nature genetics

Mutations causing medullary cystic kidney disease type 1 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 Vyleťal[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[17,18], Eric S Lander[1,18] & Mark J Daly[1,2,18]

Although genetic lesions responsible for some mendelian disorders can be rapidly discovered through massively parallel sequencing 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 did we find that each of six families with MCKD1 harbors an equivalent but apparently independently arising mutation in sequence markedly under-represented in massively parallel sequencing data: the insertion of a single cytosine 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 variable-number tandem repeat (VNTR) sequence in the *MUC1* gene encoding mucin 1. These results provide a cautionary tale about the challenges in identifying the genes responsible for mendelian, let alone more complex, disorders through massively parallel sequencing.

MCKD1 (MIM 174000) is a rare disorder characterized by autosomal dominant inheritance of tubulointerstitial kidney disease[1]. Affected individuals variably require dialysis or kidney transplantation in the third to seventh decade of life. Diagnosis of MCKD1 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 (refs. 3–7). Attempts to identify the mutated gene(s), however, have not been successful[4].

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

Linkage analysis was performed on six likely MCKD1 pedigrees (Online Methods, **Supplementary Fig. 1** and **Supplementary Table 1**),

[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, 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, Israel. [7]Institut National de la Santé et de la Recherche Médicale (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]The Commonwealth Medical College, Scranton, Pennsylvania, 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 Medical Center, 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, Sweden. [16]Office of the Clinical Director, National Human Genome Research Institute, US National Institutes of Health (NIH), Bethesda, Maryland, USA. [17]Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA. [18]These authors jointly directed this work. Correspondence should be addressed to A.J.B. (ableyer@wfubmc.edu), E.S.L. (lander@broadinstitute.org) or M.J.D. (mjdaly@atgu.mgh.harvard.edu).

Received 3 May 2012; accepted 7 January 2013; published online 10 February 2013; doi:10.1038/ng.2543

and in all families the phenotype showed perfect cosegregation with a single 2-Mb segment of chromosome 1 (**Fig. 1**). We examined the genotype data for evidence of copy number variation (CNV) in the relevant interval but found only two common copy number polymorphisms, neither of which segregated with disease. Looking at the longest stretches of allelic identity in pairwise comparisons of the phased risk-associated haplotypes from the pedigrees, we also found no obvious ancestral haplotype shared by a significant fraction of the families (beyond the background linkage disequilibrium (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-associated haplotype) across the linkage region. In protein-coding regions, we found only two rare (frequency of <1% in 1000 Genomes Project Phase 1 data[10]), non-silent point variants (SNPs or small insertions/deletions) shared by both affected individuals in any pedigree: each was in a different gene and in a different pedigree. This finding is consistent with the expected background rate of variation in 75 genes in 6 independent risk-associated 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 of other mendelian diseases, we had expected to find a gene harboring rare coding variants in multiple families. We found no such evidence.

We next examined the noncoding regions but found no regional clustering of segregating rare variants. We searched for any large structural variation (hundreds of bases or larger in size), such as deletions, insertions, duplications and inversions. All variants identified in this manner either did not segregate with disease or were found at appreciable frequencies in control populations.

On the basis of these findings, we concluded that the causal mutation(s) in MCKD1 were either located in a subregion that was recalcitrant to sequencing or represented a new mutational mechanism. We considered the possibility that MCKD1 might be caused by 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 identify 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 the 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 determined to either be non-polymorphic or to show no notable expansion in affected individuals (relative to the corresponding alleles of family members not sharing the risk-associated haplotype and Centre d'Etude du Polymorphisme Humain (CEPH) family samples) on the basis of direct assays of length by PCR.

