Supplemental Material

1. SUPPLEMENTAL METHODS

Supplemental Tables describing the Methods are embedded in the text below. Supplemental Tables of Results are in the Supplemental Results section.

1.1. Additional Cohort Information

1.1.1. FinnDiane cohort

The Finnish Diabetic Nephropathy Study (FinnDiane) is a nationwide Finnish multicenter study of individuals with T1D[1]. The participants were recruited to the study by their attending physician who performed a clinical examination and filled a questionnaire on the medical status of the individual. The participants were invited to one or more follow-up visits with a similar setting. Finnish hospital discharge registry and the individuals' medical records were used to gather additional health related information. Additional individuals were included to the FinnDiane study through collaboration with the Finnish National Institute for Health and Welfare and the Hospital Discharge Registry and medical records were used to collect health related information for these individuals. For this study, individuals were selected if the age at diabetes diagnosis was ≤ 40 and insulin treatment initiated within one calendar year from diabetes onset. In addition they were required to have macroalbuminuria which was defined by urine albumin excretion rate (AER) > 200 µg/min or > 300 mg/24h or urine albumin to creatinine ratio (ACR) > 25 mg/mmol for men and > 35 mg/mmol for women in at least two out of three consecutive overnight, 24-hour or spot urine collections. The eGFR measurements were collected starting from the first known year when macroalbuminuria was present. The follow-up was started from the first eGFR measurement in CKD stage 2 (eGFR: 60-90 ml/min/1.73m2), if available. If the individual did not have any measurements in CKD stage 2 but had measurements in CKD stage 1 (eGFR: > 90 ml/min/1.73m2), the follow-up was started from the last eGFR measurement in CKD stage 1. ESKD was defined as ongoing dialysis treatment, kidney transplantation or eGFR ≤ 10 ml/min/1.73m2. All the eGFR measurements following ESKD were removed from the analysis and eGFR was set to be 10 ml/min/1.73m2 at the time of the ESKD event. An individual was included in the analysis if the first eGFR ≥ 30 ml/min/1.73m2, and the eGFR follow-up time was ≥ 3.5 years if no ESKD was observer or ≥ 1 year if ESDK was observed. For the purposes of the genetic analysis first-degree relatives were removed.

1.2. Amplicon Design for Resequencing

Initially, 45 target inclusion regions for resequencing were identified on chr1 (*FHAD1* gene) and chr19 (*HSD17B14* gene) to include the perigenic vicinity of these two genes, coding and non-coding regions, including intronic, and the 5' or 3' flanking genome regions. The *FHAD1* data is not included in this report. One region on chrY (SRY gene) was also added for quality control to identify possible sample misclassification or mix-up through matching inferred biological sex with study specified gender. Illumina concierge

design services optimized the amplicon selection to achieve as close to the desired genomic target coverage as possible but minimizing possible multiplex cross-hybridization of primers because of sequence repeat structure in the target regions.

The final design contained 151 amplicons to tile the regions with a range of amplicon sizes from 225 to 275bp, Supplemental Methods Table 1.

Chr	Gene	N (Amplicons)	N (Contiguous Intervals)	Total Coverage (bp)
1	*FHAD1* (not reported here)	56	35	12,597
19	*HSD17B14*	91	22	16,874
Y	*SRY* (gender QC)	4	1	840

Supplemental Methods Table 1. Resequencing amplicon coverage.

Initial mapping and design of amplicons was performed on hg37 but was remapped to hg38 using the UCSC liftover utility before the sequence and bioinformatics analysis of the data.

1.3. Samples Included for Sequencing

We used the results from the prior GWAS QC analysis to identify samples to drop based on:

- Ancestry outliers
- Cryptic duplicates or known/cryptic MZ twins
- Cryptic or known close relatives (only degree 1, sibs or parent-offspring).

These dropped samples were not included in sequencing samples. One of each pair of duplicates or family of closely related samples was retained. We preferred to retain samples that had an ESKD event at a known time after enrollment to maximize study power. This applied to both prospective cohorts, or if the participant was retrospectively included in the cohort. The sample retention algorithm was to prefer to keep:

1. Sample that had a diagnosed or undiagnosed ESKD event with a time at which the event occurred after entry into the cohort. An undiagnosed ESKD event was defined as the first occurrence of GFR of 10 or less (however transient), without evidence of dialysis, transplantation, or clinically identified ESKD in the medical record.

2. Sample that had a complete data record, defined as Baseline eGFR (eGFR at time of entry into the cohort, either prospective or retrospective, study t=0), Gender; ESKD (Y/N), ESKDTIME (time after baseline when ESKD event occurred or censored time if sample did not have an event).

3. If neither 1. nor 2. distinguished the samples, prefer the sample with the longer follow-up time. eg if neither sample had an event, prefer the one with 5 years follow-up versus the one with only 2 years.

4. If 1. – 3. did not distinguish, keep the sample with higher variant call rate (more complete genotyping data).

Additional samples were included in the resequencing that did not meet the original inclusion criteria for the cohorts in this study at the time of the original array genotype-based gene aggregated exome-wide scan, but for whom DKD had subsequently progressed. For example, a T1D participant may have developed macroalbuminuria in the follow time period since the ascertainment and initial array analysis.

1.4. Amplicon Resequencing

The samples were barcoded and multiplexed into 33 separate Illumina MiSeq runs using 30 regular cartridges (15M reads), and 3 micro cartridges (4M reads). The first micro cartridge and 28 of the regular cartridges contained 95 samples, while the last two micro and four regular cartridges contained variable numbers of samples (min 10 to max 59) to sequence any remaining unsequenced samples and to increase read coverage for previous samples where the coverage was stochastically low from their first run, and where there was no evidence of poor DNA quality (very low read coverage from first run), or sample mix-up.

Sample Library Preparation Step	**N**
Total DNA library preparations (including 20 re-preps)	2456
Failed library preparations	22
Failed library re-preparations	20
Unique samples	2435
Sample Inclusion in Sequencing Runs	**N**
Samples with only 1 run	2129
Samples with 2 runs	282
Samples with 3 runs	21
Samples with 4 runs	3
Samples with <1 run	306
Unique Samples	2435
Total Sample-runs	2768

Supplemental Methods Table 2. Summary of the samples included and library preparations In the sequencing runs.

1.5. Bioinformatics Pipeline for Sequence Analysis

Gender discordant samples based on chromosome Y counts were retained in the pipeline since they were likely be the result of sample mislabeling, but still could assist variant calling if high quality samples. They were filtered after variant calling and

genotyping. Note that the processing was applied to both chr1 and chr19 amplicons together, not just for chr19 (*HSD17B14*).

1.5.1. Remove sample-runs with low read counts

Remove all FASTQC files from further processing where the total number of Fwd (R1) = Rev (R2) reads is < 3775 i.e. require minimum of mean 25x read pairs per amplicon, (151 amplicons in the set). Since all of these low sequence count sample-runs (38 in total) were repeated for each sample and had better coverage, this did not affect the overall sample inclusion.

1.5.2. Read Trimming

Trim the last base (151) from each read but do not trim further. Allow PEAR error correction to handle lower quality bases at end of reads. As expected, the Rev R2 reads were uniformly lower quality than Fwd R1.

1.5.3. Assemble Overlapping Fwd + Rev Reads into Contigs Using PEAR

Since the amplicons varied in size from 225 to 275 bp, and the read length was 150 after trimming, the Fwd and Rev reads overlapped by a minimum of 25 and a maximum of 75 base pairs. We used the Paired End Aligner (PEAR) to assemble the Fwd and Rev reads into contigs, thereby self-correcting discordant bases in the overlap regions. For reads that did not assemble with high quality, these were retained and inspected after alignment.

1.5.4. Align PEAR Contigs and Unassembled Reads to Genome

PEAR contigs and non-assembled reads were aligned genome hg38 using BWA-MEM v0.7.17. For the unassembled reads from PEAR, based on visual inspection using IGV, there were reads, both Fwd and Rev, that aligned perfectly, but their mate read did not. While Fwd reads are generally higher quality in Illumina sequencing, there were some Fwd reads that were soft or hard clipped and mapped with lower quality.

We filtered reads based on MAPQ=60 (highest BWA quality) and absence of the S (soft clip) or H (hard clip) char in the CIGAR strings. We then merged high quality assembled PEAR contigs and unassembled reads in BAM files for each sample in a run and use GATK to pre-process read group tags.

1.5.5. Merge All Read Alignments for a Sample Across Runs

We used GATK to merge the BAM files for the same sample included in more than one sequencing run. After merging all BAM, we retained 2,422 unique samples.

1.5.6. Recalibrate Base Quality Scores

We used the GATK best practices protocol step to generate BQSR recalibration table using all sample BAM files and then applied the learned BQSR table to recalibrate the base qualities in every sample BAM.

1.5.7. Generate Sample GVCF

We used the GATK HaplotypeCaller to call the GVCF for each sample from the recalibrated BAM files as per the GATK best practices.

	Variants Dropped		**Variants Remaining**	
QC Step	Multi-Allele	Bi-Allele	Multi-Allele	Bi-Allele
Initial Set	-	-	1583	1980
Genotype Level				
Set to missing any genotype: DP < 8 OR GQ <20	-	-	1583	1980
Variant level Check				
P-value (HWE) < 5 x 10^{-5} (1)	57	62		
MeanGQ < 35 (2)	4	6		
Call Rate < 0.5 (3)	249	364		
Totals (1-3)	298	418	1285	1562
GATK hard filters: MQ < 40 OR MQRS < -12.5 OR RPRS < -8 (4)	58	78		
Totals (1-4)	337	473	1246	1507

Supplemental Methods Table 3. Variant filtering QC showing variants dropped and retained at each step.

