Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines

Felix Dietlein1,2,* and Wolfgang Eschner1,2

1Center for Integrated Oncology, 2Department of Nuclear Medicine, University Hospital of Cologne, Kerpener Str. 62, Cologne 50937, Germany

Received July 10, 2013; Revised September 26, 2013; Accepted October 23, 2013

Next-generation sequencing technologies have led to profound characterization of mutation spectra for several cancer types. Hence, we sought to systematically compare genomic aberrations between primary tumors and cancer lines. For this, we compiled publically available sequencing data of 1651 genes across 905 cell lines. We used them to characterize 23 distinct primary tumor sites by a novel approach that is based on Bayesian spam-filtering techniques. Thereby, we confirmed the strong overall similarity of alterations between patient samples and cell culture. However, we also identified several suspicious mutations, which had not been associated with their cancer types before. Based on these characterizations, we developed the inferring cancer origins from mutation spectra (ICOMS) tool. On our cell line collection, the algorithm reached a prediction specificity rate of 79%, which strongly variegated between primary cancer sites. On an independent validation cohort of 431 primary tumor samples, we observed a similar accuracy of 71%. Additionally, we found that ICOMS could be employed to deduce further attributes from mutation spectra, including sub-histology and compound sensitivity. Thus, thorough classification of site-specific mutation spectra for cell lines may decipher further genome–phenotype associations in cancer.

INTRODUCTION

Over the past years, targeted therapies have tremendously affected the clinical outcome of several cancer types (1–6). The rationale of these personalized approaches is to exploit the individual biology of tumor lesions by specific kinase inhibitors (7). So far, most cancer-specific therapeutically amenable targets have been deciphered in cell culture models (8–10). Bearing in mind the enormous efforts devoted to the elucidation of mutation spectra of primary tumors (11–15), we aimed at developing a pan-cancer characterization of functionally relevant aberrations in cell lines.

Recently, sequencing data of large cohorts of patient samples [The Cancer Genome Atlas (TCGA)] (11) and cell lines [Cancer Cell Line Encyclopedia (CCLE)] (16) became publically available. We used them to establish a genomic classification, linking a gene to the cancer type, in which it is most frequently mutated. Based on this, we developed a computational tool, which infers cancer origins from mutation spectra (ICOMS). We demonstrate its application to predict multiple cell line characteristics from mutation spectra, including its primary site, sub-histology and response to specific kinase inhibitors.

This study involved sequencing data of 1651 genes across 905 cell lines derived from 23 distinct primary sites [CCLE database, (16)] (Fig. 1A). We validated our findings using a collection of 431 independent primary tumor samples [TCGA database (11–15)] (Fig. 1A).

RESULTS

A pan-cancer association between primary tumor sites and cell line mutation spectra

Our first aim was to capture the spectrum of characteristic mutations for each cancer type. To this end, we examined hybrid capture sequencing data of 905 cancer cell lines (Supplementary Material, Table S1) derived from 23 distinct primary tumor sites.
Figure 1. (A) Schematic outline of the five main steps in this study. (B) Selective (purple frame) and frequent (brown frame) biologically relevant meta-mutations of eight cancer types (histology-specific colors) are listed in mutation index descending order. Gene name, selected amino acid changes, mutation index as well as histology-specific (green) and total (red) counts are recorded.
(CCLE database). For each mutation, we recorded its recurrence in the CCLE data set, counted the number of cancer types, for which it most frequently occurred (rating), and calculated its mutation index as a novel measure to assess its feasibility for characterizing a certain primary site (Supplementary Material, Table S2 and Methods).

Altogether, we found that the CCLE data set contained 3435 groups of recurrent mutations. Of these, 2490 (72%) were selective, i.e. associated with a single primary site (Supplementary Material, Table S2, green). Further, the mutation index of 2534 mutations (74%) was high enough to serve as an acceptable genomic predictor. Both criteria were fulfilled for 1993 (58%) distinct recurrent mutations.

