Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing

Mutations causing medullary cystic kidney disease type 1 (MCKD1) lie in a large VNTR in *MUC1* missed by massively parallel sequencing

Andrew Kirby[1,2], Andreas Gnirke[1], David B. Jaffe[1], Veronika Barešová[3], Nathalie Pochet[1,4], Brendan Blumenstiel[1], Chun Ye[1], Daniel Aird[1], Christine Stevens[1], James T. Robinson[1], Moran N. Cabili[1,5], Irit Gat-Viks[1,6], Edward Kelliher[1], Riza Daza[1], Matthew DeFelice[1], Helena Hůlková[3], Jana Sovová[3], Petr Vylet'al[3], Corinne Antignac[7-9], Mitchell Guttman[1], Robert E. Handsaker[1,10], Danielle Perrin[1], Scott Steelman[1], Snaevar Sigurdsson[1], Steven J. Scheinman[11], Carrie Sougnez[1], Kristian Cibulskis[1], Melissa Parkin[1], Todd Green[1], Elizabeth Rossin[1], Michael C. Zody[1], Ramnik J. Xavier[1,12], Martin R. Pollak[13,14], Seth L. Alper[13,14], Kerstin Lindblad-Toh[1,15], Stacey Gabriel[1], P. Suzanne Hart[16], Aviv Regev[1], Chad Nusbaum[1], Stanislav Kmoch[3], Anthony J. Bleyer[17]*, Eric S. Lander[1]*, Mark J. Daly[1,2]*

[1]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [3]Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University in Prague, Czech Republic. [4]Department of Plant Systems Biology, VIB, Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [5]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [6]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel. [7]Inserm, U983, Paris, France. [8]Université Paris Descartes, Sorbonne Paris Cité, Institut Imagine, Paris, France. [9]Département de Génétique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris, Paris, France. [10]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [11]Upstate Medical University, Syracuse, New York, USA. [12]Gastrointestinal Unit, Center for the Study of the Inflammatory Bowel Disease and Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [13]Department of Medicine, Beth Israel Deaconess Med. Ctr, Boston, Massachusetts, USA. [14]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [15]Science for Life Laboratory Uppsala, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 751 23, Sweden. [16]Office of the Clinical Director, National Human Genome Research Institute, National Institutes of Health (NIH), Bethesda, Maryland. [17]Section on Nephrology, Wake Forest School of Medicine, Medical Center Blvd., Winston-Salem, North Carolina, USA. *Co-supervised the research – correspondence to ableyer@wfubmc.edu, lander@broadinstitute.org, mjdaly@atgu.mgh.harvard.edu.

While genetic lesions responsible for some Mendelian disorders can be rapidly discovered through massively parallel sequencing (MPS) of whole genomes or exomes, not all diseases readily yield to such efforts. We describe the illustrative case of the simple Mendelian disorder medullary cystic kidney disease type 1 (MCKD1), mapped more than a decade ago to a 2-Mb region on chromosome 1. Ultimately, only by cloning, capillary sequencing, and *de novo* assembly, we found that each of six MCKD1 families harbors an equivalent, but apparently independently arising, mutation in sequence dramatically underrepresented in MPS data: the insertion of a single C in one copy (but a different copy in each family) of the repeat unit comprising the extremely long (~1.5-5 kb), GC-rich (>80%), coding VNTR in the mucin 1 gene. The results provide a cautionary tale about the challenges in identifying genes responsible for Mendelian, let alone more complex, disorders through MPS.

Medullary cystic kidney disease (MCKD) type 1 (OMIM 174000) is a rare disorder characterized by autosomal dominant inheritance of tubulo-interstitial kidney disease[1]. Affected individuals variably require dialysis or kidney transplantation in the third to seventh decade of life. Diagnosis of MCKD1 in patients is complicated by the unpredictable progression of kidney disease, the absence of other specific clinical manifestations, and the high frequency of mild kidney disease in the general population[2]. Nonetheless, the disease has been compellingly and consistently mapped to a single autosomal locus at 1q21[3–7]. Attempts to identify the mutated gene(s), however, have not been successful[4].

The advent of massively parallel sequencing (MPS) technologies has made exhaustive sequencing of genomic regions a viable approach to the identification of genes responsible for rare Mendelian diseases caused by high penetrance mutations[8,9]. Yet, there is also a growing recognition that using MPS to discover disease genes is not always straightforward. Here, we report that MCKD1 is caused by an unusual class of mutations, recalcitrant to detection by MPS. The process of identifying the MCKD1 causal variation is of particular interest for human genetics, because it highlights important challenges in using current MPS for gene discovery.

Linkage analysis was performed on six likely MCKD1 pedigrees (Online Methods, **Supplementary Fig. 1** and **Supplementary Table 1**), and in all families the phenotype showed perfect co-segregation with a single 2-Mb segment of chromosome 1 (**Fig. 1**). We examined the genotype data for evidence of copy-number variation in the critical interval, but found only two common copy-number polymorphisms, neither of which segregated with disease. Looking at the longest stretches of allelic identity within pairwise comparisons of the pedigrees' phased risk-haplotypes, we also found no obvious ancestral haplotype shared by a significant fraction of the families (beyond the background LD in the general population). This result suggested that the families carried independently occurring mutations, consistent with the families' diverse ancestries.

To search for mutations, we employed whole exome-, regional-capture- and whole genome sequencing (Online Methods). We selected two affected individuals from each pedigree for sequencing, chosen, where possible, to share only a single haplotype (the risk haplotype) across the linkage region. In protein-coding regions, we found only two rare (<1% in 1000 Genomes Phase I data[10]), non-silent point variants (SNPs or small indels) shared by both of the affected individuals in any pedigree: each was in a different gene and each in a different pedigree. This finding is consistent with the expected background rate for 75 genes in six independent risk chromosomes given the presence of 100-200 rare coding variants in a typical

genome[10]. In the context of perfect segregation of the phenotype, near-complete coverage of the coding bases in the linked region and the experience with other Mendelian diseases, we had expected to find a gene harboring rare coding variants in multiple families. To our dismay, we found no such evidence.

We next examined the non-coding regions, but found no regional clustering of segregating rare variants. We searched for any large structural variation (hundreds of bases or larger) such as deletions, insertions, duplications and inversions. All variants identified in this manner either failed to segregate with disease or were found at appreciable levels in control populations.

At this point, we concluded that the causal mutation(s) in MCKD1 were either located in a subregion that was recalcitrant to sequencing or represented a novel mutational mechanism. We considered the possibility that MCKD1 might be due to expansions in a coding VNTR sequence, because recurrent mutations at coding VNTRs have been documented as the cause of many genomic disorders[11] and because massively parallel sequencing data might not readily reveal such an expansion.

We used SERV (Sequence-based Estimation of minisatellite and microsatellite Repeat Variability)[12] to identify highly variable tandem repeats (VNTRs) in or overlapping with coding regions of five genes contained within the disease-linked interval: *KCNN3, EFNA3, ASH1L*, *MEF2D and MUC1*. Candidate VNTRs in the first four genes were found either to be non-polymorphic or to show no notable expansion in affected individuals (relative to family members not sharing the risk haplotype and to CEPH family samples), based on direct assays of length by PCR.

The *MUC1* VNTR was particularly difficult to assay: it consists of many (20-125[13,14]) copies of a large repeat unit (60 bases) with very high GC-content (>80%). We ultimately assayed the VNTR by Southern blot and confirmed results with long-range PCR (Online Methods). In our patient samples, VNTR lengths were consistent with published descriptions and were not expanded on risk chromosomes, excluding VNTR length as pathogenic. *MUC1* remained particularly interesting as the only gene in the critical region displaying transcripts with kidney-specific expression, based on RNASeq data from an adult control individual (unrelated to this study). *MUC1* encodes mucin 1, a transmembrane protein expressed on the apical surface of most epithelial cells, providing (amongst other functions) a protective barrier to prevent pathogens from accessing the cell surface. The protein possesses a heavily glycosylated extracellular domain containing the VNTR and an SEA module with a cleavage site for release of the extracellular domain, which then binds noncovalently to the transmembrane domain[17,19] (**Fig. 2a**).

We considered the possibility that MCKD1 might be caused by point mutations within the *MUC1* VNTR missed due to poor sequence coverage because (i) it was excluded from whole-exome and regional-capture probes due to its low-complexity and extreme sequence composition (and also because it is rarely annotated as coding sequence) and (ii) it was dramatically underrepresented in quality-filtered data from the whole-genome sequence, likely due to its GC-richness and homopolymer content. Because the human reference sequence appeared to significantly underrepresent this region (hg19 predicts a VNTR length far smaller than the published range or that observed in any of our samples, including controls), we undertook to clone and then reconstruct the VNTR alleles of several affected individuals and a CEPH trio; we subcloned, Sanger sequenced and performed *de novo* assembly for each (Online Methods and e.g. **Fig. 2b-d**).

We found a number of point variants in the VNTR assemblies, but, with one exception, they either did not segregate with the risk haplotype or were present in the alleles of the CEPH trio and/or unaffected chromosomes. However, we found one variant consistent with disease segregation: the insertion of a single C (relative to the coding strand of *MUC1*) within a stretch of seven C's occurring at positions 53-59 in a single copy of the canonical 60-mer repeat (e.g. **Fig. 2e**). All six families carried such +C insertions, which appear to have arisen independently based on the different overall sizes of the VNTR, different local sequence contexts and different precise repeat units harboring the insertion (**Supplementary Figs 2** and **3**).

The frameshift caused by the insertion predicts a mutant protein that contains many copies of a novel repeat sequence (obtained by shifted translation of the VNTR) but which lacks, owing to a novel stop codon shortly beyond the VNTR terminus, the downstream SEA self-cleavage module and both transmembrane and intracellular domains characteristic of the normal MUC1 precursor protein (**Fig. 2a**).

Because discovery of the +C insertion required considerable labor and time, we sought to develop a simple and robust genotyping assay to enable larger population screening. We designed a probe-extension assay (Online Methods and **Fig. 3a**) capable of distinguishing reference and mutant *MUC1* VNTR repeat units, making use of MwoI (which selectively cleaves the reference sequence) to increase the stoichiometric ratio of mutant:reference repeat units.

We typed all samples collected from the six MCKD1 families used for linkage analysis, including 62 phenotypically affected and 79 unaffected relatives (**Fig. 3b-c**), and over 500 control individuals from CEU, Japanese, Chinese, Yoruba and Tuscan HapMap3 populations (**Fig. 3d**). The genotyping assay was perfectly concordant with sequencing results, and full genotyping of all family members showed that the insertion segregated perfectly with each family's risk haplotype and yet was not seen in any of the 500 HapMap samples.

Overall, the genotyping results provide strong evidence that the +C insertions are the high-penetrance genetic lesion that leads to development of MCKD1. As a statistical association, the significance of this observation can only be approximated, but it is clearly far less than the reciprocal of the number of bases in the genome (+C seen on 6/6 risk chromosomes vs. 0/1000 HapMap chromosomes). Furthermore, this observation is robust to population structure considerations since the mutations have arisen independently.

To explore the broader impact of *MUC1* mutations, we genotyped affected and unaffected individuals from 21 additional small MCKD families screened to be negative for known MCKD mutations (**Supplementary Table 1**), only one family of which had existing linkage information implicating 1q21[15]. In 13 of 21 families we found the presence of a +C insertion consistent with being a fully penetrant cause of disease, indicating a substantial role for *MUC1* in MCKD1-like phenotypes.