1.5.8. Merge Sample GVCFs

We used GATK to merge all sample GVCFs into an overall study-wide VCF file and perform joint sample calling and genotyping. We generated both Multi-allelic and Bi-allelic VCF files.

1.5.9. Variant Filtering

Because of the small number of variants called across all contiguous regions, we were unable to use the GATK best practices Variant Quality Score Recalibration method for variant filtering. This method uses a machine learning approach based on a large training set of known variants. Therefore, we defaulted to using hard filter criteria to identify the variants most likely to be poor. The GATK development team have published a set of hard filter criteria based on quality statistics from prior projects but these were not devised for a multiplexed competitive amplicon-based resequencing project with highly variable read depths and with some variants covered by very large read depths (DP=2,908 to 8.7M). (https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set). In lieu of this, we developed a pragmatic filtering process based on criteria from previous studies and GATK best practices, such that the extreme read depths would not unduly bias the variants to be filtered.

Carson et al found that even if the Broad VQSR best practices method is used, pre-filtering based on firstly, individual genotype quality, then secondarily, on variant quality, improves whole exome sequencing (WES) variant calling quality measured by genotype concordance and Ti/Tv ratio, even if VQSR is applied after these two filtering steps[2].

Since individual genotyping quality is a major concern in disease association studies, this seemed to be an important component for filtering in this study. We applied modified Carson criteria and substituted a set of hard filters for VQSR.

GATK suggested hard filter options not used:
QD < 2.0 Plot of QD against ALT allele depth shows a hyperbolic curve for many variants. This suggests that simply dividing by depth of ALT is overcorrecting for depth at higher depths. The depths vary over many orders of magnitudes compared to a typical WGS run.

FS > 60 With the enormous read depths in this targeted resequencing, even small differences in strand will become highly significant by p-value.

SOR > 3.0 Variants that are in an overlapping region of amplicons but where the variant is only covered by the Fwd or Rev read separately in the two amplicons (ie outside the overlap or stitching region) can show apparent strand bias if there are systematic differences in the number paired reads for one amplicon versus the other overlapping one. This could reflect amplicon amplification bias (one amplicon competitively versus the others) rather than allele strand bias.

1.6. Post-Sequencing Quality Control

1.6.1. Sequence Read-Estimated Gender Misclassification
Three Y chromosome amplicons were included for gender inference checks. While this was not as comprehensive as GWAS checking it was included to look for sequencing project mix-ups by comparing Y chromosome read counts.

Four samples gave clearest evidence of possible misclassification of gender, 2 in Steno, 1 in Joslin, and 1 in FinnDiane. All 4 were dropped.

1.7. Annotation Files

Annotation files were generated using VEP version 93 with the LOFTEE plugin and some post-processing to simplify. Note that VEP will give one or more predicted feature types for a variant in one or more transcripts for a gene and hence there can be multiple predicted consequences. To simplify we included only the annotation that had the most deleterious predicted effect, ie a single annotation record for each combination of variant x gene seen in the VEP output.

1.8. Statistical Genetic Models

The main statistical analysis used R 3.X versions of the burden and SKAT aggregated gene variant tests as implemented in seqMeta 1.6.7 for the individual cohort and case-control analyses.

The minimal Cox proportional hazards model used for the cohort analysis was:

Surv(ESRDTIME,ESRD) ~ BASEGFR + PC1+PC2 + ...

BASEGFR	GFR at baseline or time of entry into the cohort, either prospective or retrospective, study t=0. Since our model is based on a straight-line slope of decline to ESRD, the base GFR at t=0 was used to adjust for the starting point in the decline progress for each participant. If a participant starts from higher GFR, for all else being equal, we expect them to take longer to reach ESRD (think of the relative sizes of right-angled triangles with fixed hypotenuse angle).
ESRD	1/0 flag to indicate if the participant had an ESRD event. ESRD is defined clinically or by GFR < 10 surrogate
ESRDTIME	Time after baseline (t=0) when ESRD event occurred. If the participant did not have an event the ESRD variable will be 0 and this is the censoring time since baseline for the last follow-up observation on the participant where it was possible to determine whether the participant had had an ESRD event.
PC1+PC2+…	The minimal model required inclusion of at least two principal components, but allowed cohort analysts the flexibility to include more based on their expert local knowledge.

1.8.1 Inclusion of Principal Components in gene-testing models.

As described in the main text Methods, non-European ancestry outlier samples were identified and removed during the GWAS QC process.

Cohort Discovery Models

The minimal model required inclusion of at least two principal components, but allowed cohort analysts the flexibility to include more based on their expert local knowledge. This was predicated on several factors:

1. The well-documented much higher incidence rate of T1D in European ancestry, particularly in these study groups which recruited patients into studies typically more than 30 years ago. T1D patients in those seminal cohorts were overwhelming European ancestry.

2. The discovery study groups were cohorts with more precise longitudinal clinical diabetes phenotyping as compared to general population case-control disease studies which tend to be more permissive with concomitant higher misclassification and confounding rates.

Preliminary analyses of the screening models in the Joslin cohort, minus the genetic predictors, showed that in a Cox PH model with explanatory baseline GFR + sex + PC1 + PC2 + PC3 + PC4 + PC5, the only predictor that surpassed nominal $p<0.05$ threshold was baseline GFR. The baseline GFR was of course expected since it defines the progression to ESKD. The most significant PC was PC2 with $p=0.08$.

Of the discovery cohorts 4/5 elected to retain the first two PCs. EDC included the top 3 PCs.

Case-Control Replication Models

For the two case-control studies more detailed sensitivity analyses of the effects of different PCs were performed because of the perceived increased risk of population stratification and confounding, especially in the Joslin-Fresenius vs Joslin Medalists case-control study group with cases derived from a nationwide set of clinics versus the Boston-based Joslin Medalist controls.

1. GoKinD-GWU – Genetics of Kidneys in Diabetes, George Washington University recruitment center, extreme cases vs controls.
The GoKinD-GWU cases and controls were recruited under a single study GoKinD protocol. Logistic (GLM) modeling of the non-genetic null model of CASE ~ GFR + sex+ incremental addition of the first 5 PCs (PC1, PC1+PC2, PC1+PC2+PC3,….) showed that no PC in any of the models had a significance better than $p=0.43$. Therefore, we retained just the top 2 PCs.

2. Joslin-Fresenius cases vs Joslin Medalist controls. Logistic (GLM) modeling of the non-genetic null model of CASE ~ GFR + sex+ incremental addition of the first 15 PCs (PC1, PC1+PC2, PC1+PC2+PC3,….) showed that only PC1, PC2, PC3, PC5, PC9, PC12, PC14 had a p-value <0.1 in any of the models. Therefore, we retained just these 7 non-contiguous PCs.

1.8.2 Non-inclusion of baseline age in the gene-testing models.

We tested including baseline age in the JOSLIN null (non-genetic) model together with BASEGFR but it added little additional explanatory effect in the null model beyond BASEGFR with which it is highly confounded. A similar result would be expected for all cohorts. The potentially more serious problem is that the time to diagnosis of the underlying type 1 diabetes will be a component of this age since type 1 diabetes is a necessary condition prior to type 1 kidney function impairment and may be over-adjusting and biasing the results. This is a distinct situation from genetic association studies where the disease of interest is a primary disease and not a sequela or complication.

1.8.3 Inclusion of gender/sex in the gene-testing models.

Cohorts were given the option of including gender in their primary screening models based on their expert local cohort knowledge, but we planned a gender-stratified

secondary analysis for sensitivity analysis and importantly, to check for gender-specific effects. At the outset, we planned two exome-wide gene-based screens (burden and bidirectional SKAT) using a single putative functional variant weighting scheme to minimize the multiple testing correction and maximize power to detect any associated gene under either of the two screens. We deliberately did not include a sex-stratified screen as part of the primary screens, in order to reduce the burden of the multiple testing correction. With the most complete data set available (from the resequencing) post-hoc we found no evidence of gender-specific effects or differences of effect size in the results for *HSD17B14* (main text, Table 4). However, because we did not perform a sex-stratified primary screen, we cannot say if there are additional genes that might have reached the much more stringent level of significance for multiple whole-exome screens if we had additionally performed 4 additional sex-specific screens: (male or female) x (burden or bidirectional).

1.9. Genomic Control Adjustment of Gene-based Whole Exome Screen

We computed the genomic control parameter for each of the cohorts using the array-based screen summary results, separately for the score tests of overall variant burden and bi-directional SKAT, SM Table 4. Because the cohorts were predominantly European ancestry matching the known population disease genetics of T1D, and the inflation varied with cohort size, we hypothesized that the inflation seen was an artifact of the statistical test rather than under-adjustment for population stratification. To test this, we permuted the phenotype data in the Joslin cohort 100 times and computed the gene-based statistics across all 22 chromosomes for each permutation retaining the correlation between the gene variants, for a total of 1,402,169 individual non-monomorphic gene tests (versus ~15,000 in the actual scan). Under this global null scenario, we found an overall empirical genomic control parameter of 1.27 for the Burden tests and 1.40 for the SKAT tests. Since these were larger than the values seen in the Joslin cohort for the actual gene discovery scan (1.21, and 1.34 respectively), this strongly suggests that the inflation is largely or wholly the result of the anticonservative performance of the test with these sample sizes. Since we only had results for our lead gene from the 5-cohort discovery scan, we set the WESDR genomic control parameter to the maximum seen in any of the cohorts.

Cohort	Events/N	Burden	SKAT
Joslin	354 / 614	1.21	1.34
FinnDiane	447 / 783	1.16	1.20
Steno	132 / 414	1.49	1.03
INSERM	99 / 257	1.47	1.70
EDC	63 / 144	1.48	1.59
WESDR	20 / 160	Set to 1.49	Set to 1.70

Supplemental Methods Table 4. Empirical genomic control parameters for the type 1 DKD cohorts.

