PLOS ONE

RESEARCH ARTICLE

Discretization of Gene Expression Data Unmasks Molecular Subgroups Recurring in Different Human Cancer Types

Manfred Beleut[1]*, Robert Soeldner[1], Mark Egorov[1], Rolf Guenther[1], Silvia Dehler[2], Corinna Morys-Wortmann[1], Holger Moch[3], Karsten Henco[1], Peter Schraml[3]*

1 Qlaym Healthcare AG, Hans-Adolf-Krebs Weg 1, 37077 Goettingen, Germany, **2** Cancer Registry Zurich and Zug, University Hospital Zurich, Zurich, Switzerland, **3** Institute of Surgical Pathology, University Hospital Zurich, Schmelzbergstrasse 12, 8091 Zurich, Switzerland

* AnManfred@gmx.de (MB); Peter.Schraml@usz.ch (PS)

OPEN ACCESS

Citation: Beleut M, Soeldner R, Egorov M, Guenther R, Dehler S, Morys-Wortmann C, et al. (2016) Discretization of Gene Expression Data Unmasks Molecular Subgroups Recurring in Different Human Cancer Types. PLoS ONE 11(8): e0161514. doi:10.1371/journal.pone.0161514

Editor: Elda Tagliabue, Fondazione IRCCS Istituto Nazionale dei Tumori, ITALY

Received: January 18, 2016

Accepted: August 5, 2016

Published: August 18, 2016

Data Availability Statement: All data is publicly available from the Gene Expression Omnibus database. GSE numbers are listed in the manuscript.

Funding: Authors MB, RS, ME, RG, CM, KH are employed by the Qlaym Healthcare AG. The specific roles of individual authors are articulated in the "author contributions" section. The funders provided support in the form of salaries but did not have any additional role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This project was also sponsored by a grant of the Swiss National Science Foundation

Abstract

Despite the individually different molecular alterations in tumors, the malignancy associated biological traits are strikingly similar. Results of a previous study using renal cell carcinoma (RCC) as a model pointed towards cancer-related features, which could be visualized as three groups by microarray based gene expression analysis. In this study, we used a mathematic model to verify the presence of these groups in RCC as well as in other cancer types. We developed an algorithm for gene-expression deviation profiling for analyzing gene expression data of a total of 8397 patients with 13 different cancer types and normal tissues. We revealed three common Cancer Transcriptomic Profiles (CTPs) which recurred in all investigated tumors. Additionally, CTPs remained robust regardless of the functions or numbers of genes analyzed. CTPs may represent common genetic fingerprints, which potentially reflect the closely related biological traits of human cancers.

Introduction

The use of DNA microarray technologies enabled the generation of myriads of data, sustaining further molecular sub-classification of many previously described pathologic phenotypes with significant effects on clinical decision making and prognosis [1–6]. In particular, gene expression analysis served as an efficient cancer sub-classification tool [7], and is regarded as the most downstream signal onto which accumulated effects from different molecular layers such as genomics, proteomics or methylomics may imprint [8]. Depictions of distinct driver mutations in genes such as *BRAF*, *EGFR*, *PAK5*, *HER2*, *ALK* or hormone receptors [9, 10], all of which are embedded in individual tumor specific mutational landscapes [11–13], have also been used as prognostic or therapeutic biomarkers for further patient stratification of different cancer subtypes [14–17].

The fact that close to 75% of all genes have already been identified as being potentially cancer relevant [18], and the unique molecular make-up of each tumor [19] suggest that cancer evolution and progression are complex processes. In order to better understand the biology of tumors, functionally classifying deregulated gene candidates according to specific biologic

(grant number: 3238BO-10314) to HM. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Authors MB, RS, ME, RG, CM, KH are employed by the Qiaym Healthcare AG. This commercial affiliation does not alter the authors' adherence to PLOS ONE policies on sharing data and materials. There are no other competing interests to declare.

processes is often the practice of choice [20]. Despite the unique molecular background of each malignant tumor, the "Hallmarks of Cancer" [21] claim that all malignant neoplasms acquire similar characteristics, enabling the transformation from a normal to a cancer cell.

In line with the proposed concept of cancer hallmarks, we assumed that regardless of their highly diverse phenotypic and genotypic appearance, all tumors share common cancer traits, which may potentially be detected *via* gene expression analysis. This hypothesis is backed by our recent identification of global gene expression outputs of prognostic relevance in renal cell carcinoma (RCC) [22]. Based on this finding we evaluated whether similar gene expression patterns may exist also in other cancer types.

Unbiased by any cancer-specific marker or classification schemes currently used, we analyzed 55 published studies of gene expression data encompassing a total of 8397 patients with 13 different cancer types by means of gene-expression deviation profiling.

Materials and Methods

RCC patient data

Survival data linked to the published dataset GSE19949 was provided by the Cancer Registry Zurich and Zug and approved by the ethics committee of the Canton Zurich (KEK-ZH-Nr. 2013–0629). This data was used for the verification of the prognostic relevance (Fig 1) as proposed from our previous work using the same patient cohort [22].

Gene expression data and normalization

All data was retrieved from the GEO repository as published by the authors. An overview of all analyzed datasets is given in S1 Table. Data were pre-processed and normalized as described in S1 Text. Table A in S1 Text and S2 Table provide a systematic overview of all analyzed datasets including clinic-pathological data.

Model generation and algorithm development

The method comprised the following steps: For every normalized $\log_2$ gene probe set value, the mean expression value was calculated over all samples of the given data set. This mean value was then subtracted from the $\log_2$ expression values of all samples. The calculated difference denoted the individual deviance of the sample from the mean expression value for the respective probe set.

To those expression values which were close to mean the value 0, to those with significantly higher values than mean the value 1, and to those with significantly lower expression than mean the value -1 were assigned. The threshold for high and low was set to 43% of the standard deviation, which means that all three values (-1, 0, 1) occurred at almost the same frequency. The value 0 was assigned to those deviation values which were located between -0,43 σ and +0,43 σ. Deviation values lower than -0,43 σ were assigned with -1, those higher than +0,43 σ with +1. Additional categories were therefore automatically excluded.

The samples were then clustered using well known clustering methods such as k-means or SOM, so that samples with similar (individual) profiles were assigned to the same group A, B, or C. In order to generate the respective CTPs the average values for each probe set and group were calculated.

At the end of this procedure, a whole genome CTP profile existed for every group, represented by a vector with a length equal to the number of probe sets of the respective microarray and values ranging between -1 and 1 (Fig 2 and S1 Text).