Next, we clustered these mutations to a panel of 2717 meta-mutations (Supplementary Material, Table S3), groups of genomic alterations in the same gene that can be combined for primary site characterization. Of these, we selected 1131 (42%) meta-mutations, which were unambiguously associated with a single cancer type (Supplementary Material, Table S3, green). They constituted our genomic classification of cancer types.

The number of selective meta-mutations varied greatly between primary tumor sites (Supplementary Material, Fig. S1A), but did not correlate significantly with their a priori representation in the total cell line cohort \(P = 0.093\) (Supplementary Material, Fig. S1B).

However, we observed significant correlation \(P = 0.002\) with the average count of mutations per sample (Supplementary Material, Fig. S2).

Finally, we studied the functional relevance of the strongest genomic predictors of our primary site classification (Fig. 1B). For this, we sorted meta-mutations by mutation indices (Supplementary Material, Table S3) and compared them with previous reports in literature. Functionally prominent mutations were classified into frequent and selective aberrations; missense mutations were scored by PolyPhen (version 2) algorithm (17), which assessed their disruption of the global protein structure (Fig. 1B).

In concordance with previous studies, we associated mutations of \(K\)RAS, TP53, \(E\)GFR, STK11 (18–23) and DDR2 (24) with lung cancer, mutations of \(B\)RAF (25–27) with melanoma, mutations of \(N\)RAS (28–31) and \(M\)YC (32–34) with hematological malignancies, mutations of ARID1A (35–38) and \(P\)TEN (39–44) with endometrial cancer, mutations of \(A\)PC (14,45–47) and \(B\)MPR2 (48–50) with colorectal cancer, mutations of \(T\)SHR (51–54) and \(T\)PO (55,56) with thyroid cancer and mutations of \(V\)HL (57–61) with kidney cancer. A detailed examination can be found in the Supplementary Material, Mutation Annotation.

Of note, we also recovered typical domains of gain-of-function mutations, e.g. L858R substitutions and exon 19 deletions for \(E\)GFR (62), amino acid substitutions at G12, Q16 for \(K\)RAS (21,63–65) as well as V600E mutations for \(B\)RAF (25–27) (Fig. 1B). Additionally, we observed that loss-of-function mutations dispersed over the whole amino acid sequence of tumor suppressor genes, such as ARID1A (35–38), \(A\)PC (14,45–47) and \(V\)HL (57–61) (Fig. 1B and Supplementary Material, Table S3).

Thus, tumor suppressors could be differentiated from proto-oncogenes by analysis of meta-mutation spectra.

Owing to their specificity and high mutation index, mutations in \(B\)CL3, \(M\)LL3 (lung), \(J\)AK1, \(E\)RRB3, \(R\)BL2, \(A\)LDH7A1 and \(H\)MMR (endometrium), \(P\)2RX7, \(R\)AC1 and \(I\)RS1 (blood), \(M\)EN1, \(F\)ANCI, \(F\)ANCB and \(P\)AX8 (skin) as well as \(S\)ACS, \(R\)OCK1, \(J\)ARID2 and \(T\)ET1 (large intestine) showed up as novel promising targets (Fig. 1B and Supplementary Material, Table S3), which had, to the best of our knowledge, not been associated with their respective primary sites before.

### Computational prediction of primary tumor sites by ICOMS algorithm

We employed this classification to computationally predict primary sites of cell lines by ICOMS algorithm, a novel computational tool that we developed to infer cancer origins from mutation spectra. For each cell line, we calculated its mutation-derived coefficients according to the 23 distinct primary sites (Supplementary Material, Table S4).

Plotting mutation-derived coefficients against cancer types unraveled their strong accumulation on the diagonal (Fig. 2). This reflected their discriminatory power, on which the outcome of the ICOMS algorithm was highly dependent. For 20 of 23 histotypes, coefficients were significantly superior in the group of samples, for which pathological diagnoses matched, compared with complementary control groups (homoscedastic T-test: endometrium, \(P = 3.7 \times 10^{-14}\); blood, \(P = 5.1 \times 10^{-64}\); kidney, \(P = 9.8 \times 10^{-93}\); lung, \(P = 1.9 \times 10^{-38}\); prostate, \(P = 1.8 \times 10^{-82}\); skin, \(P = 6.4 \times 10^{-66}\); thyroid, \(P = 1.7 \times 10^{-92}\), etc.) (Fig. 2 and Supplementary Material, Table S4).

