Supplemental Information to:

Systematic Evaluation of the Prognostic Impact and Intratumour Heterogeneity of Clear Cell Renal Carcinoma Biomarkers

Gulati et al.

Contents

Literature Search

We systematically searched PubMed and Google Scholar for publications describing genetic or transcriptomic prognostic biomarkers for RCC. We restricted our search to combinations of the terms: biomarker, prognosis and renal cell carcinoma. We restricted our search to articles published until December 2013 in English language. Studies exclusively based on non-clear cell histology were excluded. Additional literature cited in identified prognostic marker publications or recent review articles [1-7] was also assessed. Only studies based on the analysis of follow-up data were included and studies which only showed an association with other poor prognosis clinical factors such as tumour stage and grade were excluded. 30 publications describing RCC genetic or gene expression prognostic biomarkers were identified in the literature search. Four biomarkers were excluded from further analysis. One biomarker [8] was based on regression coefficients devised using microarray gene expression data. This could not to be applied RNA-Seq data and was therefore excluded. The other three studies included multi-gene expression signatures, for which fewer than 70% of gene probes mapped to genes annotated in the TCGA RNA-Seq dataset [9-11]. Several publications investigating gene expression levels as potential prognostic biomarkers lacked information about how the identified genes can be applied to clinical samples in order to identify prognostically distinct subgroups. These were also excluded from further analysis.

Materials and Methods

Patient Cohort

Somatic mutation, SCNA, RNA-Seq and clinical data were available for 354 patients. Follow-up data or tumour grade were missing for four patients, leaving 350 patients, which formed our study cohort.

Statistical Methods

For the univariate analyses, patients with the field “Composite Vital Status” = “DECEASED” and “Composite Tumour Status” = “WITH TUMOR” were considered to be dead with clear cell renal cancer related causes, while those with “Composite Vital Status” = “DECEASED” and “Composite Tumour Status” = “TUMOR FREE” were considered to be dead due to other causes. Follow-up time was defined using the “Composite Days to Death” field in case of patient death, and “Composite Days to Last Contact” for patients alive at the end of study period. For the multivariate Cox regression analysis, a backwards stepwise selection process was implemented. The selection step was repeated till all the variables left in the model had $p \leq 0.05$. For all non-significant variables, the hazard ratio, 95% confidence interval (C.I.) and a p-value was generated at the step it was removed (Supplemental Table 2). Although we are not aware of a formal way to determine the number of parameters which can be tested in multivariate analysis based on the death event rate, to the best of our knowledge, we should not have more than ‘n’ number of variables in the final model [12], where n = total number of deaths from disease/10, which for our study would equal 8 variables. Our final multivariate model after stepwise selection has only 2 variables – tumour stage and the ccA/ccB gene expression signature.

Statistical analyses were performed in R (v3.0.1) [13], using the packages 'survival' [14],'gplots', 'cmprsk' and 'survcomp'. Survival graphs were generated with GraphPad Prism (v6.03).

Analysis of prognostic molecular signatures: data processing

Somatic mutations

Mutation data for each of the five tumour suppressor genes *VHL, PBRM1, BAP1, SETD2 and TP53* was obtained from the Cancer Genome Atlas ccRCC publication [15]. A non-synonymous mutation in the 'Variant classification' column assigned a patient to the mutant subgroup for each gene. CSS was assessed for patients with tumours harbouring a non-synonymous mutation in the gene vs. patient with tumours without the mutation. For *VHL*, we also tested association with survival for the subgroups with stage I-III tumours and for those with loss of function mutations (defined as frameshift and nonsense mutations).

Copy number data and SCNA profiles

The aroma R package (CRMA v2, CalMaTe "v1" algorithm & TumorBoost) [16-18] was used to obtain logR and BAF values from SNP array data which was generated on Affymetrix Genome-Wide SNP Array 6.0 platform by the TCGA, using normal samples as references. Sex chromosomes were excluded from the analysis. The ASCAT algorithm was applied to all 450 samples to obtain copy number profiles [19] as described in [20]. The SCNA data was converted to cytoband level data using the cytoband coordinates retrieved from the UCSC Genome Browser database (http://genome.ucsc.edu/) [21]. For each cytoband a weighted average copy number was obtained, and deletions and amplifications were defined as copy numbers

deviating from the ploidy, as estimated by ASCAT, by more than 0.6, similar to the original ASCAT publication [19].

CSS was compared between patients with tumours harbouring a specific SCNA vs. those with tumours without these SCNAs. Chrom22 deletion was identified as a candidate prognostic biomarker. The SNP array data did not include any probes for the Chrom22 p-arm, hence deletion of Chrom22q was used as a substitute for Chrom22 deletions.

Gene expression signatures

Reads per kilobase per million (RPKM) counts which had been normalised to the upper quartile normal counts by TCGA were used for this analysis after log2 transformation. Only genes, for which normalized RPKM counts were above 30 in at least 80% of the samples, were included in our analyses.

Classification of patients into prognostic groups

Log2 transformed expression data was used to divide the cohort into 2 groups at median values for CD31 and EDNRB expression levels and at 33rd percentile value for TSPAN7 expression levels [22].

Expression data for genes in each identified gene expression signtaures [23-27] were submitted for unsupervised consensus NMF clustering [28] analysis to the Broad Institute's GenePattern server [29]. Expression data was available for 26 out of 35 genes (74%) from [23], 220 out of 259 genes (85%) from [24], 36 out of 44 (82%) genes from [25], 103 out of 110 (94%) genes from [26] and 37 out of 48 (77%, Cluster B vs. A/C) and 21 out of 23 (91%, Cluster A vs. C) genes from the two gene panels from [27] respectively. The cluster number range was predefined from two to

10. Each clustering run returned a cophenetic correlation coefficient which measures the stability of cluster assignments, as well as consensus clustering maps. Based on both these data, we identified the optimal number of clusters for each gene expression panel. For each signature, we found the same numbers of clusters to be optimal as had been identified in the original publications.

Classification of patients based on TGFβ pathway expression signature

Using the TGFβ signature [30], we calculated a TGFβ activity score for each sample as described. RNA-Seq counts were available for 145/157 TGFβ regulated genes. In brief, the log2 expression counts of these genes are

1) Multiplied by either 1 or -1, depending on their expected regulation by TGFβ
2) These values are then averaged to give a relative TGFβ score for each sample

Using the median score of all samples as cut off, patients were divided into two cohorts as previously described.