2. SUPPLEMENTAL RESULTS

2.1 Supplemental Protein Modeling Results

As of March 2020, there are over 20 crystal structures of HSD17B14 in the Protein Data Bank (PDB) (https://www.rcsb.org), all of which clearly show that this protein forms a tetramer. All of these structures are homo-tetramers and this oligomeric assembly may contribute to the stability of the protein and its half-life in the cell or affect its enzymatic efficiency, potentially via cooperative effects. As shown in Supplemental Figure S4A and B, the 249-270 C-terminal residues play significant role in the inter-subunit contact: this fragment stretches out of the subunit of origin and for most of its length, wraps around the opposite subunit in the tetramer. Residues Val263, Pro266, and Pro269 of the C-terminal fragment form complementary hydrophobic interactions with the core of the subunit; moreover, Asp267 forms a salt bridge with Arg203 of the core (Supplemental Figure S4C). Perhaps the most important interaction is the disulfide bond between residues Cys255 of the adjacent subunits. In the structure 6EMM, the disulfide bond is clearly visible. Notably, the C-terminal fragment is disordered in many other structures, and possibly occupies a different conformation in some of them (e.g., PDB 5ICS). These differences might be a result of reducing conditions used during protein purification and crystallization that reduced the bond. The A249CfsTer55 mutant clearly has an altered inter-subunit interface in the tetramer because the mutated and elongated amino acid sequence of the C-terminal fragment is unlikely to form complimentary contact with the opposite subunit, especially the disulfide bond. As a result of this disruption, the inter-subunit contact should become weaker. In addition, the mutated HSD17B14 protein has the potential for blocking the entrance to the active site. As shown in Supplemental Figure S4D, the C-terminal fragment in the wild-type form is located near the entrance to the active site. A mutated and elongated C-terminal fragment of A249CfsTer55 is likely to hinder substrate entrance into the active site.

2.2 Comparison of Results from the Variant Risk Sets

At the outset, it is important to point out that the samples, sample sizes, variants seen, and distribution of genotype missing data differs from the sample sets used for the initial whole exome gene-based scan and therefore are not directly comparable with the results from the whole-exome array-based scan.Comparing the sequence-based burden association test results between the risk sets for the *HSD17B14* gene, Table 4, the predicted Deleterious set had a stronger and more significant association (log(HR)=-0.032, $p=3.6x10^{-4}$) than the least selective Missense set (log(HR)=-0.018, $p=5.5x10^{-3}$). Remarkably, the initial genotyping set of variants on the GWAS array resulted in an even stronger association in the sequencing data (log(HR)=-0.044, $p=1.4x10^{-5}$) than the post-hoc selected Deleterious set. This fortuitous occurrence of variants of relatively strong joint effect on the original array almost certainly aided the initial exome-wide discovery of the gene. Stratification of variants into common (MAF >0.01) and rare (MAF <0.01) risk sets showed that the set of rare variants were protective overall (beta = -0.65, Burden p=0.0033) and independently of the common variants, under a Burden model with all variants equally weighted within the MAF set. The 3 common variants

R130W, A56T, and N31D had a net non-significant association (beta=0.024, Burden p=0.65), and the results in main text Table 3 explain this result. The strong protective association of R130W was more than counteracted by the weaker but +ve (risk) direction for the other two common variants, particularly N31D with a MAF =0.25. This shows that the association and architecture of risk or protection depends on how the variants are grouped for the purposes of gene discovery or variant risk set testing.

2.3 Sensitivity of Results from the Variant Risk Sets to Possible Somatic Variants

As with all studies that sequence and report on rare variants in older populations, there remains the question of whether the putatively germline variants could be somatic in origin. For blood-specimen extracted DNA, as used in the cohorts in this study, the likelihood of the rare variant calls being somatic variants depends on the prevalence of somatic variation in hematopoiesis and the terminally differentiated cells which were the primary source of the DNA; and the sequence analysis steps that might mitigate false somatic variant calling. Since *HSD17B14* has not been identified as a driver gene in either hematologic cancers or clonal hematopoiesis (CH), coding somatic variants that were of sufficient variant allele frequency (VAF) to appear to be germline would most likely have arisen as 'hitchhiking' or 'passenger' variants in a clone that had risen to a significant percentage of the total leukocyte population as a result of selection[3], so that a sufficient proportion of sequencing reads at a *HSD17B14* SNV would be carrying the somatic variant for the software to call the rare allele as a heterozygous locus. The GATK best practices and additional variant filtering steps that we applied (Section 1.5) should have minimized the possibility that rare or very rare somatic variants were called as germline. Nonetheless, since our study used DNA that was drawn from older study participants, we conservatively tested the sensitivity of our results to exclusion of the variants that were the most likely candidates for somatic origin.

At age 80, approximately 60–240 million bases are mutated in the complete active hematopoietic stem cell pool[4]. Evolutionary modeling of CH suggests approximately 1 in 10^4 (presumably neutral) synonymous variants will hitchhike to reach a VAF of at least 10%, or up to 24,000. In a total study size of 2,000 participants, conservatively assuming that each participant has 24,000 unique higher VAF sites, 1.6% of genome sites will be sampled. While the probability of seeing any particular coding or splice site variant at elevated VAF in *HSD17B14* was quite small, the probability of seeing the site with elevated VAF independently in two participants in this study was miniscule. Hence only variants that were seen with a minor allele copy of 1 allele were considered to be candidates for somatic variants posing as germline (SVPG). Additionally, following the criteria typically used for filtering germline variants in tumor somatic variant calling, we filtered any variant that had been seen previously in gnomAD (Supplemental Results Table S6) leaving 6 variants prioritized as the mostly likely to be somatic in origin:

LOF/GOF/Splice site:	Exon 2, +1 donor; Exon 2, -1 acceptor
Missense:	L113M; T102I; P94R; A90P

Note that in Deleterious set, G9R was seen only once (MAC=1) but has been seen previously in gnomAD. We performed a sensitivity analysis, dropping these variants, and performing the same tests as described in the main text, Supplemental Results Table S7. Only 1 LOF/GOF/Splice site remained (A249CfsTer55) and was not formally tested. While the p-values attenuated slightly, the inferred beta and standard errors estimates suggested that this was primarily due to the increase in the standard error component of the test statistic. We conclude that somatic variants, if they happened to be present at a high enough VAF in the limited number of coding/splice sites specifically in the *HSD17B14* gene, are unlikely to be biasing results from the more robust germline results.

Finally, we note that the initial discovery of *HSD17B14* used array-based genotypes, with rare-variants recalled using zCall. Since zCall depends on the separation between sparse heterozygotes and the major homozygote cluster in the allele reclustering coordinate space, the ability to discriminate rare heterozygotes is much less sensitive than from deep shotgun resequencing, and will be directly dependent on the VAF. A high somatic VAF of 10% may be detectable in deep read coverage, but will display a smaller outlier separation from the main cluster than a VAF of 20% or 50% for a true germline heterozygote. Hence zCall rare genotypes are likely to be conservative against the risk of calling somatic heterozygotes. All 6 of the array-based variants had a MAC of at least 2 in the resequencing.

2.4 Expression of *HSD17B14* in public bulk RNA-seq data sets

Using three public bulk RNA-seq data sets (Supplemental Figure S5) we found that the *HSD17B14* gene was consistently robustly expressed in kidney in all cases, and kidney cortex was the 10th highest expressed tissue (53 total tissues) in the GTEx v7 data set (median TPM=35.0).

Supplemental Results Table S1. Top 10 genes for association with survival against ESKD in 5 type 1 diabetes cohorts with advanced DKD.

Chr	*Gene*	Variants	cMAF[a]	Burden p	Burden Beta(SE)[b]	SKAT p
19	*HSD17B14*	5	0.308	8.6×10^{-6}	-0.045 (0.010)	9.9×10^{-4}
17	*GPR179*	29	0.209	2.7×10^{-4}	-0.012 (0.003)	0.14
2	*GCA*	3	0.149	4.6×10^{-4}	-0.034 (0.009)	3.9×10^{-4}
17	*GEMIN4*	21	1.206	4.4×10^{-4}	-0.022 (0.006)	4.7×10^{-3}
16	*PRSS33*	1	0.0012	4.8×10^{-4}	1.106 (0.32)	9.1×10^{-4}
1	*CCDC28B*	3	0.039	0.15	-0.012 (0.008)	6.1×10^{-4}
3	*TBC1D23*	6	0.011	6.7×10^{-4}	-0.039 (0.012)	0.021
12	*TPI1*	4	0.010	4.8×10^{-3}	-0.031 (0.011)	7.0×10^{-4}
6	*LTB*	2	0.0014	7.3×10^{-4}	-0.065 (0.019)	3.7×10^{-3}
10	*ANXA11*	12	0.53	8.1×10^{-4}	0.022 (0.007)	0.032

Initial exome-wide scan for association with ESKD incidence time in 5 cohorts (Joslin, FinnDiane, Steno, INSERM, EDC), total 2,212 participants, using a Cox proportional hazards model and Beta(1,25) weighting based on minor allele frequency

Genes are ranked by decreasing minimum p-value of the Burden and SKAT results

Results were generated using only variants on the Illumina Infinium Human CoreExome Array and were corrected for genomic control by multiplying the score test null model residual SE by $\lambda^{1/2}$, where λ is the within-cohort genomic control parameter.

a cMAF – cumulative minor allele frequency > 1 is due to the presence of common variants.

b Burden and SKAT test p-value results used score tests and p-values may differ from Burden Wald test result. Betas are the joint effect of all nonsynonymous alleles, both rare and common, weighted by the Beta(MAF; 1,25) function.