The *MUC1* VNTR was particularly difficult to assay: it comprises many (20–125) copies[13,14] 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 case samples, VNTR lengths were consistent with published descriptions and were not expanded on risk-associated chromosomes, excluding VNTR length as pathogenic. *MUC1* remained particularly noteworthy as the only gene in the linked region that has transcripts with kidney-specific expression, as determined using RNA sequencing

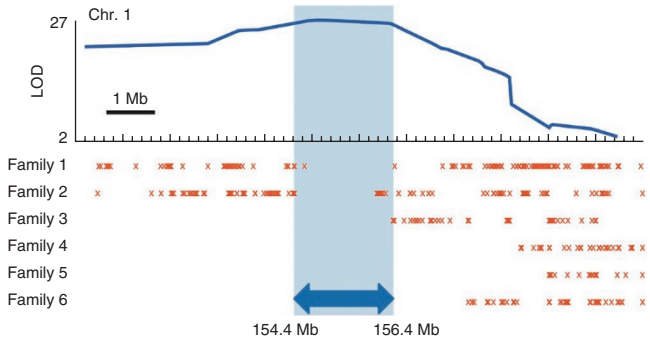

Figure 1 Linkage of six families with MCKD1 to chromosome 1. The logarithm of odds (LOD) curve shows the combined linkage score of six MCKD1 pedigrees across 12 Mb of chromosome 1, with the peak score well above the threshold of 3.6 for genome-wide significance[19]. Red X's mark the locations of opposite-allele homozygous genotype calls between affected members within each pedigree (Online Methods) and highlight regions where affected individuals de facto share no alleles identity by descent (IBD), thereby delineating genomic segments unlikely to harbor causal variation. The shaded region (chromosome 1: 154,370,020–156,439,000; hg19) was considered most likely to contain any causal mutations, bounded on each side by recombination breakpoints in two different pedigrees.

(RNA-seq) in 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 (among 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 the release of the extracellular domain, which upon release binds noncovalently to the transmembrane domain[15,16] (**Fig. 2a**).

We considered the possibility that MCKD1 might be caused by point mutations in the *MUC1* VNTR missed owing to poor sequence coverage because (i) it was excluded from whole-exome and regional-capture probes owing to its low complexity and extreme sequence composition (and also because it is rarely annotated as a coding sequence) and; (ii) it was markedly under-represented in quality-filtered data from the whole-genome sequence, likely owing to its high GC richness and homopolymer content. Because the human reference sequence seemed to substantially under-represent this region (hg19 predicts a VNTR length far smaller than the published range or that observed in any of our samples, including controls), we proceeded 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; examples shown in **Fig. 2b**,**c**).

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

The frameshift caused by the insertion is predicted to result in a mutant protein that contains many copies of a novel repeat sequence

