Supplementary Screencast:

The supplementary screencast demostrats in detail CellFinder functionality and content access at:

http://www.cellfinder.org/help/screencast/

Supplementary Methods:

Microarray data. Microarray studies were selected from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). All calculations were carried out using Bioconductor packages in the statistical language R. For Affymetrix microarrays (unless indicated otherwise), all CEL files within one study were normalized together using the rma algorithm. Present/Absent calls were calculated using the Bioconductor package "affy". Probesets were considered 'present' if the detection p-values were less than 0.05; otherwise the probesets were considered 'absent'. For microarrays originating from other platforms (*e.g*. Illumina, Agilent), normalized data (as submitted by the authors) was obtained directly from GEO unless specified otherwise. A current list of datasets can be found in Supplementary Table 1.

RNA-seq data. RNA-seq normalized data was downloaded from the RNA-seq Atlas (41) to display gene expression profiles in a set of 11 normal tissues.

Molecular datasets for comparison. Differential expression was determined for one microarray dataset and one protein dataset. The microarray dataset is a subset of StemBase (11, 48), and consist of gene expression profiles of murine stem cells and derivatives, which was used to define novel stem cell markers in the web tool MarkerServer (49). Differentially expressed genes were computed between all possible pairs of samples using the limma package. The protein dataset is from the Human Proteome Atlas (Ref HPA). Differentially expressed proteins were pre-computed between all possible pairs of samples in HPA to point to proteins with different expression according to four qualitative levels (high, medium, low, none).

Supplementary Methods:

Algorithm for determining relevance of text based search results

The following pattern-matching algorithm is applied to calculate relevance:

1. Exact filtering so that at least one "search word"* is contained within a terms name.

* Search word refers to a word within the search input; e.g.: search for "hepatic liver cell" then all terms that contain neither of the search words - "liver", "hepatic" or "cell" - are filtered out.

2. The resulting hits are then sorted by relevance according to the following algorithm:

For each label of every term we calculate a string similarity and use the highest value as the relevance.

String similarity calculation: Each word of the search input is compared to each word of the term label. If the label word contains the search word, then the similarity score is increased as follows:

If the length of the label word is equal to the reference word length:

similarity score += (((1 / Number-Label-Words + String-Length-Search-Word / String-Length-Label) /

2) + (1 / Number-Search-Words + String-Length-Search-Word / Number-Search-Words) / 2) / 2;

Else:

similarity score += (String-Length-Search-Word / String-Length-Label + (1 / Number-Search-

Words + String-Length-Search-Word / Number-Search-Words) / 2) / 2;

The similarity score is calculated for all word pairs and adds up to the final relevance score.

A similarity score of 0 means no hits at all. This does not exist on the web page because of the first filtering step (At least one word does need to match before the score is actually calculated). A score of 1 or higher means it's absolutely relevant. (Basically, 1 should be the maximum but in case the search input or the label contains duplicated words, then the score is higher. This is not the best outcome, but because searches like "liver liver liver liver liver..." are very uncommon, it does not matter in regular use cases).

As this algorithm is not state of the art, although workable at the moment, the de-facto standard Apache Lucene open source solution will be implemented in the near future.

Supplementary Table 2:
Web Application Browser compatibility list.

Screen Resolution		
Minimal *(No horizontal scrolling is needed)*	Dimension	Pixel
	Width	800
	Height	400
Devices		
CellFinder works on any device that features one of the modern browser listed below. A screen with the minimal resolution is strongly recommended, though smaller resolutions work as well.		
Browser		
Platform	Application Name	Supported versions
Desktop *(Windows, Mac and Linux)*	Google Chrome	≥ 17
	Safari	≥ 4.0.5
	Firefox	≥ 3.6
	Opera	≥ 11.52
	Internet Explorer	≥ 9
Mobile	iOS *	≥ 6
	Android Browser	≥ 4

* All browsers use the same rendering engine that’s why it technically doesn’t matter which browser is used in iOS.