Supplemental Results Table S2. Results of array gene-based testing of *HSD17B14* by cohort and case-control group including SKAT results

	Events / N	**Variants**	**Burden Beta (SE)**	**Burden P-value**	**SKAT P-value**
Discovery Cohorts					
Joslin	354 / 614	3	-0.051 (0.030)	0.091	0.20
FinnDiane	447 / 783	5	-0.045 (0.012)	$1.5x10^{-4}$	$7.1x10^{-3}$
Steno	132 / 414	3	-0.218 (0.081)	0.0072	0.013
INSERM	99 / 257	4	0.011 (0.043)	0.80	0.52
EDC	63 / 144	4	-0.078 (0.052)	0.13	0.25
Discovery Meta-analysis (n=5)	1095 / 2212	5	-0.045 (0.010)	$8.6x10^{-6}$	
Replication Cohort					
WESDR	20 / 160	3	-0.097 (0.10)	0.35	0.41
Cohort Meta-analysis (n=6)	***1115 / 2372***	***6***	***-0.046 (0.010)***	$\mathbf{6.3x10^{-6}}$	$\mathbf{7.7x10^{-4}}$
Replication Case-Control	**N Cases / Controls**		**Burden Beta (SE)**	**Burden P-value**	**SKAT P-value**
Joslin-Fresenius vs Joslin Medalists	946 / 610	5	-0.042 (0.018)	0.024	0.11
GWU-GoKinD Cases vs Controls	126 / 142	5	-0.057 (0.034)	0.095	0.55
Case-Control Meta-analysis (n=2)	***1072 / 752***	***6***	***-0.046 (0.016)***	$\mathbf{5.0x10^{-3}}$	***0.054***
Overall Meta-Analysis				$\mathbf{1.1x10^{-7}}$	$\mathbf{1.1x10^{-4}}$

Results are similar to Table 2 in the main paper and used the variants present on the Illumina Infinium Human CoreExome Bead Array. Analysis for all cohorts except WESDR used array genotyping data; WESDR genotypes were derived from resequencing (described later). Variants column shows the number of variants tested in that cohort or case-control analysis of *HSD17B14*. The number varies because of batch QC. Integers in parentheses in the Cohorts or Case-

Control groups (n) are the number of studies included. Events are ESKD diagnosed clinically or by proxy from eGFR < 10mL/min/1.73m2. Betas are the log values of the Hazard (HR) /Odd Ratios (OR) and measure the mean effect of variants tested. The HR and OR values were derived by antilog of the betas. SE are the standard errors of the beta estimates. The exome-wide gene discovery and replication was corrected for genomic control within each cohort.

Supplemental Results Table S3. Top 10 genes after meta-analysis of the type 1 diabetes cohort and case-control results

		Cohort			Case-Control			Meta	
Chr	***Gene***	**Burden Beta (SE)**	**Burden P**	**SKAT P**	**Burden Beta (SE)**	**Burden P**	**SKAT P**	**Burden P**	**SKAT P**
19	*HSD17B14*	-0.045 (0.010)	8.6E-06	0.00099	-0.045 (0.019)	0.017	0.14	4.5E-07	0.00033
6	*SFTA2*	-7.25 (3.63)	0.046	0.050	-14.02 (3.86)	0.00028	0.00043	8.0E-05	0.00012
2	*C2orf82*	0.039 (0.038)	0.30	0.32	0.089 (0.023)	0.00010	0.00018	0.00012	0.00019
9	*ZNF883*	-0.050 (0.016)	0.0019	0.0089	-0.052 (0.028)	0.060	0.28	0.00028	0.0052
19	*FKRP*	NA[a]	NA[a]	NA[a]	0.087 (0.024)	0.00029	0.00047	0.00029	0.00047
1	*C1orf172*	-0.0090 (0.070)	0.90	0.36	0.085 (0.022)	0.00014	0.00032	0.00034	0.00021
7	*KCNH2*	-0.0013 (0.023)	0.95	0.98	0.077 (0.017)	6.9E-06	8.9E-06	0.00036	0.00036
6	*LTB*	-0.065 (0.019)	0.00073	0.0037	-0.040 (0.035)	0.26	0.34	0.00046	0.0026
8	*ATAD2*	-0.032 (0.012)	0.0098	0.11	-0.045 (0.019)	0.015	0.19	0.00046	0.039
1	*FOXD3*	0.058 (0.048)	0.22	0.25	0.081 (0.025)	0.0011	0.0016	0.00054	0.00086
12	*FAM109A*	-0.033 (0.027)	0.21	0.25	0.085 (0.019)	1.2E-05	2.3E-05	0.0048	4.0E-05
6	*HLA-DQA1*	-0.011 (0.012)	0.38	0.25	0.040 (0.0095)	2.1E-05	4.4E-05	0.0057	9.3E-05
6	*SFTA2*	-7.25 (3.63)	0.046	0.050	-14.02 (3.86)	0.00028	0.00043	8.0E-05	0.00012
1	*PLEKHN1*	-0.013 (0.015)	0.38	0.041	0.056 (0.017)	0.0011	0.00050	0.14	0.00013
22	*DGCR8*	-0.0094 (0.0093)	0.31	0.36	0.030 (0.012)	0.012	6.2E-07	0.46	0.00015
2	*C2orf82*	0.039 (0.038)	0.30	0.32	0.089 (0.023)	0.00010	0.00018	0.00012	0.00019
1	*C1orf172*	-0.0090 (0.070)	0.90	0.36	0.085 (0.022)	0.00014	0.00032	0.00034	0.00021
9	*PHF2*	-0.017 (0.0099)	0.091	0.0028	-0.059 (0.020)	0.0027	0.022	0.0043	0.00022
19	*HSD17B14*	-0.045 (0.010)	8.6E-06	0.00099	-0.045 (0.019)	0.017	0.14	4.5E-07	0.00033
7	*KCNH2*	-0.0013 (0.023)	0.95	0.98	0.077 (0.017)	6.9E-06	8.9E-06	0.00036	0.00036

The top block of 10 genes is ranked by Burden test meta-analysis p-value, the bottom by SKAT test meta-analysis (p-values highlighted in blue). Note that these results do not include the extension WESDR cohort which at the time had limited resequencing-derived gene results and not whole exome, hence the results for *HSD17B14* differ slightly from those reported for the full meta-analysis in the main text.

a. Test was not performed in this gene because the variants were monomorphic.

Supplemental Results Table S4. Technical sequencing variant quality statistics of predicted LOF/GOF/splice site variants

Variant (chr19 Pos_Ref_Alt)	Ref (0) reads / Alt (1) reads	AN	Genotype Quality[a]	Total Read Depth	QD < 2	FS > 60	SOR > 10	IC < -0.8	MQ < 40	MQRS < -12.5	RPRS < -8	Genotype Likelihoods[b] 0/0, 0/1, 1/1
48813244_A_AC	107 / 77	4836	99	259345	9.7	0	0.4	0.013	60	0	-1.0	1819, 0, 2840
48835804_C_G	503 / 70 [c]	4844	99	1759740	0.5	0	0.02	$-2x10^{-3}$	60	0	4.3	333, 0, 13871
48835844_C_T	45 / 12	4844	99	1759058	2.5	0	0.08	$-2x10^{-3}$	60	0	0.6	174, 0, 1298

AN Allele Number = 2x number of participants w/genotyping data
GATK hard filtering metrics - column and filter threshold:
QD < 2 Quality by Depth
FS >60 Fisher Strand
SOR > 10 Strand Odds Ratio
IC < -0.8 Inbreeding Coefficient
MQ < 40 RMS Mapping Quality
MQRS < -12.5 Mapping Quality Rank Sum Test
RPRS < -8 Read Pos Rank Sum Test

a Conditional genotype quality in PHRED units (-log10) , 99 is the maximum possible.
b Genotype posterior likelihoods in PHRED units (-log10)
c In independent whole exome sequencing of this participant, the same variant chr19_48835804_C_G, contained reads 65 Ref / 14 Alt (FinnDiane, personal communication).

Supplemental Results Table S5. Characteristics of carriers of predicted LOF/GOF/splice site variants

Ppt	Cohort	Age T1D Diagnosis	ESKD Event	Censored Time to ESKD [a]	Variant (chr19 Pos_Ref_Alt)	rs# / consequence
1	JOSLIN	2 yrs	N	65.8 yrs	48813244_A_AC	rs758181057 / frameshift
2	FinnDiane	N/A	Y	53.5 yrs	48835804_C_G	N/A / splice donor
3	INSERM	31 yrs	Y	28 yrs	48835844_C_T	N/A / splice acceptor

Ppt: Participant number. Variant positions are given in hg38 assembly coordinates.
a If the participant had not had an ESKD event, this is the time to last determination of ESKD.

Supplemental Results Table S6. *HSD17B14* coding and splice variants in gnomAD

T1DKD		**gnomAD v2.1**			
Variant	*rsID*	*Variant Alleles*	*Total Alleles*	*Allele Frequency*	*Homozygotes*
LOF/GOF/Splice					
Ala249CysfsTer55	rs758181057	14	235,734	$5.9x10^{-5}$	0
g.48835804C>G; exon 2, +1 donor		Not seen			
g.48835844C>T; exon 2, -1 acceptor		Not seen			
Deleterious					
R130W	rs35299026	10,244	282,702	0.036	248
D62Y	rs139987974	413	282,736	0.0015	5
R27C	rs141119542	524	282,022	0.0019	0
G22E		Not seen	0	0	0
G16W	rs113246661	1,194	282,386	0.0042	3
G9R	rs565045362	1	251,008	$4.0x10^{-6}$	0
Missense					
V263M	rs202197135	15	273,304	$5.5x10^{-5}$	0
T261N	rs765538934	19	274,766	$6.9x10^{-5}$	0
R159Q*	rs762325773	30	215,534	0.00014	0
L113M		Not seen			
R108H	rs1417910002	3	282,470	$1.1x10^{-5}$	0
T102I		Not seen			
P94R		Not seen			
A90P		Not seen			
R82C	rs781248545	38	280,784	0.00014	0
V69M	rs188166617	135	282,632	0.00048	0
D62G	rs146036929	137	282,718	0.00048	1
A56T	rs116979565	2,704	282,622	0.0096	31

N31D	rs8110220	77,016	281,926	0.27	11,971

gnomAD v2.1, accessed Jan 2019.
* This variant was predicted to be present in a secondary transcript (ENST00000595764) only.