However, distribution shapes of mutation coefficients (Fig. 2 and Supplementary Material, Fig. S3) strongly varied between cancer types, e.g. coefficients of hematological and lung malignancies were broadly spread (average control groups: 1.69 and 5.83) with prominent site-specific amplitudes (coefficient peaks: 8.03 and 15.73). On the contrary, coefficient baseline (average control group: 0.01) and amplitudes (2.05) were low for kidney tumors (Supplementary Material, Fig. S3 and Table S4).

This was considered in our algorithm design both by the independent choice of upper and lower thresholds of mutation coefficients (Supplementary Material, Table S5) and by the hierarchy of cancer types according to the histology order (Supplementary Material, Fig. S4), which resolved ICOMS diagnosis conflicts. In order to optimize upper and lower mutation coefficient thresholds, we calculated two ROC analogs (66) per primary site (Supplementary Material, Fig. S3).

Naturally, strong a priori representations (e.g. breast, large intestine and ovary) allowed lower choice of both coefficients, whereas less focal distributions (e.g. blood and lung) required isolated increase of the lower threshold (Supplementary Material, Table S5).

Based on this optimal calibration, we performed the ICOMS algorithm on the whole CCLE data set (Fig. 3A,B). We calculated both the fraction of samples, for which a computational diagnosis was available (available diagnoses, i.e. sensitivity) and the portion of these, for which computational diagnosis matched CCLE cell line annotation (correct diagnoses, i.e. specificity) (Fig. 3C and Supplementary Material, Table S6).

Overall, the ICOMS algorithm achieved sensitivity and specificity rates of 58 and 71%, respectively, which strongly varied between cancer types (Fig. 3C). Surprisingly, no significant association of performance with a priori representation in the total cell line cohort \(P = 0.064\) was perceived. Particularly, breast (rank: 11) and central nervous system (rank: 13) achieved low performance results (specificity ≤65%), whereas the
weakly represented origins of endometrium (rank: 4), pancreas (rank: 3) and kidney (rank: 6) cancer had excellent accuracy rates (specificity $\geq 90\%$) (Fig. 3C).

Comparing the histology distribution in the total cell line cohort to the panel of cell lines, for which a computational diagnosis was available, confirmed these findings (Supplementary Material, Fig. S5). Lung ($P < 0.0001$) and blood ($P = 0.008$) malignancies were significantly over-represented in the former, whereas primary sites of bone ($P = 0.0003$), oesophagus ($P = 0.042$), soft tissue ($P = 0.049$), upper aerodigestive tract ($P = 0.0024$) and central nervous ($P = 0.036$) malignancies, which had achieved weak performance rates, were underrepresented. Additionally, enlarging the histology distribution of mismatches, revealed blood (36 mismatches), lung (32 mismatches), pancreas (11 mismatches) and breast (8 mismatches) to be the most frequent targets of incorrect predictions (Supplementary Material, Fig. S5).

Finally, we examined whether ICOMS prediction mode variegated specificity rate (Fig. 3D, Supplementary Material, Table S6). ICOMS derived its diagnosis either from positive histology selection (primary diagnosis, 36%) or from exclusion of residual histotypes (secondary diagnosis, 24%) (Fig. 3B,D). On the contrary, unavailability of an ICOMS diagnosis was caused either by unavailability of predictive mutations or by incompatible combination of predictive mutations (primary and secondary diagnosis conflicts); the former alternative dominated (71%) the group of unavailable diagnoses in our data set (Fig. 3D). Notably, specificity was significantly superior for primary compared with secondary diagnoses ($P < 0.0001$). This suggests that diagnosis modes can be interpreted as distinct levels of diagnosis validity (Fig. 3B,D and Supplementary Material, Table S6).

