**Supplemental material**

**Table S1** Gene ontology terms enriched in the genes differentially expressed between the two main clusters in the MDG dataset (cluster 1 and 2 in Figure 1) and the AHUS1 dataset (small and large cluster in Figure 2).

<table>
<thead>
<tr>
<th>Cluster 1 (MDG) Term</th>
<th>Count</th>
<th>FDR</th>
<th>Smaller cluster (AHUS1) Term</th>
<th>Count</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>up extracellular region part</td>
<td>115</td>
<td>1.6E-10</td>
<td>mitochondrion</td>
<td>179</td>
<td>2.5E-16</td>
</tr>
<tr>
<td>up vasculature development</td>
<td>44</td>
<td>6.4E-08</td>
<td>generation of precursor metabolites and energy</td>
<td>62</td>
<td>1.4E-07</td>
</tr>
<tr>
<td>up glucose metabolic process</td>
<td>32</td>
<td>5.2E-07</td>
<td>glucose metabolic process</td>
<td>38</td>
<td>1.8E-06</td>
</tr>
<tr>
<td>up extracellular matrix part</td>
<td>28</td>
<td>7.9E-07</td>
<td>response to hormone stimulus</td>
<td>66</td>
<td>2.2E-06</td>
</tr>
<tr>
<td>up response to hormone stimulus</td>
<td>52</td>
<td>2.4E-06</td>
<td>carboxylic acid biosynthetic process</td>
<td>37</td>
<td>9.7E-06</td>
</tr>
<tr>
<td>up regulation of lipid metabolic process</td>
<td>25</td>
<td>1.7E-05</td>
<td>triglyceride metabolic process</td>
<td>18</td>
<td>2.5E-05</td>
</tr>
<tr>
<td>up triglyceride metabolic process</td>
<td>15</td>
<td>7.9E-05</td>
<td>regulation of lipid metabolic process</td>
<td>28</td>
<td>3.5E-04</td>
</tr>
<tr>
<td>up cell fraction (membrane)</td>
<td>106</td>
<td>1.8E-04</td>
<td>regulation of lipid catabolic process</td>
<td>12</td>
<td>0.004</td>
</tr>
<tr>
<td>up response to wounding</td>
<td>58</td>
<td>0.002</td>
<td>cell fraction (membrane)</td>
<td>134</td>
<td>0.006</td>
</tr>
<tr>
<td>up plasma membrane part</td>
<td>178</td>
<td>0.004</td>
<td>Propanoate metabolism</td>
<td>14</td>
<td>0.007</td>
</tr>
<tr>
<td>up regulation of cell migration</td>
<td>26</td>
<td>0.013</td>
<td>vasculature development</td>
<td>43</td>
<td>0.009</td>
</tr>
<tr>
<td>up regulation of response to external stimulus</td>
<td>25</td>
<td>0.014</td>
<td>sulfur metabolic process</td>
<td>25</td>
<td>0.024</td>
</tr>
<tr>
<td>up positive regulation of fatty acid metabolic process</td>
<td>9</td>
<td>0.017</td>
<td>vitamin B6 binding</td>
<td>16</td>
<td>0.027</td>
</tr>
<tr>
<td>up cell motion</td>
<td>51</td>
<td>0.020</td>
<td>regulation of lipid storage</td>
<td>10</td>
<td>0.032</td>
</tr>
<tr>
<td>up blood circulation</td>
<td>27</td>
<td>0.025</td>
<td>positive regulation of fatty acid metabolic process</td>
<td>10</td>
<td>0.032</td>
</tr>
<tr>
<td>up response to extracellular stimulus</td>
<td>30</td>
<td>0.026</td>
<td>cellular respiration</td>
<td>22</td>
<td>0.046</td>
</tr>
<tr>
<td>up extracellular matrix organization</td>
<td>19</td>
<td>0.029</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>up sulfur metabolic process</td>
<td>20</td>
<td>0.034</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>up response to steroid hormone stimulus</td>
<td>27</td>
<td>0.044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>down cell junction</td>
<td>41</td>
<td>0.007</td>
<td>cell junction</td>
<td>52</td>
<td>5.8E-4</td>
</tr>
<tr>
<td>down cell-cell junction</td>
<td>22</td>
<td>0.007</td>
<td>cell-cell junction</td>
<td>26</td>
<td>0.004</td>
</tr>
<tr>
<td>down plasma membrane part</td>
<td>117</td>
<td>0.012</td>
<td>extracellular matrix</td>
<td>36</td>
<td>0.021</td>
</tr>
</tbody>
</table>
Table S2 Gene set enrichment analysis (GSEA) of cluster 1 versus cluster 2 using selected gene lists from the literature. Significant FDR-values are marked in bold.