Supplemental Results Table S7. Sensitivity of association results for survival time to ESKD by *HSD17B14* variant risk set, genotypes from resequencing

Variant Risk Set	Variant Exclusion	N	Variants Included	Burden Beta (SE)	Burden P-value
LOF/GOF/Splice	-	2239	3	-0.75 (0.49)	0.13
	Possible Somatic	2239	1	Not tested	Not tested
Deleterious	-	2239	9	-0.032 (0.009)	$3.6x10^{-4}$
	Possible Somatic	2239	7	-0.034 (0.010)	$5.1x10^{-4}$
Missense	-	2239	22	-0.018 (0.006)	$5.5x10^{-3}$
	Possible Somatic	2239	16	-0.017 (0.007)	0.010

Variant exclusion describes the set of variants that were excluded from the resequencing variant risk sets reported in the main text. Variant exclusion: - is the reference set for both males and females as reported in the main text. Possible Somatic shows the results for the corresponding variant risk set where the variants that are most likely to be somatic in origin in that set have been excluded. Similar to the main text, the variant risk sets were: From Resequencing: LOF/GOF/Splice: loss or gain or function or in a splice donor/accepter site; Deleterious: LOF/GOF variants plus those predicted by PolyPhen and SIFT to be deleterious; Missense: Deleterious variants plus any other non-synonymous variants; From Initial Genotyping: GWAS Variants present on the array, after QC applied. Results were corrected using the genomic control parameters from the whole exome scan within each cohort. Burden P-value was derived from a score test. The variant risk set LOF/GOF/Splice with possible somatic variants removed was not formally tested, since only a single variant with 1 minor allele copy (MAC) remained.

Supplemental Results Table S8. CKDGen Consortium results at rs35299026 (R130W) from publicly available GWAS data sets

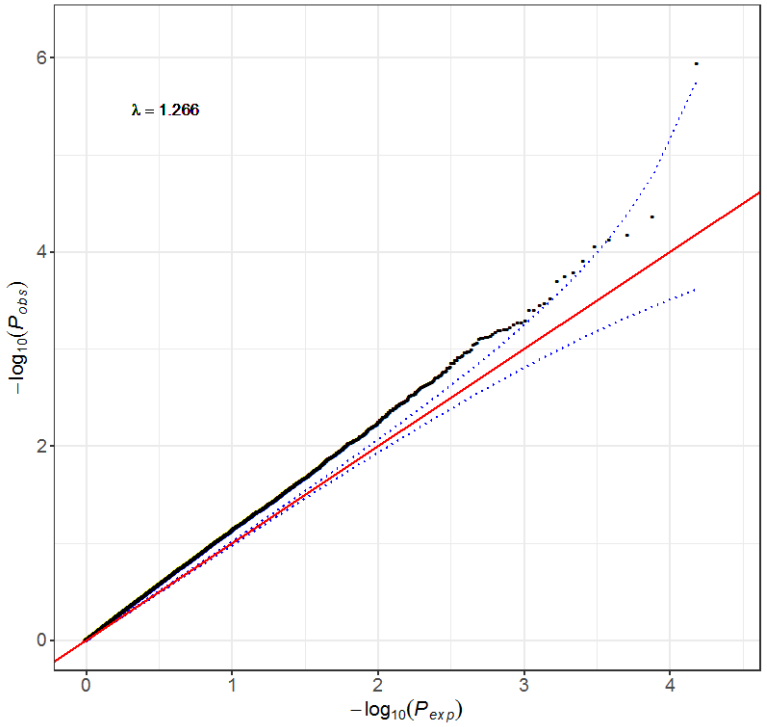
Phenotype (Q or CC)	Ancestry	N	Beta	SE	P-value
BUN (Q)	EUR	211,489	-0.0050	0.0023	0.030
CKD (CC)	EUR	388,362	-0.0088	0.023	0.70
eGFR (Q)	EUR	484,648	0.0013	0.0009	0.14
UACR (Q)	EUR	510,263	0.0033	0.0049	0.50

Q = Quantitative trait; CC = Case-control.

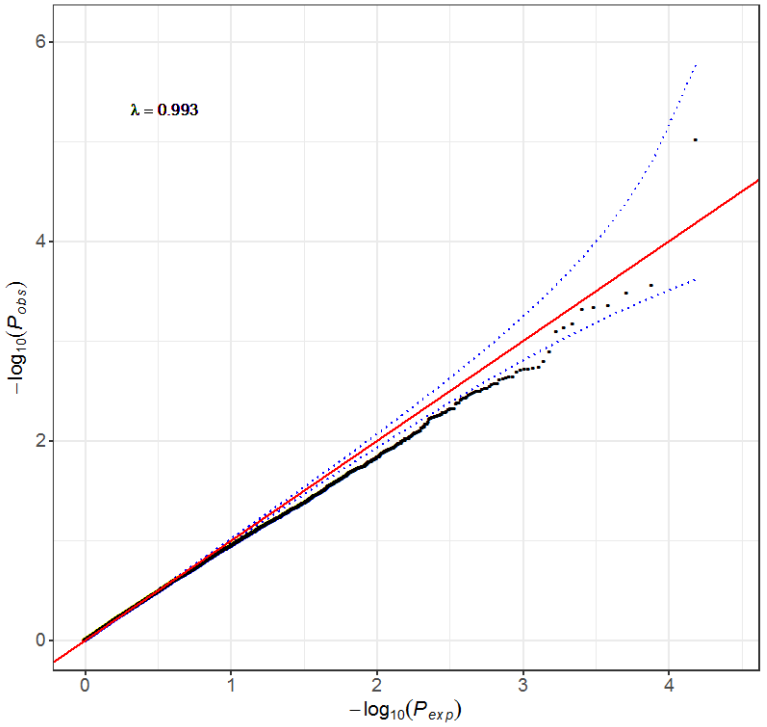
3. SUPPLEMENTAL FIGURES

Supplemental Figure S1. Quantile-quantile (QQ) plots for the 5-cohort discovery, gene-based, whole-exome scan. Both plots include the meta-analyzed results for the burden test applied to 15,449 genes. (A) shows the QQ plot for the non-genomic control corrected p-values. (B) Similar to (A), the results after applying cohort-specific genomic control correction prior to meta-analysis. No genomic control was applied the meta-analysis p-values.

A.

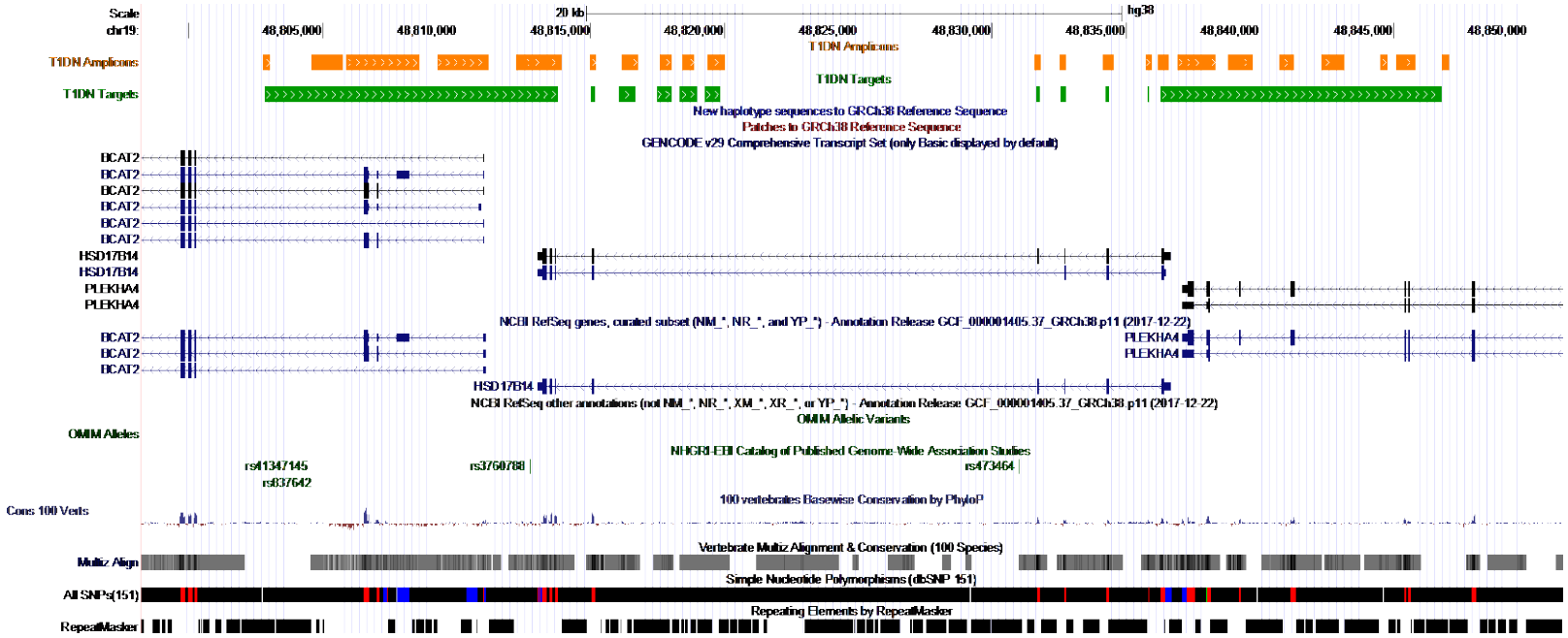


B.