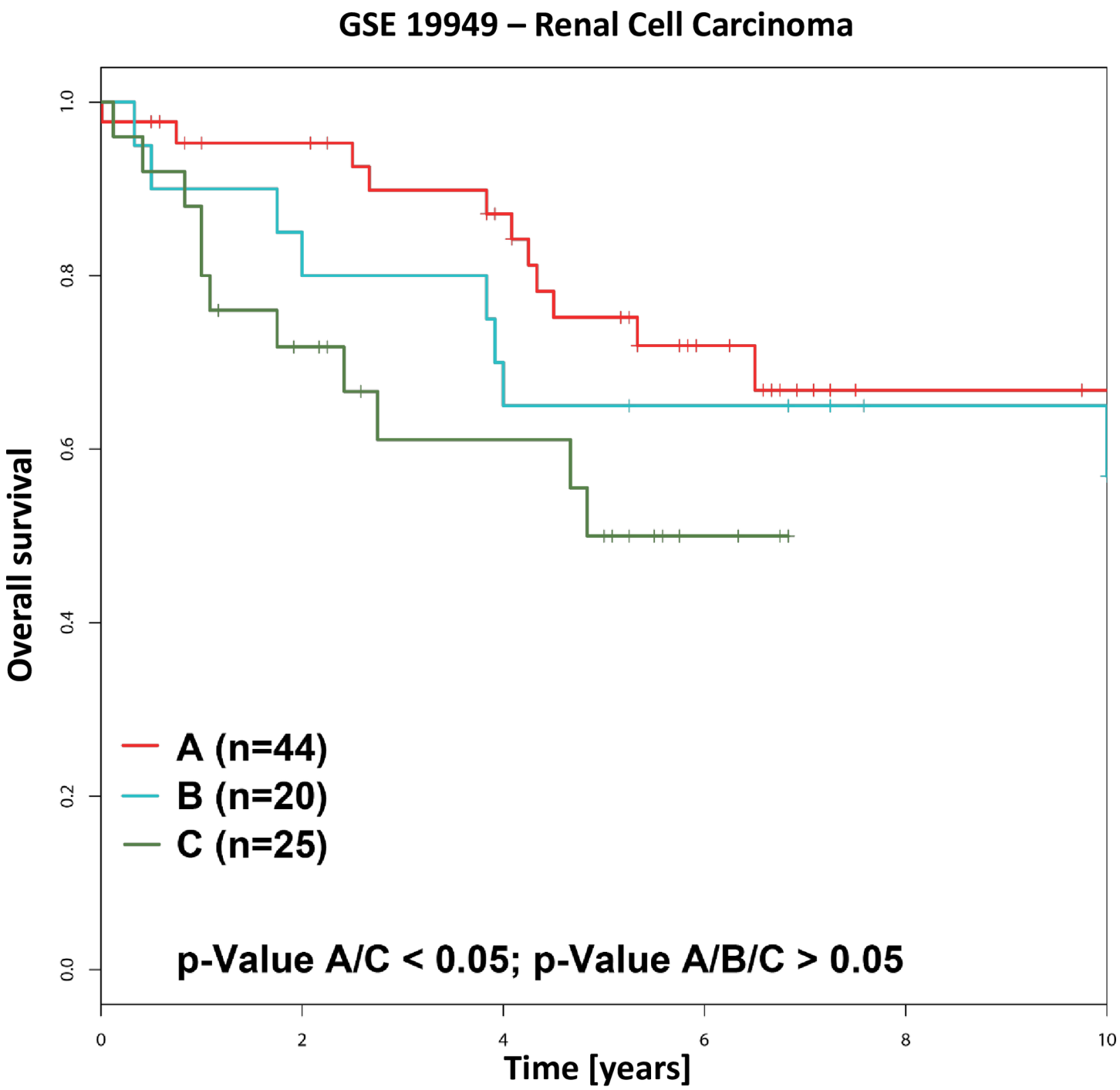

Fig 1. RCC groups A, B, and C and their correlation with patient survival. Kaplan-Meier curves for overall survival in relation to cancer transcriptome profiles (CTPs) as described for the patient cohort GSE19949 [22]. Survival data were made available by the Cancer Registry Zurich and Zug.

doi:10.1371/journal.pone.0161514.g001

For the RCC data set GSE19949 the assignments to A, B, or C groups were known *a priori* [22]. Detailed methodological descriptions and codes can further be found in S1 Text. For the interested reader, raw and calculated data is available upon request.