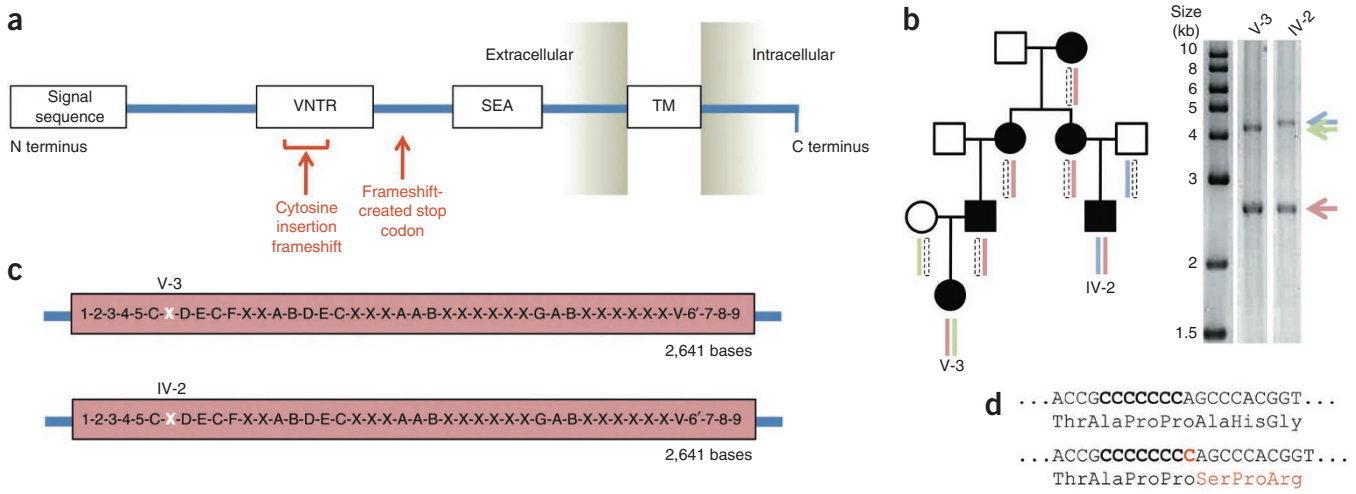

Figure 2 Discovery of a cytosine insertion in a coding VNTR of *MUC1*. (**a**) The major domains of the full-length MUC1 precursor protein are shown (TM, transmembrane domain). Based on fully and unambiguously assembled VNTR alleles, the frameshift caused by insertion of a cytosine in the coding strand is expected to introduce a premature stop codon shortly beyond the VNTR domain. (**b**) Where possible, we used knowledge of segregating phased SNP marker haplotypes to select for *de novo* VNTR sequencing and assembly those individuals sharing only a single haplotype across the region, as this aided identification of the VNTR allele segregating with the shared risk haplotype. (**c**) Independent *de novo* assembly of the shared VNTR allele in 2 individuals from family 4 shows exactly identical complete sequence, with the seventh 60-base unit (white X) out of 44 containing a cytosine insertion. The assembly is oriented relative to the coding strand of *MUC1* and covers bases 155,160,963–155,162,030 on chromosome 1 (hg19). Each unique 60-base repeat segment is represented by a different letter or number (**Supplementary Fig. 2**). (**d**) Translational impact of the cytosine insertion frameshift.

(obtained by shifted translation of the VNTR) but lacks, owing to a premature stop codon shortly beyond the VNTR terminus, the downstream SEA self-cleavage module and both the transmembrane and intracellular domains characteristic of the wild-type MUC1 precursor protein (**Fig. 2a**).

Because discovery of the cytosine insertion required considerable labor and time, we sought to develop a simple and robust genotyping assay to enable the screening of larger populations. We designed a probe-extension assay (**Fig. 3a** and Online Methods) 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 to reference repeat units.

We typed all samples collected from the six families with MCKD1 used for linkage analysis, including 62 phenotypically affected and 79 unaffected relatives (**Fig. 3b**,**c**), and over 500 control individuals from CEU (Utah residents of Northern and Western European ancestry), Japanese, Chinese, Yoruba and Tuscan HapMap 3 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-associated haplotype and yet was not seen in any of the 500 HapMap samples.

Overall, the genotyping results provide strong evidence that the cytosine insertion is the high-penetrance genetic lesion that leads to the development of MCKD1. As a statistical association, the significance