Using antibodies raised against a peptide synthesized based upon the predicted mutant VNTR sequence, we found specific intracellular staining in epithelial cells of Henle´s loop, distal tubule and collecting duct of MCKD1 patients (**Fig. 4a)**, which was absent in control kidney (**Fig. 4b**). Co-staining of patient and control tissue additionally with antibodies against normal MUC1 demonstrated the specificity of the MUC1-fs (our name for the predicted mutant protein) antibodies for the mutant protein, with diffuse and/or fine granular intracellular localization of the MUC1-fs protein in patient kidney (**Fig. 4c**), and also patchy co-localization of MUC1-fs and normal MUC1 signals on the apical membrane of collecting duct epithelial cells (**Fig. 4c** and **4d**). Detailed image analysis of patient tissue (**Fig. 4d**) compared to control tissue (**Fig. 4e**) detected no intracellular co-localization of MUC1-fs and normal MUC1 proteins in patient tissue, but

revealed *puncti* of colocalization in distinct plasmalemmal subdomains. Antibody to MUC1-fs did not stain normal kidney tissue.

This study highlights the fact that current MPS technology may not suffice to reveal disease mutations, even when linkage analysis conclusively pinpoints a critical region of a few megabases. Even if the insC event were not dramatically underrepresented in the quality-filtered MPS data and even if the reference genome assembly had been accurate in this region, it still would have been difficult to detect this particular insertion event using typical alignment and variation-detection tools due to (1) the underlying variability of VNTR size within and across individuals, (2) the inability to uniquely place reads within the VNTR, given current MPS read lengths, and (3) the fact that the mutant:reference allelic balance is skewed far from the expected 1:1 of a typical heterozygous variant.

The precise nature of the MCKD1 mutations is notable. Curiously, each independently-arising event is essentially the identical single-base insertion at the same position within one of the repeat units of the VNTR. Yet, insertions at many locations or other events (such as single-base deletions) would also result in out-of-frame translation of *MUC1* and/or novel stop codons. Possible explanations for the consistently observed mutation include: (1) this insertion event is strongly favored due to mutational mechanism, (2) other events (eg. delC) are selected against, (3) other events (eg. delC) are benign and not associated with MCKD1, and (4) other *MUC1* mutations exist but are undersampled here.

The identified mutation and the associated genotyping assay provide a screening tool for younger members of families in which MCKD1 has been previously diagnosed, as well as a diagnostic tool for sporadic cases. They also alleviate the challenge for living relative kidney donation, as potential donor family members have not known their status as unaffected or (yet-to-be) affected. Much work, however, remains to be done to elucidate the specific mechanism of pathogenesis of the MUC1 mutant protein. We note that knock-out studies indicate that the *MUC1* gene is not essential in mice[16] and support a possible dominant-negative and/or gain-of-function mode of action for the human *MUC1* mutation. Together with the dominant and late-onset nature of the disease, this raises the possibility of preventative or therapeutic approaches based on treatments that decrease expression of the *MUC1* gene or splice out its single VNTR-encoding exon.

ACKNOWLEDGEMENTS

This work was conducted as part of the Slim Initiative for Genomic Medicine, a joint U.S.-Mexico project funded by the Carlos Slim Health Institute. This research was supported in part by the Intramural Research Program of the NIH, NHGRI. S.K., H.H., J.S. and V.B. were funded by Charles University programs PRVOUK-P24/LF1/3 and UNCE 204011, and their work was supported by grants LH12015 and NT13116-4/2012 from the Ministry of Education and Ministry of Health of the Czech Republic. S.A. was supported by NIH DK34854 (The Harvard Digestive Diseases Center). N.P. is a Broad Fellow of the Broad Institute and a postdoctoral research fellow of the Fund for Scientific Research - Flanders (FWO Vlaanderen), Belgium. I.G.V. was supported by HFSP, Alon, the Israeli Centers of Research Excellence (I-CORE), and Edmond J. Safra Center for Bioinformatics at Tel Aviv University. Thanks to T. L. Hatte for reagent use. We thank David Altshuler, Todd Carter, and Johannes Schlondorff for useful discussions, and Maria Cortes, Miguel Ilzarbe, and Miguel Betancourt for helpful project management. We also thank Fran Letendre, Matthew Coole, Robert Paul Frere, Claude Bonnet, Leon Mulrain, Nyima Norbui, and Harindra Arachchi for Sanger sequencing.

AUTHOR CONTRIBUTIONS

A.B., E.L. and M.Daly jointly supervised the research. R.X., M.P. and S.A. provided study design and interpretation advice. C.A., S.J.S., P.S.H. and A.B. performed sample collection. C.Stevens managed the project. C.Sougnez and K.C. provided early genotyping and sequencing support. Linkage analysis was performed by A.K. based on prior work by P.S.H. A.K. and M.Daly developed variation-discovery and analysis methods. A.K., J.R. and R.H. analyzed structural variation. T.G. performed CNV analysis. Supervision of sequencing was by S.G. Custom-capture array design was by S.Sigurdsson and K.L.T. M.P. performed direct PCR of polymorphic-VNTR candidates selected by N.P. A.G. and D.A. performed Southern blot and long-range PCR of the *MUC1* VNTR. C.N. supervised the *MUC1*-VNTR sequencing approach. A.G. performed VNTR-allele cloning and generation of sequencing libraries. E.K., R.D., D.P. and S.Steelman performed Sanger sequencing. D.J. assembled and analyzed VNTR Sanger sequencing. M.G. provided RNAseq support. S.K. supervised immunohistochemistry and immunofluorescence work by V.B., H.H., J.S., and P.V. A.K., B.B. and M.DeFelice developed the C-insertion genotype assay. M.Z. provided informatic and sequencing consultation. A.R. provided informatic and analysis consultation. C.Y., J.R., M.C., I.G., R.H. and E.R. provided informatic support. The manuscript was written primarily by A.K., A.G., A.B., E.L. and M.Daly. The supplementary information was prepared mainly by A.K., A.G., D.J., B.B., R.H., S.Sigurdsson, S.K. and A.B.

COMPETING FINANCIAL INTERESTS

Andrew Kirby, Andreas Gnirke, Brendan Blumenstiel and Matthew DeFelice are listed as inventors on the C-insertion genotyping assay under patent review. The other authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Linkage of six MCKD1 families to chromosome 1. LOD curve shows the combined linkage-score of six MCKD1 pedigrees across 12 Mb of chromosome 1, with the peak score well above the threshold of 3.6 for genome-wide significance[17]. Red X's mark the locations of opposite-allele homozygous genotype calls between affected members within each pedigree and highlight regions where affected individuals *de facto* share no alleles IBD, thereby delineating genomic segments unlikely to harbor causal variation. The shaded region (hg19:chr1:154,370,020–156,439,000) was considered most likely to contain any causal mutations, bounded on each side by recombination breakpoints in two different pedigrees.

Figure 2. Discovery of +C insertion within *MUC1* coding VNTR. (a) The major domains of the full-length MUC1 precursor protein are shown: N-terminal signal sequence, VNTR, SEA module (where cleavage occurs), transmembrane domain, and C-terminal cytoplasmic domain. Based on fully and unambiguously assembled VNTR alleles, the frameshift caused by insertion of a C in the coding strand (as described in the main text) is expected to introduce a novel stop codon shortly beyond the VNTR domain. **(b and c)** Where possible, knowledge of segregating phased SNP-marker haplotypes was used to select for *de novo* VNTR sequencing and assembly of those individuals sharing only a single haplotype across the region, as this aided identification of the VNTR allele segregating with the shared risk haplotype. **(d and e)** Independent *de novo* assembly of the shared VNTR allele in two individuals from family 4 shows exactly identical complete sequence, with the seventh 60-base unit (red X) out of 44 containing

a +C insertion event. The assembly is oriented relative to the coding strand of *MUC1* and covers bases chr1:155,160,963-155,162,030 (hg19). Each unique 60-base repeat segment is represented by a different letter or number (**Supplementary Fig. 2**). **(e)** Translational impact of +C frameshift.

Figure 3. Detection of *MUC1* +C insertion by probe-extension (PE) assay. (a) Exemplar electropherograms for the *MUC1*-VNTR +C-insertion PE assay (Online Methods) performed on homozygous reference-allele and heterozygote samples. (**b)** Allele-intensity scatterplot for large linkage family 2. X-axis values correspond to the detected intensity at the mass of the +C PE product, while Y-axis values reflect that of the reference repeat-unit extension product. Datum coloring reflects MCKD1 diagnosis: blue = unaffected (or HapMap samples), red = affected, white = unknown. Individuals known to carry the linkage-analysis risk haplotype are represented by "+", while other family members are depicted as dots. **(c)** Allele-intensity scatterplot for all MCKD1 linkage families. Samples having log-transformed intensities below 0.25 for both alleles were excluded as failed assays. WGA and low DNA-concentration samples were also excluded for underperforming. **(d)** Allele-intensity scatterplot for HapMap samples together with selected positive controls (MCKD1 individuals known to carry the insertion).

Figure 4. Immunohistochemical and immunofluorescence studies of MUC1-fs protein. In MCKD1 patients, MUC1-fs is expressed and present in renal epithelial cells of Henle´s loop, distal convoluted tubule, and collecting duct. **(a)** Strong intracellular staining of MUC1-fs protein in MCKD1 patient, and **(b)** absence of the specific staining in control; TALH - thick ascending limb of Henle´s loop; CD – collecting duct; PT – proximal tubule. **(c)** Immunofluorescence analysis showing diffuse and/or fine granular intracellular and membrane staining of MUC1-fs protein, and its partial colocalization with normal MUC1 in collecting duct of an MCKD1 patient. MUC1-fs staining is absent in control, and colocalization with normal MUC1 is therefore not detected. The values of fluorescent signal overlaps are transformed to a pseudo-color scale shown at right bottom in the corresponding lookup table. **(d)** Immunofluorescence analysis showing different intracellular localizations and partial sub-membrane colocalization of MUC1-fs and normal MUC1 proteins in collecting duct of MCKD1 patient. Note specific staining of both forms in distinct membrane microdomains. **(e)** Absence of MUC1-fs staining and characteristic membrane localization of normal MUC1 in control.

ONLINE METHODS

Family collection and criteria for diagnosis of affected status. The six analyzed families with autosomal dominant tubulointerstitial kidney disease were among a larger group referred for evaluation. Each showed a clinical phenotype highly suggestive of MCKD1 and lacked *UMOD* or *REN* mutations. All had previously demonstrated evidence of linkage to chromosome 1. Written informed consent was obtained from all participants and the study was approved by the Wake Forest School of Medicine Institutional Review Board. Medical records were reviewed and peripheral venous blood samples were obtained for DNA isolation and laboratory determinations. Full diagnostic methods and clinical summaries are described in **Supplementary Note**.

Linkage and CNV analysis. Family members were genotyped on the Affymetrix 6.0 platform. Whole Affymetrix arrays with genotype call rates < 88% were excluded from analysis, as were

samples which yielded low OD measurements (indicating poor sample performance during laboratory steps). Further, markers were excluded for which probe sequences showed excess genomic homology or potential for significant G-quartet formation (those probe sequences for which either allele contained at least three consecutive G's).

Particularly large pedigrees (>24 bit complexity) were divided into branches where required by computational constraints. LD-independent marker maps were separately created for each pedigree/branch, choosing single, well-typed, informative markers from LD-defined bins of SNPs based on phased, population-specific HapMap data (hapmap.org, release 22). Markers which showed no-call rates > 10% or any Mendelian inheritance errors within a pedigree/branch were excluded from specific pedigree/branch analyses. Additionally, markers were required to be spaced at least 0.1 cM apart according to published sex-averaged recombination positions (affymetrix.com).

All expected intra-pedigree relationships were confirmed from pairwise IBD estimates using PLINK software[18] and similarly derived marker sets; however, markers for PLINK were selected agnostic to their being polymorphic within a pedigree/branch so as not to skew IBD calculations. Merlin software[19] was used to remove any likely genotyping errors which did not violate Mendelian inheritance rules, and then to perform parametric linkage under a rare, autosomal-dominant model using population-specific allele frequencies (affymetrix.com).