Additional genomic measures: Measures of Aneuploidy

Ploidy

Ploidy estimates are obtained from ASCAT [19], which are calculated as the average total copy number for each sample.

Weighted Genomic Instability Index (wGII)

The wGII [31] score is computed by first calculating for each chromosome the proportion of bases whose copy number deviates from the ploidy value of the sample as given by ASCAT by more than 0.6. The sample wGII score is the sum of the chromosomal scores divided by the number of analysed chromosomes (n=22).

Analysis of multi-region biopsy data: classification of tumour regions into ccA and ccB prognostic groups

Published gene expression data [32, 33] generated with Affymetrix Gene 1.0 arrays was downloaded from the GEO (datasets: GSE31610, GSE53000). Samples were normalized using the oligo R package and the RMA algorithm. Expression data was available for 107 out of 110 genes from the ccA/ccB signature [26]. We used these to classify the 63 tumour regions into either ccA/ccB expression subgroups by applying consensus NMF clustering analysis for a predefined number of clusters from two to 10. The cophenetic coefficient was highest for two clusters. Clustering was also performed using the Clearcode34 panel [34] and the same cluster assignments were obtained for 61 out of 63 (97%) regions.

Supplemental Table 1: Details of patient characteristics from studies included in analyses.

PDF File: Supplemental Table 1.pdf

N/A: Data not available. For [27], details for only the small tissue array (TMA) dataset are given, for which the survival analyses was shown. Clinical data from [24] was used in Bostrom et al. [30] to test the clinical applicability of their prognostic signature.

Supplemental Table 2: Multivariate analysis results - hazard ratios and p-values for all assessed variables ranked according to order of elimination.

All variables which failed validation are highlighted in red and final significant variables are highlighted in green.

Variable	Hazard Ratio (C.I.)	p-value
EDNRB expression **< median** **≥ median**	 1.00 (Ref) 0.98 (0.44 – 2.23)	 0.972
Beulet signature **Cluster A** **Cluster B** **Cluster C**	 1.00 (Ref) 1.52 (0.78 – 2.96) 0.95 (0.40 – 2.30)	 0.211 0.915
12 Amplification	1.00 (0.46 – 1.91)	0.882
***BAP1* non-syn mutation**	1.08 (0.56 – 2.09)	0.819
4p Deletion	1.13 (0.54 – 2.37)	0.737
Lane signature **Indolent** **Aggressive**	 1.00 (Ref) 1.13 (0.54 – 2.38)	 0.748
22q Deletion	1.24 (0. 58 – 2.67)	0.578
8q Amplification	1.27 (0.59 – 2.68)	0.536
TGFβ signature **Low expression score** **High expression score**	 1.00 (Ref) 1.25 (0.72 – 2.18)	 0.415
***TP53* non-syn mutation**	1.67 (0.54 – 5.19)	0.368
Furhmann Grade **G1/G2** **G3** **G4**	 1.00 (Ref) 1.45 (0.77 – 2.70) 1.87 (0.87 – 4.02)	 0.243 0.107
9p Deletion	1.35 (0.82 – 2.23)	0.232
20q focal Amplification	0.69 (0.40 – 1.20)	0.194
Zhao signature **Cluster 1 (good)** **Cluster 2 (poor)**	 1.00 (Ref) 1.51 (0.75 – 3.00)	 0.246
Kosari signature **Non - aggressive** **Aggressive**	 1.00 (Ref) 0.62 (0.32 – 1.16)	 0.137
TSPAN7 expression **< 33 percentile** **≥ 33 percentile**	 1.00 (Ref) 0.76 (0.43 – 1.34)	 0.341
Tumour stage **Stage I** **Stage II** **Stage III** **Stage IV**	 1.00 (Ref) 3.48 (1.20 – 10.06) 4.61 (1.93 – 11.00) 18.01 (7.89 – 41.12)	 0.022 <0.001 <0.001

Chrom 19 deletion	4.18 (1.27 – 13.69)	0.018
ccA subgroup **ccB subgroup**	1.00 (Ref) 2.99 (1.87 – 4.80)	 <0.001