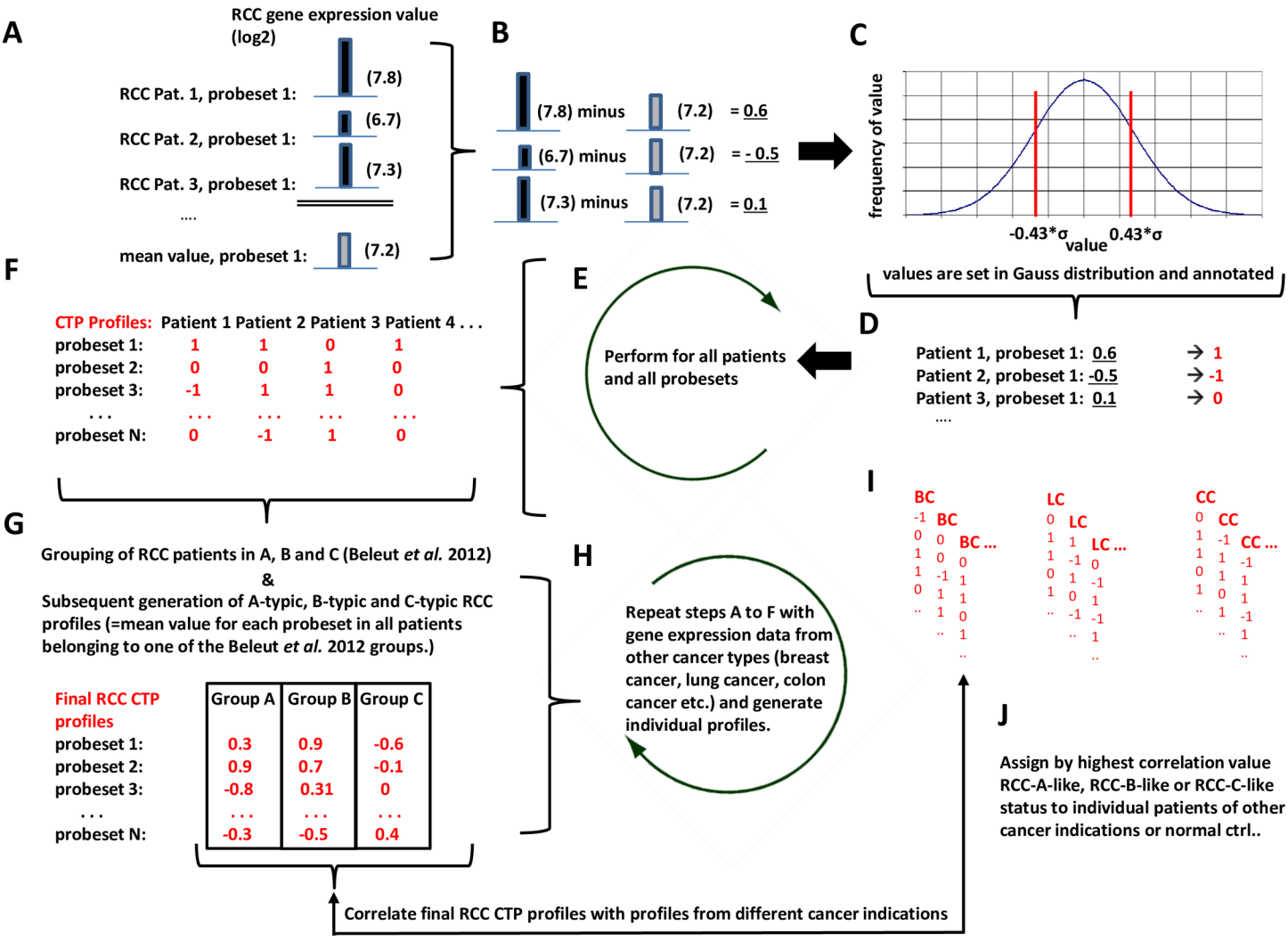

Fig 2. Workflow for generating RCC-CTPs. A. Determination of the gene expression value (log_2) of gene probe set 1 in all RCC samples tested and generation of the mean value for probe set 1 of the entire RCC tumor cohort (exemplified values are shown). **B.** Subtraction of the mean value from true expression value for probe set 1 of each patient. **C.** Distribution of the remaining deviation values from mean for probe set 1. **D.** Annotation of remaining deviation values from mean for probe set 1 as 1, -1 or 0 depending on their localization in the distribution. **E.** Steps A to D are performed for all probe sets of the gene expression microarray. **F.** Individual CTP profile for each patient given by a vector covering all expression values. **G.** Grouping of RCC patients into CTP-A, -B or -C according to Beleut *et al*. 2012; Determination of the CTP mean values of all gene probe sets for patient group A, group B and group C and generation of the final RCC CTP A, -B and -C target vectors. **H.** Gene expression data from other cancer types calculated according to steps A to F and generation of patient-specific CTP-vectors. **I.** Correlation of RCC CTP-A, -B and -C target vectors from step G with patient-specific CTP of other cancer types derived from step H. (BC) breast cancer patients; (LC) lung cancer patients, (CC) colon cancer patients. **J.** Assigning a patient or control to tumor subgroup according to the CTP with the highest correlation.

doi:10.1371/journal.pone.0161514.g002

De Finetti-like mappings

As proposed by *de Finetti* [23] ternary plots or *de Finetti* mappings are efficient means to depict percentage compositions for 3 parameters in an equilateral triangle [24, 25]. Using this method, we were able to plot the distances of individual patient samples obtained from the 3 predefined CTP centroids. Additional info to the *de Finetti* like mappings can be found in S1 Text.

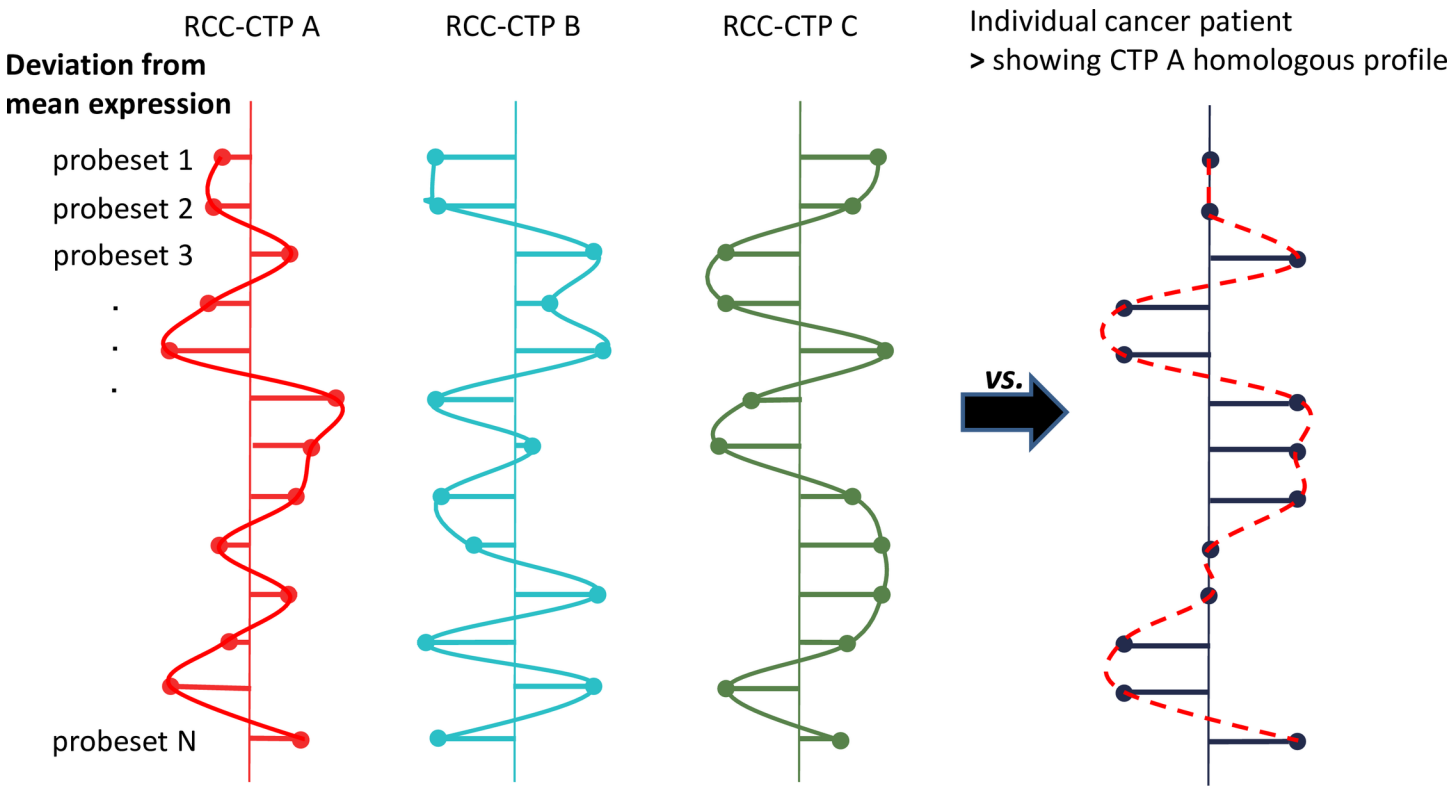

Fig 3. Graphical illustration of CTPs and patient classification. Shown is a graphical overview on the nature of generated RCC-CTPs as well as their comparison with the CTP of one patient with a different cancer type. The deviation from the mean expression for all probe sets and their relative correlation to each other (continuous curves) define the RCC-CTP target profiles. For CTP assignment, the CTP profile of an individual patient with another cancer type (dashed curve) is compared to the target RCC-CTPs.

doi:10.1371/journal.pone.0161514.g003

CTP assignment

For each sample of a new data set, the expression profile was calculated as mentioned above. As for individual samples no averaging takes place, the profiles contained only the values -1, 0, or 1. The correlations between the sample profile and the whole genome A-, B- and C-CTPs were calculated. The sample was assigned to the CTP with the highest Pearson correlation (Figs 2 and 3 and S1 Text).