<table>
<thead>
<tr>
<th>Gene list</th>
<th>FDR</th>
<th>Cluster 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>STROMA (FINAK)</td>
<td>0.00</td>
<td>Up</td>
</tr>
<tr>
<td>STROMA (CASEY)</td>
<td>0.02</td>
<td>Up</td>
</tr>
<tr>
<td>CD44+ (SHIPITSIN)</td>
<td>0.01</td>
<td>Up</td>
</tr>
<tr>
<td>BIPOLAR (VS LUM) (RAOUF)</td>
<td>0.10</td>
<td>Up</td>
</tr>
<tr>
<td>BIPOTENT (VS MYO) (RAOUF)</td>
<td>0.08</td>
<td>Up</td>
</tr>
<tr>
<td>MYOEPITHELIAL (JONES)</td>
<td>0.13</td>
<td>Up</td>
</tr>
<tr>
<td>IGS/CD44+ (LIU)</td>
<td>0.20</td>
<td>Up</td>
</tr>
<tr>
<td>EPITHELIAL (FINAK)</td>
<td>0.00</td>
<td>Down</td>
</tr>
<tr>
<td>CD24+ (VS CD44) (SHIPITSIN)</td>
<td>0.03</td>
<td>Down</td>
</tr>
<tr>
<td>IGS/CD10 (LIU)</td>
<td>0.07</td>
<td>Down</td>
</tr>
<tr>
<td>RISK (CHEN)</td>
<td>0.24</td>
<td>Down</td>
</tr>
<tr>
<td>LUMINAL (JONES)</td>
<td>0.48</td>
<td>Down</td>
</tr>
<tr>
<td>EPITHELIAL (JECHLINGER)</td>
<td>0.70</td>
<td>Down</td>
</tr>
<tr>
<td>MESENCHYMAL (JECHLINGER)</td>
<td>0.66</td>
<td>Down</td>
</tr>
</tbody>
</table>

Table S3 Comparison of cluster 1 and cell types/subtypes from published gene lists. The number of genes up- and down-regulated is given for genes characterizing each cell type and each cluster. Chi-squared test is used to illustrate the extent to which genes describing different cell types are equally regulated in the two clusters. The right column shows samples correctly identified by hierarchical clustering of the normal breast samples based on the gene list from the corresponding publication (see Supplemental file 1, Figure S2)

