SUPPORTING INFORMATION

Table S1: Details for kidney and urine samples. See Excel spreadsheet.

Table S2: Proteins observed in all three tissue/biofluid-based proteomes, sorted by NSC and providing membership in various comparison sets, Human Protein Atlas and BioGPS data, Swiss-Prot cellular localization data, and PeptideAtlas PSM counts, peptide counts, samples, and sample types. See Excel spreadsheet.

Table S3: Proteins observed in the 2013 Human All PeptideAtlas (HumanAllPA), including Human Protein Atlas and BioGPS data, Swiss-Prot cellular localization data, and PeptideAtlas PSM counts, peptide counts, samples, and sample types. See Excel spreadsheet.

Set type	Number of sets	Conditions that must hold for a Swiss-Prot identifier *i* to be in the set	Value displayed if identifier *i* is in set	Universe for GO analysis
A_c (complete mapping)	3	$NSC_{iA}>0$	NSC_{iA}	$A_c \cup B_c \cup C_c$
A_{nr} (nonredundant set)	3	Independent peptide evidence in atlas build A	NSC_{iA}	$A_c \cup B_c \cup C_c$
A AND B	3	$i \in A_{nr};\ i \in B_c$	$Avg(NSC_{iA},NSC_{iB})$	$A_c \cup B_c$
A NOT B	6	$i \in A_{nr};\ i \notin B_c$	X	A_c
A >> B	6	$i \in A_{nr};\ \log(NSC_{iA}/NSC_{iB}) > \mu+2\sigma$	NSC_{iA}/NSC_{iB}	A_c
A AND B NOT C	3	$i \in A_{nr};\ i \in B_c;\ i \notin C_c$	X	$A_c \cup B_c$
A NOT B NOT C	3	$i \in A_{nr};\ i \notin B_c;\ i \notin C_c$	X	A_c
A >> B AND C	3	$i \in A_{nr}$ $\log(NSC_{iA}/NSC_{iB}) > \mu+2\sigma$ $\log(NSC_{iA}/NSC_{iC}) > \mu+2\sigma$	NSC_{iA} / $Avg(NSC_{iB},NSC_{iC})$	A_c
A AND B >> C	3	$i \in A_{nr}$ $\log(NSC_{iA}/NSC_{iC}) > \mu+2\sigma$ $\log(NSC_{iB}/NSC_{iC}) > \mu+2\sigma$	$Avg(NSC_{iA},NSC_{iB})$ / NSC_{iC}	$A_c \cup B_c$
seen in all	1	$i \in A_{nr};\ i \in B_c;\ i \in C_c$	$Avg(NSC_{iA},NSC_{iB},NSC_{iC})$	$A_c \cup B_c \cup C_c$

Table S4: Thirty-four identifier sets compiled to support study of the relationships among the kidney, urine, and plasma proteomes. A, B, and C are each one of KidneyPA, UrinePA, or PlasmaPA. μ = mean, σ = standard deviation for distribution of logarithms of ratios of all pairs of NSC values excluding outliers. For purpose of >> (ENRICHED OVER) operation, any NSC value of zero is set to a tiny nonzero value (half the smallest NSC observed for that tissue/biofluid-based proteome). Also in Excel spreadsheet.

Table S5: Core urinary proteome (Nagaraj & Mann, JPR 2011) protein groups with no identified peptides in PeptideAtlas. See Excel spreadsheet.

Search parameters and modifications

a. X!Tandem

scoring, maximum missed cleavage sites: 2

mass tolerance: varies according to instrument

refine: yes

refine, cleavage semi: yes

All static and variable modifications were searched in both the first pass and during the refinement phase.

For ETD (electron-transfer dissociation) experiments, c and z ions were scored. For all other experiments, b and y ions were scored.

X!Tandem searches included, for each specific sample, modifications expected to be found according to the method of sample preparation. These included both static modifications (one or more of 0.984@N (deamidation), 57.0215@C (carbamidomethylation), 71.037@C (acrylamide), 227.13@C(ICAT), 414.19@C (biotin), 442.2@C(ICAT), 236.13@C(ICAT), 6.020@R (isotopic labeling), 4.025@K (isotopic labeling), 15.99@M (oxidation), 144.1@K or N-terminus(iTRAQ), 229.163@K or N-terminus (TMT6plex), 304.199@K or N-terminus(iTRAQ8plex)) and variable modifications (0.984@N(deamidation), 15.99@M (oxidation), 3.0185@L (isotopic labeling), 4.025@K (isotopic labeling), -17.0265@C (ammonia loss), 74.0366@C (acrylamide d3), 148.0@C(D5 N-ethylmaleimide+water), 236.13@C(ICAT), 7.943@K (SILAC), 5.956@L(SILAC), 79.966@T,S, or Y (phosphorylation)). In addition, all X!Tandem searches include the variable modifications -17.0265@Q (pyroglutamic acid), -18.0106@E (pyroglutamic acid), and 42.01 at protein N-terminus (acetylation).