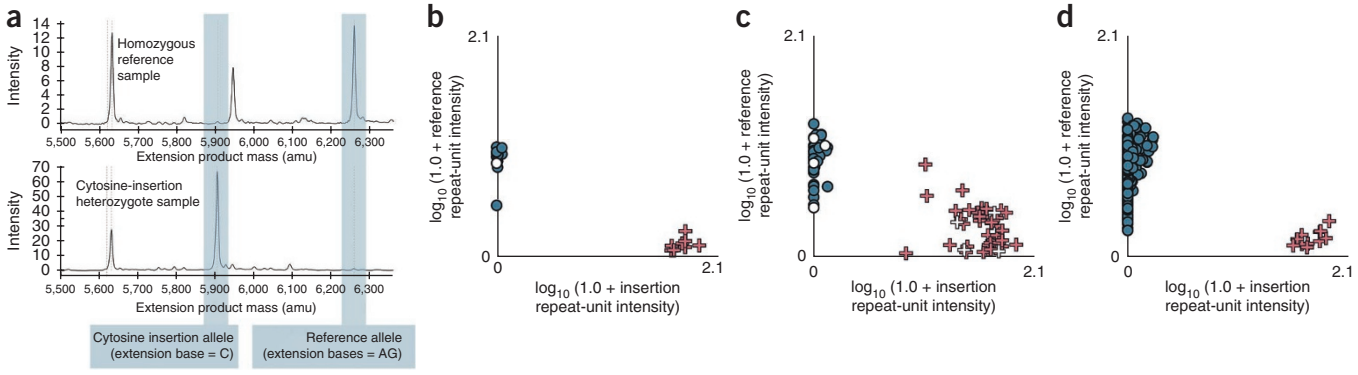

Figure 3 Detection of *MUC1* cytosine insertion by probe-extension assay. (**a**) Representative electropherograms for the probe-extension assay for the *MUC1* VNTR cytosine insertion (Online Methods) performed on homozygous reference allele and heterozygote samples. (**b**–**d**) Allele intensity scatter plots. The *x*-axis values correspond to the detected intensity at the mass of the probe-extension product for the cytosine insertion. The *y*-axis values reflect that of the extension product for the reference repeat unit. Coloring reflects MCKD1 status: blue, unaffected (or HapMap sample); red, affected; white, unknown. Individuals known to carry the risk haplotype identified by linkage analysis are represented by plus signs, whereas others are represented by circles. Samples having log-transformed intensities below 0.25 for both alleles were excluded as failed assays. Samples from whole-genome amplification and those with low DNA concentration were also excluded for underperforming. (**b**) Allele intensity scatter plot for large linkage family 2. (**c**) Allele intensity scatter plot for all families from the linkage analysis. (**d**) Allele intensity scatter plot for HapMap samples together with selected positive controls (individuals with MCKD1 known to carry the cytosine insertion).

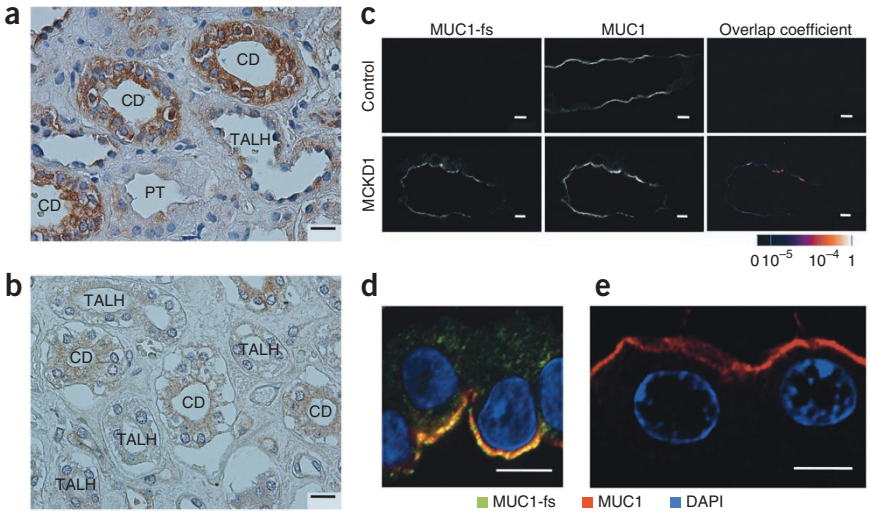

Figure 4 Immunohistochemical and immunofluorescence studies of the MUC1-fs protein. In individuals with MCKD1, MUC1-fs is expressed and present in renal epithelial cells from the Henle's loop, distal convoluted tubule and collecting duct. (**a**) Strong intracellular staining of MUC1-fs protein in an individual with MCKD1. (**b**) Absence of specific staining in a control kidney. 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 wild-type MUC1 in the collecting duct of an individual with MCKD1. MUC1-fs staining is absent in control, and colocalization with wild-type MUC1 is therefore not detected. The values for fluorescent signal overlap were transformed to a pseudocolor scale. (**d**) Immunofluorescence analysis showing the different intracellular localizations and partial submembrane colocalization of MUC1-fs and wild-type MUC1 proteins in the collecting duct of an individual with MCKD1. Note specific staining of both forms in distinct membrane microdomains. (**e**) Absence of MUC1-fs staining and characteristic membrane localization of wild-type MUC1 in control kidney. Scale bars, 25 µm in **a**,**b** and 50 µm in **c**–**e**.

