

## Supplementary Materials for **Analysis of Metagene Portraits Reveals Distinct Transitions During Kidney Organogenesis**

Igor F. Tsigelny\*, Valentina L. Kouznetsova, Derina E. Sweeney, Wei Wu, Kevin T. Bush, Sanjay K. Nigam\*

\*To whom correspondence should be addressed. E-mail: snigam@ucsd.edu (S.K.N.) and itsigeln@ucsd.edu (I.F.T.)

Published 9 December 2008, *Sci. Signal.* **1**, ra16 (2008)  
DOI: 10.1126/scisignal.1163630

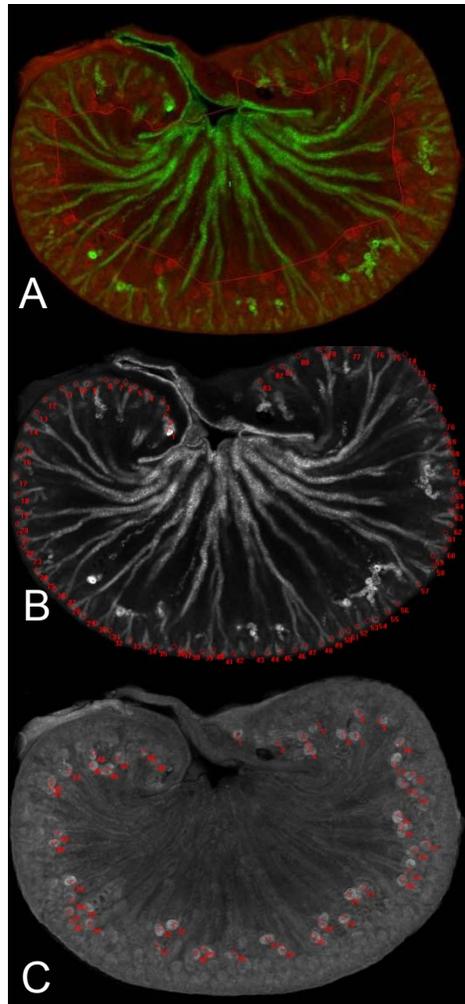
### **This PDF file includes:**

- Fig. S1. Process of obtaining measurements of glomeruli and tips from morphological images.
- Fig. S2. Three entropy profiles calculated from SOMs with different map resolutions:  $16 \times 15$ ,  $26 \times 25$ , and  $36 \times 35$  tiles.
- Fig. S3. Legend for network shapes and relationships presented in Figs. 4 and 5 (using IPA from Ingenuity Systems, [www.ingenuity.com](http://www.ingenuity.com), with permission).
- Fig. S4. Schematic of generation of SOMs, entropy calculations, correlations, and network analysis.
- Fig. S5. Robustness of SOM clustering.
- Table S1. List of genes with high correlation coefficients ( $>0.85$ ) between expression profiles and glomerular density values during kidney development.
- Table S2. Preliminary analysis of gene expression by RT-PCR.

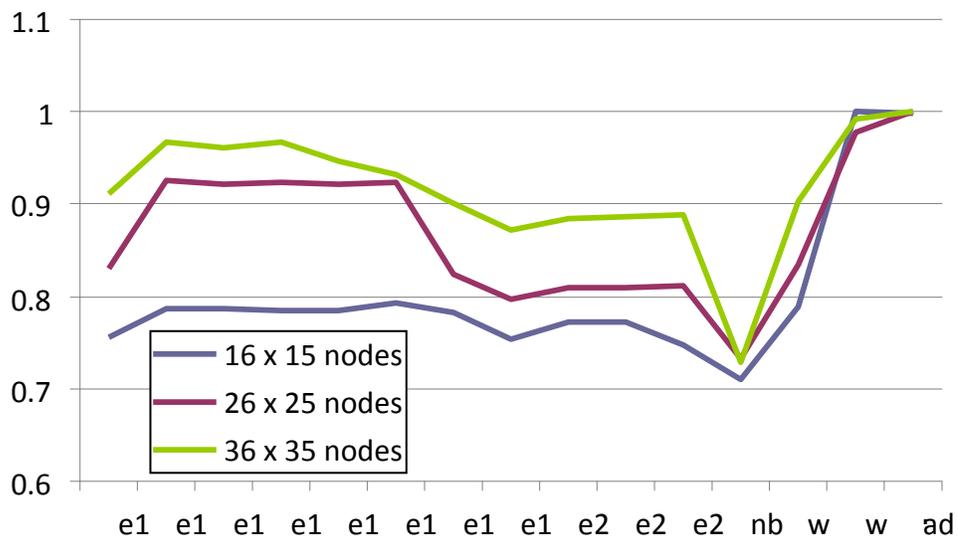
**Other Supplementary Material for this manuscript includes the following:**  
(available at [www.sciencesignaling.org/cgi/content/full/1/49/ra16/DC1](http://www.sciencesignaling.org/cgi/content/full/1/49/ra16/DC1))

Table S3. Affymetrix probe IDs of the genes found in each of the 650 metagene tiles.