Linkage mapping was performed using the Merlin package under a rare autosomal-dominant model. Scores were combined across pedigrees/branches by summing LOD values, linearly interpolating scores between marker locations as required. The consistency of the alleles carried on the segregating risk haplotype was confirmed across pedigree branches.

The boundaries of the linked region were refined by searching all well-typed markers -- including many that were dropped solely to eliminate markers in LD from the linkage calculations -- for instances where affected members within the same pedigree shared no alleles IBD (by virtue of being homozygous for opposite alleles – for example, one having genotype AA and another CC). Such markers necessarily lie outside the critical linkage interval.

Affymetrix 6.0 intensity data were used by Birdsuite software[20] to analyze copy-number variation.

Large-scale sequencing. Because the critical region contains more than 170 separate transcript annotations comprising over 75 RefSeq genes, amplicon-based resequencing of genic regions was initially not considered. Of the 12 sequenced individuals, whole-genome sequencing was performed on 11 of these individuals (~25-fold coverage on average), whole-exome sequencing on 11 individuals (~180-fold coding-sequence coverage on average) and regional-capture sequencing on 5 individuals (~220-fold coverage on average). Sequence processing is described in **Supplementary Note**. For all but three of the RefSeq genes, at least 99% of the coding bases were covered at ≥10-fold in each pedigree. Further, 98% of non-coding bases were covered at ≥10-fold in each pedigree.

As candidates for being pathogenic MCKD1 mutations, we considered any non-reference allele present in both affected individuals of any pedigree and with a population frequency ≤1%[10]. Non-coding regions were analyzed similarly.

To discover potential structural variation at the chromosome-1 locus, we ran Genome STRiP[21] on the sequenced individuals and on a control population of 32 Finnish genomes sequenced at low coverage by the 1000 Genomes Project[10] (**Supplementary Note**).

***MUC1*-VNTR Southern blot analysis.** Genomic DNA (5-8 μg) was digested with 100 u *Hin*fI (NEB). Digests were run on a 0.8% agarose gel, transferred to a BrightStar Plus Nylon membrane (Ambion) and hybridized overnight at 65˚C to a quadruply biotinylated synthetic 100mer oligonucleotide probe PS1 (**Supplementary Table 3**) (IDT) present at 2 ng/ml in SuperHyb hybridization solution (Ambion) supplemented with 100 μg/ml sonicated salmon sperm DNA (Stratagene). After a final high-stringency wash at 65˚C in 0.2x SSC and 0.1% SDS, membrane-bound biotin was detected by a BrightStar BioDetect kit (Ambion).

***MUC1*-VNTR long-range PCR.** The long-range PCR protocol was adapted from Fowler et al.[14]. Briefly, 7-μL PCR reactions contained 15 or 30 ng genomic DNA, 1.75 pmol of PS2 and PS3 primers (**Supplementary Table 3**), 5% DMSO, 625 μM of each dNTP, 1x reaction buffer with 3 mM $MgCl_2$, and 0.25 u DyNAzyme EXT DNA polymerase (Finnzymes). Thermocycling on GeneAmp 9700 instruments (ABI) was as follows: initial denaturation (90 s at 96˚C); 22 or 27 cycles (40 s at 96˚C, 30 s at 65˚C, 6 min at 68˚C) and final extension (10 min at 68˚C).

***MUC1*-VNTR sequencing and assembly.** For selected individuals, we cloned gel-purified long-range-PCR products containing the full-length VNTR. Allele sizes derived from Southern blots and long-range PCR, together with known haplotype sharing between individuals in the same pedigree, in most cases permitted the identification of which *MUC1* VNTR allele was part of the segregating risk haplotype (e.g. **Fig. 2b and c**). In a few cases, the sizes of the risk and non-risk VNTR allele were nearly the same, precluding physical separation of the two alleles prior to molecular cloning. Using transposon hopping and capillary sequencing, we then sequenced clones from each allele (**Supplementary Note**).

Because the region is exceptionally repetitive and because the read data contain both PCR errors and sequencing errors (exacerbated by the extreme GC content of the repeat), we developed a special assembly algorithm that could distinguish *bona fide* genomic differences from errors (**Supplementary Note**). Given the repetitive sequence content, not all assemblies were complete or unambiguous. Instead, some assembly frameworks suggested multiple possible resolutions across areas of uncertainty, forming full networks of possible solutions for a particular allele.

Supplementary Table 2 summarizes the key properties of the assemblies (example shown in **Figure 2d**), and **Supplementary Figures 3** and **4** provide the sequence for those unique alleles (three risk and eight non-risk) where the assembly was fully or almost fully resolved. **Supplementary Figure 5** illustrates the notation of graph assembly in a scenario where an allele could not be fully and unambiguously reconciled. We assembled each allele separately and independently. In all situations where two alleles were expected to be identical by haplotype sharing and where the assemblies were fully resolved, the assemblies were indeed identical – thus increasing our confidence that the assemblies were correct.

Genotyping of *MUC1* +C insertion event. Genomic DNA was first over-digested using restriction endonuclease MwoI which selectively cleaves the reference repeat-unit sequence (GCCCCCCCAGC), while leaving intact repeat units containing the +C insertion (GCCCCCCC*C*AGC). Tailed primers nested within the 60-bp repeat were then used to PCR amplify the remaining intact VNTR fragments, thus enriching for insertion-containing fragments over reference-sequence background. PCR products were then re-digested with MwoI for a second round of enrichment. A 20-bp probe was then designed just upstream of the insertion site, and probe extension was performed using a high fidelity DNA polymerase and a nucleotide

termination mix containing dATP, ddCTP and ddGTP. Following probe extension, reaction products were separated and sized by MALDI-TOF mass-spectrometry using the Sequenom MassArray platform. Spectra were then assessed for the presence of peaks corresponding to the mutant repeat-unit extension-product (at 5,904.83 daltons) and the reference repeat-unit extension-product (at 6258.06 daltons).

Specifically, 100 μg of genomic DNA was digested in a 25-μL reaction volume for 16 hours using 5 units of MwoI restriction endonuclease (New England Biolabs) with supplemental additions of 5 units of enzyme at hours 3 and 15. Digestion reactions were then cleaned using 50 μL AmPure beads according to manufacturers protocol (Agencourt, Beverly, MA), and digested DNA was eluted in 25 μL of nuclease-free water. Remaining intact VNTR fragments were PCR-amplified using 1X HotStart buffer, 1.0 mM $MgCl_2$ (to supplement MgCl2 already in buffers), 0.8 mM dNTPs, 0.8 units of HotStart Taq Plus (Qiagen) and 0.2 μM forward and reverse primers PS6 and PS7 (**Supplementary Table 3**) in a 25-μL reaction volume. PCR cycling conditions were: one hold at 95°C for 5 min; 45 cycles of 94°C for 30 sec, 67°C for 30 sec, 72°C for 1 min; followed by one hold at 72°C for 10 min. PCR reactions were cleaned using 50 μL AmPure beads, and amplicons were eluted in 25 μL nuclease-free water. A second round of MwoI digestion was performed again for 16 hours with 5 units of enzyme added at hours 0, 3 and 15. Digestion reactions were cleaned using 50 μL AmPure beads and product was eluted in 6.2 μL of nuclease-free water.

Using 5.2 μL of the digested eluate as template, probe extension was performed using 1X HotStart buffer, 0.6 mM $MgCl_2$ (to supplement MgCl2 already in buffers), 1.7 μL SAP buffer (Sequenom, San Diego, CA), 0.2 mM each of nucleotides ddGTP, ddCTP and dATP; 0.7 units of Thermo Sequenase DNA polymerase (Amersham) and 0.6 μM of extension probe PS8 (**Supplementary Table 3**) in a 10-μL reaction volume. Probe extension was performed on a 384-well ABI GeneAMP 9700 and cycling conditions were: one hold at 94°C for 2 min 55 cycles of 94°C for 5 sec, 52°C for 5 sec, 72°C for 5 sec; followed by one hold at 72°C for 7 min. Reactions were then de-salted by addition of a cation-exchange resin, and ~7 nL of purified extension reaction was spotted onto a SpectroChip (Sequenom) containing matrix 3-hydroxypicoloinic acid. Arrayed reactions were then analyzed by matrix-assisted laser desorbtion/ionization-time of flight (MALDI-TOF) on a Compact mass spectrometer (Sequenom/Bruker).

Assay results were clear enough to assign genotypes based on simple inspection of X-Y scatterplots depicting log-transformed reference- and mutant-repeat-unit intensities ($\log_{10}$(1.0+peak height)). Samples showing log-transformed intensities < .25 for both alleles were considered failed assays. Similarly, results from whole-genome-amplification samples or samples with low DNA concentrations were typically considered unreliable and discarded.

Antibody generation and kidney immunostaining. Immunodetection of MUC1-fs was performed with custom-prepared rabbit antibodies (PA4 302) raised against the peptide SPRCHLGPGHQAGPGLHRPP, representing the predicted mutant VNTR unit (Open Biosystems, Huntsville, AL; diluted 1:1000 in 5% BSA in PBS). The normal MUC1 protein was detected with monoclonal mouse anti-human Epithelial Membrane Antigen (EMA) mouse monoclonal antibody (DAKO, Glostrup, Denmark; diluted 1:400 in 5% BSA in PBS). Detection of bound primary antibody was achieved using either Dako EnVision + TM Peroxidase Rabbit Kit (Dako) or System-HRP labeled Polymer Anti-mouse (DAKO), for rabbit or mouse antibodies, respectively, with 3,3´-diaminobenzidine as substrate.

Paraformaldehyde-fixed human kidney biopsies were analysed. The specificity of antigen detection was always ascertained by omission of the primary antibody-binding step.

For immunofluorescence analysis, PA4 302 antibody was diluted 1:200 in 5% BSA in PBS and EMA antibody was diluted 1:10 in 5% BSA in PBS. Fluorescence detection used species-specific secondary antibodies. Alexa Fluor® 488 goat-anti rabbit IgG and Alexa Fluor® 568 goat-anti mouse IgG (Molecular Probes, Invitrogen, Paisley, UK). Nuclei were stained with 4',6-diamidino-2-phenylindole (DAPI). Prepared slides were mounted in Immu-Mount fluorescence mounting medium (Shandon Lipshaw, Pittsburgh, PA) and analyzed by confocal microscopy.

XYZ images sampled according to Nyquist criterion were acquired using a TE2000E C1si laser scanning confocal microscope, Nikon PlanApo objective (40x, N.A.1.30), 488 nm and 543 nm laser lines and 515 +/-15 nm and 590 +/-15 nm band pass filters. Images were deconvolved using the classic maximum likelihood restoration algorithm in Huygens Professional Software (SVI, Hilversum, The Netherlands). Colocalization maps employing single pixel overlap coefficient values ranging from 0-1 were created using Huygens Professional Software. The resulting overlap coefficient values are presented as pseudo-color (scale is shown in corresponding figure lookup tables).