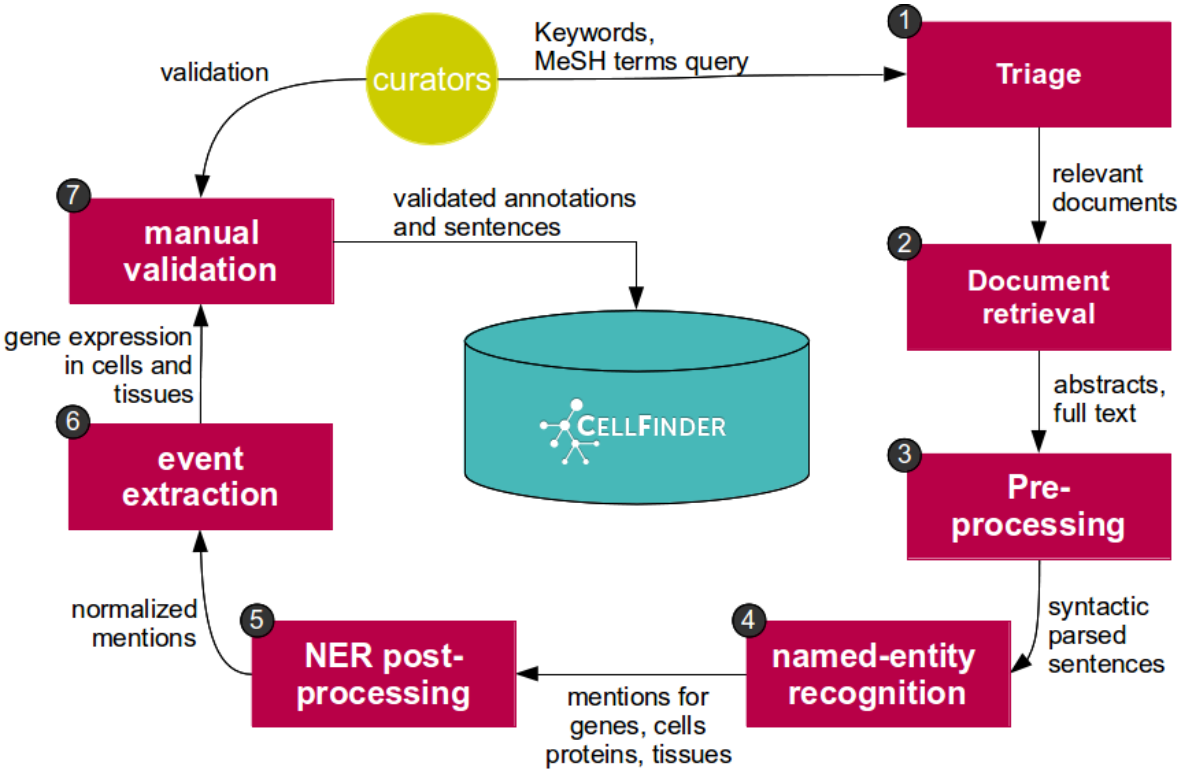
Supplementary Table 3

Current roadmap for CellFinder development and updating schedule. The implementation of organs is planned in a stepwise fashion and priority level based on user community feedback, which will be re-evaluated annually.

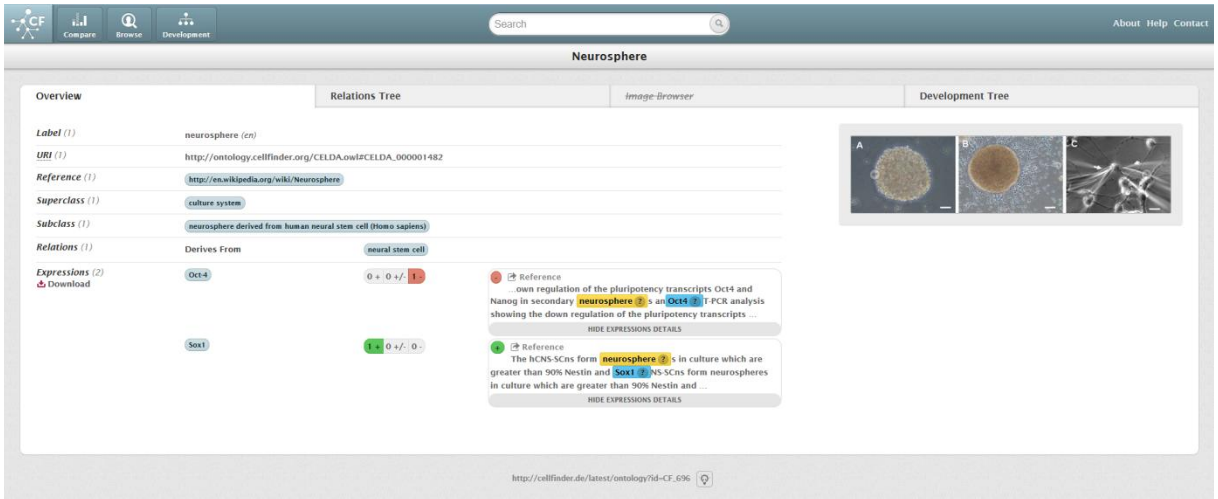
Priority for Implementation of organs and species		Priority
Solid organs	Kidney, liver	Implemented
	Spleen, thymus	High
	Pancreas,	Medium
	Digestive tract	Medium
	Eyes, ears	Low
Organ systems with systemic features	Hematopietic system	Very high
	Cardiovascular system	Very high
	Musculoskeletal system and	High
	Skin (incl. e.g. breast)	High
Nervous system	Peripheral nervous system	High
	Central nervous system	Medium
Species		
	Human	Very high
	Mouse	High
	Zebrafish	Medium