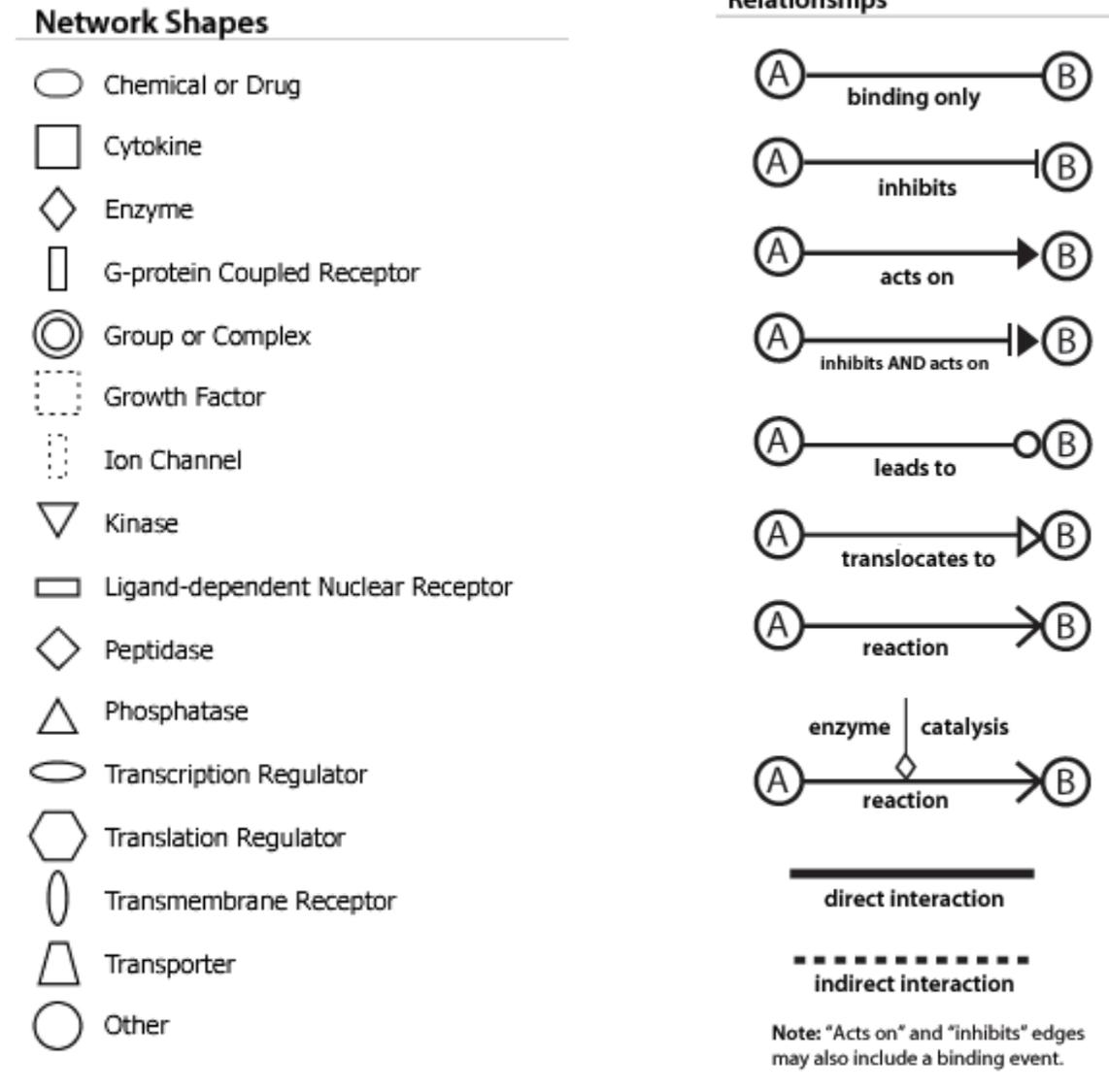
**Fig. S1.** Process of obtaining measurements of glomeruli and tips from morphological images. Image of 200 $\mu$ m section from an e19 kidney stained with FITC-labeled Dolichos biflorus agglutinin (green) and TRITC labeled Peanut agglutinin (red) (**A**). Tips were counted from the DBA stained image of tissue (**B**), while glomeruli were counted by the PNA staining (**C**).



**Fig. S2.** Three entropy profiles calculated from SOMs with different map resolution:  $16 \times 15$ ,  $26 \times 25$ , and  $36 \times 35$  tiles. The overall shapes of entropy profiles during development are similar for all three resolutions but, for reasons discussed in the text, the  $26 \times 25$  map was deemed most suitable.