b. SpectraST

```
indexRetrievalUseAverage = false
indexRetrievalMzTolerance = 3.0
detectHomologs = 4
expectedCysteineMod = CAM, or parameter omitted, depending on data
ignoreSpectraWithUnmodCysteine = false
ignoreChargeOneLibSpectra = false
ignoreAbnormalSpectra = false
hitListTopHitFvalThreshold = 0.0
hitListLowerHitsFvalThreshold = 0.45
hitListShowHomologs = true;
hitListOnlyTopHit = true
hitListExcludeNoMatch = true
peakScalingMzPower = 0.0
peakScalingIntensityPower = 0.5
peakBinningNumBinsPerMzUnit = 1
peakBinningFractionToNeighbor = 0.5
peakScalingUnassignedPeaks = 0.1filterAllPeaksBelowMz = 520
filterMinPeakCount = 6
filterCountPeakIntensityThreshold = 2.01
filterRemovePeakIntensityThreshold = 2.01
filterRemoveHuge515Threshold = 0.0
filterMaxPeaksUsed = 150
filterMaxDynamicRange = 1000
```

By the nature of spectral searching, SpectraST searches covered only the modifications included in the spectral library searched. The NIST spectral library used for most searches includes 69,000 spectra for carbamidomethyl, 47,000 spectra for oxidation, and <10,000 for each of acetyl, pyroglutamic acid, and pyro-carbamidomethyl modifications (http://peptide.nist.gov/browser/lib_stats.php?organism=human&description=IT). For phosphorylation experiments, we searched against an in-house library of identified spectra from previous phosphorylation experiments. For each of the workflows iTRAQ, ICAT, and SILAC we used SpectraST to create an artificial spectral library from the NIST library by simulating the effects of the modifications on the NIST spectra. Finally, for one plasma experiment that included both iTRAQ and phosphorylation modifications, it was not possible to create a spectral library including the modifications so no SpectraST searching was conducted.

Finding commonalities between two proteomics protein sets

```
Q12345             IQTCRLITPAEGPVVTKNSLARAQYECLGCVHPISTKSPDLEPVLRYAIQYFNNNTSHSHLFDLKE
P69890             IQTCRLITPAEGPVVTKNSLARAQYECLGCVHPISTKSPDLEPMKTEGSTTVSLPHSAMSPVQDEERDSGKEQ
tryptic peptides
seen in urine           LITPAEGPVVTK     AQYECLGCVHPISTK
tryptic peptides
seen in plasma          LITPAEGPVVTK                           TEGSTTVSLPHSAMSPVQDEER
```

Atlas build	Identifiers in complete mapping (c)	Nonredundant identifiers (nr)	Reason identifier chosen for nonredundant set
urine	Q12345 (2 peps) P67890 (2 peps)	Q12345	same peptide evidence so choice is arbitrary
plasma	Q12345 (1 pep) P67890 (2 peps)	P67890	more peptide evidence

Set intersection	Identifiers
plasma$_{c}$ ∩ urine$_{c}$	Q12345, P67890
plasma$_{nr}$ ∩ urine$_{nr}$	<empty>
plasma$_{nr}$ ∩ urine$_{c}$	P67890
urine$_{nr}$ ∩ plasma$_{c}$	Q12345

A tiny invented example illustrates various methods for taking the intersection between two proteomics protein sets. Here, two peptides are observed in urine and two in plasma; one is seen in both. The shared peptide maps to two protein sequences; therefore, both protein sequences are included in the complete mapping for both atlases. The second urine peptide also maps to both, but the second plasma peptide maps to only one.

In a protein list formed by complete mapping, both atlas builds contain the same two protein identifiers and thus 100% overlap. When redundancy is removed from each list, the two builds show zero overlap, even though they do share an observed peptide and it is quite possible that the same isoform(s) was/were observed for both atlases. When the nonredundant set for one is intersected with the complete mapping for the other, we get only one sequence. This is the approach used in the current work. Note that a different result is obtained depending on whether the nonredundant list for plasma is intersected with the complete mapping for urine, or vice versa.