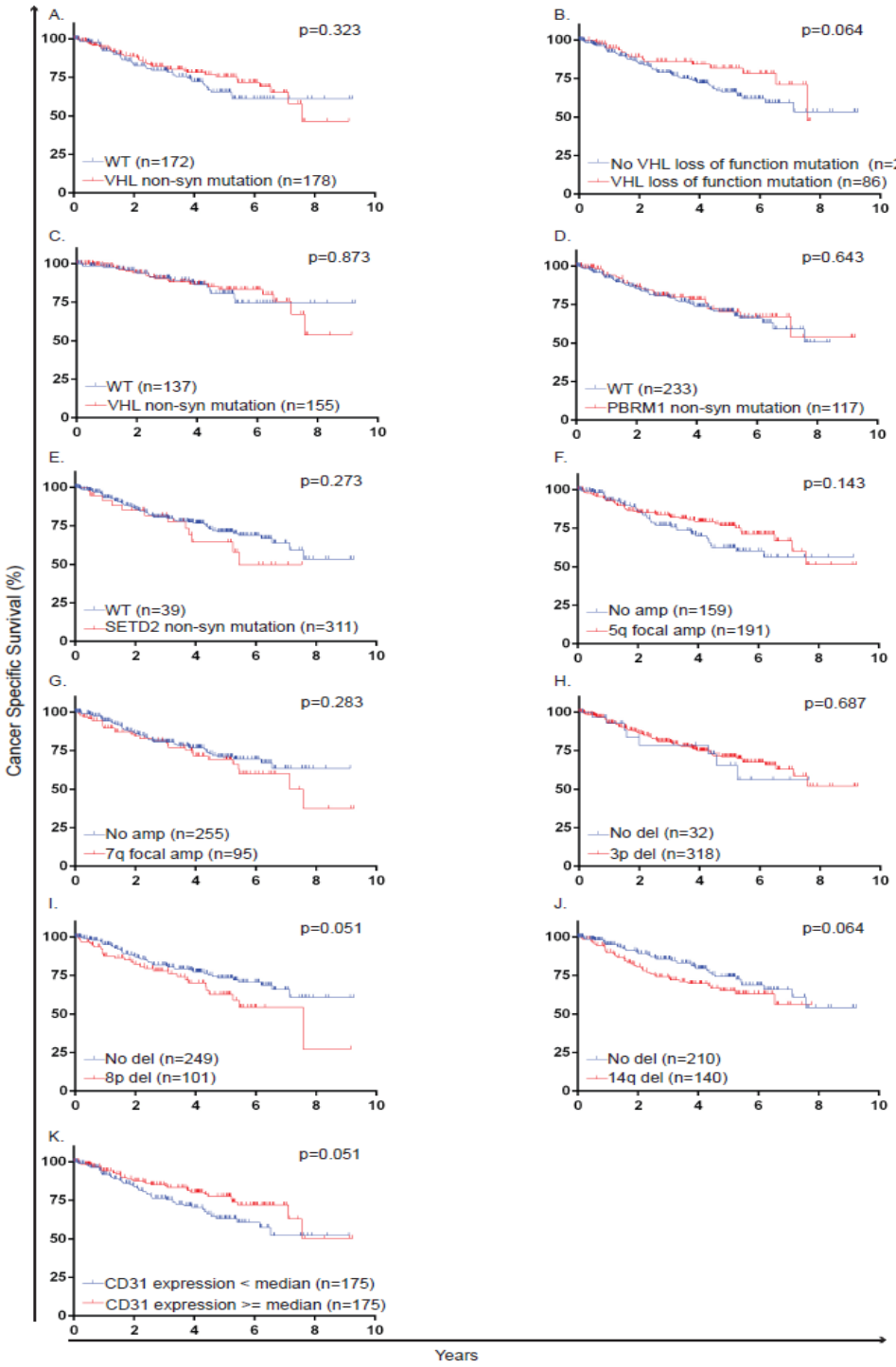
Supplemental Table 3: Multivariate analysis with SSIGN score

Variable	Hazard Ratio (C.I.)	p-value
SSIGN score **0-1** **2-4** **5-6** **7-9** **≥10**	 1.00(Ref) 2.69 (0.64 – 11.29) 8.28 (2.28 – 30.06) 13.23 (3.92 – 44.61) 34.73 (10.29 – 117.20)	 0.175 0.001 <0.001 <0.001
ccA subgroup **ccB subgroup**	1.00 (Ref) 2.24 (1.38 – 3.64)	 0.001

Supplemental Table 4: Multivariate analysis with ClearCode34 signature

Variable	Hazard Ratio (C.I.)	p-value
Tumour stage **Stage I** **Stage II** **Stage III** **Stage IV**	 1.00 (Ref) 3.92 (1.36 – 11.32) 4.86 (2.51 – 13.90) 19.32 (8.44 – 44.21)	 0.012 <0.001 <0.001
ClearCode34 **ccA subgroup** **ccB subgroup**	 1.00 (Ref) 2.23 (1.39 – 3.60)	 <0.001

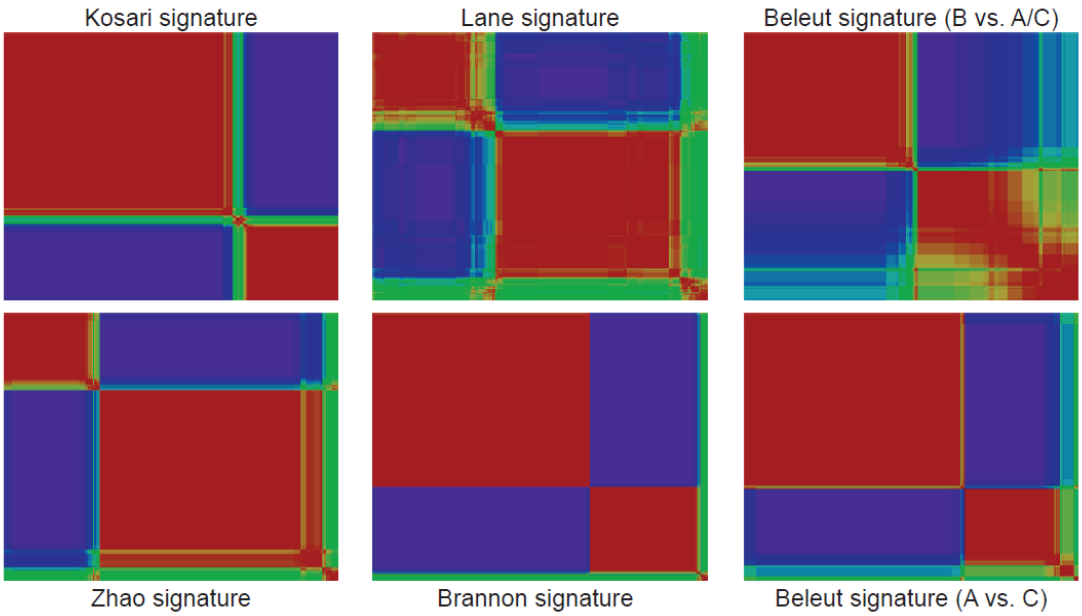
Supplemental Fig. S1: Kaplan-Meier cancer specific survival estimates for genetic alterations which failed univariate validation.



A. *VHL* non-synonymous (non-syn) mutations, all cases, B. *VHL* loss of function OF mutations, all cases, C. *VHL* non-syn mutations, stage I-III cases only, D. *PBRM1* non-syn mutations, E. *SETD2* non-syn mutations, F. Chromosome 5q focal amplification (amp) status, F. Chromosome 7q focal amp status, G. Chromosome 3p deletion (del) status, H. Chromosome 8p del status, J. Chromosome 14q del status, K. CD31 expression levels. WT=wild type.