<table>
<thead>
<tr>
<th>Publication</th>
<th>Cell type</th>
<th>Method</th>
<th>up in cluster 1</th>
<th>down in cluster 1</th>
<th>cluster 1 resembles</th>
<th>Samples correctly identified by clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shipitsin,</td>
<td>Epithelial</td>
<td>CD24+</td>
<td>132</td>
<td>152</td>
<td>CD44+</td>
<td>12/12</td>
</tr>
<tr>
<td>2007</td>
<td>Stem cell-like</td>
<td>CD44+</td>
<td>394</td>
<td>89</td>
<td></td>
<td>66/67</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>0</td>
<td>106</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>526</td>
<td>347</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>χ²</td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td>2.2E-16</td>
</tr>
<tr>
<td>Jechlinger,</td>
<td>Mesenchymal</td>
<td>Before and after TGFbeta-induced EMT</td>
<td>61</td>
<td>17</td>
<td>Mesenchymal</td>
<td>11/12</td>
</tr>
<tr>
<td>2003</td>
<td>Epithelial</td>
<td></td>
<td>38</td>
<td>32</td>
<td></td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>0</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>99</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>χ²</td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td>6.9E-14</td>
</tr>
<tr>
<td>Publication</td>
<td>Cell type</td>
<td>Method</td>
<td>up in cluster 1</td>
<td>down in cluster 1</td>
<td>cluster 1 resembles</td>
<td>Samples correctly identified by clustering</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>-----------------------------</td>
<td>-----------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Raouf, 2008</td>
<td>Luminal</td>
<td>CD49flow/CD10low</td>
<td>521</td>
<td>442</td>
<td>12/12</td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD49fhigh/CD10high</td>
<td>259</td>
<td>162</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unchanged</td>
<td>343</td>
<td>338</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum</td>
<td>1123</td>
<td>942</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>χ² p-value</td>
<td></td>
<td></td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>Raouf, 2008</td>
<td>Myoepithelial</td>
<td>CD49flow/CD10high</td>
<td>139</td>
<td>88</td>
<td>12/12</td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD49fhigh/CD10high</td>
<td>232</td>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unchanged</td>
<td>156</td>
<td>171</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum</td>
<td>527</td>
<td>387</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>χ² p-value</td>
<td></td>
<td></td>
<td>2.40E-05</td>
<td></td>
</tr>
<tr>
<td>Liu, 2007</td>
<td>Stem cell-like</td>
<td>CD44+ CD10+</td>
<td>50</td>
<td>17</td>
<td>11/12</td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td>Epithelial</td>
<td>CD49fhigh/CD10high</td>
<td>34</td>
<td>58</td>
<td>Stem-like</td>
<td>Invasive cells</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td>CD49flow/CD10low</td>
<td>36</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum</td>
<td>120</td>
<td>102</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>χ² p-value</td>
<td></td>
<td></td>
<td>1.30E-05</td>
<td></td>
</tr>
<tr>
<td>Jones, 2004</td>
<td>Myoepithelial</td>
<td>MUC1+ CD10+</td>
<td>65</td>
<td>91</td>
<td></td>
<td>12/12</td>
</tr>
<tr>
<td></td>
<td>Luminal</td>
<td>CD49fhigh/CD10high</td>
<td>21</td>
<td>47</td>
<td>Low risk</td>
<td>66/67</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td>CD49flow/CD10low</td>
<td>30</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum</td>
<td>116</td>
<td>166</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>χ² p-value</td>
<td></td>
<td></td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Chen, 2009</td>
<td>IDC-like normal</td>
<td>Based on gene expression</td>
<td>11</td>
<td>81</td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td></td>
<td>Other normal c</td>
<td>CD49fhigh/CD10high</td>
<td>17</td>
<td>4</td>
<td>11/12</td>
<td>62/67</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td>CD49flow/CD10low</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum</td>
<td>28</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>χ² p-value</td>
<td></td>
<td></td>
<td>1.3E-09</td>
<td></td>
</tr>
<tr>
<td>Finak, 2006</td>
<td>Stroma</td>
<td>Micro-dissection</td>
<td>457</td>
<td>91</td>
<td></td>
<td>12/12</td>
</tr>
<tr>
<td></td>
<td>Epithelial</td>
<td>CD49fhigh/CD10high</td>
<td>51</td>
<td>317</td>
<td>Stroma</td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td>CD49flow/CD10low</td>
<td>59</td>
<td>92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum</td>
<td>567</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>χ² p-value</td>
<td></td>
<td></td>
<td>2.2E-16</td>
<td></td>
</tr>
<tr>
<td>Publication</td>
<td>Cell type</td>
<td>Method</td>
<td>up in cluster 1</td>
<td>down in cluster 1</td>
<td>cluster 1 resembles</td>
<td>Samples correctly identified by clustering</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Casey, 2008</td>
<td>Fibroblasts</td>
<td>Micro-dissection</td>
<td>330</td>
<td>140</td>
<td>Fibroblasts</td>
<td>11/12</td>
</tr>
<tr>
<td></td>
<td>Epithelial</td>
<td></td>
<td>119</td>
<td>134</td>
<td></td>
<td>64/67</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>224</td>
<td>228</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>673</td>
<td>502</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>p-value</td>
<td>$1.9E-12$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Villadsen, 2007</td>
<td>Stem-like K19+/K14+</td>
<td></td>
<td>71</td>
<td>104</td>
<td>Lineage restricted</td>
<td>7/12</td>
</tr>
<tr>
<td></td>
<td>Lineage restricted</td>
<td>K19+/K14-</td>
<td>46</td>
<td>30</td>
<td>progenitor K19+/K14-</td>
<td>29/67</td>
</tr>
<tr>
<td></td>
<td>progenitors K19-/K14-</td>
<td></td>
<td>48</td>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>169</td>
<td>220</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>p-value</td>
<td>$0.008$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asztalos, 2010</td>
<td>Nullipara Micro-dissection</td>
<td></td>
<td>8</td>
<td>2</td>
<td>Post-pregnant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postpregnant</td>
<td></td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>13</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fisher exact p-value</td>
<td></td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sørlie, 2001</td>
<td>Basal-like HER2-enriched</td>
<td>Based on gene</td>
<td>62</td>
<td>66</td>
<td>-</td>
<td>12/12</td>
</tr>
<tr>
<td></td>
<td>HER2-enriched</td>
<td>gene expression</td>
<td>72</td>
<td>69</td>
<td></td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td>Luminal A</td>
<td></td>
<td>73</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Luminal B</td>
<td></td>
<td>70</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Normal-like</td>
<td></td>
<td>80</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>357</td>
<td>366</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>p-value</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hersch-kowitz, 2007</td>
<td>Basal-like Claudin-low</td>
<td>Based on gene</td>
<td>56</td>
<td>79</td>
<td>claudin-low</td>
<td>12/12</td>
</tr>
<tr>
<td></td>
<td>Claudin-low HER2-enriched</td>
<td>expression</td>
<td>121</td>
<td>33</td>
<td></td>
<td>67/67</td>
</tr>
<tr>
<td></td>
<td>HER2-enriched</td>
<td>gene expression</td>
<td>76</td>
<td>98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Luminal</td>
<td></td>
<td>70</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Normal-like</td>
<td></td>
<td>76</td>
<td>67</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
<td></td>
<td>163</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td>562</td>
<td>488</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>p-value</td>
<td>$1.52E-12$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chen et al, 2008
IDC-like normal cells vs other normal cells
$\chi^2$ p-value: $1.3\times 10^{-9}$
Bonferroni: $1.7\times 10^{-8}$