### Systematic characterization of overlaps between meta-mutation spectra

Our cancer origin classification followed a winner-take-all strategy (67), which made it robust against minor changes in low-frequency histology-mutation associations as well as inadequate primary site annotations in the training set (Supplementary Material, Methods). Hence, merely the most probable primary site of each mutation was considered for prediction (Fig. 1B and Supplementary Material, Table S3), so that we assessed whether strong overlaps of mutation spectra might negatively affect ICOMS accuracy (Fig. 3C,D).

First, we systematically characterized the overlaps of specific meta-mutations (Supplementary Material, Fig. S6A and Table S7) by generalizing the definition of the rating $g_i$ (Supplementary Material, Fig. S6B and Methods). We detected nine inter-histologic overlaps of mutation spectra. In concordance with previous
studies, we found activating mutations of \textit{BRAF}, apart from their high frequency in melanoma (25–27), also in cell lines derived from colorectal (14,26,68) or thyroid cancer (69,70) (Supplementary Material, Fig. S6A and Table S7). Additionally, both in endometrium (39–44) and brain cancer cell lines, frameshift mutations of \textit{PTEN} were present (Supplementary Material, Fig. S6A and Table S7), which have been characterized as the second most frequent mutation of glioblastoma multiforme (71–73) (prevalence: \(\approx 20\%\)).

Most strikingly, though, we found that malignancies of lung, endometrium, large intestine and prostate were strongly cognate from a genomic perspective. In total, 20 selective mutations associated with all four cancer types; 27 and 33 further mutations were related to three or two of these histotypes, respectively (Supplementary Material, Fig. S6A and Table S7), which have been characterized as the second most frequent mutation of glioblastoma multiforme (71–73) (prevalence: \(\approx 20\%\)).

Independent validation of ICOMS algorithm accuracy

Next, we sought to confirm the accuracy of ICOMS algorithm for disjoint choices of training and validation cohorts. We used training sets to derive genomic classifications and validation sets to assess the ICOM performance. This strategy was implemented by two orthogonal approaches.

First, we compiled 500 random segregations of the CCLE data set (16) into non-overlapping training (95%) and validation (5%) cohorts. From each 95%-training set, we derived a classification
of cancer types; subsequently, we employed it to perform ICOMS on the complementary 5%-validation set with unaltered choices of thresholds. For each partition, we recorded ICOMS performance (availability and correctness) both on the total validation set as well as on each histology subset (Supplementary Material, Figs. S8 and S9).

As expected, ICOMS performance rates varied strongly, in particular for cancer types, which were weakly represented in the total cell line cohort. However, average rates were close to our previous prediction outcomes on the total CCLE data set ($D_{\text{total}} \sim 2\%, D_{\text{max}} = 6.2\%$) for 19 of 23 cancer types (Supplementary Material, Figs. S8 and S9). This provided a first confirmation of the general validity of our approach on cancer cell lines.

Secondly, we rated the performance of ICOMS on an independent validation panel of 431 patient tumor samples [TCGA database (11–15)], which comprised equal portions of cancers originating from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8). As the genes, which were captured during propagation, originated from lung, kidney, breast and large intestine (Supplementary Material, Table S8).

With this calibration, ICOMS achieved average sensitivity (58%) and specificity (71%) rates on TCGA samples, which were comparable with cell lines (Fig. 3E and Supplementary Material, Fig. S10). Hence, mutation-derived coefficients of TCGA samples calculated on the CCLE classification were naturally lower (e.g. large intestine, $P = 2.9 \times 10^{-8}$, lung, $P = 3.7 \times 10^{-4}$) than coefficients of CCLE cell lines, so that mutation index thresholds had to be adapted for primary tumors (Supplementary Material, Table S9).

With this calibration, ICOMS achieved average sensitivity (58%) and specificity (71%) rates on TCGA samples, which were comparable with cell lines (Fig. 3E and Supplementary Material, Fig. S10). Hence, mutation-derived coefficients of TCGA samples calculated on the CCLE classification were naturally lower (e.g. large intestine, $P = 2.9 \times 10^{-8}$, lung, $P = 3.7 \times 10^{-4}$) than coefficients of CCLE cell lines, so that mutation index thresholds had to be adapted for primary tumors (Supplementary Material, Table S9).