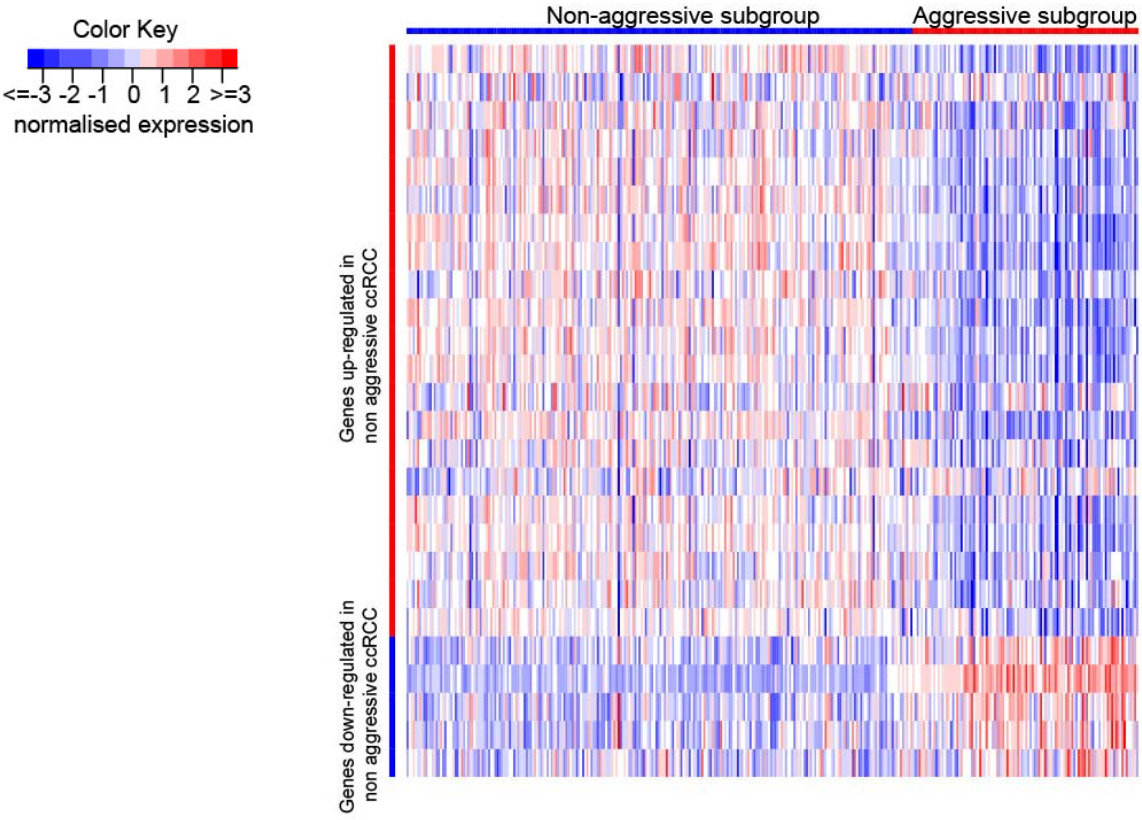
Supplemental Fig. S2: Consensus NMF clustering and expression heatmaps for all gene expression panels included in study.

A. Ordered Consensus NMF clustering maps for k=2 for gene expression classifying panels included in the study

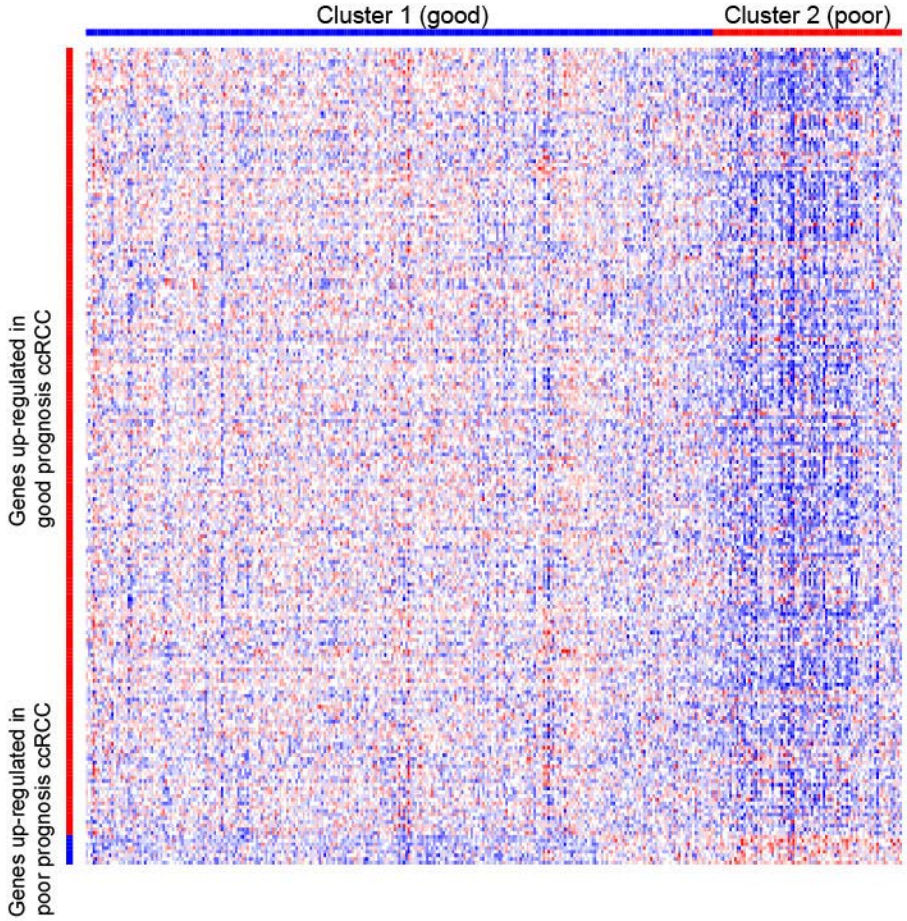


Ordered Consensus maps for k=2. Each heatmap depicts the stability of consensus clustering assignment for two clusters.

