1 The Model

The marginal distributions are of the form:

$$y_{g,s} \sim NBin(\mu_{g,s}, \sigma_{g,s}),$$

with

$$log(\mu_{g,s}) = log(nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g} \times sd_s) + gd_g + se_{s,g}$$

and

$$log(\sigma_{g,s}) = \alpha_g + \beta_g \times log(nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g} \times sd_s).$$

The expression y of gene g in spot s, $y_{g,s}$, depends on the level of the mean $\mu_{g,s}$ on nc_s, the number of cells in spot s, $\boldsymbol{w_s^T}$, the cell type proportions of spot s, $\boldsymbol{m_g}$, the mean contributions of the cell types to the expression of gene g, sd_s, a spot detection efficiency for the experiment, gd_g, a (log) gene detection efficiency for the experiment, and $se_{s,g}$ the (log) spatial effect of spot s on the expression of gene g. The term $nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g}$ can be thought of as a theoretical mean, without taking into account technical errors in the experiment and a spatial effect on gene expressions beyond that which is due to spatial distribution of the cell types. The technical errors eluded to above are encompassed in sd_s and gd_g. The marginal distributions are combined to spot level distributions with a Gaussian copula. For more details see the actual master thesis.

2 Some distributions of estimated parameters for the mouse kidney dataset

The estimated number of cells per spot can be inspected in figure 1. Towards the middle of the sample the estimated number of cells are higher. The spot detection efficiencies (figure 2) have no spatial structure which is in accordance with the assumptions of the model.

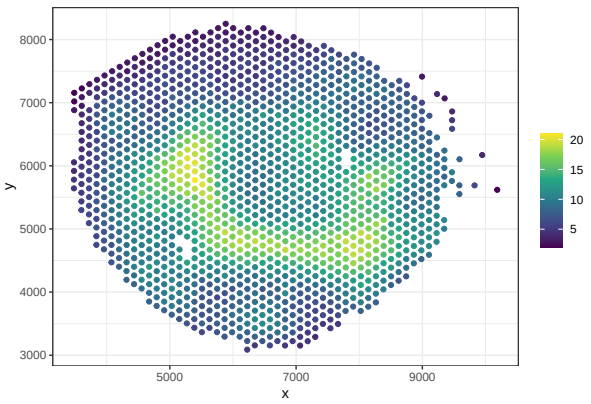

Figure 1: Estimated number of cells per spot for the mouse kidney ST data

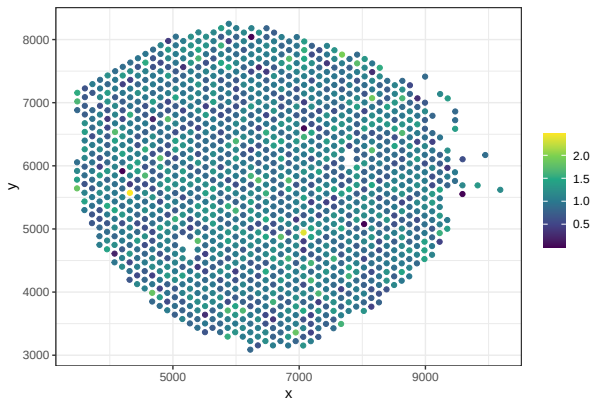

Figure 2: Estimated spot detection efficiencies for the mouse kidney ST data

For the two genes Kap and Gpx3 the domain effects on the mean of their expressions are shwon in figure 3 for gene Kap and figure 4 for gene Gpx3. The domain effects are set to give a geometric mean of 1.

Fitted means and standard deviations per spot for gene Tmsb4x can be found in figure 5 and figure 6.

Figure 7 shows estimated gene detection efficiencies for the mouse kidney dataset. These are mostly near zero, because the spot detection were set to be near 1 and the generally low efficiency of the VISIUM platform on which the spatial transcriptomic data was gathered compared to single cell RNA sequencing technologies.