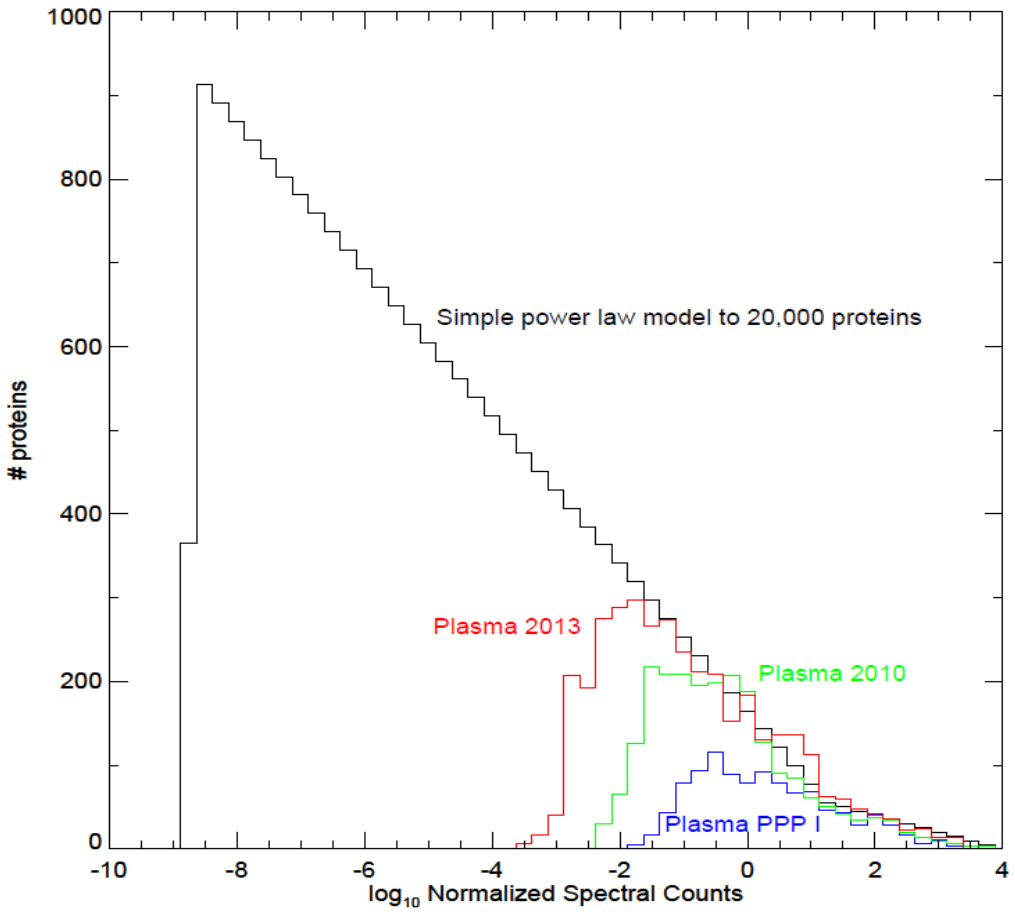


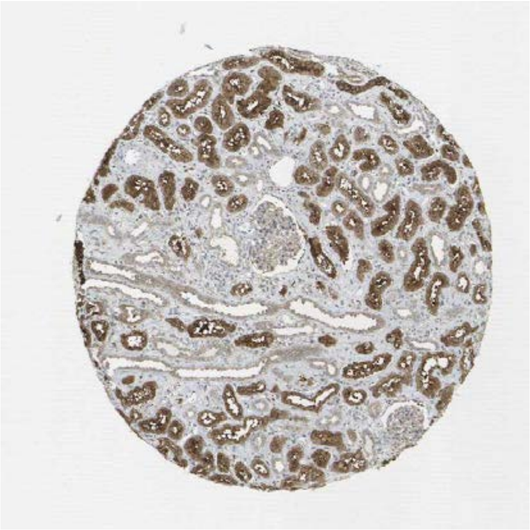
Figure S1: The number of proteins per log10 NSC bin for the current Plasma PeptideAtlas build as well as historical builds from 2010 and from the Plasma Proteome Project I data in 2005 overlaid. These histograms depict how the Plasma PeptideAtlas has pushed its coverage to lower abundances. In each case the low abundance limit is tapered due to the sensitivity drop off for shotgun mass spectrometry. There is no evidence that any fundamental limit in the true plasma abundance scale is being reached. For scale, a simple power law model (with a single inflection at log10 NSC = 1) aligned to the high abundance end and extrapolated to 20,000 proteins is shown.