Concordantly, also primary tumors of large intestine, lung and kidney reached excellent outcomes (sensitivity ≥64%, specificity ≥68%), whereas prediction performance was still unsatisfactory for breast cancer (Fig. 3E, Supplementary Material, Fig. S11 and Table S10).

Clinical application of ICOMS algorithm

Several recent studies (74–79) have endeavored enormous technical efforts to isolate circulating tumor cells (CTCs), which have sloughed off cancerous tissue, from peripheral patient blood. As typically only few tumor cells circulate in patient blood [5–50 CTCs per teaspoon of blood (79)], current diagnostic techniques are limited.

Although the sensitivity and specificity rates, which ICOMS reached on CCLE and TCGA data sets, were far too low for safe clinical application, we exemplarily tested clinical feasibility of ICOMS for this diagnostic purpose (Supplementary Material, Fig. S12). We analyzed sequencing data (68 cancer-associated genes) of CTCs, which were collected from peripheral blood of two colorectal cancer patients (80) according to our CCLE classification (Supplementary Material, Table S3). Both samples were correctly allocated to large intestine. While diagnosis of Patient 1 was unambiguous (secondary diagnosis), a diagnosis conflict (large intestine versus lung) had to be resolved for Patient 2 (Supplementary Material, Fig. S12).

Further application of mutation index classifications and ICOMS prediction algorithm

Finally, we tested two further potential applications of our classification and prediction procedures (Fig. 4 and Supplementary Material, Fig. S13). Parallel to classifications of general cancer types, we first used our algorithms to render cancer sub-histologies. For lung cancer, we found unique annotations for 148 of 172 cell lines (Supplementary Material, Table S1), from which we derived a genomic lung cancer classification (Fig. 4). Again, we found that several strong genomic predictors were in concordance with literature (Fig. 4A). Mutations in KRAS (18,20,21,23), ALK (3,81,82), EGFR (1,2,18,20,21) and STK11 (18–20) were associated with the adenocarcinoma subtype, whereas alterations of RBL (83,84) and PTEN (83,85) predicted the small-cell (SCLC) histology (Fig. 4A). Mutation indices of large-cell ($P = 4.37 \times 10^{-7}$, LCLC) and squamous-cell ($P = 1.93 \times 10^{-8}$, SQLC) carcinomas differed significantly from residual lung tumors. They have been reported to exhibit higher spontaneous mutation rates (83), resulting in less recurrent and characteristic mutations, which could be used for sub-histology classification.

Nevertheless, mutation indices still showed strong accumulation on the diagonal (Fig. 4B) and discriminatory powers of mutation indices were significantly strong for small-cell ($P = 6.91 \times 10^{-29}$), large-cell ($P = 2.89 \times 10^{-18}$), squamous-cell ($P = 1.29 \times 10^{-8}$) and adenocarcinomas ($P = 5.77 \times 10^{-17}$). Accordingly, ICOMS performance was satisfactory on lung cancer in general (sensitivity: 81.1%, specificity: 95.0%) (Fig. 4C). The same held true for all subtypes except for large-cell carcinoma, which was erroneously mapped to small-cell and adenocarcinomas in 33% of all diagnoses (Fig. 4C). On the TCGA lung cancer cohort (104 adenocarcinomas), ICOMS reached performance rates of 64.4% (sensitivity) and 97.0% (specificity), respectively (Fig. 4D).