REFERENCES

1. Bleyer, A. J., Hart, P. S. & Kmoch, S. Hereditary interstitial kidney disease. *Semin. Nephrol.* **30**, 366–373 (2010).
2. Castro, A. F. & Coresh, J. CKD surveillance using laboratory data from the population-based National Health and Nutrition Examination Survey (NHANES). *Am. J. Kidney Dis.* **53**, S46–55 (2009).
3. Christodoulou, K. *et al.* Chromosome 1 localization of a gene for autosomal dominant medullary cystic kidney disease. *Hum. Mol. Genet.* **7**, 905–911 (1998).
4. Wolf, M. T. F. *et al.* Medullary cystic kidney disease type 1: mutational analysis in 37 genes based on haplotype sharing. *Hum. Genet.* **119**, 649–658 (2006).
5. Serafini-Cessi, F., Malagolini, N. & Cavallone, D. Tamm-Horsfall glycoprotein: biology and clinical relevance. *Am. J. Kidney Dis.* **42**, 658–676 (2003).
6. Vylet'al, P. *et al.* Alterations of uromodulin biology: a common denominator of the genetically heterogeneous FJHN/MCKD syndrome. *Kidney Int.* **70**, 1155–1169 (2006).
7. Scolari, F. *et al.* Uromodulin storage diseases: clinical aspects and mechanisms. *Am. J. Kidney Dis.* **44**, 987–999 (2004).
8. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19096–19101 (2009).
9. Al-Romaih, K. I. *et al.* Genetic diagnosis in consanguineous families with kidney disease by homozygosity mapping coupled with whole-exome sequencing. *Am. J. Kidney Dis.* **58**, 186–195 (2011).
10. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
11. Gemayel, R., Vinces, M. D., Legendre, M. & Verstrepen, K. J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
12. Legendre, M., Pochet, N., Pak, T. & Verstrepen, K. J. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* **17**, 1787–1796 (2007).

13. Horne, A. W. *et al.* MUC 1: a genetic susceptibility to infertility? *Lancet* **357**, 1336–1337 (2001).
14. Fowler, J. C., Teixeira, A. S., Vinall, L. E. & Swallow, D. M. Hypervariability of the membrane-associated mucin and cancer marker MUC1. *Hum. Genet.* **113**, 473–479 (2003).
15. Auranen, M., Ala-Mello, S., Turunen, J. A. & Järvelä, I. Further evidence for linkage of autosomal-dominant medullary cystic kidney disease on chromosome 1q21. *Kidney Int.* **60**, 1225–1232 (2001).
16. Spicer, A. P., Rowse, G. J., Lidner, T. K. & Gendler, S. J. Delayed mammary tumor progression in Muc-1 null mice. *J. Biol. Chem.* **270**, 30093–30101 (1995).
17. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).
18. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
19. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
20. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
21. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).

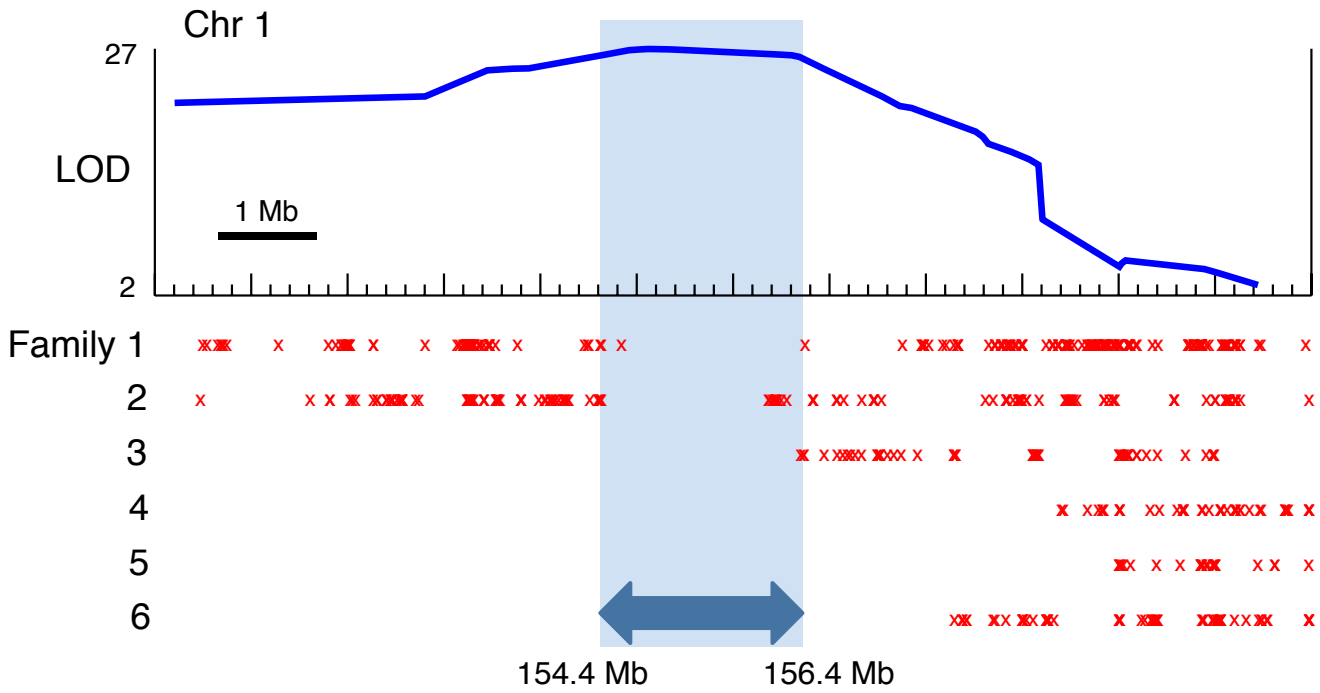


FIGURE 1

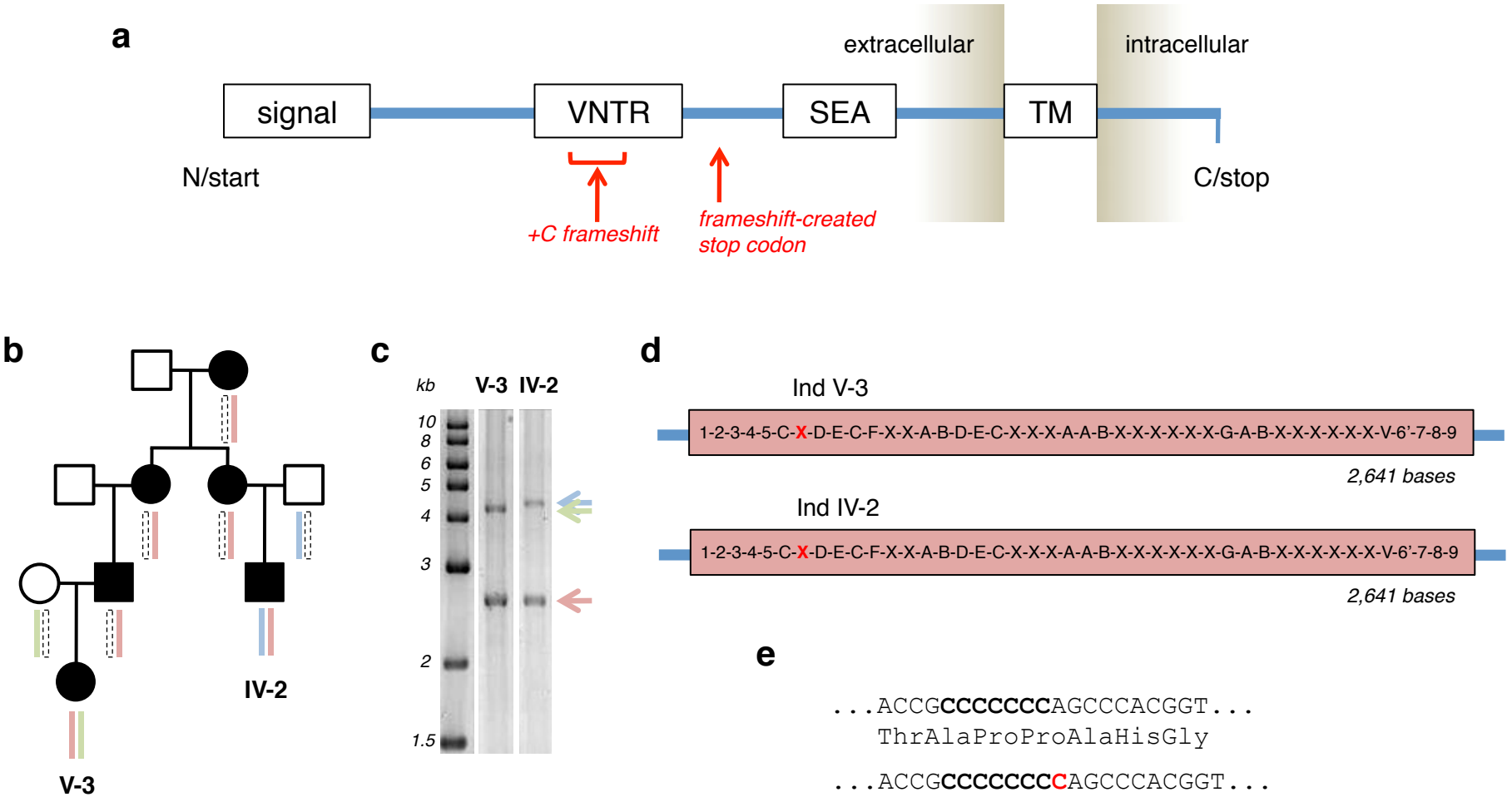

a
extracellular
intracellular
signal
VNTR
SEA
TM
N/start
C/stop
+C frameshift
frameshift-created
stop codon
b
IV-2
V-3
c
kb
V-3
IV-2
10
8
6
5
4
3
2
1.5
d
Ind V-3
1-2-3-4-5-C-X-D-E-C-F-X-X-A-B-D-E-C-X-X-X-A-A-B-X-X-X-X-X-X-G-A-B-X-X-X-X-X-X-V-6'-7-8-9
2,641 bases
Ind IV-2
1-2-3-4-5-C-X-D-E-C-F-X-X-A-B-D-E-C-X-X-X-A-A-B-X-X-X-X-X-X-G-A-B-X-X-X-X-X-X-V-6'-7-8-9
2,641 bases
e
...ACCGCCCCCCCAGCCCACGGT...
ThrAlaProProAlaHisGly
...ACCGCCCCCCCCAGCCCACGGT...
ThrAlaProProSerProArg