Results

Genome wide expression analysis confirms three subgroups in two independent RCC patient cohorts

By analyzing gene expression profiles in a RCC patient cohort (GSE19949) we recently identified three subgroups (termed groups A, B and C), which were not significantly associated with pathological prognostic parameters such as nuclear differentiation grade and tumor stage (Beleut *et al.*, see additional file 12, table S8 [22]). As survival data was sparse at that time point for this patient cohort, we analyzed a second RCC patient cohort using tissue microarrays and immunohistochemistry. By correlating survival data and expression levels of proteins whose genes were highly expressed in the three groups, we found an association between identified subgroups and patient clinical outcome [22].

As survival data were meanwhile also available for 89 patients of the first cohort, we could confirm the result obtained from the second tissue microarray patient cohort. As shown in Fig 1 the

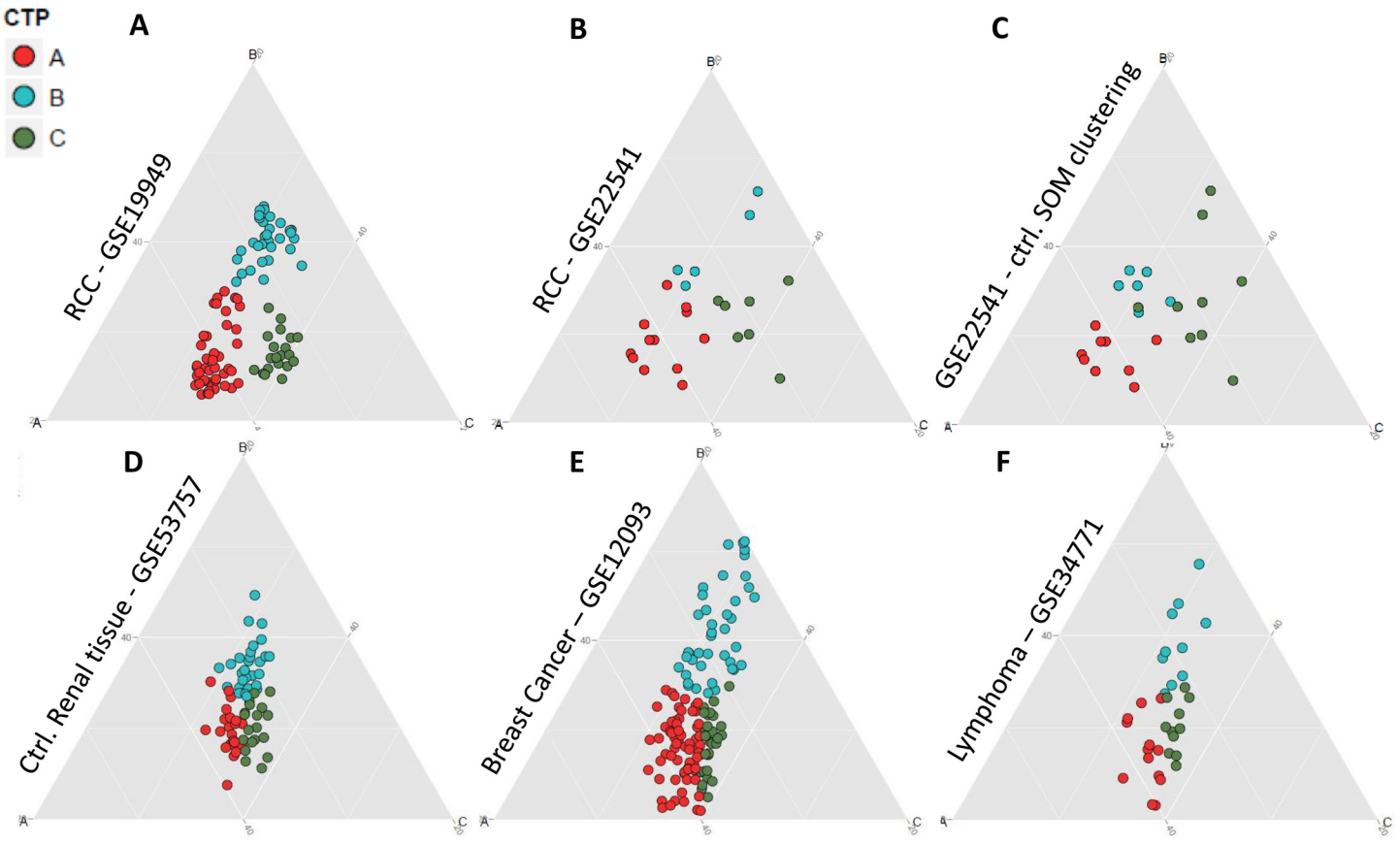

Fig 4. Visualization of CTPs by *de Finetti* like mapping. A. *De Finetti* like diagram illustrating RCC classification of GSE19949 into the three CTP groups as previously defined [22]. Each dot represents one patient, colors code for the distinct groups, as indicated. **B.** *De Finetti* like diagram illustrating the sub-classification of an independent ccRCC dataset (GSE22541) according to the Beleut *et al*. 2012 rules. The color code per patient defines its CTP assignment as identified in A. **C.** Control *de Finetti* like diagram, in which Self Organizing Maps (SOM) have been applied in classifying GSE22541 ccRCC primary tumors for comparison. The color code per patient defines the "CTP" to which the respective tumor would belong according to SOM. **D.** *De Finetti* like diagram illustrating the classification of normal renal tissues of GSE53757 according to identified CTPs. Note the weak correlation of individual samples with profile vectors, leading to mostly a central clustering of analyzed samples. **E.** *De Finetti* like mapping of a breast cancer dataset (GSE12093) and **F.** a Lymphoma dataset (GSE34771) according to RCC-CTPs. Note the increased scattering of individual patients in E and F when compared to control.

doi:10.1371/journal.pone.0161514.g004

survival rate was highest for patients belonging to group A followed by group B and group C, with statistically significant differences between group A versus group C (p-value = 0.040).

Algorithmic Cancer Transcriptomic Profile (CTP) model for subgrouping RCC

To exclude that those RCC gene-expression signatures are biased due to non-cancerous effects, we first aimed at designing a comprehensive algorithmic Cancer Transcriptomic Profile (CTP) classification model using the RCC gene expression data. Requirements for the model included automated reassignment of GSE19949 RCC tumors to the known CTP groups A, B and C. Additionally, the algorithms should be independent from biases derived from different normalization techniques as well as tissue type artefacts. Finally, the model should consider the entire gene expression profile resulting from the microarray chip.