Integrated Analysis of Genomic Variation and Protein Detection in Kidney, Urine, and Plasma: Seeking Clues for New Biomarker Candidates

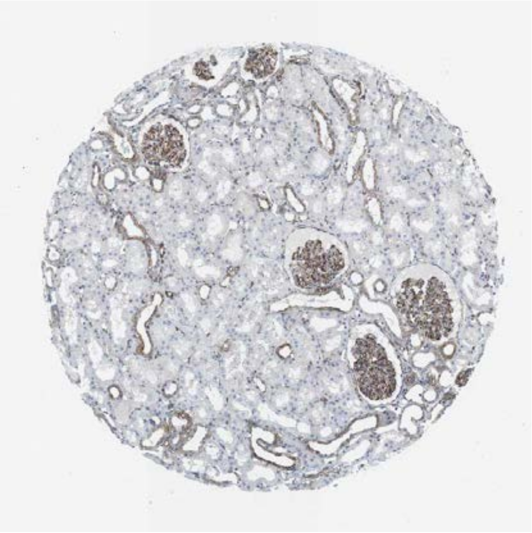
Kottgen, et al.[52] reported confirmation of five loci and discovery of 16 new loci statistically associated with decline in kidney function (glomerular filtration rate, measured with creatinine or cystatin c); they also found 7 loci associated with creatinine production and secretion. We examined each of the 11 loci for which the SNP variant was actually in the named gene, rather than somewhere nearby (their Table 2), as well as thereof the creatinine-associated loci; these 14 included all three that produced non-synonymous amino acid substitutions in the corresponding protein.

The most interesting protein is DAB2, a cytoplasmic adaptor protein expressed in renal proximal tubular cells and physically linking megalin and non-muscle myosin heavy polypeptide 9 (MYH9). DAB2 (P98082) was detected in KidneyPA in glomeruli, cortex, and mixed (but not medulla). It was not detected in UrinePA at all, but at a very low level in PlasmaPA. The two linked proteins were detected at higher levels. Megalin (low-density lipoprotein-related protein 2, LRP-2) is observed in all types of kidney specimens at about 4 times the level of DAB2; megalin has 552 KDa MW and a 25-residue signal peptide. Probably it is secreted into urine, as the UrinePA NSC of 67 is by far the highest among all of these gene products and ten times that for megalin in KidneyPA, yet it has not been detected in PlasmaPA. The myosin polypeptide MYH9, known to be primarily expressed in glomerulus [55], has far higher expression in KidneyPA, being detected in all samples studied, with a spectral count of 194 (30 times that of megalin). Lacking a signal peptide for secretion, it is detected in UrinePA at a level 1/10th of that of megalin, and it has been detected at a low level in PlasmaPA (0.29, which is, nevertheless, the highest value among all of the 14 genes in this special analysis). We attempted to confirm the colocalization of these three molecules via immunohistochemistry (SI Figure S2); while the patterns for DAB2 and megalin are similar, the pattern for MYH9 is quite different.

DAB2



MYH9



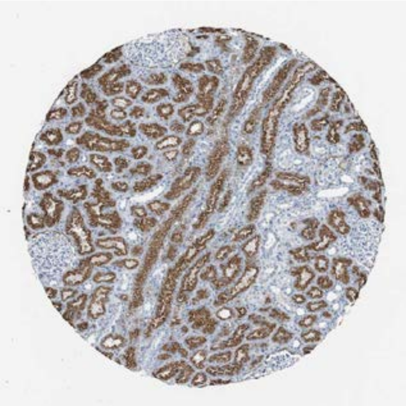
LRP2 (Megalin) 

Figure S2: Immunohistochemistry images for DAB2, MYH9, and LRP2 (megalin) in kidney (www.proteinatlas.org/ENSG00000144035/normal/kidney).

