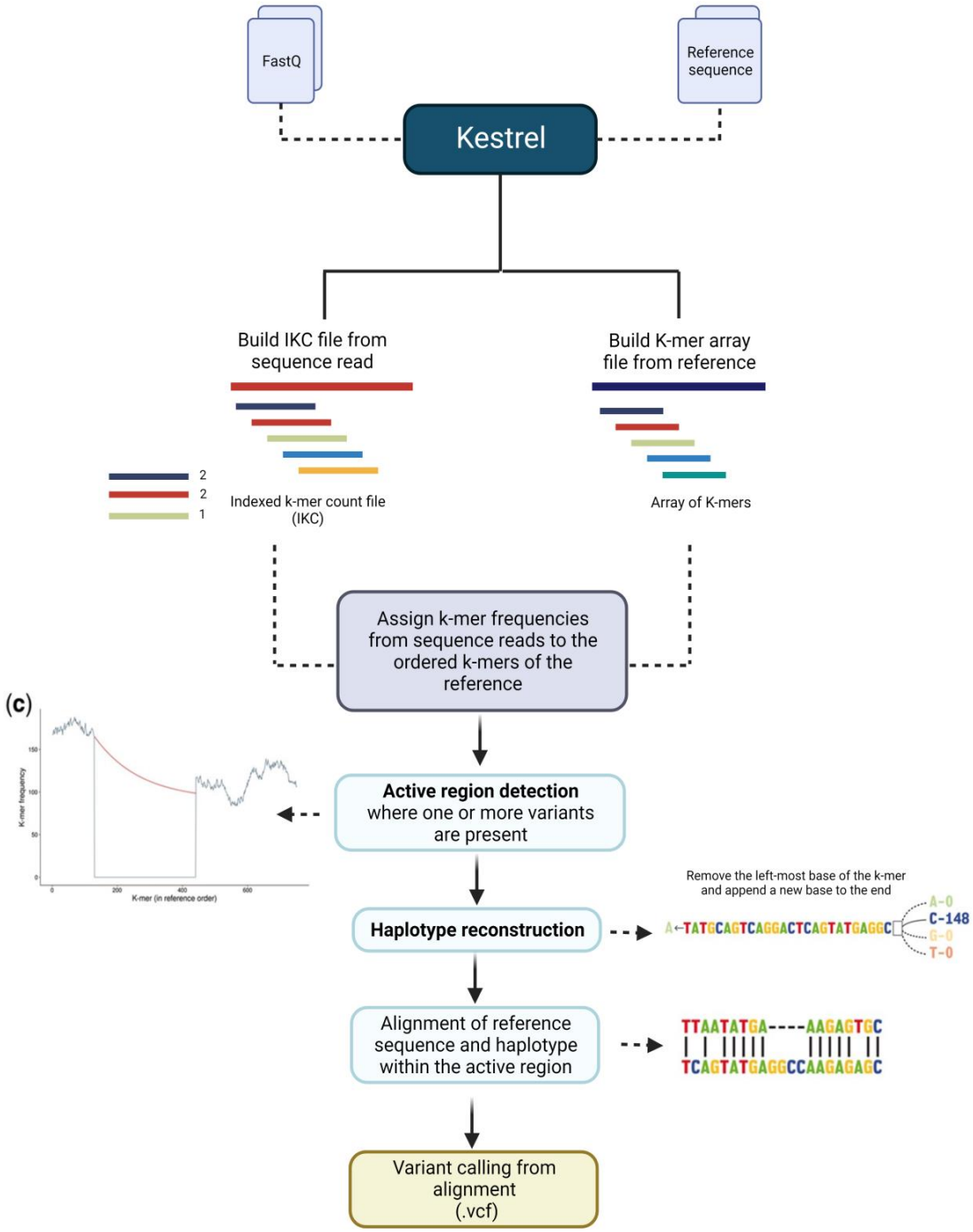

Figure S2. The overview of the Kestrel algorithm for mapping-free variant calling, related to STAR methods. The Kestrel algorithm takes fastq file(s) and the reference sequence and convert them to the indexed k-mer count file (IKC) and k-mer frequencies respectively. It allocates the sequence reads' k-mer frequencies to the array of ordered k-mers from the reference sequence. A sharp decline in k-mer frequency represents the active region where one or more variants exist. The algorithm then rebuilds the real sequence from the sample (haplotype), similar to the local assembly of k-mers with minimal resource consumption. In the last step it aligns the haplotype to the reference sequence using Smith-Waterman approach and variant calls are extracted from the active region. The output vcf files contain all true and false variants, and annotation and processing of the output vcf files separate the informative variants from the mass of non-informative calls.