In order to further investigate whether RCCs could be sub-classified by our proposed CTP-model, we utilized the following approach. We classified the GSE19949 tumors into the three known groups by implementing the algorithmic workflow shown in Fig 2A–2G.

Fig 3 provides a graphical overview on the nature and comparison of CTPs within RCC (Fig A in S1 Text). Highest Pearson correlation with one of the three RCC-CTPs assigned a patient either to CTP group A, B or C.

We visualized achieved results by means of a *de Finetti* like mapping (23, 25) (Fig 4A) to better highlight the potential distances for each patient from the respective cluster centroids defined as A, B and C. Next, we classified an independent set of clear cell RCCs (ccRCC) (GSE22541), resulting in three similar CTP clear cell RCC cohorts as shown in Fig 4B.

To further solve the question whether other clustering technologies could also have been used as a potential starting point for CTP identification, we utilized Self Organizing Maps (SOM) [26] for clustering of GSE22541 and compared achieved results with the classification shown in Fig 4B (Fig 4C). Result overlap of both technologies was 75%, suggesting that CTPs can be identified independently of the data set GSE19949 which was the starting point of our observations and calculations. The similar outcome with the two clustering technologies used suggests the presence of distinct CTPs in different sets of RCC tumors.

As control, we applied this classification rule on a set of normal renal tissues (GSE53757) but also on other healthy tissues derived from different organs or anatomic body parts as annotated in GSE1133 and GSE2361. In contrast to RCC, healthy controls remained grouped around the center of the *de Finetti* like mapping (Fig 4D and Fig B in S1 Text) confirming RCC specificity of identified CTPs.

To better investigate cancer specificity in general, we also strictly applied identified RCC-CTPs on one dataset of breast cancer (Fig 4E) and one dataset of lymphoma (Fig 4F), respectively. Despite potential cancer-specific background, resulting *de Finetti* like mappings still illustrate an increased scattering of individual patients as compared to control mappings, thus strengthening the assumption of general cancer specificity rather than RCC specificity of proposed CTPs.

Additionally, we investigated the average correlation of all patients assigned for a particular CTP (S1 Text). For every CTP of a data set we collected all contributing patients and their correlation values with the respective CTP and calculated the average. The results may be interpreted as "cluster diameter" estimate. For RCC, the correlation was highest with 0.439; 0.476; 0.413 for CTP-A, CTP-B and CTP-C, respectively. For the other tumor types, these values ranged between 0.14 and 0.18, whereas for control normal tissues these values ranged between 0.08 and 0.14, much lower than for tumors. This difference between normal and tumor tissue was highly significant (p-value = $2^{*}10^{-6}$). The direct comparison of the normal and tumor tissue samples of the renal data set GSE53757 (0.08 vs. 0.14, respectively) was also significant (p-value = $5^{*}10^{-6}$). We intentionally omitted the SOM clustered data from GSE22541 from this averaging since the independent SOM clustering introduced additional positional scatter. Detailed results for each GSE dataset and corresponding CTPs are listed (Table E in S1 Text).

Transferring the CTP model established for RCC to other cancer types

By transferring this RCC based model to gene expression data of other cancer types, we aimed to regroup other cohorts according to conditions homologous to RCC CTP groups A, B and C. The CTP model and its transfer to other cancer types occurred as illustrated in Fig 2H–2J. Resulting CTP subgroups of different cancer types were also correlated with associated clinical data where available. We calculated respective CTPs for all cancer types from published studies as denoted in S1 Table and compared them with reference RCC-CTPs. Highest Pearson correlation with one of the three RCC-CTPs assigned a patient either to CTP group A, B or C.

CTPs are not RCC-specific and commonly exist in cancer transcriptomes

In order to test whether or not CTPs are exclusively related to RCC, we performed a reverse transfer of the CTP concept (S1 Text). Data from the breast cancer study GSE2603 was used to generate breast cancer-specific CTPs. Samples from this study were clustered with standard unsupervised clustering methods (k-means, SOM) into three clusters, and the respective Kaplan-Meier curves were calculated. The resulting groups were used for transfer to the RCC study GSE19949 with the same procedure utilized for transfer to other cancer types (reverse transfer). The resulting distribution into three groups was 65% identical with the original RCC grouping [22] suggesting a nonrandom finding and a general presence of CTPs detectable in all human cancers.

CTPs seem to be robust irrespective from gene function and number

Our efforts for the identification of our CTP concept in human cancer types always included complete sets of genes present on microarray chips. In order to investigate whether indeed all genes or only specific subsets of genes contributed to CTP vector definition, we chose limited numbers of either randomly or functionally defined gene subsets [27] for the generation of CTP target vector profiles using RCC as starting point. We then transduced resulting confined RCC-CTPs to the expression data of the same genes in all other tumor types and compared achieved CTP affiliation of each tumor. We observed for all tumors and tumor types that 700 randomly chosen genes led to an overall similarity of 86 ± 6.6%. This held also true when 716 tumor suppressor genes and 690 oncogenes were chosen for CTP calculation (80 ± 7.7% and 81 ± 8.1% similarity, respectively). Specific results for the individual studies are depicted in S3 Table. Our overall finding suggests that CTPs are non-random and measurable irrespective of gene expression deviations relatively to each other, distinct gene types, gene functions or gene numbers (Fig 5, S4 and S5 Tables). The data shown in S4 and S5 Tables give examples how gene sets with different functions (tumor suppressor genes and oncogenes) are classified into 3 CTPs. The relationship of the CTP groups among different cancer types can be visualized by sorting the values (ascending or descending) obtained from one gene set of one CTP group.

As measuring gene expression is always accompanied by noise, we investigated its impact by performing a simulation. We asked if the separation between CTP-A, -B, and -C as observed in Fig 1 may be entirely or to large part due to noise. In the simulation, we thus assumed that for all patients the deviation from the average expression value of each probe set is just the result of a random fluctuation. The distribution of the deviations is assumed to be normal for every probe set. The subsequent discretization of the values, calculation of the CTP profiles and the ensuing assignment of the virtual patients to a particular CTP was performed according to the protocol, as well as the calculation of the "virtual cancer" specific profile. A random subset of 716 genes was finally picked to mimic the effect of specific gene selection as in the tumor suppressor or oncogene examples.

For the simulation, we generated 14 virtual cancer patient groups with 50 patients each, similar to the experimental situation.

Correlating all CTP-A profiles with each other across different virtual cancer types, and similarly for CTP-B and -C, yielded average correlations of less than 0.02. Calculating the same average correlations for the experimental tumor suppressor genes or oncogenes yielded average correlations between 0.7 and 0.76. The probability that these high correlations between CTPs across different cancer types were generated by random is infinitesimally small (p-value $< 10^{-100}$). Between the two experimental sets, however, the p-values were much higher (0.04–0.98), indicating that the null hypothesis (generated from the same underlying distribution and not by noise) held as expected.