Supplemental Figure S2. Amplicon design in the HSD17B14 genomic region for the resequencing project.
(A) shows the contiguous regions able to be tiled by the designed amplicons (orange intervals) and the original targeted regions (green intervals). (B) shows the intervals covered by the designed amplicons. FHAD1 data is not reported in this manuscript.

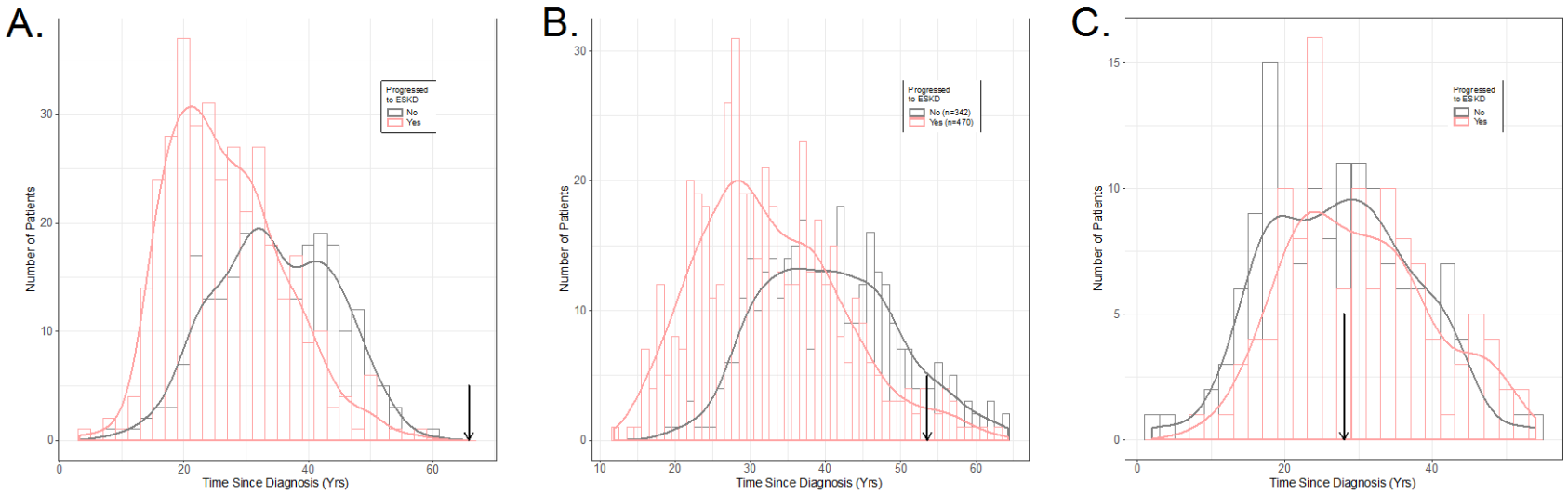
A.



B.

Chromosome	*Gene Locus*	Amplicons	Contiguous Intervals	Total Coverage
Chr1	*FHAD1*	56	35	12,597 kbases
Chr19	*HSD17B14*	91	21	16,608 kbases
ChrY	*SRY (Q/C)*	4	1	839 bases

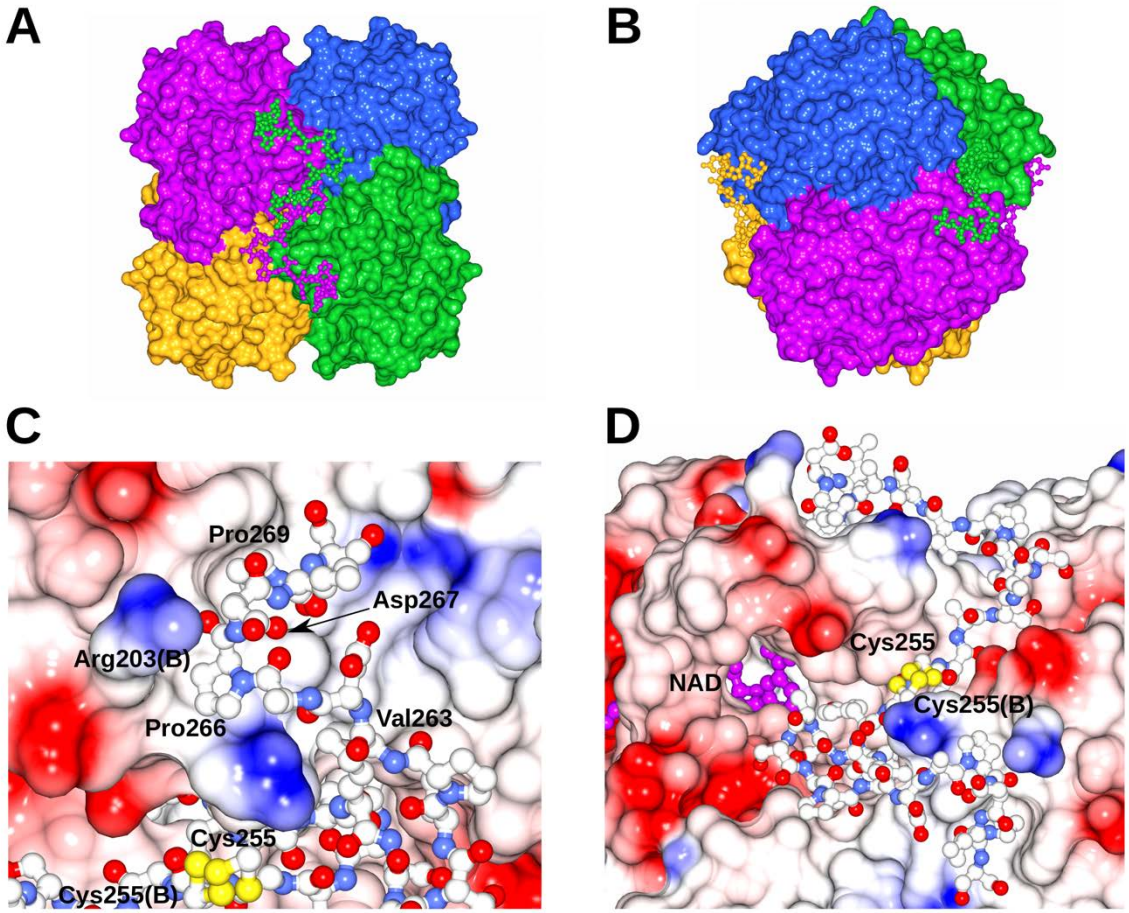
Supplemental Figure S3. Distribution of maximum duration since diagnosis of T1D cohort carrying LOF variants. The cohorts are Joslin (A), FinnDiane (B), and INSERM (C), stratified into two groups - those patients with incident ESKD (salmon pink color) or censored last clinical visit with serum creatinine measurement for GFR estimation (grey color). The duration for the single patient carrying the LOF/GOF/Splice mutant allele in each cohort is shown by the black arrow. (A) is identical to Figure 3A except for plotting color and layout differences.



Chr19 Pos	Ref	Alt	Location	Consequence	Cohort	MAC	N	Beta (SE)	P-val
48813244	A	AC	Exon 9, Ala249	Terminal frameshift + elongated protein	Joslin	1	604	-1.42 (0.92)	0.12
48835804	C	G	Exon 2, +1 donor splice site	Exon 2 splicing?	Finn Diane	1	812	-0.73 (0.63)	0.24
48835844	C	T	Exon 2, -1 acceptor splice site	Exon 2 splicing?	INSERM	1	254	1.03 (1.49)	0.49

Supplemental Figure S4. Tetrameric organization of the wild-type HSD17B14 protein.

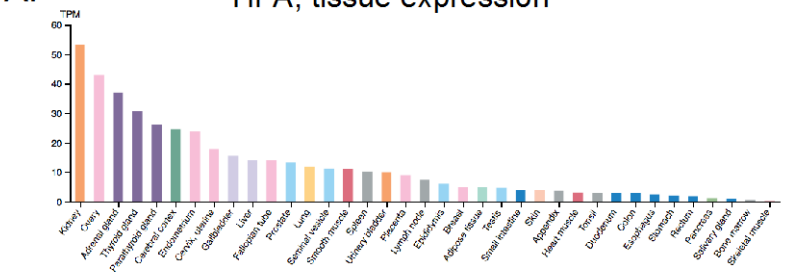
Each subunit is colored in a distinct color. The core of each subunit is shown in molecular surface representation; the C-terminal fragments 249-270 are shown in ball-and-stick representation. C-terminal fragments of only two subunits are visible on this panel; the C-terminal fragments of other two subunits are located on the opposite side of the tetramer. (B) A lateral view of the tetramer. (C) Interactions between the C-terminal fragment 249-270 and the tetramer core. The tetramer core is shown with the electrostatic potential of its surface: blue – positive, red – negative, white – hydrophobic. The C-terminal fragment is shown in ball-and-stick representation with carbon atoms in white (to match the hydrophobic character of the core) and sulfur in yellow. Residues mentioned in the text are labelled, those from the adjacent subunit are labelled with B in parenthesis. Cys255 has two conformations, resulting in for possible positions for the two sulfur atoms. (D) The entrance to the active site in one of the subunits of the tetramer. The NAD molecule, shown in magenta ball-and-stick model, can be seen through the entrance.