Figure S3. Code-adVNTR method, related to STAR methods. Code-adVNTR is the second algorithm included in the pipeline to compare discordant results. This tool can detect motif count variations and small nucleotide mutations within a target list of coding VNTRs. Briefly, the algorithm initially recruits reads from alignment files harbouring repeat units (RUs) for a specific VNTR. The reads are then aligned using a modified version of the Viterbi algorithm to the multi-motif profile HMMs. After variation detection, the likelihood of the observed mutation and the chance of observing indel transition due to the sequencing errors was calculated by the binomial distribution. The statistical analysis using the log-likelihood ratio, was used to calculate the probability that the variant is legit.

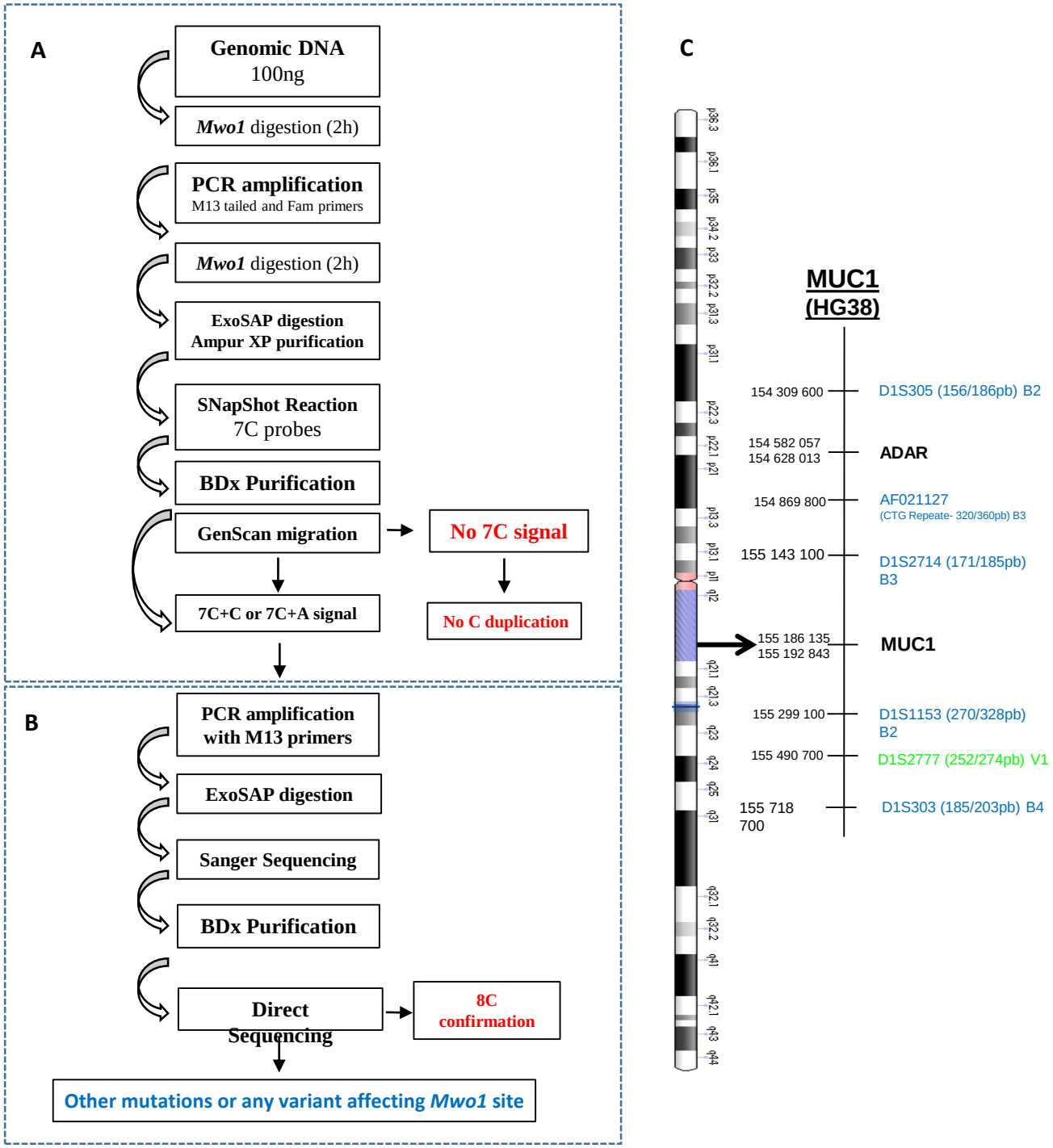

Figure S4. SNaPshot step-wise protocol for identification of dupC, dupA and SNPs in the *Mwo1* restriction site and markers used for linkage analysis, related to STAR methods. A) In brief, 100ng of genomic DNA is digested once with *Mwo1*. Remaining intact VNTRs are amplified using more extended primers, tagged with M13 sequences to increase the PCR size. The second digestion with Mwo1 is performed in one step, followed by two-step purification: ExoSAP, to remove primers, single stranded PCR product, and dNTP not incorporated, and Ampure XP beads, for size selection. **B)** Longer PCR products enabled us to Sanger sequence SNaPshot products to validate variations in the *Mwo1* site. **C)** The markers used for linkage analysis in the MUC1 positive families are shown.