Secondly, we sought to infer cell line response to specific kinase inhibitors from mutation spectra (Supplementary Material, Fig. S13). To this end, we involved screening data of three compounds (Erlotinib, PLX4720 and LBW242) across large panels of cancer cell lines ($n = 450$; 443; 450) (Supplementary Material, Fig. S13A,B,C). In concordance with other studies (1,2), mutant EGFR was identified as strongest sensitivity marker for the EGFR-inhibitor Erlotinib (Supplementary Material, Fig. S13D). As reported previously (4,5,86), the BRAF-inhibitor PLX4720 displayed strongest potency in BRAF-mutant lines (Supplementary Material, Fig. S13E). The RSI1 gene, known to modulate BRAF addiction (8), was also associated with PLX4720 sensitivity. In contrast, the TP53 gene emerged as strong predictor for PLX4720 resistance, as the majority of sensitive lines (72%) were derived from melanoma, which is known to exhibit low TP53 mutation frequency (88–90) (Supplementary Material, Fig. S13E). More genomic diversity was observed for the IAP-inhibitor LBW242 (Supplementary Material, Fig. S13F). Here, mutations in FBXW7 and NFKB2 pointed to alterations in pathways, which have been associated with the antineoplastic effect of IAP inhibitors in combination with steroids (91–94).

DISCUSSION

Our study developed a comprehensive framework of mathematical approaches, which systematically capture the information about primary tumor sites that are conveyed by the mutation spectra of cell lines (Fig. 1A). For this purpose, we developed a novel mutation index, by which we established a classification of site-specific genomic aberrations for each cancer type (Fig. 1B...
and Supplementary Material, Table S3). Based on this, we introduced ICOMS (Fig. 3A,B), a computational tool to infer the origin of a cell line merely from its mutation spectrum (Supplementary Material, Tables S6 and S10). We demonstrated that the range of potential ICOMS applications might be extended, including the discrimination of sub-histologies (Fig. 4) as well as the pre-selection of a suitable therapy regimen (Supplementary Material, Fig. S13). In summary, we derive the following three main conclusions from our analyses:

(1) Strong ICOMS performance on an independent cohort of primary tumors (Fig. 3E and Supplementary Material, Table S10) suggested that this approach may amend molecular-pathologic diagnostic procedures of tumor classification in ambiguous cases. We exemplarily demonstrated its clinical feasibility to sequencing data of circulating tumor cells (Supplementary Material, Fig. S12).

(2) Tumor-specific mutation spectra differed strongly in size, recurrence and specificity between cancer types (Fig. 1B and Supplementary Material, Table S3). For this reason, cell lines of certain primary sites could be easier allocated by ICOMS than others (Fig. 3C,E). This aspect should be considered for further design and interpretation of cell culture studies; it provides a way to identify those cell lines, which are genomically marked-off and carry mutations that are characteristic of their respective origin.

(3) We employed mutation indices to characterize spectra of genomic aberrations that are characteristic of a particular cancer type. Thorough comparison to literature revealed that these spectra were mainly in concordance with previous studies (Fig. 1B and Supplementary Material, Mutation Annotation) (11–15,18–61). However, we also identified some novel promising aberrations, which had not been associated with the respective histology before. Yet, most of these novel associations between cancer origin and mutation are still on an unconfirmed level and require further validation by an independent sequencing method.

MATERIALS AND METHODS

Sequencing data and annotation of cell lines and patient tumor samples

We downloaded hybrid capture sequencing data of 905 cell lines (Supplementary Material, Table S1) as MAF-files from Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) Download
Portal (http://www.broadinstitute.org/ccle/home) (16). Additionally, we used CCLE cell line response data (half maximal growth inhibitory concentrations, GI₅₀) to the specific kinase inhibitors Erlotinib, PLX4720 and LBW242 (Supplementary Material, Fig. S1A,B,C). Further cell line annotation data (primary tumor site and sub-histology type) were compiled from the CCLE database (16) or, if not annotated, directly from the cell line provider (ATCC, DSMZ, HSRRB and ECACC) (Supplementary Material, Table S1).

Mutation calling data of 431 patient samples (MAF-files) were obtained from TCGA (11) Data Portal (Data Matrix Access, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) (Supplementary Material, Table S8). These data had been submitted by Broad Institute (Lung Adenocarcinoma and Kidney Renal Clear Cell Carcinoma) (12,13), Baylor College of Medicine (Colon Adenocarcinoma) (14) and Washington University School of Medicine (Breast Invasive Carcinoma) (15). For each of these cancer types, we selected ~100 random samples for download (Supplementary Material, Table S8). Our analysis further included sequencing data (68 cancer-associated genes) of circulating tumor cells (Supplementary Material, Fig. S12), collected from peripheral blood of two colorectal cancer patients (80).

