The Landscape of Precision Cancer Combination Therapy: A Single-Cell Perspective

Saba Ahmadi[1,6*], Pattara Sukprasert[2,7*], Natalie Artzi[3,4], Samir Khuller[2,7], Alejandro A. Schäffer[5^], Eytan Ruppin[5^]

[1] Dept. of Computer Science, University of Maryland, College Park MD 20742 USA

[2] Dept. of Computer Science, Northwestern University, Evanston IL 60208 USA

[3] Dept. of Medicine, Engineering in Medicine Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02139 USA

[4] Broad Institute of Harvard and MIT, Cambridge, MA 02139 USA

[5] Cancer Data Science Laboratory, National Cancer Institute, Bethesda, MD 20892 USA

[6] Part of this research done while at Dept. of Computer Science, Northwestern University, Evanston IL 60208 USA

[7] Part of this research done while at Dept. Computer Science, University of Maryland, College Park MD 20742 USA

[*] Equally contributing first authors

[^] Equally contributing corresponding authors

Correspondence should be addressed to alejandro.schaffer@nih.gov and eytan.ruppin@nih.gov.

Physical address: Cancer Data Science Laboratory, National Cancer Institute, Bldg. 10, Room 1-5140 Bethesda, MD 20892 USA

Abstract

The availability of single-cell transcriptomics data opens up new opportunities for designing combination cancer treatments. Mining such data, we employed combinatorial optimization to explore the landscape of optimal combination therapies in solid tumors (including brain, head and neck, melanoma, lung, breast and colon cancers), assuming that each drug can target one of 1269 genes encoding cell surface receptors and deliver a toxin into the cell. We also identified optimal combinations among a subset of 58 genes for which targeted treatments have been already tested *in vitro*. We first study a `personalized' treatment objective, identifying optimal combinations for each patient, aimed at killing most tumor cells while sparing most non-tumor cells. We find that a single-digit number of targets is sufficient for killing at least 80% of the tumor cells while killing at most 10% of the non-tumor cells in each patient. However, with more stringent killing requirements, the number of targets required may rise sharply in some cancer types. We additionally study a 'fair' objective, identifying an optimal treatment basket for a multi-patient cohort while bounding the number of extra treatments given to each patient. Encouragingly, we find that optimally fair combinations usually require at most three extra treatments compared to the personalized combinations. Targets appearing in many optimal solutions include *PTPRZ1* (especially for brain cancer), *CLDN4*, *CXADR*, *EPHB4*, *NTRK2*, *EGFR*, *SLC2A*, *ERBB3*, *IL17RD and EDNRB*. This multi-disciplinary analysis provides the first systematic characterization of combinatorial targeted treatments for solid tumors and uncovers promising cell-surface targets for future development.

Introduction

Personalized oncology offers hope that each patient's cancer can be treated based on its genomic characteristics (Von Hoff et al. 2010; Schütte et al. 2017). Several trials have suggested that it is possible to collect genomics data fast enough to inform treatment decisions (e.g., Jameson et al., 2014; Saulnier Sholler et al. 2015; Byron et al. 2018). Meta-analysis of Phase I clinical trials completed during 2011-2013 showed that on the whole, trials that used molecular biomarker information to influence treatments gave better results than trials that did not (Schwaederle et al. 2016). However, most precision oncology treatments utilize only one or two treatments, and resistant clones frequently emerge, emphasizing the need to deliver personalized medicine as multiple treatments combined together (Arnedos et al. 2014; Schwaederle et al. 2016; Nikanjam et al., 2017; Rebollo et al. 2017; Sureda et al. 2018; Sicklick et al. 2019).

Here we design and implement a computational approach to identify *optimal precision combination treatments*. Using single-cell data from tumor cells and non-tumor cells of the same patient's tumor, we formulate and systematically answer two basic questions. First, how many targeted treatments are needed to selectively kill some fraction (ideally close to 1) of its tumor cells while sparing most of the non-tumor cells? Second, if we have a cohort of patients, how many distinct single-target treatments need to be prepared beforehand so that there is a combination that kills at least a specified proportion of the tumor cells of each patient?

We focus our analysis on genes encoding protein targets that are on the cell surface, as these may be precisely targeted by various technologies: e.g, by antibody or nanoparticle technologies, spanning immunotoxins ligated to antibodies (Bjorn et al 1985; Pastan et al. 1986) or ligated to mimicking peptides (Gray and Brown 2014) , conventional chemotherapy ligated to nanoparticles (Liu et al. 2017), degraders associated with ubiquitin E3 ligases (Fisher and Phillips 2018) and designed ankyrin repeat proteins (DARPins) (Plückthun 2015; Sokolova et al. 2019). These treatments are all "modular", including one part that specifically targets the tumor cell via one gene/protein and other parts that deliver the lethal toxin. Notably, several of these modular treatment technologies rely on receptor-mediated endocytosis (RME) via the receptor to enter the cell (Říhová 1998; Tortorella and Karagiannis 2014), but do not necessarily downregulate the target receptor itself.