Updating schedule		Frequency
Ontology		Quarterly
Web application	With every newly developed functionality	Usually every 3-6 months
Gene/protein expression data	in line with curation of these data	Continuous
Bug fixing		Immediately

Supplementary Figure 1: Overview of the literature curation pipeline for CellFinder: (1) triage of relevant documents, **(2)** retrieval of abstracts and full texts, **(3)** preprocessing (sentence splitting, tokenization and parsing), **(4)** named-entity recognition (NER; genes, proteins, cell lines, cell types, organs, tissues, expression triggers), **(5)** named-entity post-processing (acronym resolution, ontology mapping), **(6)** gene expression events extraction, **(7)** manual validation of the results. Manual intervention from the biocurators is only necessary for the triage and the final validation of the results.



Supplementary Figure 2: Visualization of text mining data in CellFinder. Text mining was performed after manual annotation of a literature corpus of human embryonic stem cells and kidney development for training and subsequent mining of literature followed again by manual validation. Currently, only gene and protein expression data are shown in CellFinder. The data are provided through the cell and tissue specific information as shown for the term 'neurosphere', an in vitro generated structure consisting mainly of neural stem cells visualized using sentence-based syntax highlighting.



Neurosphere

Label (1)	neurosphere (en)		
URI (1)	http://ontology.cellfinder.org/CELDA.owl#CELDA_000001482		
Reference (1)	http://en.wikipedia.org/wiki/Neurosphere		
Superclass (1)	culture system		
Subclass (1)	neurosphere derived from human neural stem cell (Homo sapiens)		
Relations (1)	Derives From	neural stem cell	
Expressions (2) Download	Oct-4	0 + 0 +/- 1 -	Reference ...own regulation of the pluripotency transcripts Oct4 and Nanog in secondary neurosphere ? s an Oct4 ? T-PCR analysis showing the down regulation of the pluripotency transcripts ... HIDE EXPRESSIONS DETAILS
	Sox1	1 + 0 +/- 0 -	Reference The hCNS-SCns form neurosphere ? s in culture which are greater than 90% Nestin and Sox1 ? NS-SCns form neurospheres in culture which are greater than 90% Nestin and ... HIDE EXPRESSIONS DETAILS