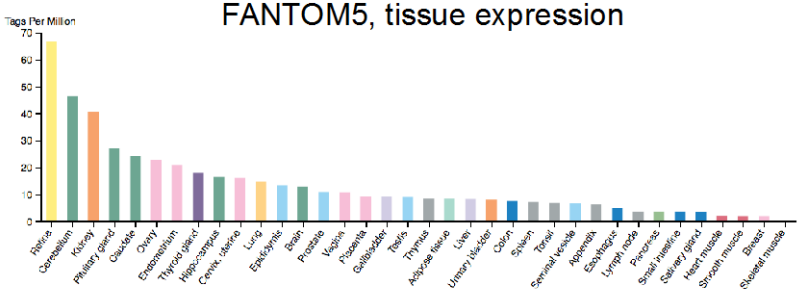
Supplemental Figure S5. Bulk non-diseased tissue expression of *HSD17B14* in multiple human public data sets.

(A) Human Protein Atlas (HPA) tissue gene expression by RNA-Seq, n=9 samples[5], kidney has highest expression of tissues studied (orange bar, mean TPM=53.5). (B) FANTOM5 Cap Analysis of Gene Expression (CAGE) in Tags per Million of reads that align with the transcription start site of the gene[6,7]. Kidney n=1 sample (orange bar, 40.7 TPM) has the third highest expression and upper 10%ile of all tissues in FANTOM5 data set. (C) Version 7 of GTEx bulk RNA-Seq data (https://gtexportal.org). Kidney cortex (red font) is ranked 10th by expression of the 53 tissues included. Tissues in salmon/red font have median transcripts per million (TPM) > 30, bootstrap 95% confidence interval for the median shown. (D) In the Human Protein Atlas, the primary transcript (ENSEMBL ENST00000263278, #201, evidence level 1 (TSL-1), also based on RefSeq NM_016246.2, has the highest median expression (n=9 samples). Secondary, putatively alternatively spliced transcripts (#202, #203, #204) with lower levels of evidence (respectively TSL-5, TSL-5, TSL-3) were estimated to have lower expression levels, but no full-length mature mRNA has been assessed as evidence for their sequence and structure at time of writing.

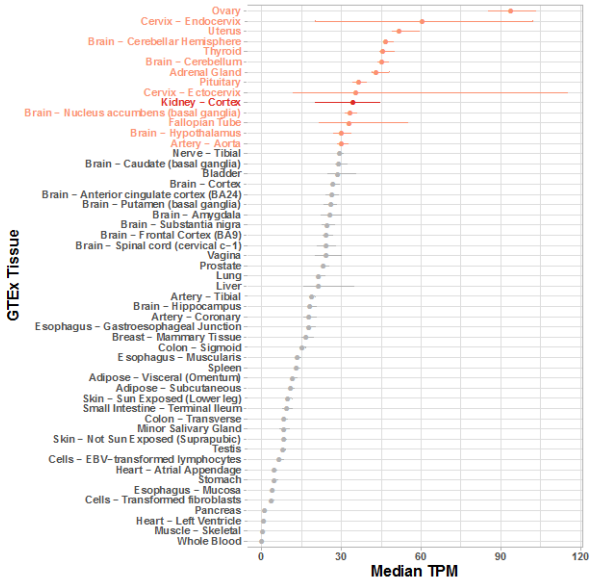
A.


HPA, tissue expression
TPM

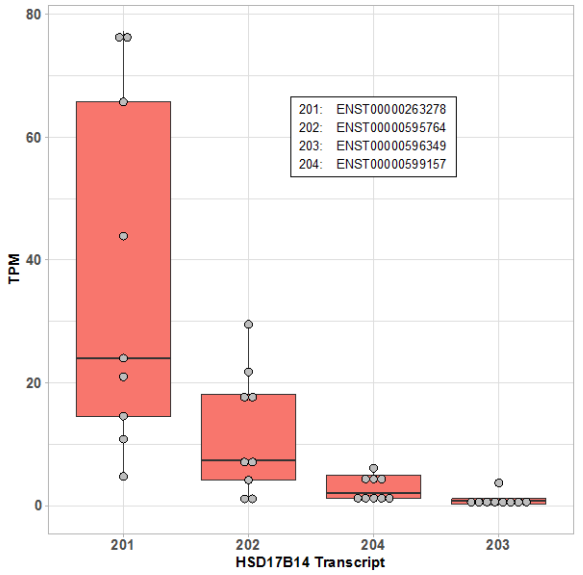
B.


FANTOM5, tissue expression
Tags Per Million

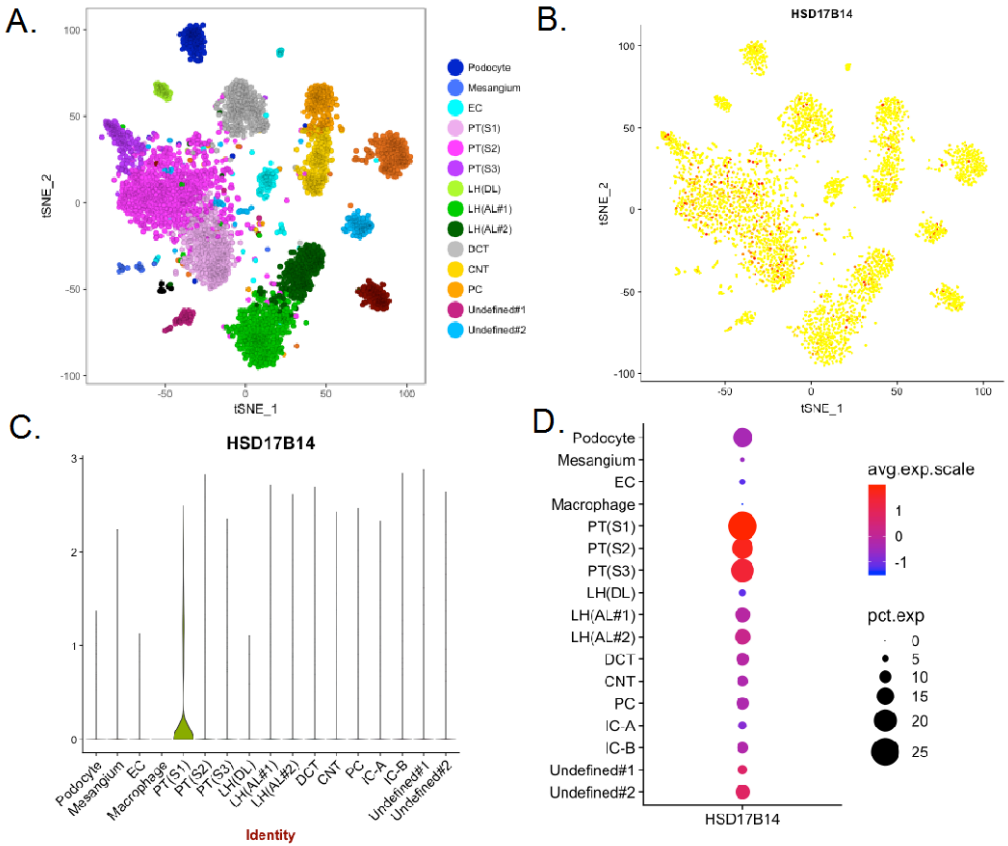
C.


GTEx Tissue
Median TPM
Ovary
Cervix – Endocervix
Uterus
Brain – Cerebellar Hemisphere
Thyroid
Brain – Cerebellum
Adrenal Gland
Pituitary
Cervix – Ectocervix
Kidney – Cortex
Brain – Nucleus accumbens (basal ganglia)
Fallopian Tube
Brain – Hypothalamus
Artery – Aorta
Nerve – Tibial
Brain – Caudate (basal ganglia)
Bladder
Brain – Cortex
Brain – Anterior cingulate cortex (BA24)
Brain – Putamen (basal ganglia)
Brain – Amygdala
Brain – Substantia nigra
Brain – Frontal Cortex (BA9)
Brain – Spinal cord (cervical c-1)
Vagina
Prostate
Lung
Liver
Artery – Tibial
Brain – Hippocampus
Artery – Coronary
Esophagus – Gastroesophageal Junction
Breast – Mammary Tissue
Colon – Sigmoid
Esophagus – Muscularis
Spleen
Adipose – Visceral (Omentum)
Adipose – Subcutaneous
Skin – Sun Exposed (Lower leg)
Small Intestine – Terminal Ileum
Colon – Transverse
Minor Salivary Gland
Skin – Not Sun Exposed (Suprapubic)
Testis
Cells – EBV-transformed lymphocytes
Heart – Atrial Appendage
Stomach
Esophagus – Mucosa
Cells – Transformed fibroblasts
Pancreas
Heart – Left Ventricle
Muscle – Skeletal
Whole Blood
0
30
60
90
120