FIGURE 2

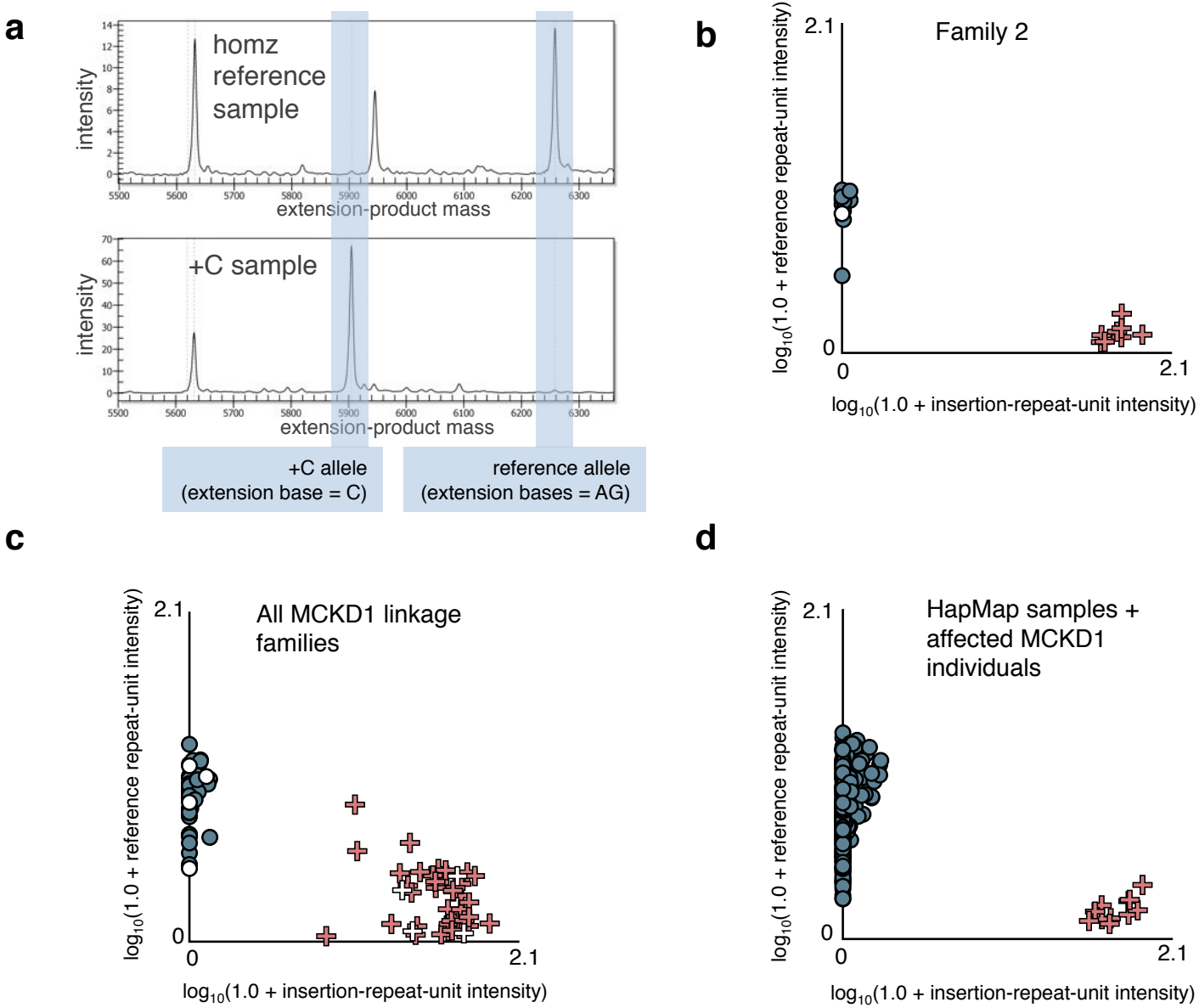

a
homz reference sample
intensity
extension-product mass
+C sample
intensity
extension-product mass
+C allele (extension base = C)
reference allele (extension bases = AG)
b
Family 2
log10(1.0 + reference repeat-unit intensity)
log10(1.0 + insertion-repeat-unit intensity)
c
All MCKD1 linkage families
log10(1.0 + reference repeat-unit intensity)
log10(1.0 + insertion-repeat-unit intensity)
d
HapMap samples + affected MCKD1 individuals
log10(1.0 + reference repeat-unit intensity)
log10(1.0 + insertion-repeat-unit intensity)

FIGURE 3

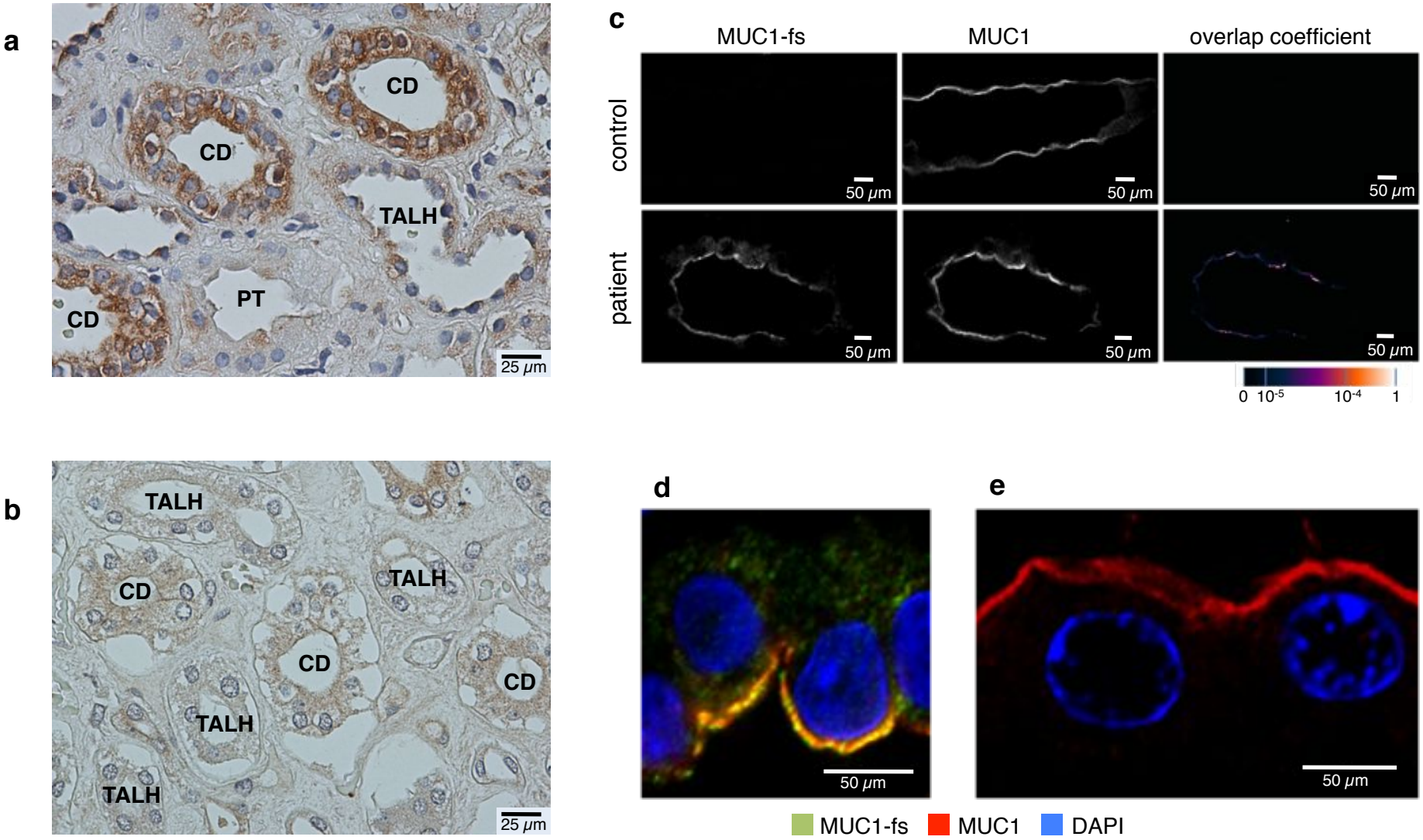

a
CD
CD
TALH
PT
CD
25 μm
b
TALH
CD
TALH
CD
CD
TALH
TALH
25 μm
c
MUC1-fs
MUC1
overlap coefficient
control
patient
50 μm
50 μm
50 μm
50 μm
50 μm
50 μm
0 10-5 10-4 1
d
50 μm
e
50 μm
MUC1-fs
MUC1
DAPI

FIGURE 4

Mutations causing medullary cystic kidney disease type 1 (MCKD1) lie in a large VNTR in *MUC1* missed by massively parallel sequencing

Andrew Kirby[1,2], Andreas Gnirke[1], David B. Jaffe[1], Veronika Barešová[3], Nathalie Pochet[1,4], Brendan Blumenstiel[1], Chun Ye[1], Daniel Aird[1], Christine Stevens[1], James T. Robinson[1], Moran N. Cabili[1,5], Irit Gat-Viks[1,6], Edward Kelliher[1], Riza Daza[1], Matthew DeFelice[1], Helena Hůlková[3], Jana Sovová[3], Petr Vylet'al[3], Corinne Antignac[7-9], Mitchell Guttman[1], Robert E. Handsaker[1,10], Danielle Perrin[1], Scott Steelman[1], Snaevar Sigurdsson[1], Steven J. Scheinman[11], Carrie Sougnez[1], Kristian Cibulskis[1], Melissa Parkin[1], Todd Green[1], Elizabeth Rossin[1], Michael C. Zody[1], Ramnik J. Xavier[1,12], Martin R. Pollak[13,14], Seth L. Alper[13,14], Kerstin Lindblad-Toh[1,15], Stacey Gabriel[1], P. Suzanne Hart[16], Aviv Regev[1], Chad Nusbaum[1], Stanislav Kmoch[3], Anthony J. Bleyer[17]*, Eric S. Lander[1]*, Mark J. Daly[1,2]*

[1]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [3]Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University in Prague, Czech Republic. [4]Department of Plant Systems Biology, VIB, Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [5]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [6]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel. [7]Inserm, U983, Paris, France. [8]Université Paris Descartes, Sorbonne Paris Cité, Institut Imagine, Paris, France. [9]Département de Génétique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris, Paris, France. [10]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [11]Upstate Medical University, Syracuse, New York, USA. [12]Gastrointestinal Unit, Center for the Study of the Inflammatory Bowel Disease and Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [13]Department of Medicine, Beth Israel Deaconess Med. Ctr, Boston, Massachusetts, USA. [14]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [15]Science for Life Laboratory Uppsala, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 751 23, Sweden. [16]Office of the Clinical Director, National Human Genome Research Institute, National Institutes of Health (NIH), Bethesda, Maryland. [17]Section on Nephrology, Wake Forest School of Medicine, Medical Center Blvd., Winston-Salem, North Carolina, USA. *Co-supervised the research – correspondence to ableyer@wfubmc.edu, lander@broadinstitute.org, mjdaly@atgu.mgh.harvard.edu.

Supplementary Note

Supplementary Figures

Supplementary Tables

References

Diagnostic methods and clinical summaries

Family collection and phenotyping

As previously described in families with MCKD1, these families shared the following characteristics: absent or low grade proteinuria with bland urinary sediments; slowly progressive kidney dysfunction; absence of causative findings on renal ultrasound; and absence of other associated signs or symptoms of systemic disease. Hypertension tended to occur only after the onset of chronic kidney failure.

Given these clinical characteristics, the only quantifiable relevant phenotype for analysis was estimated glomerular filtration rate (GFR). Often serum creatinine is used to measure kidney function, as serum creatinine levels are reciprocally related to glomerular filtration rate. However, serum creatinine levels are also affected by other factors, including protein intake and muscle mass, making it an inaccurate renal function marker in people with preservation of 70%-100% of normal kidney function. The MDRD formula estimates kidney function based on race, gender, age, and serum creatinine. Unfortunately, this formula is also inaccurate for patients with 70%-100% of normal kidney function. To further complicate matters, kidney function normally decreases with age, with some individuals experiencing significant loss of function – this is especially the case in African Americans and Native Americans.

Therefore, we considered the age and GFR in arriving at clinical diagnoses. We did not include in the analysis individuals less than 18 years of age, as even affected individuals under 18 could have a normal GFR. Individuals with significantly abnormal kidney function for their age were considered affected. Individuals were considered unaffected if kidney function was normal or if they were older and still had relatively preserved kidney function. In most of the families, affected individuals initiated renal replacement therapy (started dialysis) between 40 and 60 years of age.

Accordingly, individuals were considered affected if they required renal replacement therapy, had biopsy-proven interstitial kidney disease, or had an estimated GFR $\geq$ 2 standard deviations below the mean, adjusted for age and race. Individuals were considered to be unaffected if they were greater than 25 years of age and their estimated GFR was considered significantly higher than expected for affected family members. Approximately one-third of family members were excluded from initial genetic analysis due to indeterminate renal phenotypes.

Clinical summaries

All six families were linked to Chromosome 1 and had very consistent and similar presentations (**Supplementary Table 1**). For all affected individuals in the families, urinalysis results revealed minimal proteinuria and no hematuria. Pathology from kidney biopsies consistently revealed tubulo-interstitial fibrosis with no or minimal inflammatory infiltrate. Renal ultrasounds occasionally revealed some cortical cysts, but there were no medullary cysts identified. In most families, the disease was present in several generations and was always consistent with autosomal dominant inheritance. The age of onset of end-stage kidney disease varied somewhat between families. One family had associated bipolar disease, which was not present in any other families.