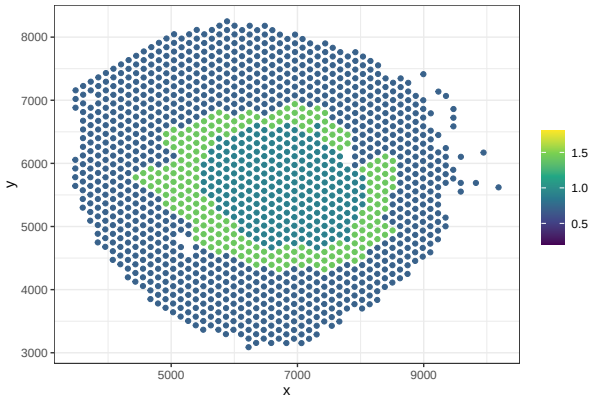

Figure 3: Domain effects of gene Kap for the DLPFC ST data

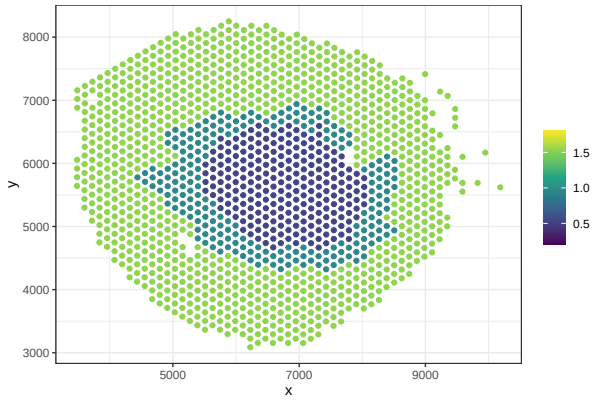

Figure 4: Domain effects of gene Gpx3 for the DLPFC ST data

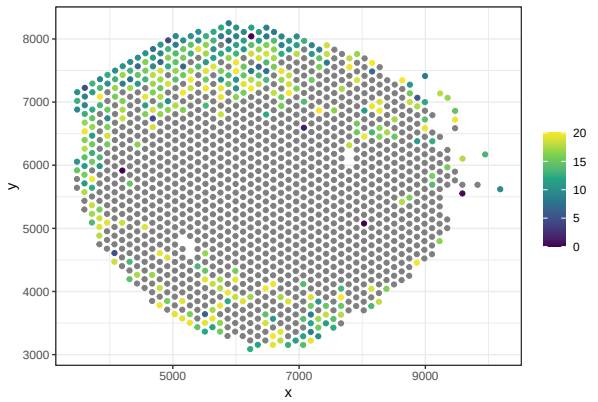

Figure 5: Fitted mean of gene Tmsb4x for the DLPFC ST data

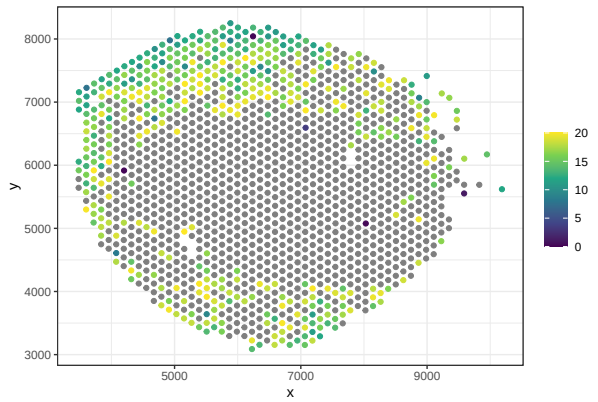

Figure 6: Fitted standard deviation of gene Kap for the DLPFC ST data

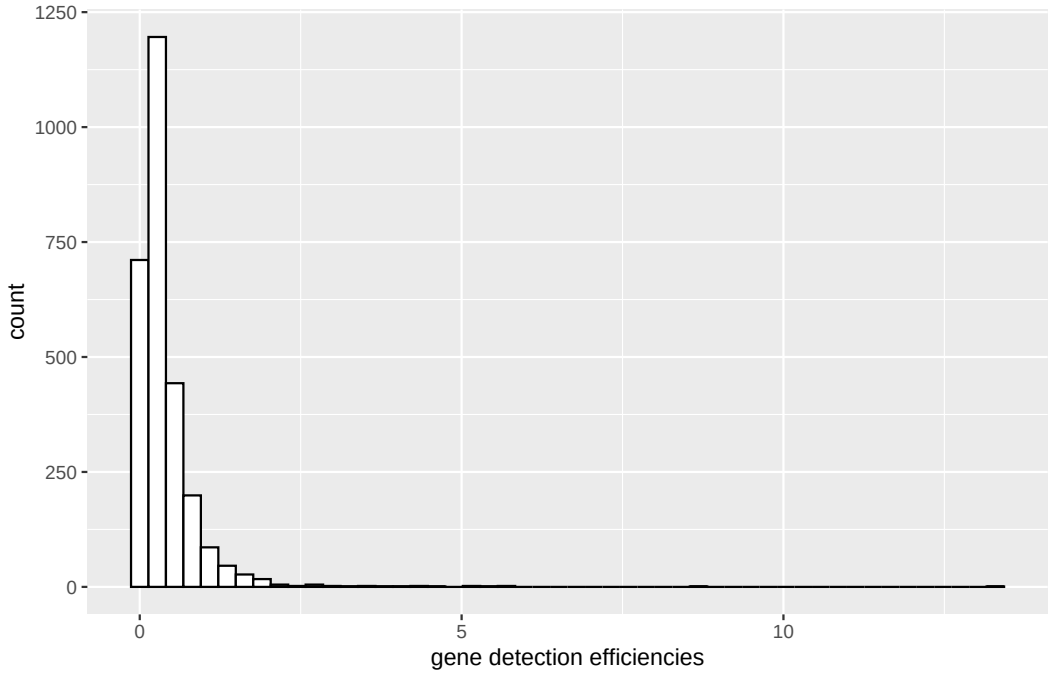

Figure 7: Histogram of the gene detection efficiecnies for the mouse kidney data. Most of the genes have a low estimated detection efficiency.

3 Some plots to evaluate the simulate data

Gene-wise means and standard deviations are well preserved in the simulated data (figure 8 and figure 9). The spot-wise means and standard deviations are less well preserved, because the marginal expressions were fit for genes, not for spots (figure 9 and figure 11).