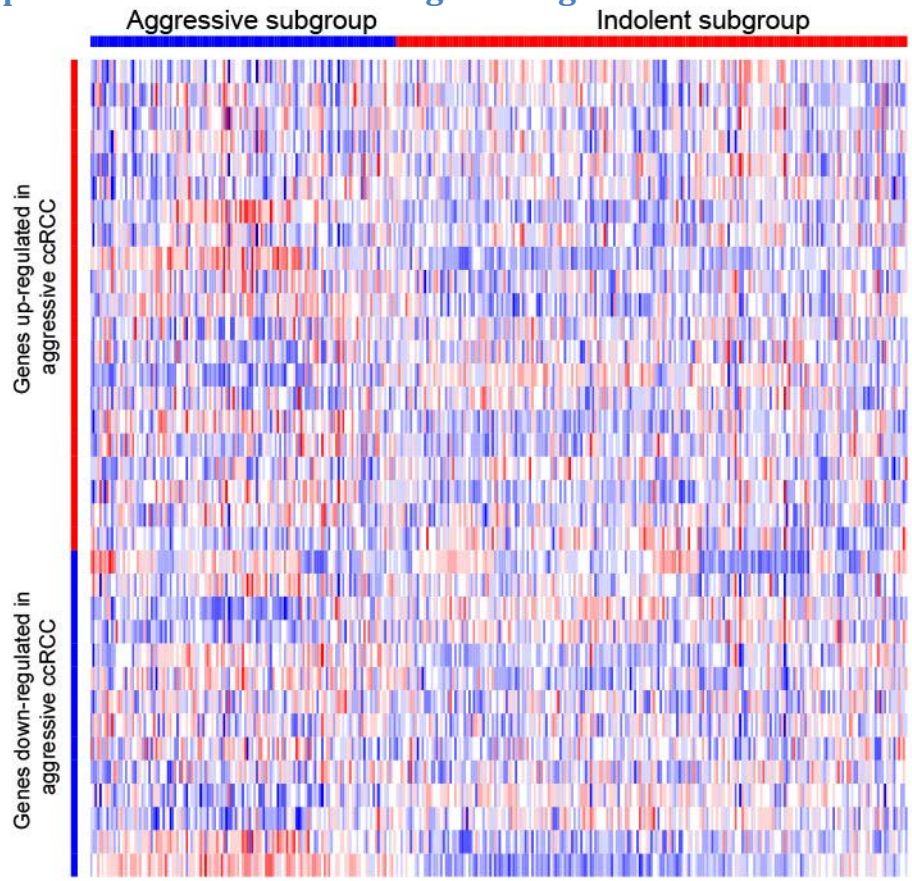
B. Heatmap showing consensus NMF clustering analysis based on gene expression data of 26 Kosari signature genes.



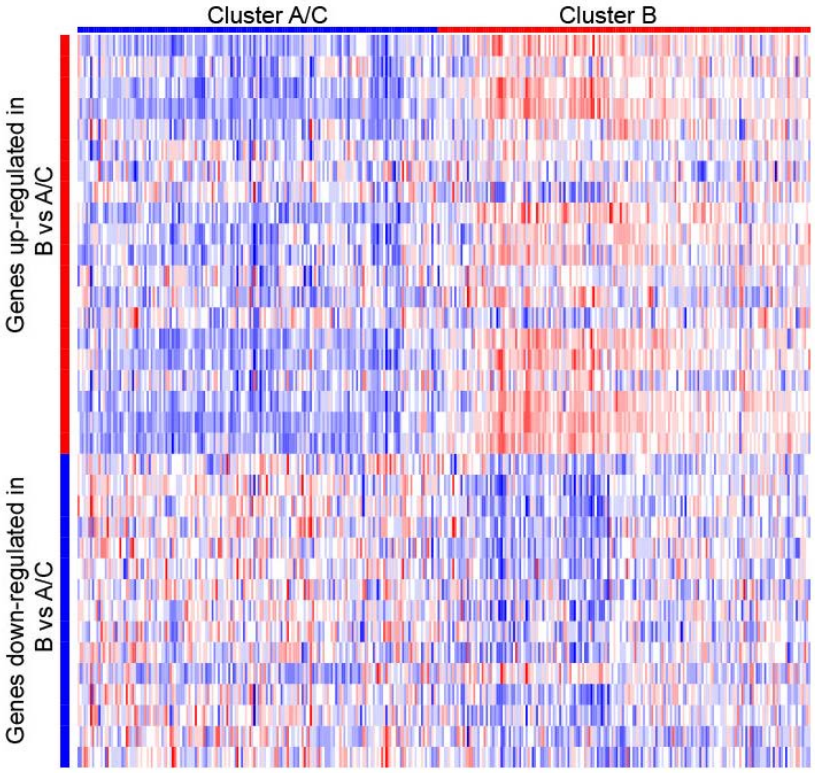
C. **Heatmap showing consensus NMF clustering analysis based on gene expression data of 220 Zhao signature genes.**



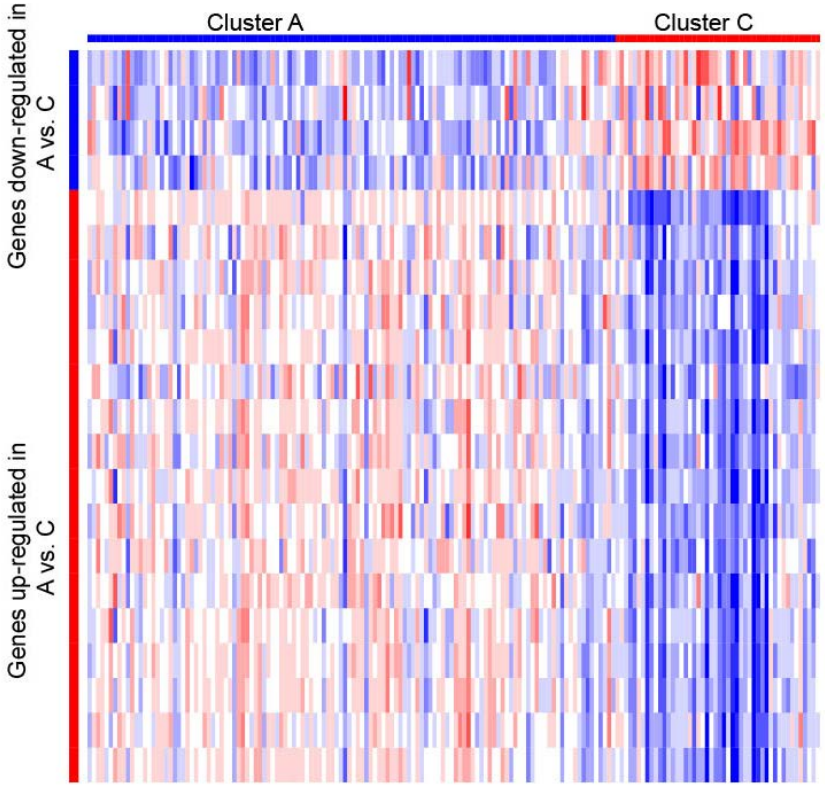
D. Heatmap showing consensus NMF clustering analysis based on gene expression data of 35 Lane signature genes.