of this observation can only be approximated, but its *P* value is clearly far less than the reciprocal of the number of bases in the genome (the cytosine insertion was seen on 6/6 risk-associated chromosomes versus 0/1,000 HapMap chromosomes). Furthermore, this observation is robust to population structure considerations, as the mutations have arisen independently.

To explore the broader impact of *MUC1* mutations, we genotyped affected and unaffected individuals from 21 additional small families with MCKD shown to be negative for known MCKD-related mutations in screening (**Supplementary Table 1**), only one family of which had existing linkage information implicating the 1q21 region[17]. In 13 of 21 families, we found the presence of a cytosine insertion consistent with this mutation being a fully penetrant cause of disease, indicating a substantial role for *MUC1* in MCKD1-like phenotypes.

Using antibodies raised against a peptide synthesized to correspond to the predicted mutant VNTR sequence, we found specific intracellular staining in epithelial cells from the Henle's loop, distal tubule and collecting duct of individuals with MCKD1 (**Fig. 4a**), which was absent in control kidney (**Fig. 4b**). Additionally, costaining of tissues from affected individuals and controls with antibody against wild-type MUC1 showed the specificity of the antibody to the mutant MUC1 protein (termed MUC1-fs), with diffuse and/or fine granular intracellular localization of the MUC1-fs protein in kidney from affected individuals (**Fig. 4c**) along with patchy colocalization of MUC1-fs and wild-type MUC1 signals on the apical membranes of collecting duct epithelial cells (**Fig. 4c**,**d**). Detailed image analysis of tissue from affected individuals (**Fig. 4d**) compared to that from controls (**Fig. 4e**) detected no intracellular colocalization of MUC1-fs and wild-type MUC1 proteins in tissue from affected individuals but showed puncta of colocalization in distinct plasmalemmal subdomains. Antibody to MUC1-fs did not stain normal kidney tissue.

This study highlights the fact that current massively parallel sequencing technology may not always suffice to identify disease-relevant mutations, even when linkage analysis conclusively pinpoints a region of a few megabases in size. Even if the cytosine insertion event were not considerably under-represented in the quality-filtered massively parallel sequencing 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 owing to (i) the underlying variability of VNTR size within and across individuals, (ii) the inability to uniquely place reads within the VNTR given current massively parallel sequencing read lengths and (iii) the fact that the mutant/reference allelic balance is skewed far from the expected 1:1 ratio of a typical heterozygous variant.

The precise nature of the MCKD1-causing mutations is notable. Each independently arising event is essentially an 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 the introduction of a premature stop codon. Possible explanations for the consistently observed mutation include the possibilities that (i) this insertion event is strongly favored owing to mutational mechanism; (ii) other events (for example, single-cytosine deletion) are selected against; (iii) other events (for example, single-cytosine deletion) are benign and do not cause MCKD1; and (iv) other *MUC1* mutations exist but are undersampled here.

The identified mutation and the associated genotyping assay provide a tool for screening younger members of families in which MCKD1 has previously been diagnosed, as well as a diagnostic tool for sporadic cases. They also alleviate the challenge of 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 elucidate the specific mechanism of pathogenesis of the MUC1 mutant protein. We note that knockout studies in mice indicate that the *Muc1* gene is not essential[18], supporting 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 the expression of the *MUC1* gene or splice out its single VNTR-encoding exon.