Family 1 includes 4 generations with autosomal dominant inheritance of interstitial kidney disease. Ages of end-stage kidney disease ranged from 29 to 69. Urinalyses from three affected family members revealed no blood or protein, and historically, there were no family members in whom hematuria or proteinuria was noted. A kidney biopsy performed on one family member and revealed focal interstitial fibrosis with tubular atrophy affecting approximately 30 to 40% of the tubule-interstitial areas.

Family 2 includes 4 generations of affected individuals with the autosomal dominant inheritance of progressive interstitial kidney disease. Kidney biopsy was performed on one individual who was being considered as a potential kidney donor and had normal kidney function (serum creatinine 0.9 mg/dl). Results of the biopsy showed patchy, mild interstitial fibrosis with associated mild tubular atrophy and a variable mixed inflammatory infiltrate including occasional eosinophils. Four renal ultrasounds in different affected family members revealed three with cortical cysts and two with small kidneys bilaterally. Urinalyses in affected individuals revealed no protein or blood. Age of kidney failure requiring dialysis ranged from 25 to 67 years, with some family members in their 60's with stage III kidney disease.

Family 3 has been extensively described[1]. Affected family members span 5 generations, with ages of onset of kidney failure ranging from 36 to 67 years. Characteristics of disease have included a steady, progressive decline in kidney function. Urinary examination in affected individuals revealed the absence of hematuria or proteinuria. Kidney biopsy results from four individuals[1] revealed tubular atrophy and interstitial fibrosis without a glomerular lesion. In the study of Kiser, renal ultrasound in 9 individuals revealed small bilateral cysts at the corticomedullary junction in one patient, 3 patients with acquired cystic disease, 3 with increased echogenicity, and 2 normal ultrasounds. In our analysis of this family, there were five renal ultrasounds, in which one individual had small cortical cysts, and three of the five ultrasounds revealed small, echogenic kidneys. None of the patients in the previous report had glomerulonephritis on biopsy or proteinuria or symptoms of other kidney diseases. In our evaluation,

there was one patient with proteinuria and diabetes who appeared to have diabetic nephropathy; mutational analysis in this patient was negative for the *MUC1* gene mutation, and in linkage studies, this patient had been labeled as having an uncertain diagnosis.

Family 4 was identified from referral of a 44-year-old female with chronic kidney disease whose urinalysis revealed no blood or protein. A kidney biopsy revealed widespread distention of tubules and chronic active interstitial nephritis with small microcystic changes to the tubules. The patient's sister underwent a kidney biopsy at age 27 years which showed moderate patchy interstitial fibrosis with tubular atrophy. Inheritance of kidney disease was documented as autosomal dominant over 5 generations, with family members starting dialysis between the ages of 45 and 79 years.

Family 5 suffers from the autosomal dominant inheritance of both interstitial kidney disease and bipolar disease, both having been linked to chromosome 1[2]. There were no other associated physical findings in this family, and no other families studied suffered from bipolar disease or other psychiatric disorders. Age of onset of kidney failure was younger than in other families, with the need for renal replacement therapy occurring at ages 22 to 33 over three generations. There was no kidney biopsy material available. Urine studies in this family were negative for blood and protein. All clinical findings were very consistent with MCKD1.

Family 6 is of European descent from the mid-West of the United States has previously been described (Kindred B in the investigation of Lindeman et al.[3]). The autosomal dominant inheritance of interstitial kidney disease was noted over 5 generations. In one individual, at autopsy results revealed marked atrophy of the tubules and marked interstitial fibrosis; no cysts were present. No glomerular changes were noted. Age of onset of end-stage kidney faiure ranged from 27 to 44 years. Renal ultrasound in one individual revealed normal sized kidneys and occasional cysts. Urinalysis results in family members did not reveal blood or protein.

Supplementary Table 1 also presents data on families that were tested for the C insertion in the VNTR of the *MUC1* gene. Except for Family 6[4], which had previously been linked to the area of interest on Chromosome 1, linkage studies had not been done in these families. All families demonstrated findings consistent with MCKD1: autosomal dominant inheritance was present; urinalysis in affected individuals were bland without hematuria and with minimal proteinuria; renal biopsies revealed interstitial fibrosis, and there were no associated clinical abnormalities of other organ systems. In two families, kidney biopsy material was stained with the antibody to the mutant MUC1 protein, with results similar to those documented in the figure in the main manuscript.

Large-scale regional sequencing

Whole-genome and whole-exome sequencing

Whole-genome and whole-exome libraries were sequenced on either Illumina HiSeq 2000 or Illumina GAIIX with the use of 101-bp paired-end reads for whole-genome sequencing and 76-bp paired-end reads for whole-exome sequencing.

For a subset of samples, starting with 3 μg of genomic DNA, library construction was performed as described by Fisher et al[5]. Another subset of samples, however, was prepared using the Fisher et al. protocol with some slight modifications: initial genomic DNA input into shearing was reduced from 3 μg to 100 ng in 50 μL of solution.

Exomes were captured using the Agilent SureSelect v2. For a subset of whole-genome samples, size selection was performed using gel electrophoresis, with a target insert size of either 340 bp or 370 bp +/- 10%. Multiple gel cuts were taken for libraries that required high sequencing coverage. For another subset of whole-genome samples, size selection was performed using Sage's Pippin Prep.

Following sample preparation, libraries were quantified using quantitative PCR (kit purchased from KAPA biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2 nM and then denatured using 0.1 N NaOH using Perkin-Elmer's MultiProbe liquid handling platform.

Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using either Genome Analyzer v3, Genome Analyzer v4, HiSeq 2000 v2, or HiSeq v3 cluster chemistry and flowcells. For a subset of samples, after cluster amplification, SYBR Green dye was added to all flowcell lanes, and a portion of each lane visualized using a light microscope, in order to confirm target cluster density. Flowcells were sequenced either on Genome Analyzer IIX using v3 or v4 Sequencing-by-Synthesis Kits, then analyzed using RTA v1.7.48; or on HiSeq 2000 using HiSeq 2000 v2 or v3 Sequencing-by-Synthesis Kits, then analyzed using RTA v1.10.15. or RTA v.1.12.4.2.

Custom capture-array

A custom sequence-capture microarray was designed to perform target enrichment of chromosome 1 152-156 Mb (hg18), including the critical MCKD1-linked region. The four megabase region was tiled with probes using Nimblegen in-house repeat masking algorithm[6]. Each array contains 385,000 probes tiled

with ~5 bp overlap, excluding the repetitive sequence. The probe set covered 70% of the 4-Mb target region, or 85% if 100 bp offset around probes is used.

For each sample we fragmented 5 μg of genomic DNA, using a water bath sonicator (Bioruptor, Diagenode) on high power. The sonication program was cyclical, 30 seconds on and 30 seconds off for a total of 15 minutes to obtain fragments of 200 to 800 bp. The size of the fragments was evaluated on DNA1000 chips (Agilent 2100 Bioanalyzer, Agilent). Library preparation was performed using the Illumina Paired end kit according to the manufacturer's protocol. In short: DNA fragments were blunt ended with T4 and Klenow polymerases and T4 polynucleotide kinase with 10 mM dNTP. A 3' adenosine overhang was added using Klenow exo fragment and 1 mM dATP followed by ligation of Illumina sequencing adapters with Quick ligase. The agarose gel electrophoresis size-selection step was omitted. Ligated fragments were enriched using 11-15 cycles of linker mediated (LM) PCR using a high-fidelity polymerase (Phusion, Finnzymes). 1 μg of cleaned PCR product was hybridized on the Roche Nimblegen Sequence Capture array following the manufacturer's protocol. After 3 days of hybridization, slides were washed and the enriched DNA was eluted. Eluted DNA was amplified with 14-18 cycles of LM-PCR, cleaned with QIAquick PCR purification kit (Qiagen) and quantified. Sequence capture efficiency was verified using real-time PCR amplification of three primer sets within the target region and three primer sets outside of the target region. Amplification of primer sets inside and outside the target region was compared using both sequence capture-enriched DNA and pre-capture DNA. An eight-to-ten fold difference in rt-PCR amplification was deemed to be successful enrichment.

Sequence analysis

Sequence data was processed with Picard (http://picard.sourceforge.net/)[7] and BWA[8] for mapping reads. SNPs and small indels were called using GATK, which utilizes base quality-score recalibration and local realignment at known indels[7,9]. The analyzed variable sites were restricted to those that pass GATK standard filters to eliminate events with strand-bias, low quality for the depth of sequencing achieved, homopolymer runs, and SNPs near indels.

We ran Genome STRiP both with the default parameters and also with relaxed parameter settings to attempt to increase sensitivity. In normal use, Genome STRiP looks only for signatures of large deletions indicated by unusual spacing or orientation of read pairs. For this application, we additionally enabled Genome STRiP to perform read pair clustering across all possible read pair orientations to generate a set of 2,208 loci that might potentially contain some form of structural variation.

We reviewed the evidence for each potential SV locus, looking for evidence of structural rearrangement that was inherited and segregated with disease but was not found at appreciable frequency in the Finnish control samples or in other published data sets (1000 Genomes, Database of Genomic Variants). We reviewed all loci that passed initial screening manually using IGV[10]. Although we found evidence for a number of structural polymorphisms (as would be expected in any population of this size), none met the criteria expected for causal mutations. Additional small-indel and structural-variation detection was performed using an in-house tool (unpublished).

MUC1-VNTR sequencing and assembly

Cloning and Sanger sequencing

PCR reactions were run on 0.8% agarose gels. *Bona fide* full-length PCR products were excised, cleaned-up by QiaQuick gel-extraction kits (Qiagen) and TOPO-TA cloned in pCR-4-TOPO vector in TOP10 cells (Invitrogen). After electroporation, kanamycin-resistant transformants (typically 8 clones per gel-purified PCR product) were analyzed by *Eco*RI digestion and long-range PCR (see above). *Bona fide* full-length plasmid clones, typically 2x2 clones for each allele in a given individual (2 independent PCR reactions, 2 clones each, typically 8 clones per individual) were subjected to *in vitro* transposition with EZ-Tn5 <TET-1> (Epicentre). For each clone, 384 triple-resistant (50 μg/ml Ampicillin + 50 μg/ml Kanamycin + 10 μg/ml Tetracycline) EC100 TransforMAX (Epicentre) transformants were robotically picked, grown up, miniprepped and sequenced with TET-1 PS4 and TET-1 PS5 primers (**Supplementary Table 3**). Sequencing reactions (5 μL) contained 0.75 μL 5x sequencing buffer, 0.4 μL BigDye Terminator v3.1, 0.1 μL dGTP BigDye Terminator v3.0 (all ABI), 3.5% DMSO, 1.8 pmol sequencing primer and 1.5 μL miniprepped plasmid. Sequencing reactions were thermocycled as follows: initial denaturation (1 min at 96˚C) then 40 cycles (30 s at 96˚C, 15 s at 50˚C, 4 min at 60˚C). Extension products were cleaned up by ethanol precipitation and run on 3730xl DNA Analyzers (ABI).

Sequence assembly

The exceptionally repetitive nature of the region, as well as the presence in the read data of both PCR errors and sequencing errors (exacerbated by the extreme GC content of the repeat), required a special assembly method that could distinguish *bona fide* genomic differences from errors. The method included three key conceptual components:

(1) The ability to distinguish between base calls supported by multiple reads from only a single clone, and those base calls supported by multiple clones.