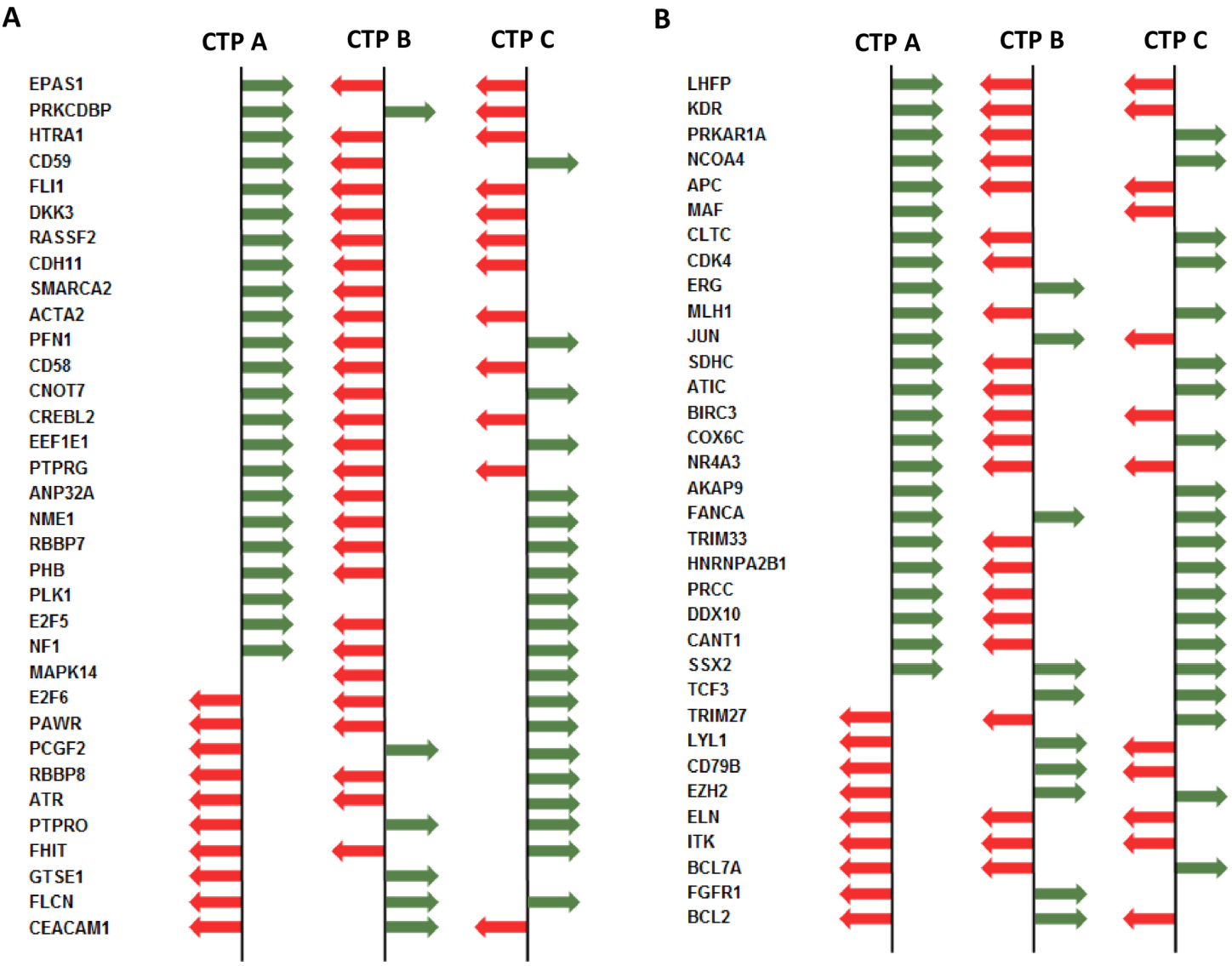

Fig 5. CTP profiles of the different cancer types using randomly picked tumor suppressor or oncogenes. Shown are best descriptors of gene expression deviations from the mean of randomly selected tumor suppressor genes **(A)** and oncogenes **(B)** encompassing different cancer types. For a detailed overview illustrating all expression deviations including genes showing no deviations from the mean, see S4 and S5 Tables. **Red:** relative gene expression deviation is lower than the mean expression, **Green:** relative gene expression deviation is higher than the mean expression. CTPs differ by the overall relative gene expression deviations. The entirety of expression deviations from the mean of genes, not that of one single gene is relevant for affiliating one distinct tumor to CTP-A, -B or -C.

doi:10.1371/journal.pone.0161514.g005

A similar result was obtained when the different CTPs within a cancer type were compared. As expected, experimental correlation values between different CTPs showed values different from 1, i.e. -0.38 (AB), -0.5 –-0.6 (AC), and -0.37 –-0.51 (BC). Again, these correlations were stable across the tumor suppressor and oncogenes platform (p-values ranging between 0.25 and 0.98), while the simulation yielded average correlations between 0.008 and 0.21. This differed significantly from the experimental values (p-value < 0.001).

In summary, these simulations demonstrate that random effects like mere fluctuations of expression values could not explain the high correlations between CTPs across and within different tumor types.

Discussion

The hallmarks of cancer imply that all cancers follow specific biological concepts which form the basis of malignant behavior. Therefore, we speculated that such biological concepts among cancers relate closely to each other and may be visualized by gene expression patterns. Our hypothesis originated from the detection of three gene expression outputs in renal cell carcinoma (RCC), which appeared to be independent from pathological parameters and correlated with patient outcome [22]. In an attempt to further support this hypothesis we designed an algorithmic model, which enabled us to classify Cancer Transcriptomic Profiles (CTP). The initial experiment on RCC was performed on Affymetrix arrays. In order to transduce the algorithmic rule on other cancer types in the first step, we focused on studies performed using Affymetrix microarrays and screened GEO for surgically treated tumor cohorts with associated progression data and identified 55 studies. Being aware of additional useful meta-analysis on different platforms, we noticed that all these platforms lack a proper key to translate their annotations to each other or to Affymetrix or Illumina or even present entirely different technologies such as RNAseq. By focusing on a single platform, we could exclude all additional factors and potential artifacts due to the chip related technical differences. In one instance, however, we compared the Affymetrix outcome with results obtained from Illumina data to demonstrate that the observed results are independent of the gene expression platform chosen. We selected Illumina for this comparison as this platform has about 6000 probes in common with Affymetrix with equal annotations, thus enabling the partial transfer as performed in our text appendix. The generalizations found across tumor boundaries could thus be safely attributed to the underlying tumor biology.