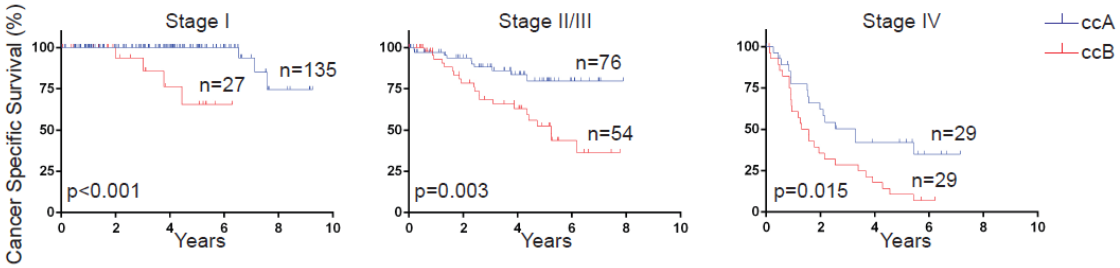
E. Heatmap showing consensus NMF clustering analysis based on gene expression data of 37 Beleut signature (Cluster B vs. A/C) genes.



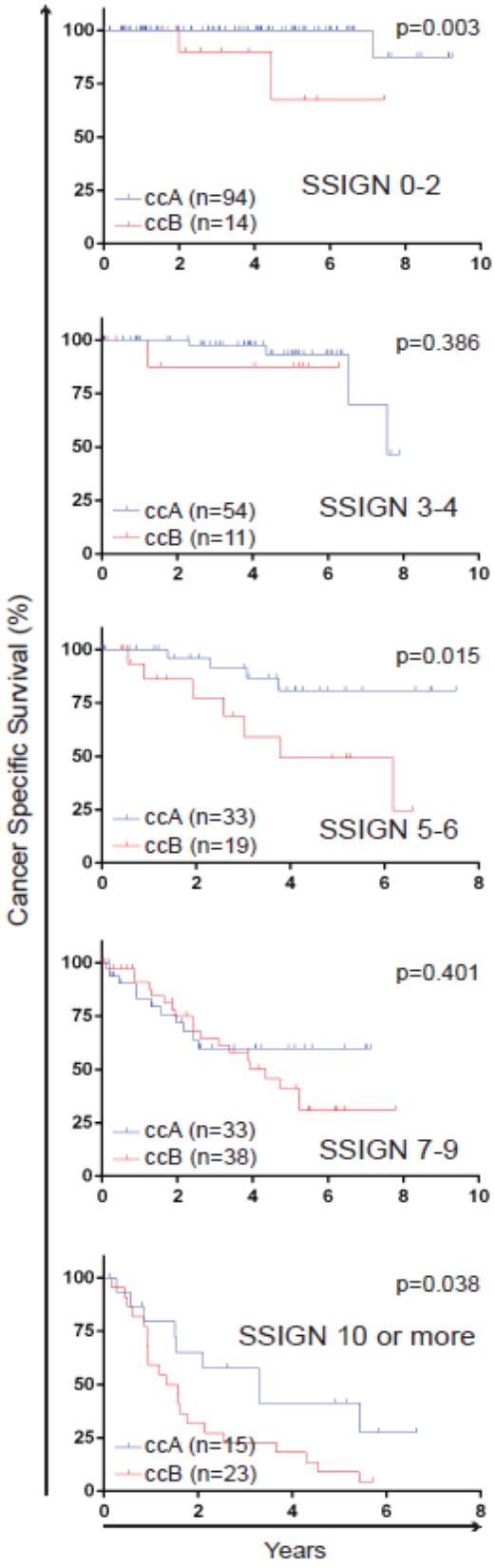
F. Heatmap showing consensus NMF clustering analysis based on gene expression data of 21 Beleut signature (Cluster A vs. C) genes.