(2) Sensitive error detection in stacks of reads determined to belong to the same genomic region as the result of an initial, less sensitive round of error correction.

(3) Allele construction by gluing reads together along long, perfect overlaps. Because of the repetitive sequence content, not all assemblies were complete or unique. Instead, some assembly frameworks suggested multiple possible resolutions across areas of ambiguity, forming entire/full networks of possible solutions for a particular allele.

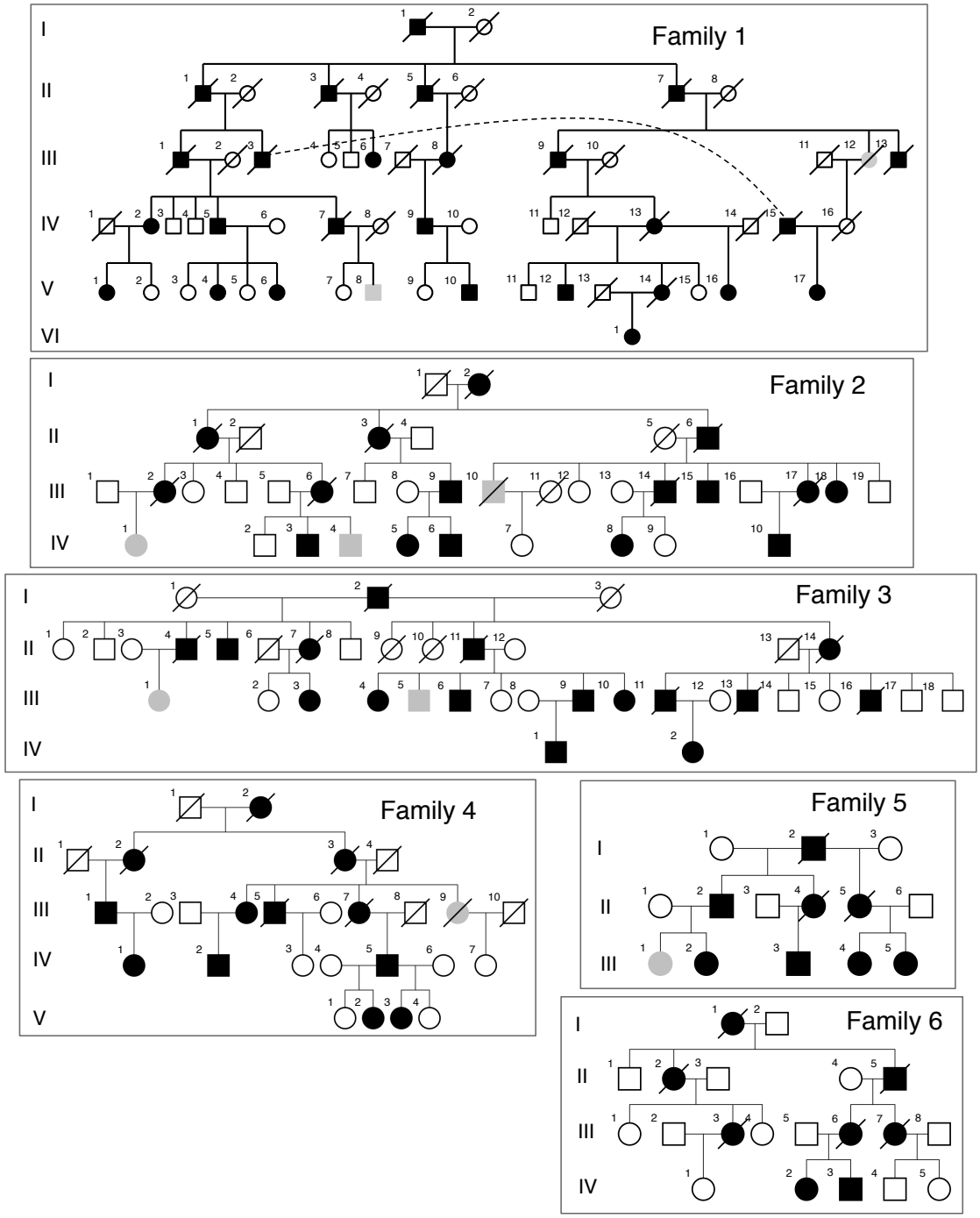
The steps performed were:

1. Combining transposon pairs. The two reads of a pair were merged if we could find an eight-base perfect overlap between them at the expected position on the reads. Otherwise, the two reads were used as input to the next step but not joined. At this stage we also trimmed off read tails having low quality scores and removed vector sequence.

2. Error correction. We developed a special error correction algorithm to account for PCR errors in a clone.

   (a) We found all L-mers in the reads, L = 21, discarding those for which the quality of the middle base was below 30. We then associated a count to the L-mer, namely the total number of instances in the reads, minus those in reads from the clone having the most instances. If the L-mer had a reduced count of at least two, we called it *good*.
   (b) Now again traverse the reads, looking at all L-mers that are not *good*. Consider the three possible L-mers obtained by changing the middle base. Suppose that only one of these is *good*, and that its reduced count is at least five. Then we made a 'recommendation' that the middle base be changed.
   (c) Traverse the recommendations. Carry them out, except in cases where two are within L/2 of each other. Set the quality score of changed bases to zero.
   (d) Now being a second phase of error correction by aligning the reads to each other, using only alignments for which the sum of the two longest perfect match lengths was at least 500.
   (e) For a given read, edit it, using the stack of reads aligned to it.

3. Joining error-corrected reads. Reads were formed into a graph by gluing them together along a minimum overlap of K=544. We then deleted all material in the graph that could not be found on a path from the first PCR primer to the second. Finally, we simplified the graph to reflect known limitations on the size of the PCR product.



Supplementary Figure 1. Linkage pedigrees. Depicted individuals are those who were (1) linkage-analyzed, (2) sequenced, (3) genotyped, or (4) otherwise required for pedigree connectivity of other included family members. Filled symbols indicate affected individuals. Unfilled symbols indicate unaffected individuals. Grey symbols indicate unknown affection status. To protect patient privacy we have altered some pedigree drawings by adding or removing up to two unaffecteds and assigning sex randomly. This does not matter for the analysis.

AAGGAGACTTCGGCTACCCAGAGAAGTTCAGTGCCCAGCTCTACTGAGAAGAATGCTGTG 1
AGTATGACCAGCAGCGTACTCTCCAGCCACAGCCCCGGTTCAGGCTCCTCCACCACTCAG 2
GGACAGGATGTCACTCTGGCCCCGGCCACGGAACCAGCTTCAGGTTCAGCTGCCACCTGG 3
GGACAGGATGTCACCTCGGTCCCAGTCACCAGGCCAGCCCTGGGCTCCACCACCCCGCCA 4
GGACAGGATGTCACCTCGGTCCCAGTCACCAGGCCAGCCCTGGGCTCCACCACCCCaCCA 4'
GCCCACGATGTCACCTCAGCCCCGGACAACAAGCCAGCCCCGGGCTCCACCGCCCCCCCA 5

GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCCCCCCA X
GCCCACGGTGTCACCTCGGCCCCGGAgAgCAGGCCGGCCCCGGGCTCCACCGCgCCCgCA A
GCCCACGGTGTCACCTCGGCCCCGGAgAgCAGGCCGGCCCCGGGCTCCACCGCCCCCCCA B
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCCCCCaA C
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCcGCCCCGGGCTCCACCGCCCCCCCA D
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCcGCCCCGGGCTCCACCGCgCCCgCA E
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCCCCaCA F
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCgCCCgCA G
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCgCCCCA H
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCgCCCCCA I
GCCCACGGTGTCACCTCGGCaCCGGAgAgCAGGCCGGCCCCGGGCTCCACCGCgCCCgCA J
GCCCACGGTGTCACCTCGGCCCCGGAgAgCAGGCCGGCCCtGGGCTCCACCGCCCCCCCA K
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCaCCCCCA V
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCCCCCCg W

GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCGGGCCCCGGGCTCCACCCCGGCCCCG 6
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCcGGCCCCGGGCTCCACCCCGGCCCCG 6'
GGCTCCACCGCCCCCCCAGCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCG 7
GGCTCCACCGCCCCCCCAGCCCATGGTGTCACCTCGGCCCCGGACAACAGGCCCGCCTTG 8
GGCTCCACCGCCCCTCCAGTCCACAATGTCACCTCGGCCTCAGGCTCTGCATCAGGCTCA 9

Supplementary Figure 2. Common 60-mer units found within and near the *MUC1* VNTR. Oriented relative to the coding strand of *MUC1* (negative strand of hg19), 60-mer units 1-5 (including variant 4') appear near the beginning of the VNTR, whereas 6-9 (and variant 6') appear near the end. The rest are in the middle, and are very similar to the 'canonical' unit X, with variant bases shown in lower case. The underlined base in structure 9 corresponds to hg19 chr1 position 155,160,963, and the underlined base in structure 1 corresponds to hg19 chr1 position 155,162,030.

(a) F4:IV-2 (identical to F4:V-3 assembly)

1-2-3-4-5-C-**X**-D-E-C-F-X-X-A-B-D-E-C-X-X-X-A-A-B-X-X-X-X-X-X-G-A-B-X-X-
X-X-X-X-V-6′-7-8-9

(b) F6:IV-3

1-2-3-4′-**5**-C-X-D-X-A-A-A-A-B-X-X-X-X-B-A-A-B-X-X-X-X-X-X-X-G-A-B-X-X-X-
X-X-X-X-X-6-7-8-9

(c) F2:IV-3

1-2-3-4-5-C-X-D-E-C-X-H-X-A-A-**B**-D-E-C-X-X-X-A-A-B-X-X-X-X-X-E-C-X-X-X-
A-A-B-**X-X-X-X-X-X-X-X-X-X-X-X-X-X**-V-X-A-J-B-X-X-X-X-X-X-V-6′-7-8-9

Supplementary Figure 3. Complete assemblies of risk alleles from three families. For families 2, 4, and 6 the structure of the risk *MUC1*-VNTR allele was determined almost exactly, and we are thus able to determine the position of the mutant repeat unit and its sequence context, both of which are different in all three cases. Each assembly is depicted as a series of 60-mer units (**Supplementary Fig. 2**) covering hg19 chr1 positions 155,160,963 to 155,162,030 (inclusive), and oriented relative to the *MUC1* coding strand (hg 19 negative strand). Units shown in red contain the insertion of an extra C into the C^7 sequence appearing at positions 53-59. For (a) and (b), the exact allele structure is shown, whereas for (c), the structure is completely determined except for the exact length of the stretch X-X-X-X-X-X-X-X-X-X-X-X-X-X (14 copies, shown in blue), whose predicted length from the gel size is 13.8 copies, and which, given limitations in gel measurement accuracy, could in fact be 13, 14 or 15 copies.