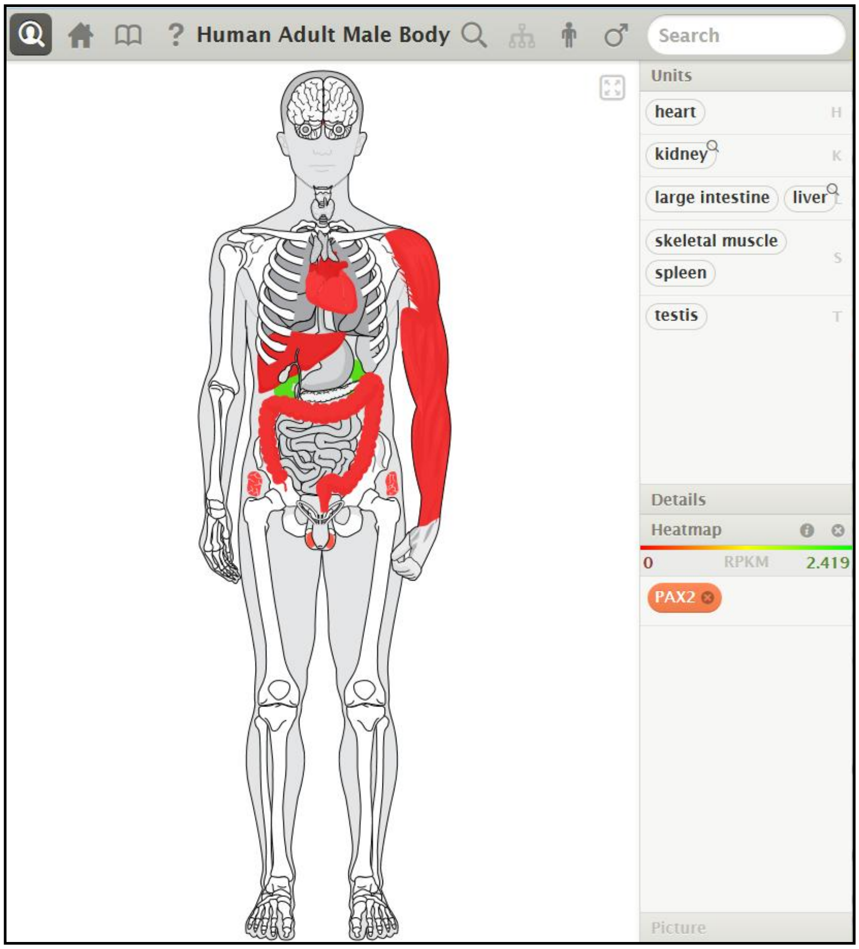
http://cellfinder.de/latest/ontology?id=CF_696

Supplementary Use Case 1:

Characterization of cell derived by stem cell differentiation

CellFinder provides reference data and tools that aid in the development of cell based regenerative therapies. Here the major challenge is to direct stem cell differentiation into the direction of a specific organ to derive cells that have a regenerative potential for example when transplanted into a diseased organ.

One example is that a user wants to derive kidney cells from pluripotent stem cells. To direct stem cell differentiation he tested different stimuli and derives at the end several cell populations. To test which population is closest to the desired cell type the user measures their gene expression (e.g. by PCR or microarrays) which results in a list of differentially expressed genes compared to the undifferentiated stem cells. Now these genes can be tested separately or in combination using the heatmap feature of the semantic body browser (see figure below). The displayed heatmaps will indicate the organ to which the tested cell population is related to.