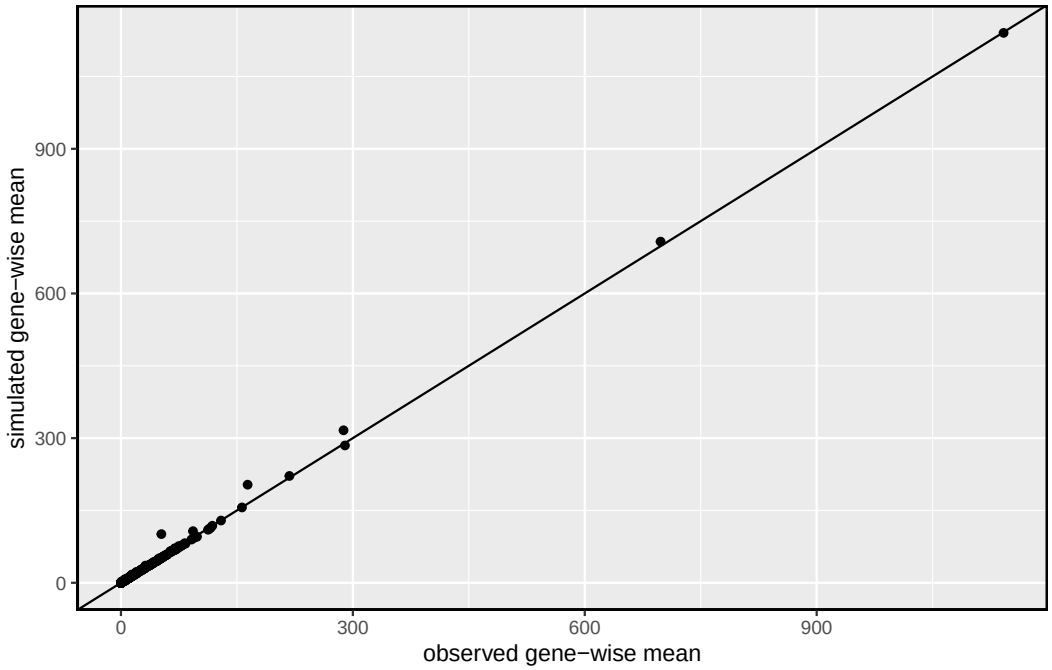

Figure 8: Gene-wise means of observed vs simulated data. The means are preserved in the simulated data.

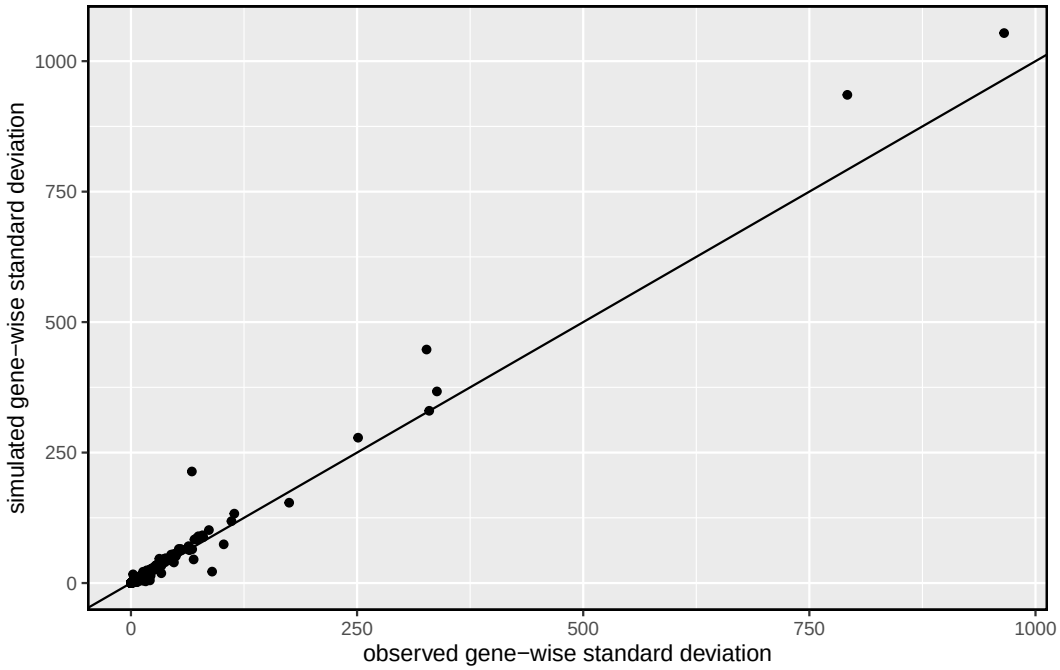

Figure 9: Gene-wise standard deviations of observed vs simulated data. For most genes, the standard devations are similar, but there are outliers for which the standard deviation in the observed data is very low, whereas it is high in the simulated data.