Characterization of mutations by recurrence, rating and mutation index

We called a mutation recurrent, if at least one independent mutation was tractable in the data set, so that the affected amino acids of the two aberrations were displaced by at most one position (Supplementary Material, Table S2).

For each recurrent mutation, we determined its rating and mutation index. The rating denoted the number of cancer types, for which the mutation was predominantly detected (mutation associated types) (Supplementary Material, Methods). The mutation index lₓ served as overall measure for the qualification of a mutation to discriminate between tissues. We derived the latter from a weighted product of conditional occurrence and histology likelihood, incorporating histology-specific occurrence (mutation frequency in mutation-associated histotypes), mutation-specific occurrence (conditional probability of a mutation-associated histology upon mutation detection) and mutation rendering ratio (ratio of frequency in mutation-associated histotypes and frequency in the gₓ next most common histologies) (Supplementary Material, Table S2 and Methods).

Based on this, we termed a mutation selective, if exactly one histology was discriminated from the others, i.e. gₓ = 1 (Supplementary Material, Table S2, green). We considered a mutation as qualified for characterizing its associated cancer types, if lₓ ≥ 1.5. The latter threshold was derived from overall analysis of histology-wise specificity-sensitivity-patterns (receiver operating characteristic (ROC) curves (66), data not shown).

Choice of meta-mutations

For each cancer type, we compiled a characteristic list of pairwise disjoint meta-mutations (Supplementary Material, Table S3). The latter are collections of genomic alterations in the same gene that can be combined for predicting a primary site. We obtained them by maximizing a cardinality-penalized version of the mutation index lₓ (Supplementary Material, Methods). We included all optimized meta-mutations, which associated with a single cancer type, in our genomic classification (Supplementary Material, Table S3, green). For each combination of gene and primary site, this resulted in one of the following four relation types:

No Association the gene bears no meta-mutations, which are selective for the respective cancer type. Amino Acid Specific Association the gene bears a collection of at most three recurrent aberrations, which are characteristic of the respective primary site. Dissimilar alterations of the gene can still be associated with other sites. Mutation Recurrence Association the group of recurrent mutations detected in the gene is linked to the respective tissue. Complete Mutation Spectrum Association all mutations in the gene, including sporadic alterations, are linked to the respective tissue (Supplementary Material, Table S3 and Methods).

Inferring cancer origins from mutation spectra (ICOMS)

The ICOMS prediction procedure is composed of three main steps (Fig. 3A,B) and uses sample-wise mutation spectra, two coefficient thresholds δᵢ ≥ δᵧ (Supplementary Material, Tables S5 and S9) for each histotype H (upper threshold δᶇ, lower threshold δᵧ), as well as a partial histology order (Supplementary Material, Fig. S4) as input.

Given a mutation spectrum, we computed histology-specific mutation coefficients m₁ (Fig. 2 and Supplementary Material, Table S4) (step 1) and compared them with upper and lower thresholds δᵢ (Supplementary Material, Table S5). An H was associated as primary histotype, if the corresponding mᵤ exceeded δᵢ, whereas histologies, for which mᵤ undercuts δᵢ, were excluded from further analysis of the mutation spectrum (excluded histotype) (Fig. 3A). For the residual histotypes, we did not draw any conclusion (secondary histotype) in this step.

Finally, these assessments were combined to derive histology predictions (step 3) (Fig. 3B, Supplementary Material, Tables S6 and S10). If a unique histology had been primarily associated in step 2, it was chosen as computational diagnosis (primary diagnosis). In case more than one existed, we tested whether the set of primary associations gained a unique maximum with respect to the partial histology order defined in the input. If so (resolvable primary diagnosis conflict), the maximum was chosen as final prediction; if no, no diagnosis was available (unresolvable primary diagnosis conflict). In case no primary histotypes were available, we studied the set of secondary associations further. If the latter was unique, this sole histology was selected as prediction (secondary diagnosis); if the set was either empty (no mutations available) or not unique, then no diagnosis was available (secondary diagnosis conflict) (Fig. 3B,D and Supplementary Material, Methods).