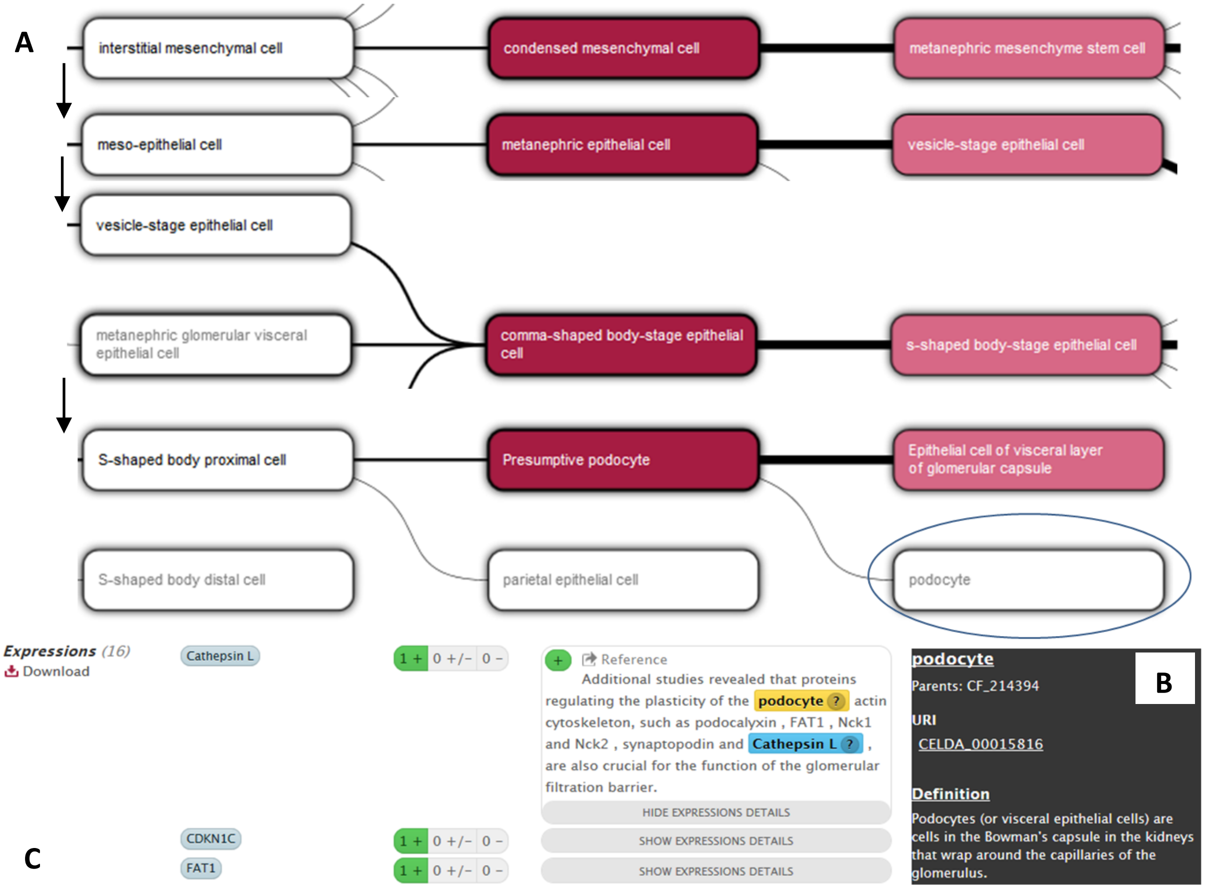
Example of a heatmap of the expression of the gene PAX2: The green coloring of the kidney and the red coloring of the other organs indicates a high specificity of PAX2 expression in renal tissue. Therefore PAX2 can also be used as a Marker for kidney cells.

Supplementary use case 2:

Differentiation pattern of iPS-derived renal precursor cells and related gene expression

CellFinder provides developmental trees for cell types, identifying origin and differentiation fate of a given cell. This user case is interested in the fate of a metanephric mesenchymal stem cell, which can be derived from a pluripotent stem cell. Furthermore, the user is interested in the intermediate stages that the tissue specific stem cell passes through while differentiating towards a terminally differentiated podocyte. Finally, the genes or proteions that are expressed and characterize the intermediates and terminal cell types are queried.

The figure below provides (**A**) screenshot sections of the developmental steps, leading from a metanephric mesenchymal stem cell (top right pink box) to a podocyte (bottom right circled box). Only parts of the tree are shown. In addition to the forward looking cell fate, the tree also provides information on the origin of the metanephric mesenchymal stem cell. (**B**) For each cell, a window displays CELDA derived infomration, including the CELDA ID. (**C**) This ID, or the cell term is linked further information about the selected cell (e.g. podocyte) such as description, relations, gene and protein expressions, including those identified by text mining.


A
interstitial mesenchymal cell
condensed mesenchymal cell
metanephric mesenchyme stem cell
meso-epithelial cell
metanephric epithelial cell
vesicle-stage epithelial cell
vesicle-stage epithelial cell
metanephric glomerular visceral epithelial cell
comma-shaped body-stage epithelial cell
s-shaped body-stage epithelial cell
S-shaped body proximal cell
Presumptive podocyte
Epithelial cell of visceral layer of glomerular capsule
S-shaped body distal cell
parietal epithelial cell
podocyte
Expressions (16)
Download
Cathepsin L
1 + 0 +/− 0 −
+
Reference
HIDE EXPRESSIONS DETAILS
CDKN1C
1 + 0 +/− 0 −
SHOW EXPRESSIONS DETAILS
FAT1
1 + 0 +/− 0 −
SHOW EXPRESSIONS DETAILS
C
B

Additional studies revealed that proteins regulating the plasticity of the podocyte ? actin cytoskeleton, such as podocalyxin , FAT1 , Nck1 and Nck2 , synaptopodin and Cathepsin L ? , are also crucial for the function of the glomerular filtration barrier.

podocyte

Parents: CF_214394

URI

CELDA_00015816

Definition

Podocytes (or visceral epithelial cells) are cells in the Bowman's capsule in the kidneys that wrap around the capillaries of the glomerulus.