URLs. International HapMap Project, http://hapmap.org/; sex-averaged recombination positions and population-specific allele frequencies, http://www.affymetrix.com/support.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank T.L. Hatte for reagent use. We thank D. Altshuler, T. Carter and J. Schlondorff for useful discussions and M. Cortes, M. Ilzarbe and M. Betancourt for helpful project management. We also thank F. Letendre, M. Coole, R.P. Frere, C. Bonnet, L. Mulrain, N. Norbui and H. Arachchi for Sanger sequencing. This work was conducted as part of the Slim Initiative for Genomic Medicine, a joint United States–Mexico project funded by the Carlos Slim Health Institute. This research was supported in part by the Intramural Research Program of the US NIH, National Human Genome Research Institute (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 the Ministry of Health of the Czech Republic. S.L.A. was supported by US NIH grant 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 the Human Frontier Science Program, Alon, the Israeli Centers of Research Excellence (I-CORE) and the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

AUTHOR CONTRIBUTIONS

A.J.B., E.S.L. and M.J.D. jointly supervised the research. R.J.X., M.R.P. and S.L.A. provided study design and interpretation advice. C.A., S.J.S., P.S.H. and A.J.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. on the basis of previous work by P.S.H. A.K. and M.J.D. developed variation discovery and analysis methods. A.K., J.T.R. and R.E.H. analyzed structural variation. T.G. performed CNV analysis. S.G. supervised the sequencing. S. Sigurdsson and K.L.-T. designed the custom capture array. M.P. performed direct PCR of the 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.B.J. assembled and analyzed VNTR Sanger sequencing. M.G. provided RNA-seq support. S.K. supervised the immunohistochemistry and immunofluorescence work performed by V.B., H.H., J.S. and P.V. A.K., B.B. and M.D. developed the C-insertion genotype assay. M.C.Z. provided informatic and sequencing consultation. A.R. provided informatic and analysis consultation. C.Y., J.T.R., M.N.C., I.G.-V., R.E.H. and E.R. provided informatic support. The manuscript was written primarily by A.K., A.G., A.J.B., E.S.L. and M.J.D. The supplementary information was prepared mainly by A.K., A.G., D.B.J., B.B., R.E.H., S. Sigurdsson, S.K. and A.J.B.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/ng.2543.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

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–S55 (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. Fuchshuber, A. *et al.* Refinement of the gene locus for autosomal dominant medullary cystic kidney disease type 1 (MCKD1) and construction of a physical and partial transcriptional map of the region. *Genomics* **72**, 278–284 (2001).
6. 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).
7. Wolf, M.T. *et al.* Telomeric refinement of the MCKD1 locus on chromosome 1q21. *Kidney Int.* **66**, 580–585 (2004).
8. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* **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. 1000 Genomes Project Consortium. 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. Brayman, M., Thathiah, A. & Carson, D.D. MUC1: a multifunctional cell surface component of reproductive tissue epithelia. *Reprod. Biol. Endocrinol.* **2**, 4 (2004).
16. Levitin, F. *et al.* The MUC1 SEA module is a self-cleaving domain. *J. Biol. Chem.* **280**, 33374–33386 (2005).
17. 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).
18. 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).
19. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).

ONLINE METHODS

Family collection and criteria for diagnosis of affected status. The six analyzed families with autosomal dominant tubulointerstitial kidney disease were selected from among a larger group referred for evaluation. Each showed a clinical phenotype highly suggestive of MCKD1 and lacked *UMOD* and *REN* mutations. All had previously demonstrated evidence of disease 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 the **Supplementary Note**.

Linkage and CNV analysis. Family members were genotyped on the Affymetrix 6.0 platform. Whole Affymetrix arrays with samples with genotype call rates of <88% excluded from analysis, as were samples that yielded low optical density measurements (indicating poor sample performance during laboratory steps). Further, markers were excluded for which probe sequences showed excess genomic homology or potential for significant guanine-quartet formation (those probe sequences for which either allele contained at least three consecutive guanines).

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 or branch, choosing single, well-typed, informative markers from LD-defined bins of SNPs on the basis of phased, population-specific HapMap data (release 22). Markers that showed no-call rates of >10% or any mendelian inheritance errors within a pedigree or branch were excluded from specific pedigree or branch analyses. Additionally, markers were required to be spaced at least 0.1 cM apart according to published sex-averaged recombination positions.