Parameter calibration for ICOMS algorithm

In order to optimize upper and lower mutation coefficient thresholds (Supplementary Material, Tables S5 and S9), we calculated two ROC analogs (66) per primary site (Supplementary Material, Fig. S3 and Methods). More precisely, we continuously varied for each cancer type its associated mutation coefficient threshold. We determined the fraction of samples, which exceeded this threshold, separately for cell lines of the respective
cancer type \( (n_1, \text{matched curve}) \) as well as the complementary subset \( (n_2, \text{control curve}) \). Given \( e(y_1) \) and \( e(y_2) \) as coordinates of matched and control curves, choosing \( e \) as upper threshold results in \( y_1 n_1 \) correct and \( y_2 n_2 \) incorrect primary associations. Hence, the ratio \( y_1 n_1/y_2 n_2 \) served as optimization term of the upper threshold. Analogously, given \( (\delta y_1) \) and \( (\delta y_2) \) as coordinates of matched and control curves, setting the lower cutoff to \( \delta \) produces 1 \(-\) \( y_1 \) incorrect and 1 \(-\) \( y_2 \) correct exclusions of cancer types from diagnosis, so that the quotient \( (n_2 - y_2 n_2)/(n_1 - y_1 n_1) \) served as optimization term of the lower threshold (Supplementary Material, Fig. S3).

Also the choice of a primary partial order (Supplementary Material, Fig. S4) was carried out by a greedy algorithm technique, iteratively selecting a new relation by maximizing the difference between correct and incorrect primary conflict solutions. More precisely, starting from an empty set, a relation \( H_1 \succ H_2 \) was added, if the ratio of correct and incorrect primary conflict solutions derived from the additional respective relation was at least 3. If multiple alternatives for the subsequent choice existed, we selected the relation, for which the difference between correct and incorrect solutions was maximal. Finally, the primary partial order was defined as convex hull of the relations selected (Supplementary Material, Fig. S4 and Methods).

**Inferring compound sensitivity from mutation spectra**

We dichotomically clustered cell lines into resistant and sensitive groups, according to their response (half maximal growth inhibitory concentrations, GI50) to the specific kinase inhibitors Erlotinib (EGFR), PLX4720 (BRAF) or LBW242 (IAP) (Supplementary Material, Fig. S13A,B,C). As expected, resistant groups were strongly over-represented (\(-97\%) in the total cell line cohort. For each compound, we therefore compiled 50 random cell line panels, which included all sensitive lines as well as an equal number of resistant cell lines. For each random panel, we determined a genomic sensitivity classification, comprising meta-mutations and their mutation indices. We employed them to calculate weighted averages of mutation indices, where resistant and sensitive groups were equally assessed with contrary signs (Supplementary Material, Fig. S13D,E,F). Similarly, we calculated the weighted fraction of subsets (selective fraction), which rendered a meta-mutation as predictor for either of the two groups. Finally, we sorted the meta-mutations of each group by mutation indices, in order to decipher the strongest genomic markers for compound sensitivity or resistance (Supplementary Material, Fig. S13D,E,F).

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**ACKNOWLEDGEMENTS**

We thank Drs. Markus Dietlein, Alexander Drzezga, Florian Malchers, Martin Peifer, H. Christian Reinhardt and Roman K. Thomas for support and fruitful discussions.

**FUNDING**

Felix Dietlein is supported by the Deutsche Krebshilfe through Mildred-Scheel-Doktorandenprogramm (Grant 110770).

**Conflict of Interest statement.** None declared.

**REFERENCES**


[Downloaded from https://academic.oup.com/hmg/article-abstract/23/6/1527/733856 by guest on 29 January 2019]