(a) CEPH mother (NA12892), short allele = CEPH child short allele

1-2-3-4-5-C-X-D-E-C-F-X-X-A-B-D-E-C-X-X-W-A-A-B-X-X-G-A-B-X-X-X-X-X-X-V-6′-7-8-9

(b) CEPH mother (NA12892), long allele

1-2-3-4-5-C-X-D-E-C-X-H-X-A-B-D-E-C-X-X-X-A-A-B-X-X-X-X-X-X-X-X-I-X-A-J-B-X-X-X-X-X-V-6′-7-8-9

(c) CEPH father (NA12891), short allele

1-2-3-4-5-C-X-D-E-C-F-X-X-A-B-D-E-C-X-X-X-A-A-B-X-X-X-X-G-A-B-X-X-X-X-X-V-6′-7-8-9

(d) CEPH father (NA12891), long allele = CEPH child long allele

1-2-3-4-5-C-X-D-E-C-X-X-X-A-A-B-X-X-X-X-X-A-A-A-B-X-X-X-X-A-A-X-X-X-X-X-X-X-A-A-B-X-X-X-X-B-B-X-X-A-A-B-X-X-X-B-A-A-X-X-X-X-X-X-V-6-7-8-9

(e) F1:IV-2 = F5:IV-5 = F6:IV-3= F6:IV-4 non-risk alleles

1-2-3-4-5-C-X-D-E-C-F-X-X-A-B-D-E-C-X-X-X-A-A-B-X-X-X-X-X-X-G-A-B-X-X-X-X-X-X-V-6′-7-8-9

(f) F1:V-6

1-2-3-4-5-C-X-D-E-C-X-X-X-A-A-B-X-X-X-X-X-X-G-A-B-X-X-X-X-X-X-G-A-B-X-X-X-X-X-X-V-6-7-8-9

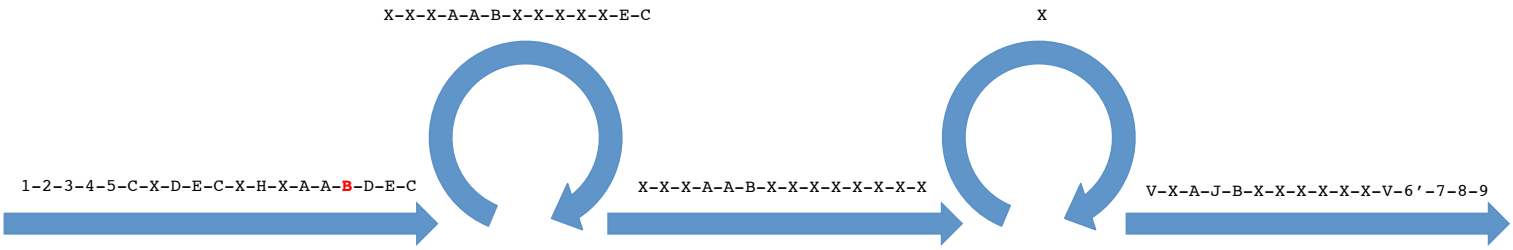
(g) F2:IV-5

1-2-3-4-5-C-X-D-E-C-X-H-X-A-A-K-D-E-C-X-X-X-A-A-A-B-X-X-X-X-V-6′-7-8-9

(h) F5:IV-1

1-2-3-4-5-C-X-D-E-C-X-X-X-A-A-B-X-X-X-X-X-A-A-A-B-X-X-X-X-A-A-X-X-X-X-X-X-X-A-A-B-X-X-X-X-B-X-X-A-A-B-X-X-X-X-B-A-A-X-X-X-X-X-G-A-B-X-X-X-X-X-X-V-6-7-8-9

Supplementary Figure 4. Complete assemblies of non-risk alleles. In several instances, the structure of the non-risk *MUC1*-VNTR allele was determined exactly. Each assembly is depicted as a series of 60-mer units (**Supplementary Fig. 2**) covering hg19 chr1 positions 155,160,963 to 155,162,030 (inclusive), and oriented relative to the *MUC1* coding strand (hg 19 negative strand).



Supplementary Figure 5. Ambiguous assembly of a risk allele. For individual F2:IV-3's risk allele, the assembly of amplicon sequence data yielded a graph as shown, with an edge count of 5. The semantics of such a graph is that the true allele should be represented by some path through the graph from beginning to end. There are infinitely many paths through this particular graph, depending on how many times each loop is traversed, however almost all are inconsistent with the approximate amplicon size as measured from the long-range PCR gel (4.1 kb). Indeed we reasoned that the only probable paths are those that traverse the first loop exactly 1 time, and the second loop between 5 and 7 times. This gives the same result as the assembly of the risk allele as depicted in **Supplementary Fig. 2 (c)**.

	Nationality/Ethnicity	*MUC1* mutation identified	Renal histopathology	Age of onset of end-stage kidney disease in affected family members	Generations affected	Number of affected individuals sequenced in whom a mutation was found	Number of unaffected individuals sequenced in whom a mutation was found
Initial scan: families with linkage to chromosome 1							
1	Middle Eastern	Yes	Focal interstitial fibrosis	29,32,33,34,37,40 ,54,60,63,69	5	10/10	0/18
2	African American	Yes	Patchy, mild interstitial fibrosis	43,43,45,57,60,61 ,64	4	8/8	0/10
3	Native American	Yes	Tubular atrophy and interstitial fibrosis	36,38,38,42,43,45 ,49,56,65,67	5	12/12	0/11
4	European American	Yes	Widespread tubular distension and tubular microcystic dilatation	45,50,50,60,65,79	4	4/4	0/5
5	European American	Yes	No biopsy available	22,22,23,23,33,33	4	5/5	0/2
6	European American	Yes	Marked tubular atrophy	27,31,32,40,44	6	4/4	0/7
Mutational analysis of additional families/cases (none with linkage data except Finnish family 6[4])							
1	European American	Yes	Chronic tubulo-interstitial fibrosis	32	2	3/3	0/0
2	European American	Yes	Patchy tubular atrophy and interstitial fibrosis	75,75	5	3/3	0/0
3	European American	Yes	Mild interstitial fibrosis	43,44	4	1/1	0/0
4	European American	Yes	No biopsy	27,32,31,27,39	5	1/1	0/0
5	African American	Yes	No biopsy	26,30,39,49	3	2/2	0/0
6	Finnish	Yes	Tubulo-interstitial disease	52,52,53,53	5	1/1	0/0
7	European American	Yes	No biopsy	21,27,30,34,34,35 ,36	4	1/1	0/0
8	European American	Yes	No biopsy	28,35	2	1/1	0/0
9	European American	Yes	Severe chronic tubulo-interstitial disease	39	2	1/1	0/0
10	European American	Yes	No biopsy		2	1/1	0/0
11	Australian	Yes	No biopsy	16,23,25,28,31,31	3	3/3	0/0
12	European	Yes	Tubulo-interstitial disease, positive staining for mutant *MUC1*		1	1/1	0/0
13	European	Yes	Tubulo interstitial disease, positive staining for mutant *MUC1*	30,44,50	3	1/1	0/0
14	European American	No	No biopsy	22,55,56	2	0/1	0/0
15	European American	No	Chronic tubulo-interstitial disease	40,40,50	2	0/1	0/0
16	European American	No	Moderate patchy interstitial fibrosis	37,45,55	3	0/2	0/0
17	Canadian	No	No biopsy	45,54,64,65,79,85	3	0/1	0/0
18	European American	No	No biopsy		1	0/1	0/0
19	European American	No	No biopsy	64	2	0/1	0/0
20	European	No	No biopsy		1	0/1	0/0
21	European	No	No biopsy		1	0/1	0/0

Supplementary Table 1. Clinical characteristics of studied families.

Family	ID	Affected	Relationship	Linkage sharing across VNTR	Linkage haplotypes across VNTR	VNTR size in bases	Edge count in VNTR assembly	+C found in sequence data	+C found by genotype assay
1	IV-2	Y	aunt/ nephew	IBD1	F1-a	2,640	1	N	Y
					F1-b (risk)	(4.8kb)	12	Y	
	V-6	Y			F1-c	2,700	1	N	Y
					F1-b (risk)	(4.8kb)	50	Y	
2	IV-5	Y	second cousins	IBD1	F2-a	2,100	1	N	Y
					F2-b (risk)	(4.1kb)	19	Y	
	IV-3	Y			F2-c	(2.2kb)	13	N	Y
					F2-b (risk)	(4.1kb)	5	Y	
3[1]	III-13	Y	full sibs	IBD1	F3-a (risk)	(4.8kb)	13	Y	Y
					F3-b	(4.8kb)	n/s?	?	
	III-16	Y			F3-c	(4.0kb)	47	N	Y
					F3-a (risk)	(4.8kb)	39	Y	
4	V-3	Y	first cousins once removed	IBD1	F4-a (risk)	2,641	1	Y	Y
					F4-b	(4.3kb)	n/a	?	
	IV-2	Y			F4-a (risk)	2,641	1	Y	Y
					F4-c	(4.5kb)	n/a	?	
5[2]	IV-5	Y	half first cousins	IBD1	F5-a	2,640	1	N	Y
					F5-b (risk)	(4.7kb)	33	Y	
	IV-1	Y			F5-c	4,680	1	N	Y
					F5-b (presumed risk)	(4.7kb)	n/s?	?	
6[3]	IV-3	Y	first cousins	IBD0	F6-a	2,640	1	N	Y
					F6-b (presumed risk)	2,641	1	Y	
	IV-4	N			F6-c	2,640	1	N	N
					F6-d	(4.6kb)	12	N	
CEPH	mom	N			CEPH-a	2,400	1	N	N
					CEPH-b	2,940	1	N	
	dad	N			CEPH-c	2,580	1	N	N
					CEPH-d	4,320	1	N	
	child	N			CEPH-a	2,400	1	N	N
					CEPH-d	4,320	1	N	

Supplementary Table 2. Summary of *MUC1*-VNTR assemblies. Knowledge of IBD sharing between/among sequenced individuals and the segregation of different phased SNP-marker haplotypes across the VNTR region were used to assign the sequenced MUC1 alleles to the different observed categorical linkage haplotypes, where possible. Furthermore, where able, we have assigned VNTR alleles to a pedigree's risk haplotype. The reported size of each allele's assembly covers hg19 chr1 155,160,963 to 155,162,030 (inclusive). Numbers in parentheses are estimated sizes derived from long-range PCR gels in those cases where the allele was not assembled or did not assemble into a single unambiguous solution with an edge count of 1. Alleles with edge counts of "n/a" were not assembled, and alleles with edge counts of "n/s?" are believed to have not been assembled due to inadequate separation from the individual's other allele prior to molecular cloning.

Name	Method	Comment	Primer sequence
PS1	*MUC1*-VNTR Southern blot	biotinylated	/52-Bio/ CAGCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGC CCCGGGC /iBiodT/ CCACCGCCCCCCCCAGCCCACGGTGTCACC /iBiodT/ CGGCCCCGGACACCAGGCCGGC
PS2	*MUC1*-VNTR long-range PCR	forward	GGAGAAAAGGAGACTTCGGCTACCCAG
PS3		reverse	GCCGTTGTGCACCAGAGTAGAAGCTGA
PS4	Sanger sequencing of clones	forward	GGGTGCGCATGATCCTCTAGAGT
PS5		reverse	TAAATTGCACTGAAATCTAGAAATA
PS6	C-insertion genotype assay	forward	CTGGGAATCGCACCAGCGTGTGGCCCCGGGCTCCACC
PS7		reverse	CGTGGATGAGGAGCCGCAGTGTCCGGGGCCGAGGTGACA
PS8		extension	CGGGCTCCACCGCCCCCCC

Supplementary Table 3. Primer sequences.

References

1. Kiser, R. L. *et al.* Medullary cystic kidney disease type 1 in a large Native-American kindred. *Am. J. Kidney Dis.* **44**, 611–617 (2004).
2. Kimmel, R. J. *et al.* Cosegregation of bipolar disorder and autosomal-dominant medullary cystic kidney disease in a large family. *Am J Psychiatry* **162**, 1972–1974 (2005).
3. Kenya, P. R., Asal, N. R., Pederson, J. A. & Lindeman, R. D. Hereditary (familial) renal disease: clinical and genetic studies. *South. Med. J.* **70**, 1049–1051 (1977).
4. Auranen, M., Ala-Mello, S., Turunen, J. A. & Järvelä, I. Further evidence for linkage of autosomal-dominant medullary cystic kidney disease on chromosome 1q21. *Kidney Int.* **60**, 1225–1232 (2001).
5. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
6. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
7. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
8. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
10. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).