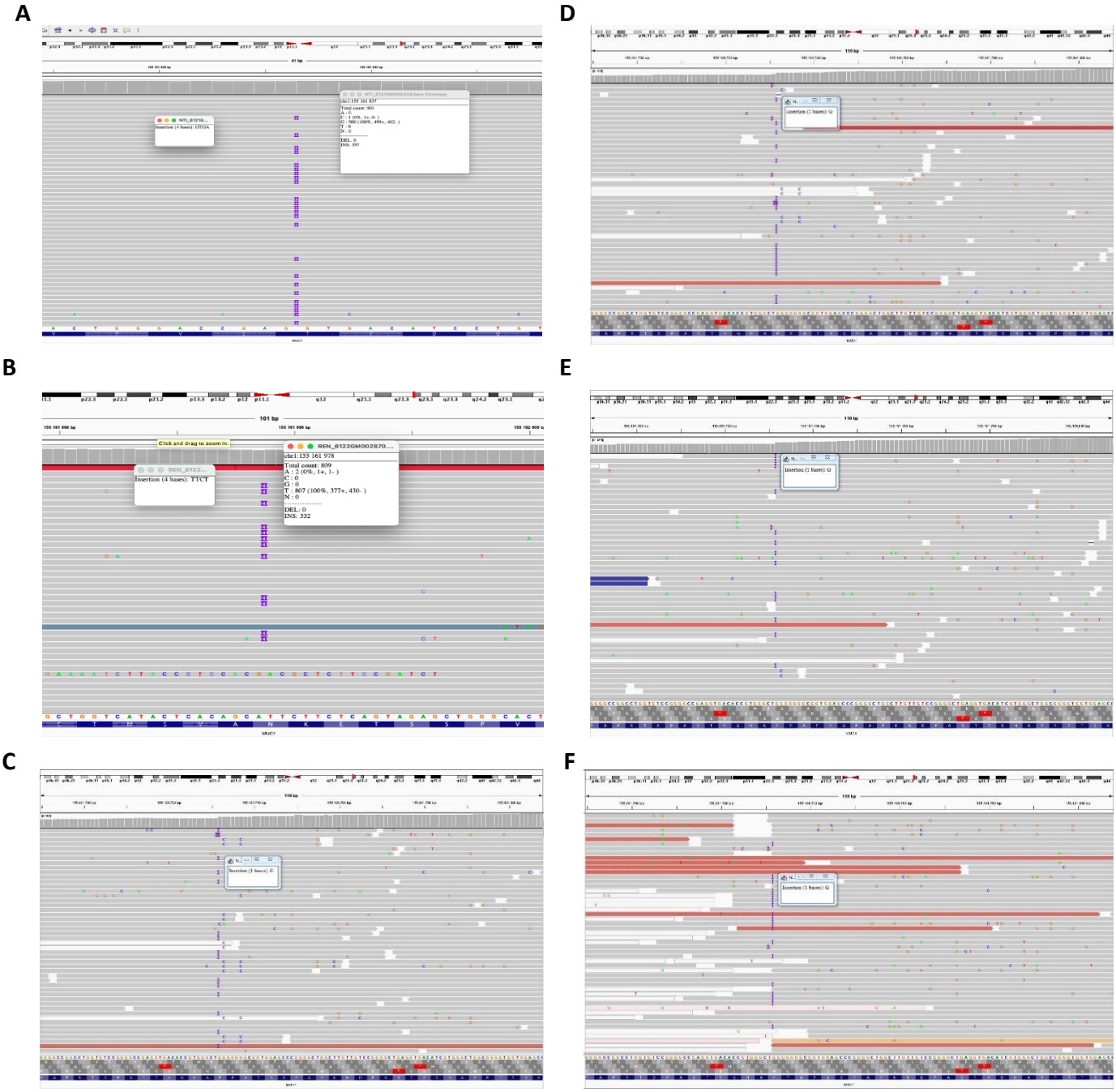

Figure S5. IGV visualization of bam files from patients with *MUC1* pathogenic variations in the early conserved motifs, related to Figure 6. A) Represents the 4bp (GTGA) duplication in motif 3 in a patient (NTI_6121GM005428). **B)** Another patient (REN_6122GM002870) with 4bp duplication of TTCT in motif 1 was detected by both VNtyper and standard pipeline. **C to F)** A dupC variation had been found in motif 5 in four unrelated patients: NPH1908593, NTI_6121GM003097, NPH2001245, NTI1129.

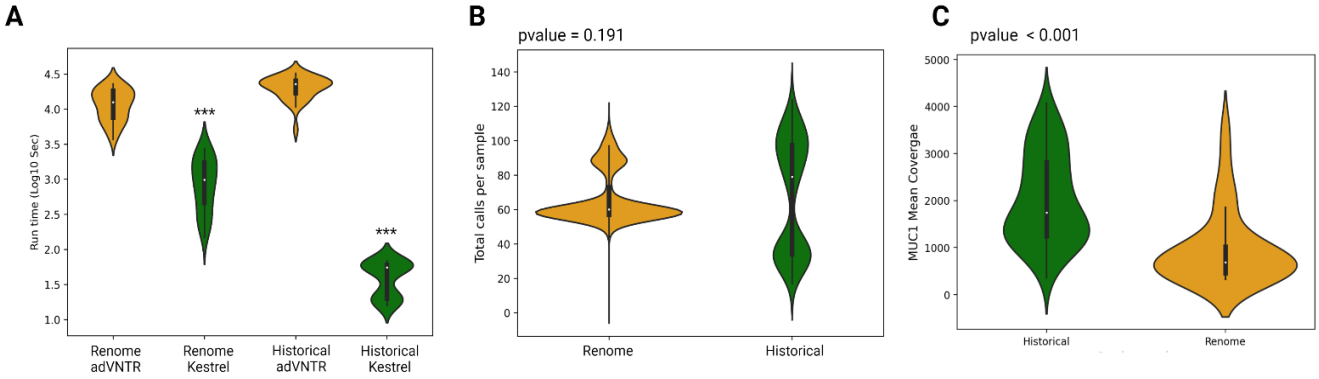

Figure S6. Coverage, run time and total variant count comparison between samples of the two NTI and renome cohorts, related to STAR methods. A) In both historical and renome analyses, the Kestrel method ran much faster than code-adVNTR (pvalue= 2.154E-20 for renome cohort and pvalue =2.516E-49 for historical cohort, n=30 in each group). The median run time for Kestrel for historical panel was 1.32 minutes and for renome panel it was 16.34 minutes. The median runtime for code-adVNTR for renome was 3.48 hours and for historical cohort was 6.33 hours (n=30, with Intel® Xeon® Processor X5550 2.67GHz CPU (8 cores and 16 threads)). Kestrel performs a rapid analysis of the input file because it utilizes all of the server's CPU threads. The computation load on the CPU is not high during the running time and it uses the maximum capacity whenever hard computation is needed. Therefore parallel analysis of multiple cases is possible using GNU parallel. **B)** Total number of variants (including SNPs and indels) called by the Kestrel method did not differ significantly across the historical and renome cohorts. **C)** The mean converge of the *MUC1* gene was investigated using the Sambamba tool and the "chr1:155158300-155162706" region on the hg19 assembly. There was a significant difference in mean coverage between our historical and renome cohort (pvalue < 0.001).