Another gene of interest is NAT8 (Q9EHE5), which is detected at the same level as DAB2 (NSC=1.78) in kidney, but not in urine and in plasma. It was found with 5 distinct peptides and 144 observations in cortex and medulla, but not in the study of isolated glomeruli. NAT8 is known to be expressed primarily in tubules, though it was not detected in tubules in KidneyPA. In the Human Protein Atlas the protein is seen in proximal tubules via immunohistochemistry. According to Kottgen et al, NAT8 is expressed only in kidney and liver; its SNPs have been associated with systolic blood pressure and kidney function previously. A nearby gene is ALMS1, in which mutations cause Alstrom syndrome; we found no evidence for this protein in kidney, urine, or plasma.

Gene ID	Swiss-Prot Accession	Total potential extramembrane tryptic peptides	SRM tryptic peptides	HPA Annotation
SLC34A1	Q06495	24	9	HPA positive in kidney and in 53/78 cell types
SLC7A9	P82251	6	7	HPA positive in kidney tubules and small intestine
SLC22A2	O15244	11	10	HPA positive in kidney and in 39/80 cell types
SLC6A13	Q9NSD5	9	9	Not yet analyzed in HPA

Table S6: Solute carrier proteins associated with declining renal function. *SRM tryptic peptides* refers to the peptides for which SRM assays are included in the Human SRMAtlas[56-57].

Fascinating results emerged for the solute carrier (SLC) proteins, which are highly hydrophobic membrane-embedded proteins. See Table S6 for a summary of the tryptic peptides and HPA observations for the four SLC proteins under discussion. SLC7A9 (P82251) is an amino acid transporter expressed in renal proximal tubule cells; mutations cause the important metabolic kidney disorder cystinuria type B, with urinary tract stones, interstitial fibrosis, and chronic renal insufficiency; a mouse model mimics the whole disorder. HPA shows strong positivity for renal tubules, yet we found no evidence of this protein in KidneyPA, UrinePA, PlasmaPA, HumanAllPA, or in the Muraoka et al exhaustive analysis enriched for membrane proteins[58] (JPR 2013). This protein has 12 trans-membrane domains, and very limited numbers of lysine and arginine residues to permit tryptic digestion, even for the limited exposed sequences. Such proteins surely contribute to the “missing proteins” of the Human Proteome Project. SRM assays for multiple peptides in each of these proteins have been developed and are presented in the

Human SRMAtlas[56-57] ; see Table S6. Another factor is the difficulty of solubilization from tissues for MS sample preparation.

Another SLC protein SLC22A2 (O15244) was detected in kidney (NSC=3.1) in cortex and medulla, but not in urine or plasma. The kidney has the highest level of transcript expression for this gene. Like SLC7A9, this protein of 555 amino acids has 12 transmembrane domains; the two tryptic peptides detected in kidney (44 total observations) come from the only two long extra-membrane topological domains with peptide 53-62 in a 107-residue sequence and peptide 331-340 in a 64-amino acid sequence. Yet another creatinine-associated protein SLC6A13 was unseen in KidneyPA, UrinePA, and PlasmaPA, and not yet analyzed in HPA. The Human-All PeptideAtlas does contain two peptides (both in extracellular domains) with 14 observations from brain and cancer cell lines.

A few final observations: DACH1 was detected in kidney (NSC 0.32), but not in urine or plasma, with 2 peptides and 4 observations in glomeruli. PRKAG2 similarly had NSC of 0.081 in KidneyPA, but all its observed peptides appear in a related protein with more evidence (PRKAG1) . And SHROOM3 had NSC of 0.15 in kidney with 5 peptides, 11 observations, in glomeruli, and no detection in urine or plasma. All the other proteins listed in Table 4 of the main text were undetected in the 3 HKUP builds.

Going from Genome-Wide Association Studies (GWAS) of 67,000 people to monitoring for disease risk in individuals requires protein biomarkers. In fact, the 23 loci described in

Kottgen, et al.[52] account for only 1.4% of the attributable risk for chronic kidney dysfunction. Here we have demonstrated the use of PeptideAtlas and other proteomics databases to guide the development of potential disease biomarkers. From this whole list selected from Kottgen et al, only the DAB2/megalin-LRP-2/MYH9 complex of proteins, all of which were detected in KidneyPA and two at NSC>5, and all of which were detected in UrinePA and/or PlasmaPA, emerge as striking protein candidates for renal disease studies. It is possible that these proteins may change long before declining glomerular filtration rate leads to increases in creatinine or blood urea nitrogen, the standard clinical assays. It is well known that 75 percent of glomerular function must be lost before the standard clinical analytes creatinine and BUN begin to rise above the normal range.