In this study, we analyzed gene expression data obtained from 8397 samples. Sixteen of 55 datasets accounting for 4177 out of the 8387 tumors were from breast cancer, which currently presents the most frequent tumor type in the GEO repository. We detected three different Cancer Transcriptomic Profiles (CTPs) that are recurring in 13 different cancer types. These CTPs suggest molecular relationships among human cancers.

Absolute gene expression values with subsequent associated gene enrichment technologies have been used in many gene expression studies. In keeping with this, we first used a similar approach and detected 3 groups in RCC by two-way non-supervised hierarchical clustering [22]. Our subsequent goal was to find a mathematic approach, an algorithm, for depicting those 3 groups. According to our opinion, this was best achieved by discretization of gene expression data we describe here. The method we have chosen is *per se* not novel, but to our knowledge, its conceptual application to characterize mathematically the 3 CTPs has not been used and published so far. A closer look at the expression profiles of all genes (Fig A in S1 Text) clearly shows that the expression levels of the genes in each CTP group are equally distributed. CTP groups are not characterized by expression patterns of a specific set of genes (e.g. high *versus* low) which differ, for example, between two organs, tissue/cell types or healthy/diseased tissue. As illustrated in Fig 3, it is the composition of the "expression profiles" yielded from the expression deviations from the mean values from all gene probe sets which defines a specific CTP. According to our opinion one would hardly be able to identify those 3 CTPs by using absolute gene expression levels or a more sophisticated "barcoding" model in which gene expression measurements are banalized to 0 (not expressed) *vs*. 1 (expressed) [28, 29].

Furthermore, a binary approach would only distinguish between expression levels which are either low or high in two groups. With the binary approach one focuses only on low (underexpressed) and high (overexpressed) expressed genes but exclude those genes which are more or less equally expressed. Our model includes also genes with no or small expression level changes. A closer look on the 3 CTPs clearly demonstrates that these genes provide similar

contributions to the CTP profiles as the high differentially expressed genes. As a result, CTPs yielded with a 2-bin system would introduce substantial distortions compared to our CTPs.

These three CTP groups first detected in RCC and confirmed by the discretization method, were transferable to 12 additional cancer types. We developed the algorithm applied to address various challenges encountered in transcription analysis, so that individual expression fluctuations per gene or probe set were neutralized. Subsequent to the standard GCRMA normalization step for Affymetrix gene expression chips (S1 Text), we normalized and scaled the expression values of the entire patient set gene-wise. Subsequently we discretized the scaled expression values into 3 levels (S1 Text). By using our -1,0,1 approach of discretization, which aimed at turning the continuous and multi-parametric data into this finite number of discrete elements, we were able to not only facilitate ensuing computations, but also reduce data noise. Clustering the discretized data of the RCC study GSE19949 for 3 Cancer Transcriptome Profiles (CTP-A, -B, -C) with different methods (e.g., k-means, SOM) and an independent classification based on histological parameters [22], yielded almost identical results. Elements of our approach relate to methods applied earlier, such as Linear Discriminant Analysis, Significance Analysis of Microarrays [30] or shrunken centroids [31]. One of the most relevant advantages is that this method is invariant against differences in expression levels resulting from different tissues of origin. It thus may present a novel method enabling the detection of pan-cancer signatures.

According to our mathematical approach, CTPs are a reappearing pattern throughout the cancer transcriptomes. Notably, different clustering technologies, such as k-means or SOM [26, 32], applied after discretization generated similar results. The robust nature of a defined CTP in a tumor was also demonstrated when we investigated the average correlation of all patients assigned for a particular CTP. Despite potential biases which may be caused when calculating with data sets generated from different patient cohorts in different laboratories, the difference of the correlation values obtained from normal and tumor tissue was highly significant. Even the reverse transfer of breast cancer-specific CTPs to RCC resulted in 3 groups which were 65% identical with the original RCC subgrouping further supporting the existence of CTPs in different cancers.

It is of note that due to the very limited availability of samples from normal (control) tissue published in GEO, it is currently not possible to define the particular threshold between normal and tumor tissue. We noticed that the studies mostly consist of disease sample collections with no corresponding healthy counterparts. Therefore, a healthy/disease threshold could not be yielded, unless by including matched pairs of affected and non-affected tissue samples in a sufficiently large amount. The *de Finetti*-like visualizations demonstrate, however, that normal samples located closer to the triangle center than tumor samples. At the center, the correlation with any of the CTP was lowest, pointing to a lower CTP differentiation in normal tissue samples. Being limited to the present situation, however, we can only state, as shown in the results part, that the defined calculated differences are highly significant (p-value = $2^{*}10^{-6}$) and are not occurring due to randomness.

Finally, CTPs still remained stable with sets of only several hundred genes, whether randomly or non-randomly selected. Therefore, we conclude that i. the expression status of every single gene is important to contribute to a CTP and ii. the expression profile of a minimum set of genes is required to yield a CTP.

As the CTPs of our RCC indicated correlation with patient outcome, we calculated Kaplan Meier survival plots using all survival data sets from different cancer types that were available in the GEO database. The results were, however, difficult to interpret (data not shown). Some survival plots for sarcoma, breast and lung cancer were similar to those obtained from RCC. Other survival plots showed no associations or, as in the case of ovarian cancer, even

contrasting patterns. This data strongly suggest that the analysis of survival data sets from different patient cohorts (S1 and S2 Tables) generated from different research groups, as well as the use of different assay and standardization methods may cause dataset bias and inter-dataset noise. The GEO database presents to our knowledge the major resource for researchers for getting access to clinical survival data. Based on our experience its use for analyzing the prognostic value of molecular markers in one specific tumor type or in different cancers is rather limited. However, this tool is ideally suited to perform comprehensive analysis to detect potential associations between molecular signatures and cancer.

Conclusions

We present a novel model that can be applied to identify cancer-specific gene expression profiles. It has been widely accepted that each tumor has developed its own individual molecular landscape. The entirety of molecular events leading to a malignant tumor affects always the same biological concepts described in the hallmarks of cancer. We believe that the CTPs identified by us represent the molecular outputs which exist in all human cancers. However, more in-depth investigations with larger and better defined cancer patient cohorts are needed to support our CTP concept for cancer biology but also as possible additional tool for cancer prognosis.

Supporting Information

S1 Table. All studies considered for CTP affiliation.
(XLSX)

S2 Table. Clinico-pathological parameters per CTPs.
(XLSX)

S3 Table. CTP affiliation and similarity by considering randomly chosen genes.
(XLSX)

S4 Table. CTPs and Tumor Suppressor Genes in different cancer types.
(XLSX)

S5 Table. CTPs and Oncogenes in different cancer types.
(XLSX)

S1 Text. Additional methods, results and figures.
(DOCX)

Acknowledgments

We thank HS Lifesciences GmbH for supporting this project, Andreas Brede-Buchenau, Martin Wiesenfeldt and Ruediger Goetz for critical proofreading and valuable comments.