All expected intrapedigree relationships were confirmed from pairwise IBD estimates using PLINK software[20] and similarly derived marker sets; however, markers for PLINK were selected regardless of whether they were polymorphic within a pedigree or branch so as not to skew IBD calculations. Merlin software[21] was used to remove any likely genotyping errors that did not violate mendelian inheritance rules and then to perform parametric linkage under a rare, autosomal dominant model using population-specific allele frequencies.

Linkage mapping was performed using the Merlin package under a rare, autosomal dominant model. Scores were combined across pedigrees or branches by summing LOD values, linearly interpolating scores between marker locations as required. The consistency of the alleles carried on the segregating risk-associated 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 IBD alleles (by virtue of being homozygous for opposite alleles; for example, one having genotype AA and another CC). Such markers necessarily lie outside the disease-relevant linkage interval.

Affymetrix 6.0 intensity data were analyzed for CNVs using Birdsuite software[22].

Large-scale sequencing. Because the linkage region contained more than 170 separate transcript annotations comprising over 75 RefSeq genes, amplicon-based resequencing of genic regions was not initially considered. Of the 12 sequenced individuals, whole-genome sequencing was performed on 11 (~25-fold coverage on average), whole-exome sequencing was performed on 11 (~180-fold coding-sequence coverage on average) and regional-capture sequencing was performed on 5 (~220-fold coverage on average). Sequence processing is described in the **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 noncoding bases were covered at ≥10-fold in each pedigree.

We considered as a candidate for a pathogenic MCKD1 mutation any non-reference allele present in both affected individuals of any pedigree and also seen at a population frequency of ≤1% (ref. 10). Noncoding regions were analyzed similarly.

To discover potential structural variation at the chromosome 1 locus, we ran Genome STRiP[23] 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 HinfI (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 100-mer oligonucleotide probe PS1 (IDT; **Supplementary Table 3**) 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 0.2× SSC (30 mM NaCl, 3 mM sodium citrate, pH 7.0) and 0.1% SDS, membrane-bound biotin was detected with the BrightStar BioDetect kit (Ambion).

***MUC1* VNTR long-range PCR.** The long-range PCR protocol was adapted from a previous study[14]. Briefly, 7-µl reactions contained 15 or 30 ng of genomic DNA, 1.75 pmol of PS2 and PS3 primers (**Supplementary Table 3**), 5% DMSO, 625 µM of each dNTP, 1× reaction buffer with 3 mM $MgCl_2$ and 0.25 U DyNAzyme EXT DNA polymerase (Finnzymes). Thermocycling on GeneAmp 9700 instruments (ABI) included initial denaturation (90 s at 96 °C), 22 or 27 cycles (40 s at 96 °C, 30 s at 65 °C and 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-associated haplotype (examples shown in **Fig. 2b**). In a few cases, the sizes of the risk and non-risk VNTR alleles were nearly the same, precluding physical separation of the two alleles before molecular cloning. Using transposon hopping and capillary sequencing, we then sequenced clones from each allele (**Supplementary Note**).

Because the VNTR 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 true 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.

The key properties of the assemblies (examples shown in **Fig. 2c**) are summarized in **Supplementary Table 2**, and the sequences for the unique alleles (three risk and eight non-risk) where the assembly was fully or almost fully resolved are provided in **Supplementary Figures 3** and **4**. The notation for graph assembly in a scenario where an allele could not be fully and unambiguously reconciled is shown in **Supplementary Figure 5**. 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 the *MUC1* cytosine insertion event. Genomic DNA was first overly digested using restriction endonuclease MwoI, which selectively cleaves the reference repeat-unit sequence (GCCCCCCCAGC), while leaving intact repeat units containing the cytosine insertion (GCCCCCCCC**C**AGC, where the inserted cytosine appears in bold). 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 digested again with MwoI for a second round of enrichment. A 20-bp probe was designed corresponding to a sequence 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. After probe extension, reaction products were separated and sized by matrix-assisted laser desorbtion/ionization–time of flight (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 6,258.06 daltons).