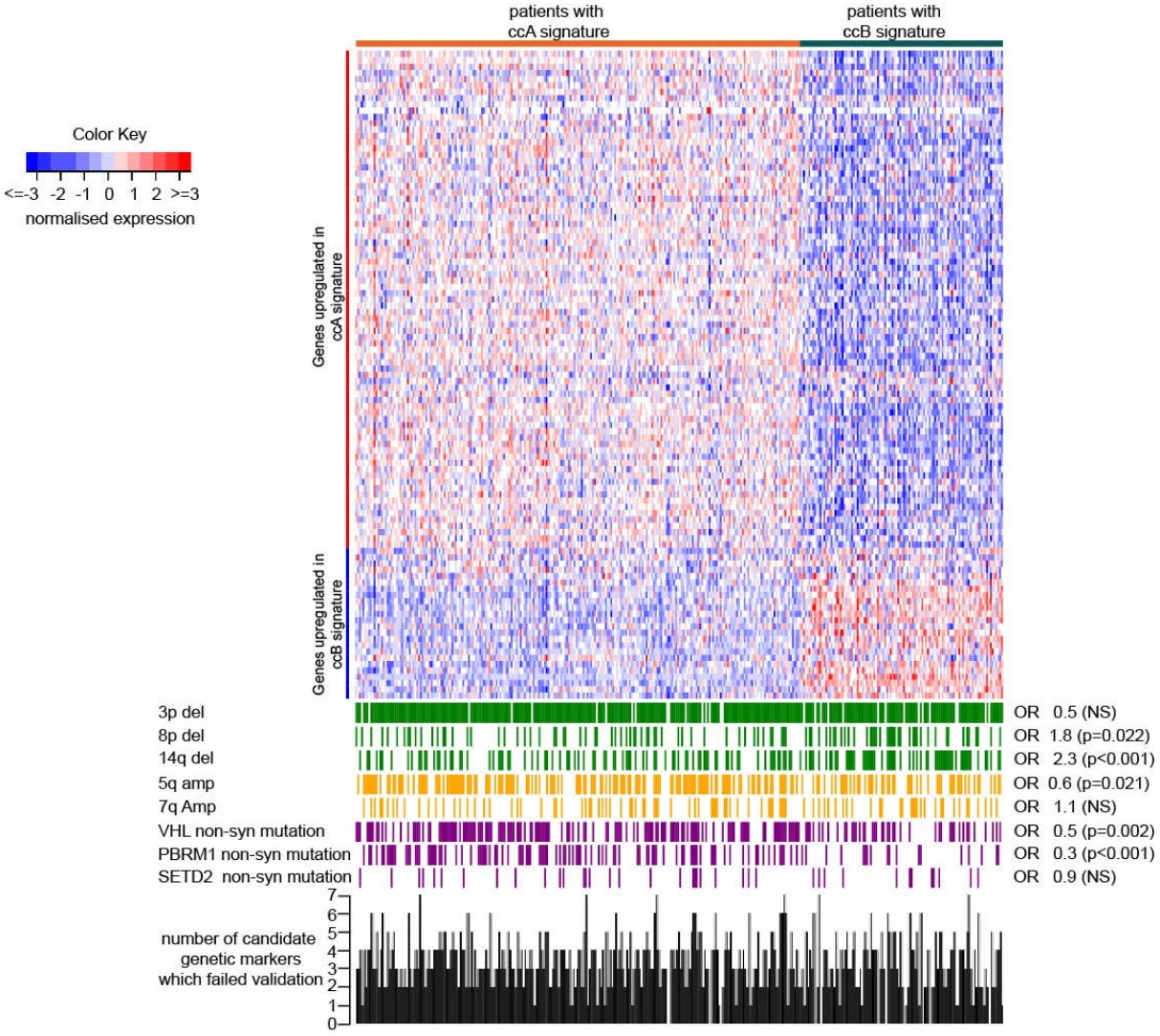
Supplemental Fig. S3: Kaplan-Meier cancer specific survival estimates for ccA/ccB subgroups split by tumour stage.



Supplemental Fig. S4: Kaplan-Meier cancer specific survival estimates for ccA/ccB subgroups split by SSIGN score classes (n=334).

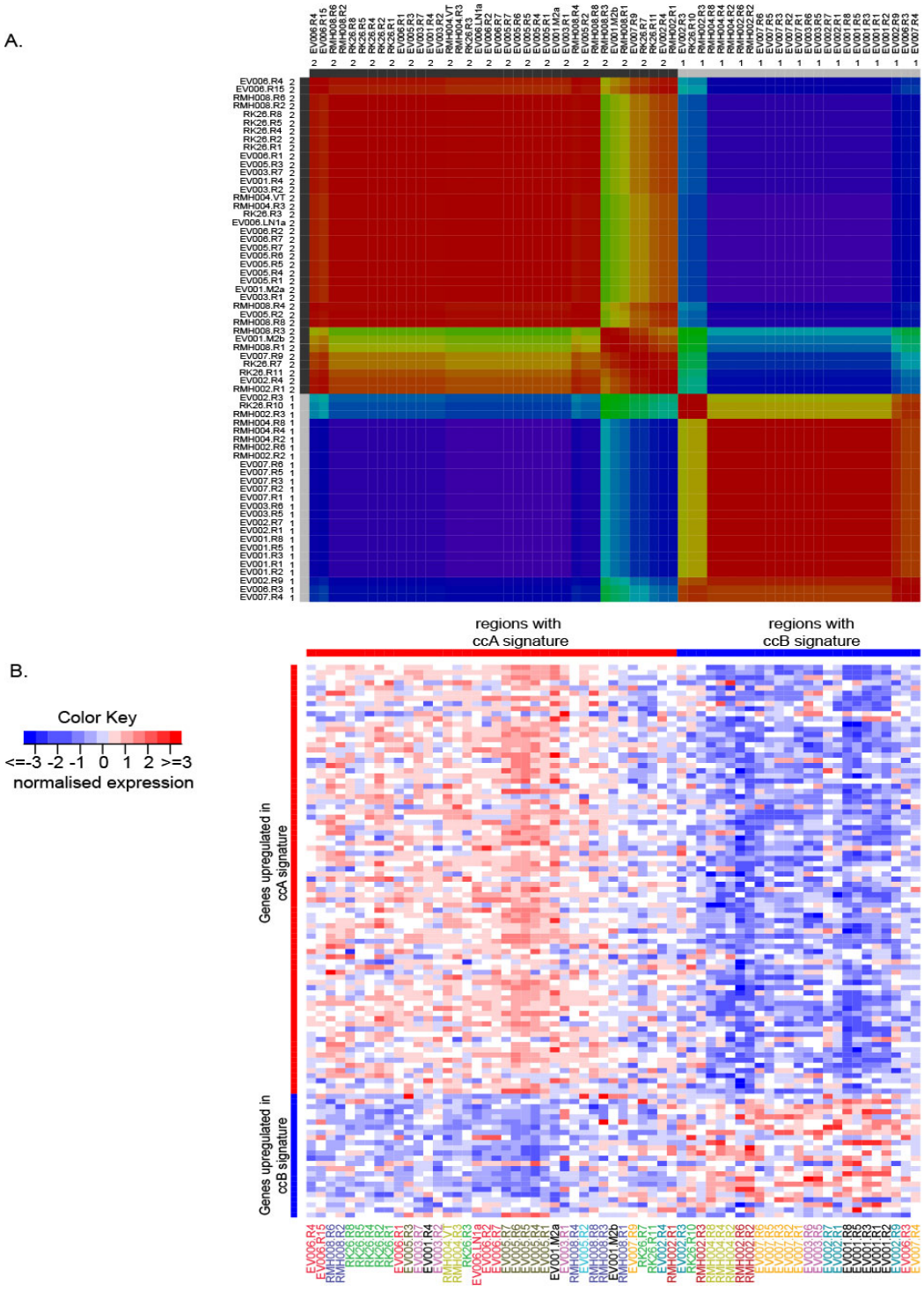


Supplemental Fig. S5: Heatmap showing consensus NMF clustering analysis based on gene expression data of 103 ccA/ccB signature genes and the enrichment of genetic markers which failed univariate validation



Patient assignment to ccA and ccB prognostic subgroups is indicated by coloured bars at the top of the heatmap. Coloured bars below the heatmap depict the presence of genetic aberrations alterations which failed univariate validation. The bar chart at the bottom of the figure represents the number of these genetic aberrations per patient.