Author Contributions

Conceptualization: MB PS.

Data curation: RS ME RG CM.

Formal analysis: RS ME RG.

Funding acquisition: KH HM.

Methodology: MB RS ME RG.

Resources: SD KH HM.

Supervision: MB.

Validation: MB RS ME RG CM.

Visualization: MB ME RG CM.

Writing - original draft: MB PS.

Writing - review & editing: MB ME RG CM PS.

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747–52. doi: 10.1038/35021093 PMID: 10963602.
2. Burgess DJ. Gene expression: colorectal cancer classifications. Nature reviews Cancer. 2013; 13 (6):380–1. doi: 10.1038/nrc3529 PMID: 23640209.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286 (5439):531–7. PMID: 10521349.
4. West L, Vidwans SJ, Campbell NP, Shrager J, Simon GR, Bueno R, et al. A novel classification of lung cancer into molecular subtypes. PloS one. 2012; 7(2):e31906. doi: 10.1371/journal.pone.0031906 PMID: 22363766; PubMed Central PMCID: PMC3283716.
5. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine. 2002; 347(25):1999–2009. doi: 10.1056/NEJMoa021967 PMID: 12490681.
6. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769):503–11. doi: 10.1038/35000501 PMID: 10676951.
7. Rhodes DR, Chinnaiyan AM. Integrative analysis of the cancer transcriptome. Nature genetics. 2005; 37 Suppl:S31–7. doi: 10.1038/ng1570 PMID: 15920528.
8. Sohn KA, Kim D, Lim J, Kim JH. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. BMC systems biology. 2013; 7 Suppl 6:S9. doi: 10.1186/1752-0509-7-S6-S9 PMID: 24521303; PubMed Central PMCID: PMC3906601.
9. Banin Hirata BK, Oda JM, Losi Guembarovski R, Ariza CB, de Oliveira CE, Watanabe MA. Molecular markers for breast cancer: prediction on tumor behavior. Disease markers. 2014; 2014:513158. doi: 10.1155/2014/513158 PMID: 24591761; PubMed Central PMCID: PMC3925609.
10. Zhou JX, Yang H, Deng Q, Gu X, He P, Lin Y, et al. Oncogenic driver mutations in patients with non-small-cell lung cancer at various clinical stages. Annals of oncology: official journal of the European Society for Medical Oncology / ESMO. 2013; 24(5):1319–25. doi: 10.1093/annonc/mds626 PMID: 23277484.
11. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446(7132):153–8. doi: 10.1038/nature05610 PMID: 17344846; PubMed Central PMCID: PMC2712719.
12. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010; 463 (7278):191–6. doi: 10.1038/nature08658 PMID: 20016485; PubMed Central PMCID: PMC3145108.
13. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239):719–24. doi: 10.1038/nature07943 PMID: 19360079; PubMed Central PMCID: PMC2821689.
14. Liu YJ, Shen D, Yin X, Gavine P, Zhang T, Su X, et al. HER2, MET and FGFR2 oncogenic driver alterations define distinct molecular segments for targeted therapies in gastric carcinoma. British journal of cancer. 2014; 110(5):1169–78. doi: 10.1038/bjc.2014.61 PMID: 24518603; PubMed Central PMCID: PMC3950883.
15. Fawdar S, Edwards ZC, Brognard J. Druggable drivers of lung cancer. Oncotarget. 2013; 4(9):1334–5. PMID: 23963079; PubMed Central PMCID: PMC3824536.
16. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150(2):251–63. doi: 10.1016/j.cell.2012.06.024 PMID: 22817889; PubMed Central PMCID: PMC3600117.

17. Bollag G, Tsai J, Zhang J, Zhang C, Ibrahim P, Nolop K, et al. Vemurafenib: the first drug approved for BRAF-mutant cancer. Nature reviews Drug discovery. 2012; 11(11):873–86. doi: 10.1038/nrd3847 PMID: 23060265.

18. Huret JL, Ahmad M, Arsaban M, Bernheim A, Cigna J, Desangles F, et al. Atlas of genetics and cytogenetics in oncology and haematology in 2013. Nucleic acids research. 2013; 41(Database issue):D920–4. doi: 10.1093/nar/gks1082 PMID: 23161685; PubMed Central PMCID: PMC3531131.

19. Ogino S, Fuchs CS, Giovannucci E. How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. Expert review of molecular diagnostics. 2012; 12(6):621–8. doi: 10.1586/erm.12.46 PMID: 22845482; PubMed Central PMCID: PMC3492839.

20. Chittenden TW, Howe EA, Culhane AC, Sultana R, Taylor JM, Holmes C, et al. Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. Genomics. 2008; 91 (6):508–11. doi: 10.1016/j.ygeno.2008.03.002 PMID: 18434084; PubMed Central PMCID: PMC2492759.

21. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144(5):646–74. doi: 10.1016/j.cell.2011.02.013 PMID: 21376230.

22. 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. doi: 10.1186/1471-2407-12-310 PMID: 22824167; PubMed Central PMCID: PMC3488567.

23. de Finetti. La Prévision: Ses Lois Logi ques, ses Sources Subjectives. Annales de l'Institut Henri Poincaré 7. 1937:pages 1–68.

24. West D. Ternary Equilibrium Diagrams 2nd. edition: Springer; 2013.

25. Cannings C, Edwards AW. Natural selection and the de Finetti diagram. Annals of human genetics. 1968; 31(4):421–8. PMID: 5673165.

26. Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics. 1982; 43(1):59–69. doi: 10.1007/BF00337288

27. Cao Q, Zhou M, Wang X, Meyer CA, Zhang Y, Chen Z, et al. CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic acids research. 2011; 39(Database issue):D968–74. doi: 10.1093/nar/gkq997 PMID: 20972221; PubMed Central PMCID: PMC3013814.

28. Zilliox MJ, Irizarry RA. A gene expression bar code for microarray data. Nature methods. 2007; 4 (11):911–3. doi: 10.1038/nmeth1102 PMID: 17906632; PubMed Central PMCID: PMC3154617.

29. McCall MN, Jaffee HA, Zelisko SJ, Sinha N, Hooiveld G, Irizarry RA, et al. The Gene Expression Barcode 3.0: improved data processing and mining tools. Nucleic acids research. 2014; 42(Database issue):D938–43. doi: 10.1093/nar/gkt1204 PMID: 24271388; PubMed Central PMCID: PMC3965035.

30. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98 (9):5116–21. doi: 10.1073/pnas.091062498 PMID: 11309499; PubMed Central PMCID: PMC33173.

31. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(10):6567–72. doi: 10.1073/pnas.082099299 PMID: 12011421; PubMed Central PMCID: PMC124443.

32. Hartigan JA. Clustering algorithms. 99 ed: John Wiley & Sons; 1975.