Shiptisin et al, 2007
Luminal (CD24+) vs stem-cell like (CD44+) cells
$\chi^2$ p-value: $2.2\times 10^{-16}$
Bonferroni: $2.8\times 10^{-15}$

Jechlinger et al, 2003
Epithelial cells before and after TGFbeta-induced EMT
$\chi^2$ p-value: $6.9\times 10^{-14}$
Bonferroni: $8.9\times 10^{-13}$

Raouf et al, 2008
Luminal (MUC1+) vs bipotent (CD49f+, CD10+) cells
$\chi^2$ p-value: $0.001$
Bonferroni: $0.013$

Raouf et al, 2008
Myoepithelial (CD10+) vs bipotent (CD49+, CD10) cells
$\chi^2$ p-value: $2.4\times 10^{-5}$
Bonferroni: $3.1\times 10^{-4}$

Liu et al, 2007
Stem-cell like (CD44+) vs luminal (CD10+) cells
$\chi^2$ p-value: $1.3\times 10^{-5}$
Bonferroni: $1.7\times 10^{-4}$

Jones et al, 2004
myoepithelial (MUC1+) vs luminal (CD10+) cells
$\chi^2$ p-value: $0.06$
Bonferroni: $0.78$

Chen et al, 2008
IDC-like normal cells vs other normal cells
$\chi^2$ p-value: $1.3\times 10^{-9}$
Bonferroni: $1.7\times 10^{-8}$
Hierarchical clustering of gene expression from 79 samples from breasts of healthy women. The samples are clustered based on gene lists from the literature, describing different cell types. The two last panels are clustered based on gene lists used to identify breast cancer subtypes. Cluster 1-samples are marked light blue and cluster 2-samples dark blue. The dendrogram colors represent the two main clusters in the clustering performed based on the gene list in question.
Figure S2 Biopsies from healthy women (MDG) clustered with two unpublished datasets from Akershus University Hospital (AHUS) A) AHUS1 with breast biopsies from mammoplasty reductions (yellow) and tumor adjacent (red) tissue and B) AHUS2 with breast biopsies containing different known proportions of fat tissue and C) a dataset previously published by Nicoalu et al [1] with breast biopsies from mammoplasty reductions (yellow) and tumor adjacent (red) breast tissue. In all cases, the two datasets are merged by use of Distance Weighted Discrimination (DWD). This resulted in datasets with A) 8520 genes, B) 10078 and C) 3555 genes. Hierarchical clustering with Euclidean distance and Ward linkage was performed as described. Ward linkage was performed as described in Materials and methods.

Reference list

Figure S3

Unsupervised hierarchical clustering of 79 samples from healthy individuals as shown in Figure 1. Phenotypes with tests for significant difference in values between cluster 1 (blue) and cluster 2 (red). Continuous variables are categorized for the illustration, but significance tested as continuous variables. P-values from two-sided t-tests assuming equal variance for continuous variables (*) and chi-squared tests (**) for categorical variables are given. The numbers along the y-axis denotes the number of genes. The claudin-low subtype is found by use of a predictor as described in the Methods section. PAM50 is used to estimate subtypes for the non-claudin-low samples.

Age= Age at time of inclusion. BMI: Body mass index. HT: Use of hormone therapy. MD: Mammographic density.