doi:10.1038/ng.2543

Specifically, 100 µg of genomic DNA was digested in a 25-µl reaction volume for 16 h using 5 U of MwoI restriction endonuclease (NEB) with supplemental additions of 5 U of enzyme at hours 3 and 15. Digestion reactions were then cleaned using 50 µl of AMPure beads according to the manufacturer's protocol (Agencourt), and digested DNA was eluted in 25 µl of nuclease-free water. Remaining intact VNTR fragments were PCR amplified using 1× HotStart buffer, 1.0 mM $MgCl_2$ (to supplement the $MgCl_2$ already in the buffers), 0.8 mM dNTPs, 0.8 U 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 included one hold at 95 °C for 5 min, 45 cycles of 94 °C for 30 s, 67 °C for 30 s and 72 °C for 1 min, and one hold at 72 °C for 10 min. PCR products were cleaned using 50 µl of AMPure beads, and amplicons were eluted in 25 µl of nuclease-free water. A second round of MwoI digestion was performed, again for 16 h with 5 U of enzyme added at hours 0, 3 and 15. Digestion reactions were cleaned using 50 µl of AMPure beads, and products were eluted in 6.2 µl of nuclease-free water.

Using 5.2 µl of the digested eluate as template, probe extension was performed using 1× HotStart buffer, 0.6 mM $MgCl_2$ (to supplement the $MgCl_2$ already in the buffers), 1.7 µl of SAP buffer (Sequenom), 0.2 mM each of ddGTP, ddCTP and dATP nucleotides, 0.7 U of Thermo Sequenase DNA polymerase (Amersham) and 0.6 µM 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 included one hold at 94 °C for 2 min, 55 cycles of 94 °C for 5 s, 52 °C for 5 s and 72 °C for 5 s, and one hold at 72 °C for 7 min. Reactions were then desalted by the addition of a cation-exchange resin, and ~7 nl of purified extension reaction was spotted onto a SpectroChip (Sequenom) containing matrix 3-hydroxypicolinic acid. Arrayed reactions were analyzed by MALDI-TOF on a Compact mass spectrometer (Sequenom/Bruker).

Assay results were clear enough to assign genotypes on the basis of simple inspection of *x-y* scatter plots depicting log-transformed reference and mutant repeat-unit intensities ($\log_{10}$ (1.0 + peak height)). Assays for samples showing log-transformed intensities of <0.25 for both alleles were considered to have failed. Similarly, results from samples from whole-genome amplification or those with low DNA concentrations were typically considered unreliable and were 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; diluted 1:1,000 in 5% BSA in PBS). Wild-type MUC1 protein was detected with monoclonal mouse antibody to human epithelial membrane antigen (EMA; Dako, M0613; diluted 1:400 in 5% BSA in PBS). Detection of bound primary antibody was achieved using either the EnVision+ 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 analyzed. 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 antibody to EMA was diluted 1:10 in 5% BSA in PBS. Fluorescence detection used species-specific secondary antibodies: Alexa Fluor 488–conjugated goat secondary antibody to rabbit IgG and Alexa Fluor 568–conjugated goat secondary antibody to mouse IgG (Molecular Probes, Invitrogen). Nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI). Prepared slides were mounted in Immu-Mount fluorescence mounting medium (Shandon Lipshaw) 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 (40×, numerical aperture (NA) of 1.30), laser lines of 488 nm and 543 nm and band-pass filters of 515 ± 15 nm and 590 ± 15 nm. Images were deconvolved using the classic maximum-likelihood restoration algorithm in Huygens Professional Software (SVI). Colocalization maps employing single-pixel overlap coefficient values ranging from 0–1 were created using Huygens Professional Software.

20. 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).
21. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardonl, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
22. 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).
23. 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).