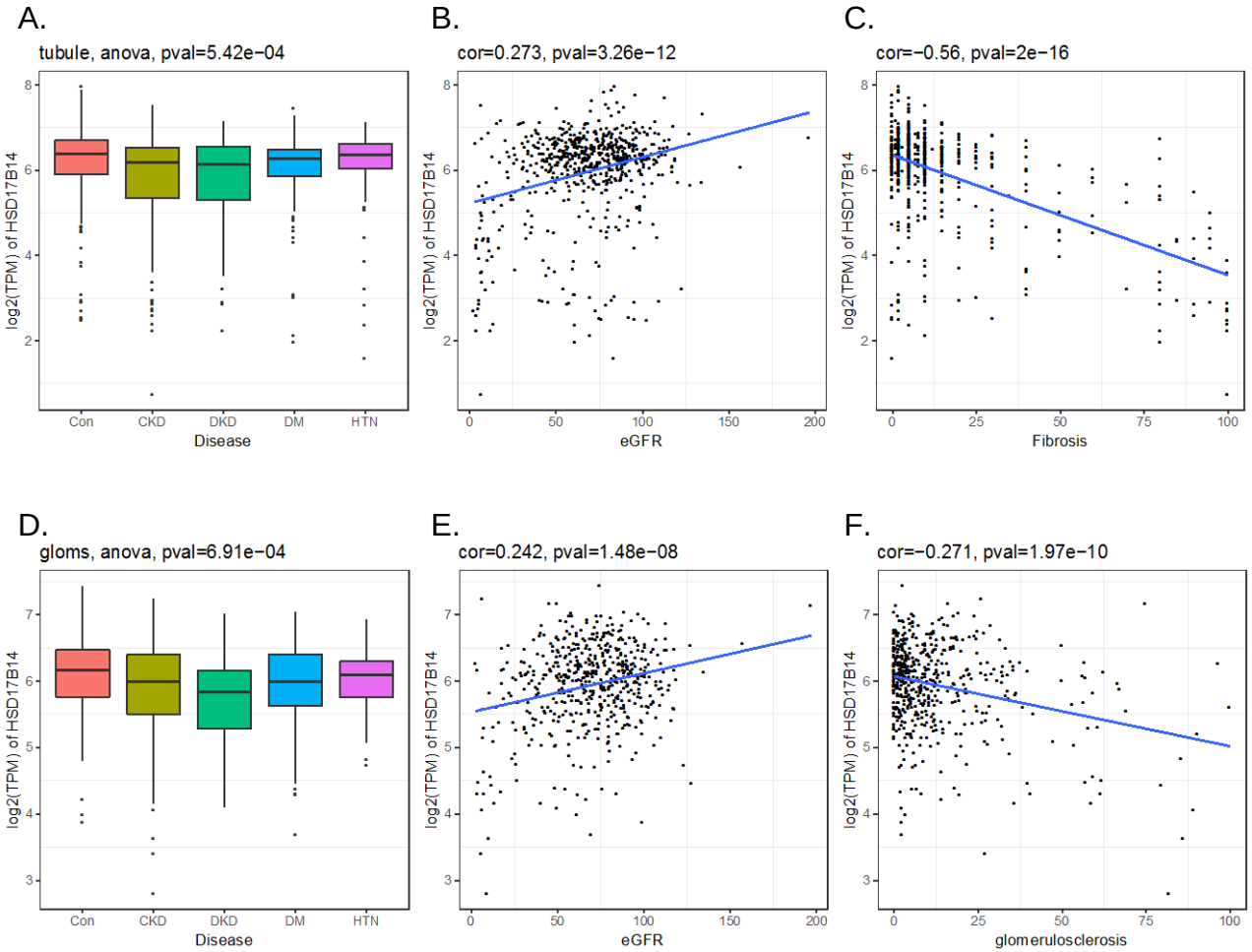
D.


TPM
201: ENST00000263278
202: ENST00000595764
203: ENST00000596349
204: ENST00000599157
201
202
204
203
HSD17B14 Transcript

Supplemental Figure S6. Single nucleus RNA-seq results for *HSD17B14* gene expression in normal, undiseased kidney tissue from a single adult nephrectomy.
(A) TSNE cluster plot of cells color coded by inferred cell type. Proximal tubule segments PT(S1-S3) are colored in pink-purple shading on the left of the panel. (B). Corresponding to cluster in (A) the individual cell expression of *HSD17B14*, with yellow dots indicating no expression in that cell, and red detected expression with red intensity a measure of relative expression levels. (C) Violin plots of the distribution of *HSD17B14* expression in cells in (C) by cell type. (D) Dot plot of the cellular expression of *HSD17B14* by cell type showing the % of cells ('dot' size) and average relative expression level from red (highest) to blue (lowest). Figure was generated from the http://humphreyslab.com. web site and used Seurat software.



Supplemental Figure S7. Variation in *HSD17B14* expression in kidney tissue from patients in 4 human disease states, and undiseased controls, measured by RNA-seq.
(A) Bulk RNA-seq expression of *HSD17B14* in proximal tubules (log(TPM)), in 5 disease states: Con: Control, non-diseased; CKD: Chronic Kidney Disease, eGFR<60; DKD: Diabetic Kidney Disease from Type 2 patients; DM: Type 2 diabetes without DKD; HTN: Hypertensive. Panels (A-C) are repeated from the main text Figure 5D-5F for ease of visual comparison. (D-F) are the analogous figures for microdissected glomeruli from the same study for comparison except panel (F) shows that % glomerulosclerosis instead of fibrosis.

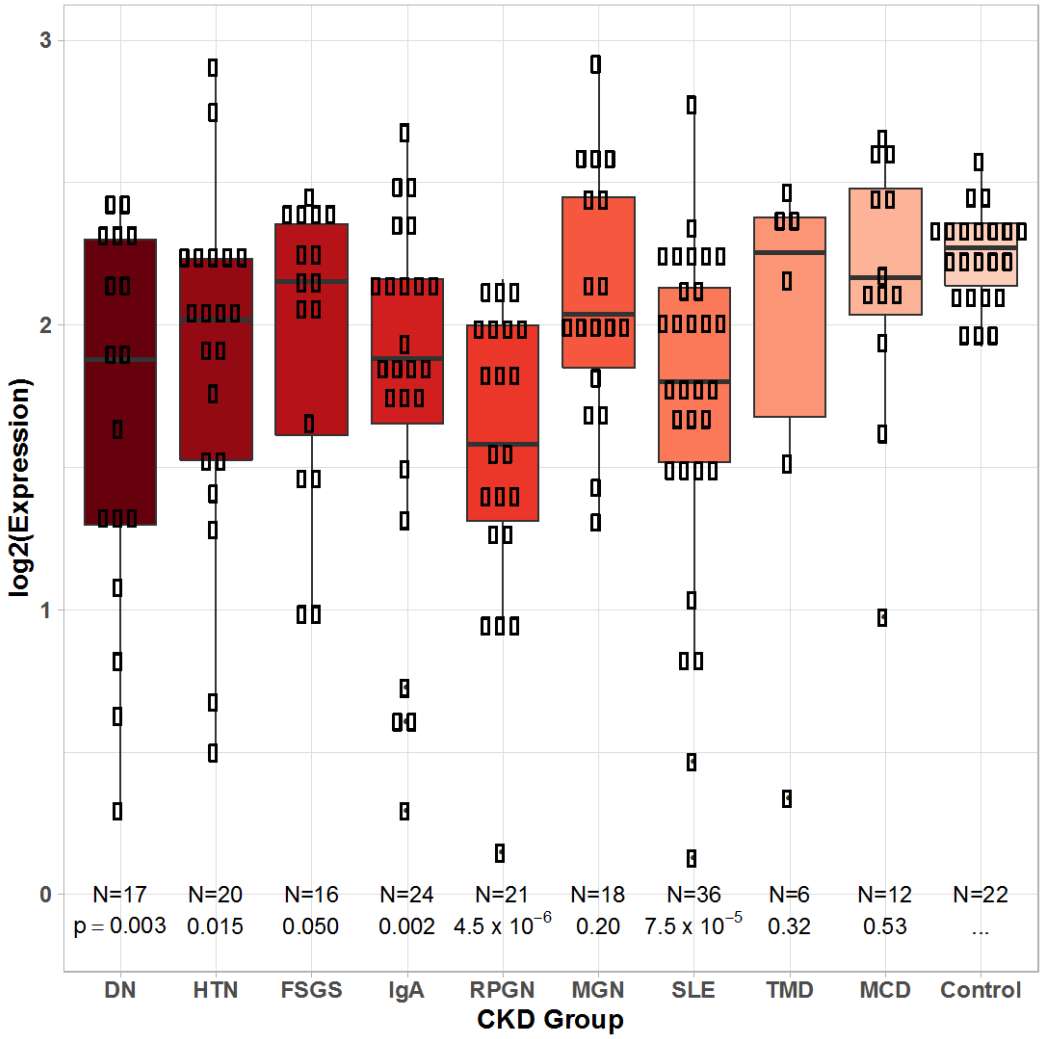


Supplemental Figure S8. *HSD17B14* comparative gene expression in multiple chronic kidney disease pathologies.

Differences in gene expression of *HSD17B14* (log2 scale) in chronic kidney disease associated with other pathologies. Data is from Affymetrix U133 arrays, http://www.nephroseq.org.

DN: Diabetic Nephropathy
HTN: Hypertension
FSGS: Focal Segmental Glomerulosclerosis IgA: IgA Nephropathy
RPGN: Rapidly Progressing Glomerulonephritis
MGN: Membranous Glomerulonephritis
SLE: Systemic Lupus Erythematosus
TMD: Thin Basement Membrane
MCD: Minimal Change Disease
Con: Control, CKD-free.

P-values are for t-tests of means of each CKD group against the Controls, sample size (N) and p-value shown.



REFERENCES

1. Thorn LM, Forsblom C, Fagerudd J, Thomas MC, Pettersson-Fernholm K, Saraheimo M, et al.: Metabolic syndrome in type 1 diabetes: Association with diabetic nephropathy and glycemic control (the FinnDiane study). *Diabetes Care* 28: 2019–2024, 2005
2. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen JB, et al.: Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 15: 125, 2014
3. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al.: The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* 367: 1449–1454, 2020
4. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al.: Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep* 25: 2308-2316.e4, 2018
5. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al.: Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 13: 397–406, 2014
6. Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, et al.: FANTOM5 CAGE profiles of human and mouse samples. *Sci Data* 4: 2017
7. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al.: Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* 16: 2015