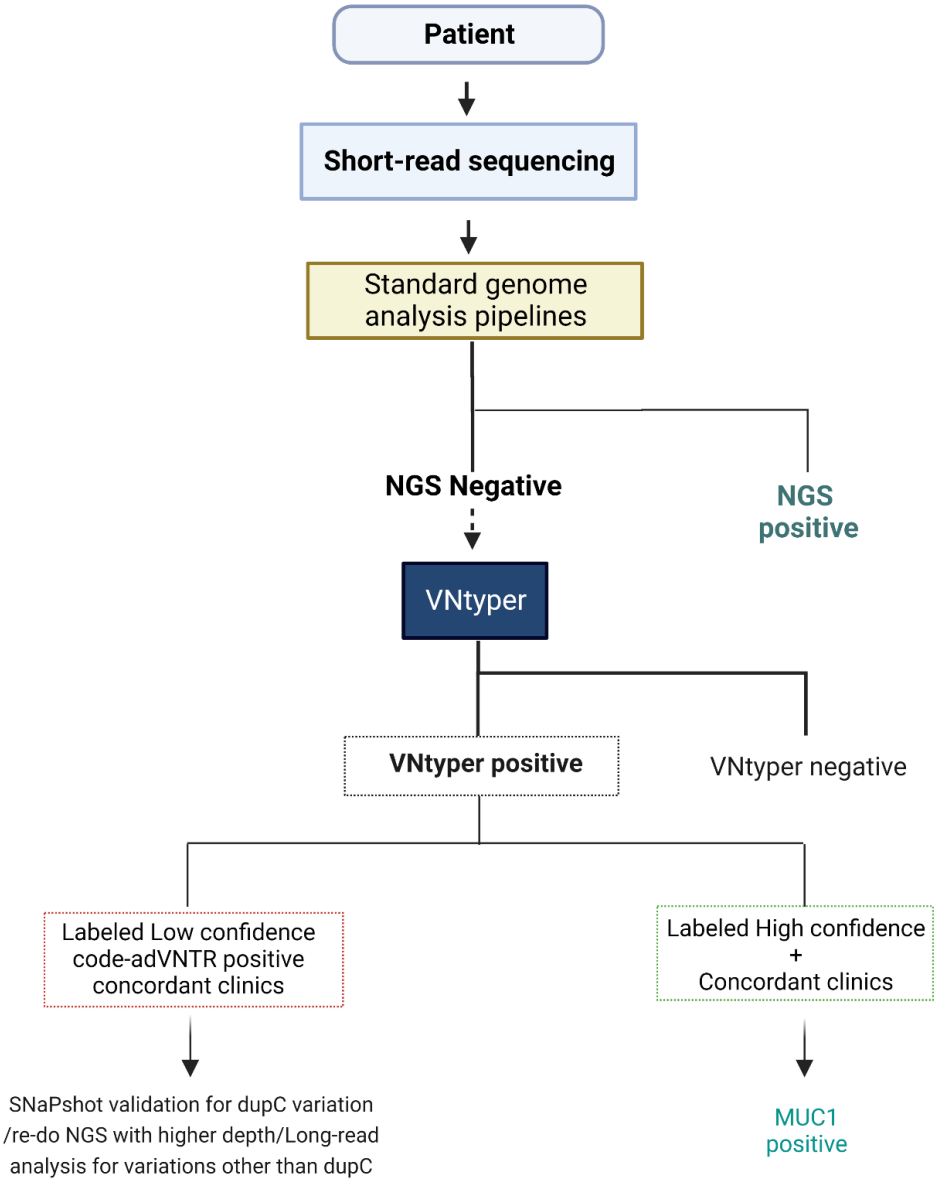

Figure S7. Diagnostic algorithm for suspected ADTKD patients with VNtyper pipeline, related to Figure 2. Using this algorithm, patients meeting the ADTKD inclusion criteria might be studied. See **Figure 2A** for inclusion and exclusion criteria in detail. Short-read sequencing and data analysis utilizing standard pipelines could uncover disease-causing variants in ADTKD-related genes or other disease-causing genes. In cases of negative NGS, VNtyper could analyse the alignment file (bam). The pipeline examine *MUC1* VNTR and discover pathogenic variations if exist. The variant reported could be interpreted as follows: (i) any variant reported with high confidence (Depth-score above 10% of the threshold (0.00515) and AltDepth > 20) could be considered as *MUC1* positive, (ii) variants labeled low confidence should be further studied by NGS with higher quality, the SNaPshot method for *MUC1* dupC, dupA, or long-read sequencing for any variants that do not alter the *Mwo1* restriction site.