Supplemental Fig. S6: Consensus NMF clustering analysis for multi-region biopsy dataset.



A. Consensus NMF clustering matrix for multi-region biopsy dataset for two clusters (obtained from http://genepattern.broadinstitute.org/), B. Heatmap showing consensus NMF clustering analysis for the multi-region biopsy dataset using gene expression data of 107 ccA/ccB signature genes. Tumour regions assigned to the ccA or ccB prognostic subgroups is indicated by coloured bars at the top of the heatmap.

References

[1] Brannon AR, Rathmell WK. Renal cell carcinoma: where will the state-of-the-art lead us? Current oncology reports. 2010;12:193-201.
[2] Jonasch E, Futreal PA, Davis IJ, Bailey ST, Kim WY, Brugarolas J, et al. State of the science: an update on renal cell carcinoma. Molecular cancer research : MCR. 2012;10:859-80.
[3] Tang PA, Vickers MM, Heng DY. Clinical and molecular prognostic factors in renal cell carcinoma: what we know so far. Hematology/oncology clinics of North America. 2011;25:871-91.
[4] Eichelberg C, Junker K, Ljungberg B, Moch H. Diagnostic and prognostic molecular markers for renal cell carcinoma: a critical appraisal of the current state of research and clinical applicability. European urology. 2009;55:851-63.
[5] Junker K, Ficarra V, Kwon ED, Leibovich BC, Thompson RH, Oosterwijk E. Potential role of genetic markers in the management of kidney cancer. European urology. 2013;63:333-40.
[6] Arsanious A, Bjarnason GA, Yousef GM. From bench to bedside: current and future applications of molecular profiling in renal cell carcinoma. Molecular cancer. 2009;8:20.
[7] Oosterwijk E, Rathmell WK, Junker K, Brannon AR, Pouliot F, Finley DS, et al. Basic research in kidney cancer. European urology. 2011;60:622-33.
[8] Yao M, Huang Y, Shioi K, Hattori K, Murakami T, Sano F, et al. A three-gene expression signature model to predict clinical outcome of clear cell renal carcinoma. International journal of cancer Journal international du cancer. 2008;123:1126-32.
[9] Takahashi M, Rhodes DR, Furge KA, Kanayama H, Kagawa S, Haab BB, et al. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. Proceedings of the National Academy of Sciences of the United States of America. 2001;98:9754-9.
[10] Sultmann H, von Heydebreck A, Huber W, Kuner R, Buness A, Vogt M, et al. Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. Clinical cancer research : an official journal of the American Association for Cancer Research. 2005;11:646-55.
[11] Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, et al. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. Proceedings of the National Academy of Sciences of the United States of America. 2003;100:6958-63.
[12] Zwiener I, Blettner M, Hommel G. Survival analysis: part 15 of a series on evaluation of scientific publications. Deutsches Arzteblatt international. 2011;108:163-9.
[13] R Development Core Team. R: A Language and Environment for Statistical Computing. 2013.
[14] Therneau TM. A Package for Survival Analysis in S. 2014.
[15] Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499:43-9.
[16] Bengtsson H, Neuvial P, Speed TP. TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC bioinformatics. 2010;11:245.
[17] Bengtsson H, Wirapati P, Speed TP. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics. 2009;25:2149-56.
[18] Ortiz-Estevez M, Aramburu A, Bengtsson H, Neuvial P, Rubio A. CalMaTe: a method and software to improve allele-specific copy number of SNP arrays for downstream segmentation. Bioinformatics. 2012;28:1793-4.
[19] Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America. 2010;107:16910-5.
[20] Martinez P, Birkbak NJ, Gerlinger M, McGranahan N, Burrell RA, Rowan AJ, et al. Parallel evolution of tumour subclones mimics diversity between tumours. The Journal of pathology. 2013;230:356-64.

[21] Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research. 2013;41:D64-9.
[22] Wuttig D, Zastrow S, Fussel S, Toma MI, Meinhardt M, Kalman K, et al. CD31, EDNRB and TSPAN7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases. International journal of cancer Journal international du cancer. 2012;131:E693-704.
[23] Kosari F, Parker AS, Kube DM, Lohse CM, Leibovich BC, Blute ML, et al. Clear cell renal cell carcinoma: gene expression analyses identify a potential signature for tumor aggressiveness. Clinical cancer research : an official journal of the American Association for Cancer Research. 2005;11:5128-39.
[24] Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD. Gene expression profiling predicts survival in conventional renal cell carcinoma. PLoS medicine. 2006;3:e13.
[25] Lane BR, Li J, Zhou M, Babineau D, Faber P, Novick AC, et al. Differential expression in clear cell renal cell carcinoma identified by gene expression profiling. The Journal of urology. 2009;181:849-60.
[26] Brannon AR, Reddy A, Seiler M, Arreola A, Moore DT, Pruthi RS, et al. Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns. Genes & cancer. 2010;1:152-63.
[27] Beleut M, Zimmermann P, Baudis M, Bruni N, Buhlmann P, Laule O, et al. Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient outcome. BMC cancer. 2012;12:310.
[28] Beroukhim R, Brunet JP, Di Napoli A, Mertz KD, Seeley A, Pires MM, et al. Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. Cancer research. 2009;69:4674-81.
[29] Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nature genetics. 2006;38:500-1.
[30] Bostrom AK, Lindgren D, Johansson ME, Axelson H. Effects of TGF-beta signaling in clear cell renal cell carcinoma cells. Biochemical and biophysical research communications. 2013;435:126-33.
[31] Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller MC, Shaikh N, et al. Replication stress links structural and numerical cancer chromosomal instability. Nature. 2013;494:492-6.
[32] Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine. 2012;366:883-92.
[33] Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature genetics. 2014;46:225-33.
[34] Brooks SA, Brannon AR, Parker JS, Fisher JC, Sen O, Kattan MW, et al. ClearCode34: A Prognostic Risk Predictor for Localized Clear Cell Renal Cell Carcinoma. European urology. 2014.