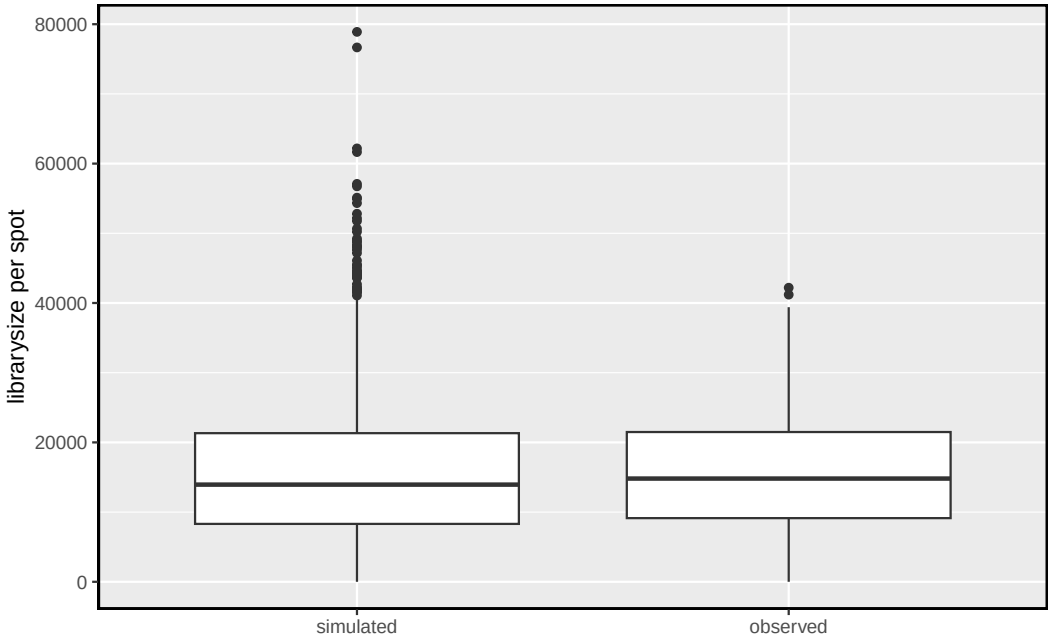

Figure 10: Observed vs simulated librarysizes per spot. The distribution of librarysizes per spots stays roughly the same in the simulated data.

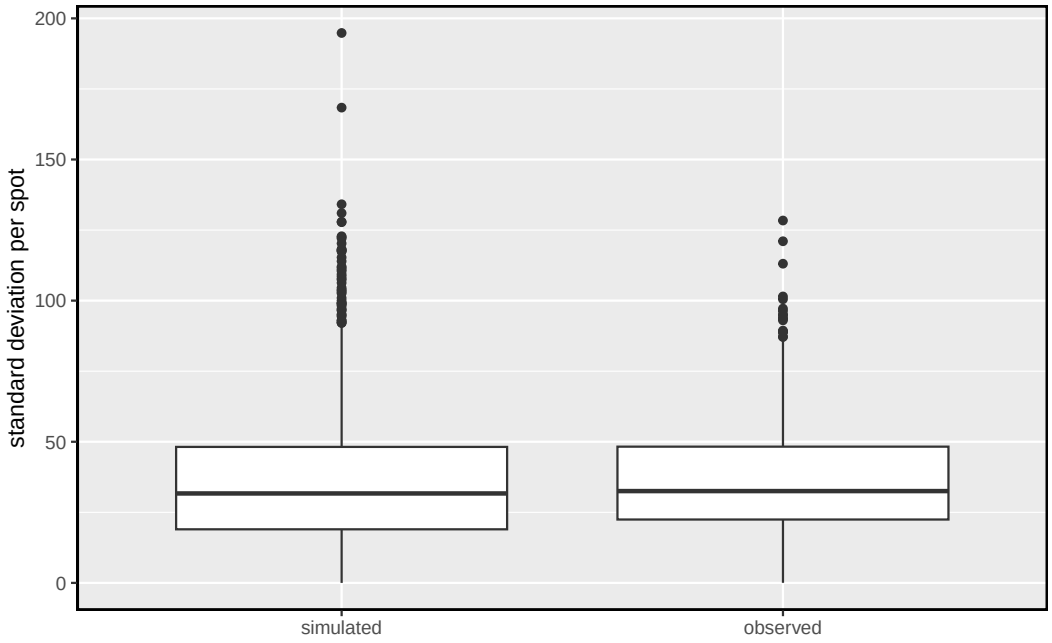

Figure 11: Observed vs simulated standard devations of expressions per spot. The distribution of standard deviations per spots stays roughly the same in the simulated data.