**Fig. S3.** Legend for network shapes and relationships presented in Figures 4 and 5 (using IPA from Ingenuity Systems, [www.ingenuity.com](http://www.ingenuity.com), with permission).



**Fig. S4.** Schematics of generation of SOMs, entropy calculations, correlations, and network analysis.

**(Step 1)**

**(a) Data preparation.** Three samples were collected for each time point (day of development), using GeneChip Rat Genome 230 2.0 Arrays (Affymetrix Inc., 2007). Data preparation includes statistical preprocessing and data-format conversion: (a) checking for outliers and removal if necessary; (b) for every time point, calculation of average value for each probe; (c) quantile normalization for each probe across the time series with the formula  $NE_{i,j} = E_{i,j} / E_{max,j}$ ;  $i \in \{1, \dots, 15\}$ ;  $j \in \{1, \dots, 31099\}$ , where  $NE_{i,j}$  and  $E_{i,j}$  are a normalized and a measured expression for a specified sample for each probe, correspondingly, and  $E_{max,j}$  is the maximum expression for each probe.

In microarray experiments there are a number of “missing” expression values. In many cases, these values are missed because of some technical problems or because they are assigned by the intrinsic manufacturer’s program as absent (“A”) and are excluded. Clusterization methods, including *K*-means clustering, SOMs, etc., may not always be robust in the context of missing data. To solve this problem, a set of methods for data substitution has been developed. The missed values are, for example, substituted by (1) zeros, (2) average values, and (3) values defined by methods based on comparison with expression profiles of the other genes, e.g., weighted *K*-nearest neighbors (KNNimpute) [Troyanskaya *et al.* (2001)]. In this case, we could not use the imputing methods based on “neighboring” gene profiles because we aimed to also study the clustering with SOMs, and adding similar profiles for genes might add artifacts for the map self-organization process. Therefore, we used only average substitutions and compared unfiltered data to data filtered in this manner.

Thus, we use two strategies: (1) initial normalized set of unfiltered data was used for generation of SOMs with the GEDI program. The resulting genes included in each metagene set were filtered for “A” values and their expression values were evaluated; (2) initial normalized set of genes was filtered for genes having all absent calls or genes having only a single present or marginal call through the time series. These genes were excluded, and for the remaining genes

data values with absent calls were substituted with the average values of the data points having present calls for each gene in the time series separately. Figures S5A and S5B show that the SOM is a robust method and the results for defining stages of development are quite similar for both data preparation strategies (unfiltered and filtered).

**(b) SOM generation.** For Experiments 1 and 2, we used the Gene Expression Dynamic Inspector (GEDI). vector. In Experiment 1, GEDI generated 15 SOMs for each time point studied, as well as a density map, and allowed for export a centroid-value list for each map. For Experiment 2, the second-level clustering, map centroids were used to create metametagenes whose normalized expressions were converted to input GEDI format and fed into GEDI. For the purpose of classification of stages of organ development, we used the density map. The closest adjusted sectors were colored similarly (except the week 1 after birth with week 4 after birth and adult because further analysis showed that their hierarchical clustering and Tsallis entropies differ significantly).

## **(Step 2)**

Output of **Experiment 1** was used for Tsallis entropy calculation. First, 15 histograms were generated. Then each histogram was quantized for five color-intensity groups, and the quantity of points in each group was counted and probability for each group was calculated. Five samplings for histogram were used because further study performed indicated no significant difference in entropy profiles between five and higher-value samplings. Then Tsallis entropy values were calculated using the formula embedded in the flowchart figure. In the equation,  $p$  denotes the probability distribution of the mosaic, and  $q$  is a real parameter that is related to non-extensivity. When a parameter is “extensive” it depends on the size of the system, for example, mass. When a parameter is “intensive,” i.e., “non-extensive,” it does not depend on the size of the system, for example, density or temperature. The developing organ has both properties: It is growing, changing its size, and simultaneously developing some structures with different topology.

Tsallis entropy was chosen because it has been proposed that this method may be well suited to calculate non-extensive or mixed entropy. Parameter  $q$  is thought to quantify the degree of departure from extensivity: if  $q = 0$ , the system is intensive, if  $q = 1$ , the system is extensive, and its entropy is equivalent to Boltzmann–Gibbs or Shannon entropy, which are both extensive entropies. The Tsallis entropy for each mosaic was calculated and the profile chart was created.

### (Step 3)

**Morphometric measurements** for the developing kidney were obtained, including kidney section area, kidney section perimeter, kidney section aspect ratio, kidney section major and minor axes, kidney roundness, cortex area, medulla area, Feret whole kidney, cortex, and medulla areas, numbers of tips and glomeruli, and tip and glomerular densities. Spearman rank correlation coefficients between entropy and these measurements were calculated.

The morphometric parameters with the highest correlation were glomerular density, cortex area, and medulla area.

Because increasing entropy could be an indicator of decrease of organization and organization seems to increase during organ development, to obtain positive correlation, “reversed” entropy was calculated, multiplying entropy values by  $(-1)$ . This parameter correlated with morphologic measurements. For convenience values were normalized between 0 and 1.

After highly correlated morphologic parameters were found, the Spearman rank correlation coefficients between each gene in the microarray and the selected morphologic parameter were calculated. Genes with top-correlated expression profiles were mapped back and were found in the compact central area that is outlined for clarity. There is one tile in the center that is excluded because no highly correlated gene fell within this tile.

The output of this calculation is the expanded list of genes shown in Table S1.

#### (Step 4)

**Gene network analysis** was done using IPA software (Ingenuity® Systems, [www.ingenuity.com](http://www.ingenuity.com)).

There were two steps of this analysis:

1. The network is generated from genes that are highly correlated with morphological parameters (see Table S1). The IPA creates multiple networks with different scores. The network with the highest score is presented in Figure 4B.
2. The input of network shown in Figures 5B and 5C is the same set of genes selected from the outlined area of SOMs (see Figure 5A). The same set of genes but with expressions corresponding each day of development (e13, e14, e15, e16, and e17) was imputed to the IPA. For each day of development the IPA generated at least four networks. The networks for days of development e13-e16 show similar configurations and are overlaid with the close values of gene expressions (more than a half of the genes are down regulated). Because of that the network of the day e16 was shown in Fig. 5B as a representative of this stage. The network for e17 has slightly different configuration and is overlaid with up-regulated expression values. Figures 5B and 5C represent the networks with the highest scores assigned by IPA.

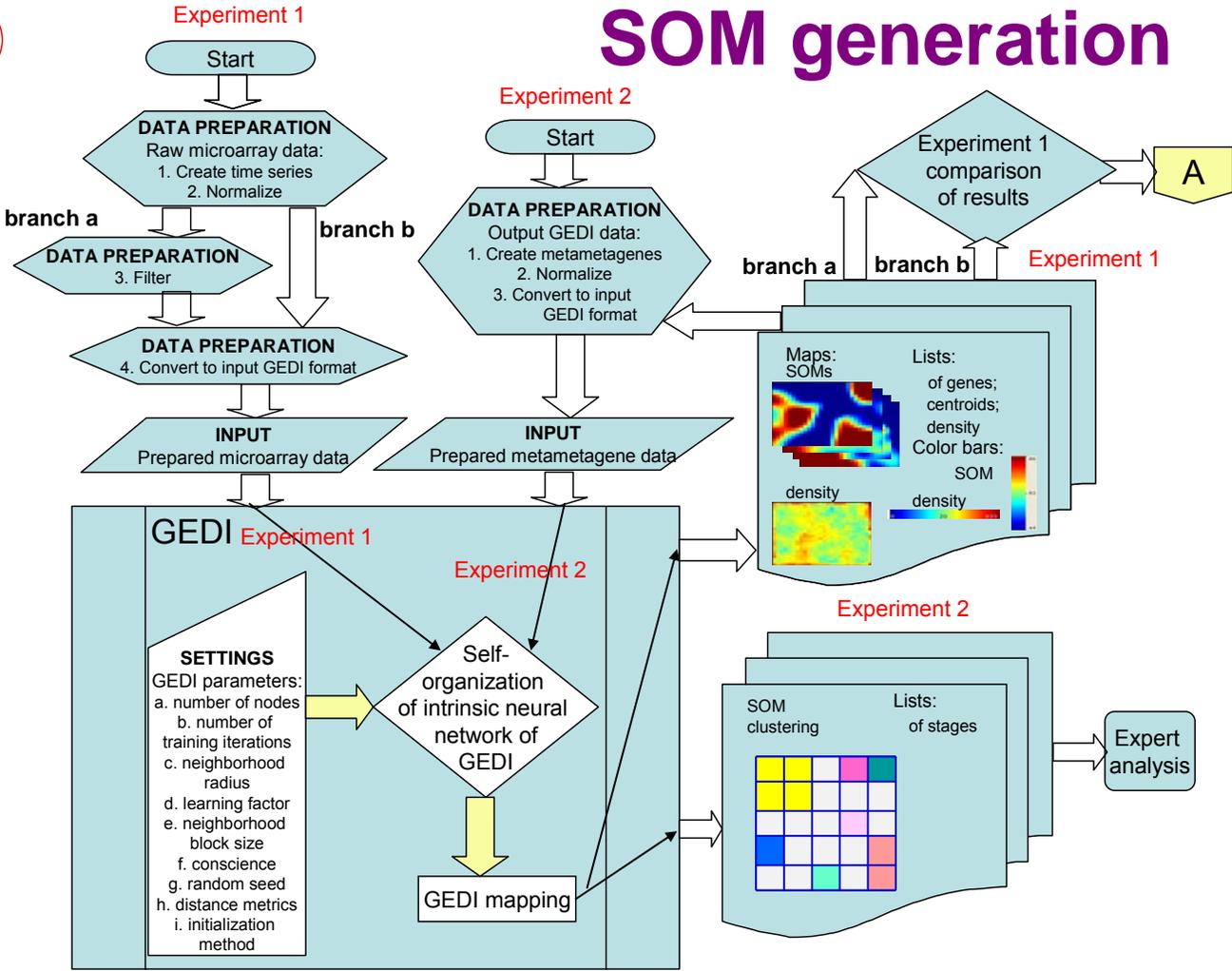
Expression of several genes was confirmed by Real-Time PCR.

#### **References.**

1. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).

1

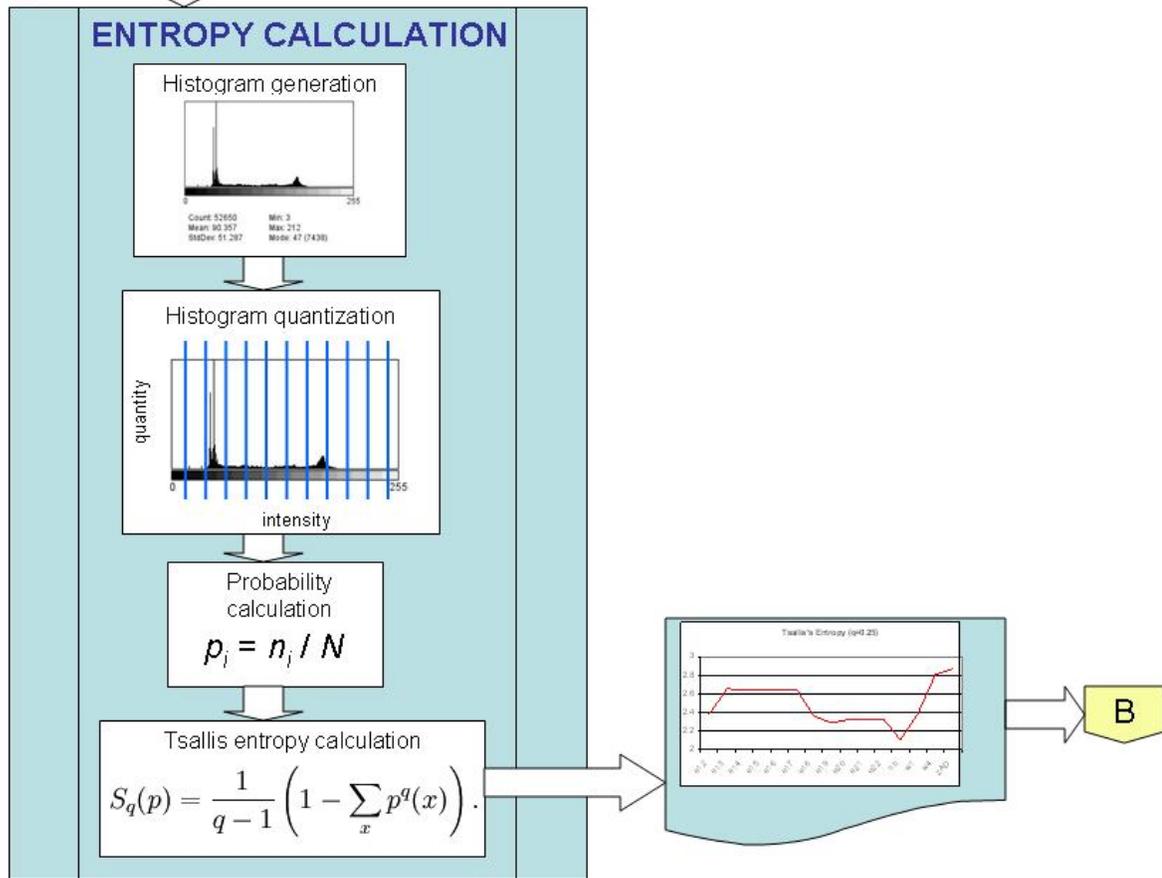
# SOM generation



2

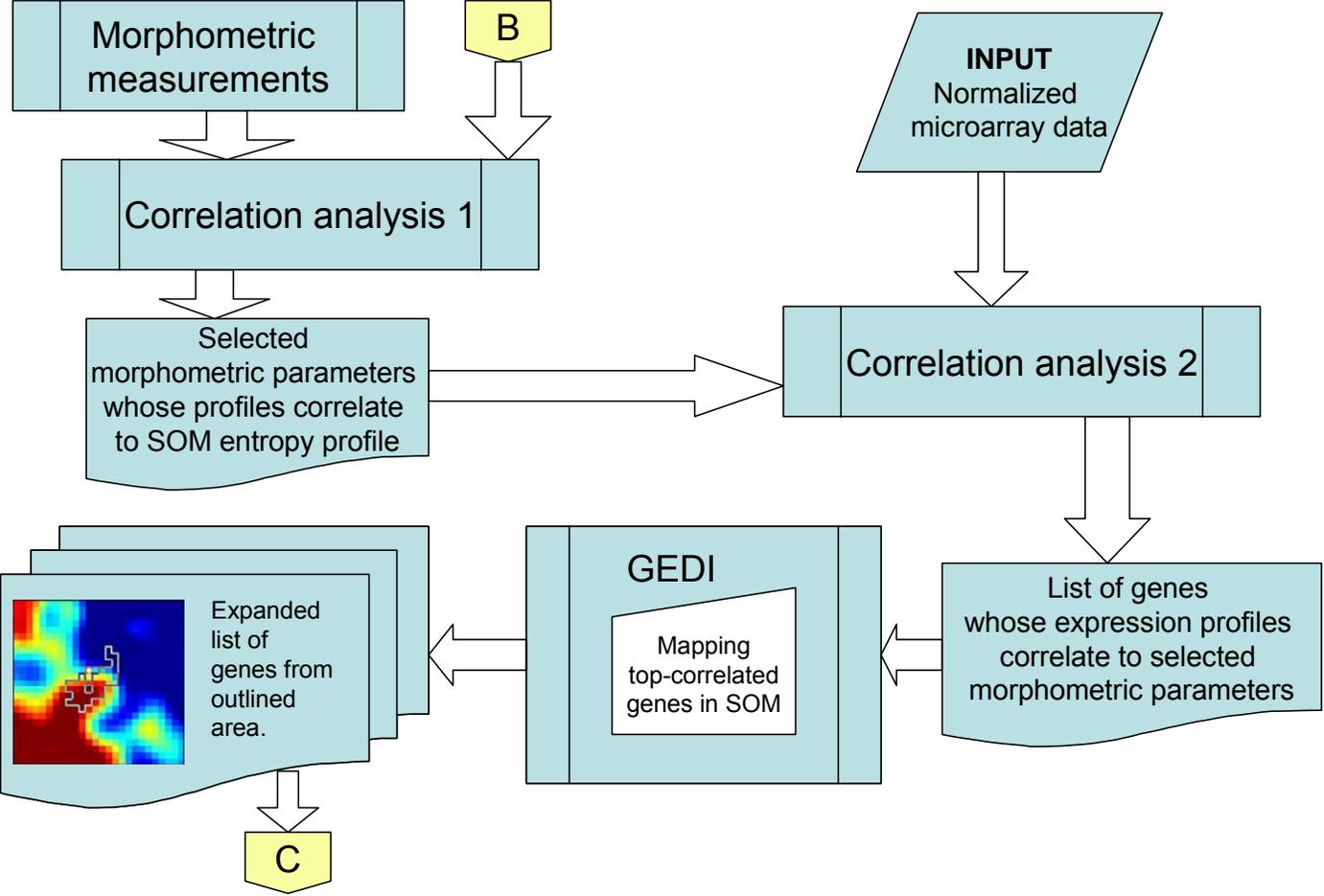
A

# SOM entropy calculation

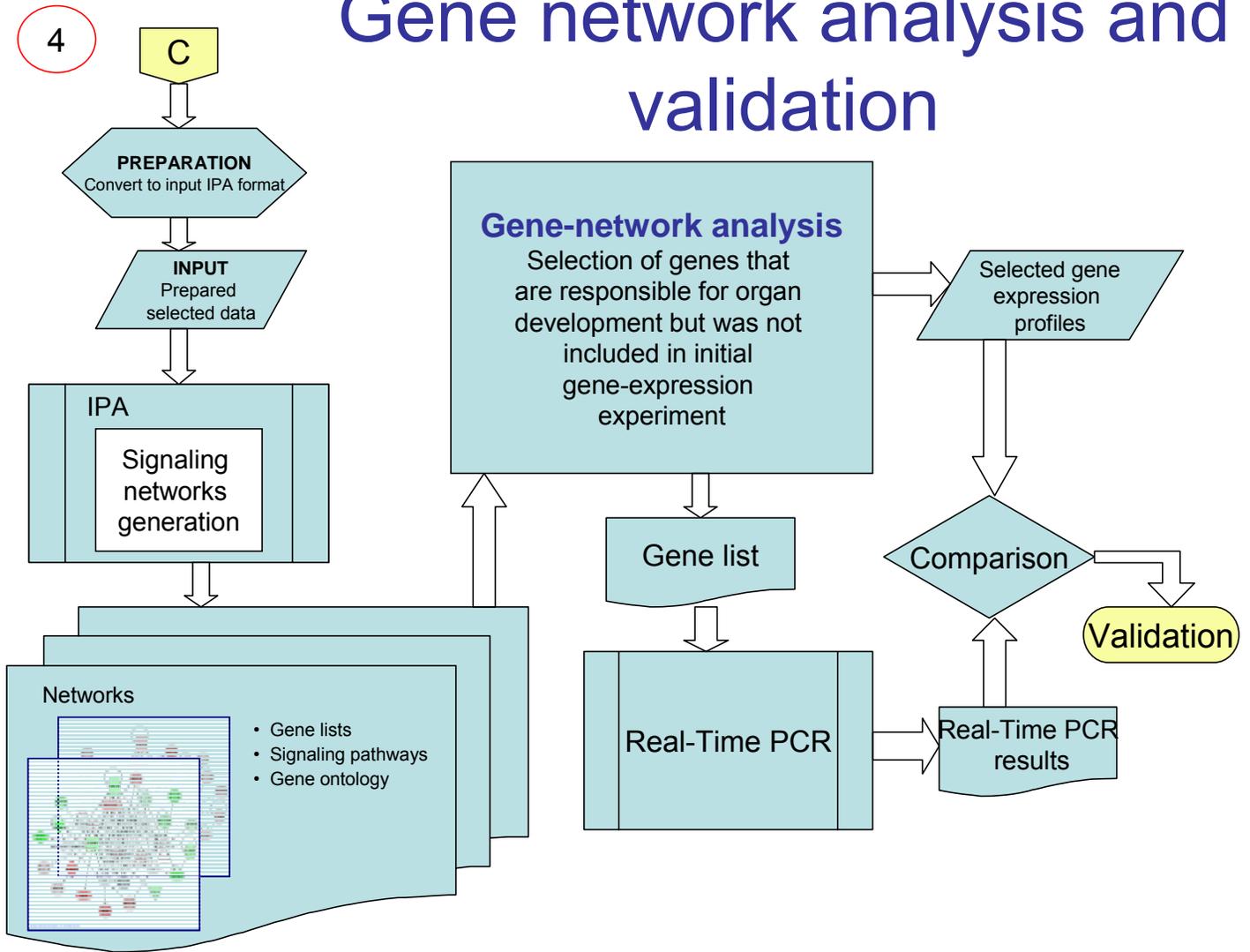


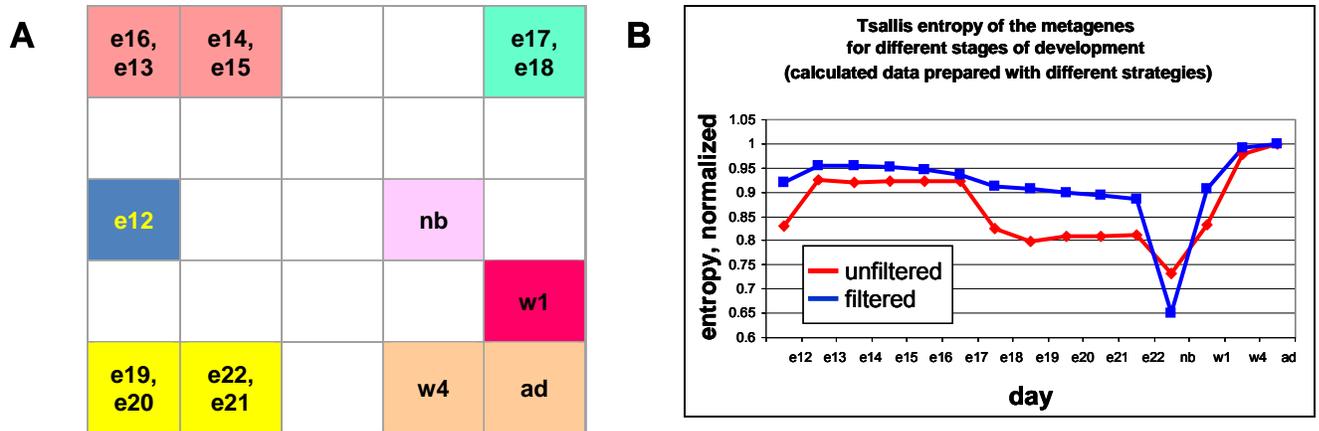
3

# Correlation between SOMs, morphometric parameters, and gene expression



# Gene network analysis and validation





**Fig. S5.** Robustness of SOM clustering. The microarray dataset was filtered the following way: Data values were normalized through each time series and the genes having all absent calls or/and the genes having a single present call and the rest absent calls through the time series were excluded. In the remaining data, gene-expression values having absent calls were substituted for the average values of the data points having present calls for each gene in time series separately.

(A) Clusterization results of kidney developmental stages. The clusterization method is similar to that shown in Fig. 2A. Although the stages are differently situated in this second-level SOM, their number and grouping remains the same. (B) Tsallis entropy profiles calculated from SOMs with  $26 \times 25$  metagene grid, using two different strategies of data preparation: Red curve is calculated on the basis of data from the entire normalized unfiltered microarray dataset. Blue curve is calculated on the basis of the filtered microarray dataset described above. Spearman rank correlation between these two entropy profiles is 0.92.

**Table S1.** List of genes with high correlation coefficients (>0.85) between expression profiles and glomerular density values during kidney development.

Genes	Corr	Genes	Corr	Genes	Corr	Genes	Corr	Genes	Corr
Cox4i2	0.9451494	Col4a1	0.9075759	Zfp346_predicted	0.8942287	Mcam	0.878355	Egfl7	0.8617828
Npr1	0.9403069	Vamp1	0.9075755	Em13_predicted	0.8942176	Rasd1	0.8779989	Rhbdf1	0.8605113
Sparc	0.9348365	4-Sep	0.9072753	Scin	0.8931662	Qscl6	0.8768528	Podxl	0.8594388
Ncstn	0.9346081	Col4a2_predicted	0.9072514	Hspb7	0.8930977	Phldb1	0.8765086	Naga	0.8585008
Col18a1	0.9281299	Mrgpr1	0.9072506	Dag1	0.8915806	Cdwr92	0.8758889	Slc44a2_predicted	0.8580257
Pofut2_predicted	0.9258626	Stap2	0.9070598	Asb8_predicted	0.8913566	Col1a1	0.8746371	Tnnc1	0.8574948
Ndrp4	0.9245606	Dpep2	0.9064534	Upk1b	0.891285	Olfm1	0.8739733	Tpm1	0.8569697
Cdh1	0.9225536	Thbs2	0.9051934	Nphs1	0.8907958	Numbl	0.8739186	Impa2	0.8566211
Mgst2_predicted	0.9177333	Col6a2	0.9022767	Atp5s	0.8888159	Plscr1	0.8736973	Hsd3b7	0.8558939
S100a9	0.9162048	Ralb	0.9016633	Fhl2	0.8885019	Sema4g_predicted	0.872961	Acta2	0.8554283
Tgfb1	0.9158145	Heyl_predicted	0.9002183	Gga2	0.8884444	Ltbp4	0.870723	Ptgis	0.8544333
Pcp4	0.9154757	Galt	0.8993685	Tcfap2b_predicted	0.8874332	Cpt1b	0.8697296	Dapk1_predicted	0.8540845
Col4a1	0.9145593	Fgf13	0.8990253	Hod	0.8862416	Slahbp1	0.8688832	Lcn2	0.8539896
Alkbh3	0.9134476	Papss1_predicted	0.899021	Sumf2	0.8847608	Aqp9	0.8679087	Aph1b	0.8538776
Col1a1	0.9133456	Pdgfra	0.8989432	Upk3a_predicted	0.8838573	Adora2a	0.8678172	Tnfrsf1a	0.85316
Il3ra	0.9119605	Ppap2b	0.898494	Dapk1_predicted	0.883794	Krt1-19	0.8677332	St6galnac3	0.8525651
Pld2	0.9114148	Adcy6	0.8983796	Sh3glb2	0.881923	Ppp1r14a	0.8664892	Ptpro	0.8500485
Pigt_predicted	0.9105552	Cldn1	0.897153	Adams8_predicted	0.8817734	Nid2	0.8663584		
Tspan8	0.9102256	Vat1	0.8971444	Slc27a1	0.8817674	Kidins220	0.8658676		
Kcnj8	0.9101354	Tpo1	0.8968284	Lox	0.881546	Cldn1	0.8636269		
S100a8	0.910109	Evpl_predicted	0.8963297	Mmp23	0.8812255	Cldn1	0.8629992		
Zplp2	0.9095394	Tagln2	0.8961216	Mylk_predicted	0.8804149	Itm2c	0.8629779		
Gm2a	0.9081852	ORF19	0.8956117	Baalc	0.880348	Bmp3	0.8629617		
Plac9_predicted	0.9077896	Glx1	0.898494	Atf3	0.8796184	Col6a1_predicted	0.8623223		

**Table S2.** Validation of gene expression by real-time PCR.

**Trends of Gene Expression During  
Rat Kidney Development by Real-Time PCR**

<b>Gene Name</b>	<b>e13.5 to e15.5</b>	<b>e15.5 to e17.5</b>
<i>FKBP8</i>	Down	Up
<i>Bcl2</i>	Down	Up
<i>RTKN</i>	Down	Up
<i>Gopc/PIST</i>	Down	Up
<i>Lin7b</i>	Even	Up

Gene expression was determined by Real-Time PCR in metanephroi at e13.5, e15.5 and e17.5. All listed genes were expressed.