As tumors typically have considerable intra-tumor heterogeneity (ITH) (Marusyk and Polyak 2010; McGranahan and Swanton 2015), each single modular agent is likely to select only a fraction of the tumor cells, raising the need to target a tumor with a cocktail of treatments using the same toxic mechanism, such that each tumor cell is covered by at least one treatment whose target is overexpressed. This is the classical "hitting set problem" in combinatorial algorithms (Karp et al. 1972), which is formally defined in the Methods (see also Supplemental Materials 1). Figure 1A shows a small schematic example in which there are alternative hitting sets of sizes two and three. One would prefer the hitting set of size two because the patients would need to receive only two distinct treatments rather than three treatments.

Results

The Data and the Combinatorial Optimization Analysis Framework

We focus our analysis on nine single cell RNAseq data sets from publicly available databases that include tumor cells and non-tumor cells from at least three patients (Methods; Table 1). Those datasets include four brain cancer datasets and one each from breast, melanoma, colon, lung and head and neck cancers. Most analyses were done for all datasets, but for clarity of exposition, we focus in the main text on four datasets from four different cancer types (brain, head and neck, melanoma, lung); the results on the other five datasets are provided in Supplemental Materials.

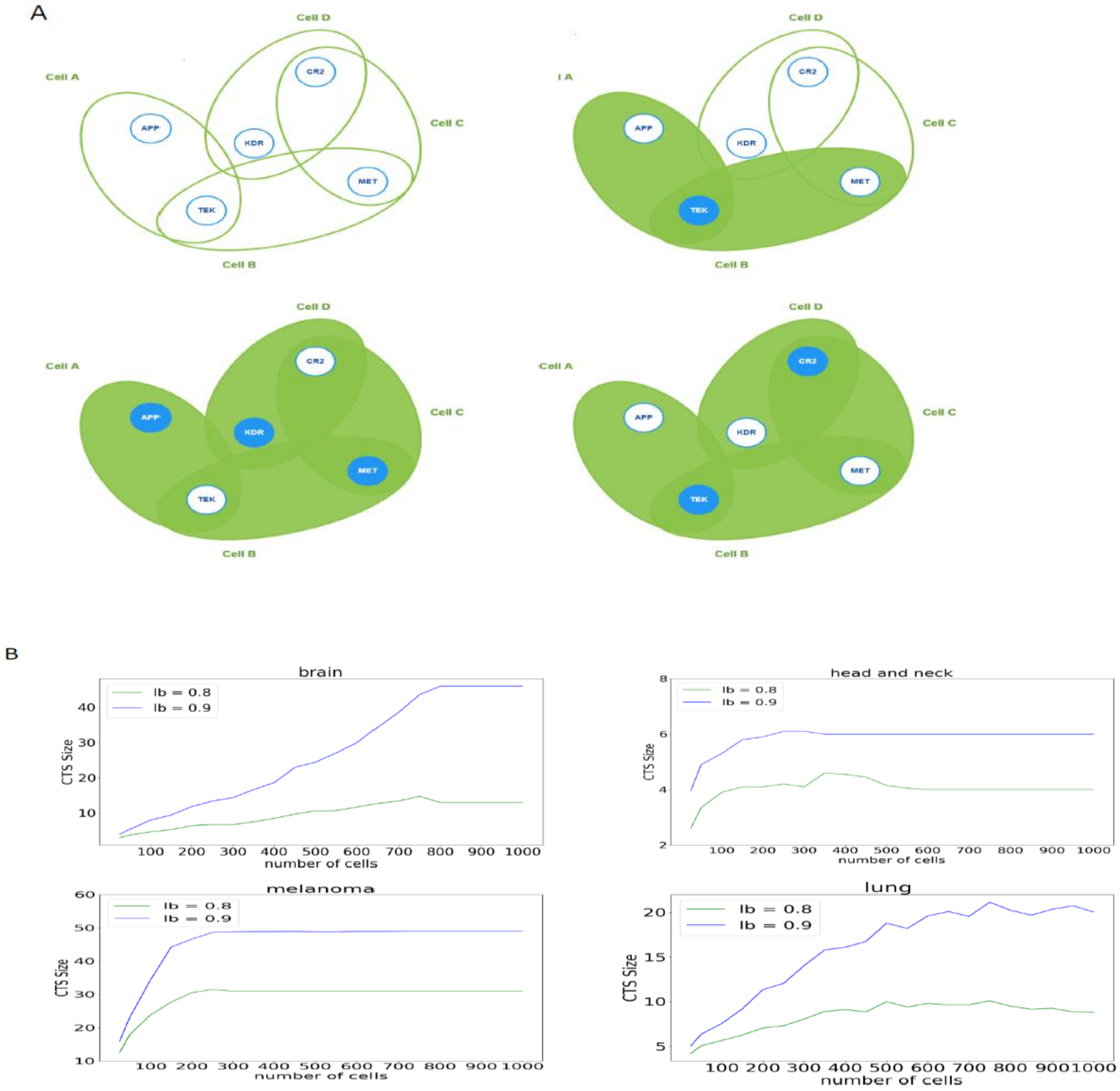

Figure 1. (A) A schematic small example of killing a multi-cell tumor. The tumor has four cells (A,B,C,D, portrayed by large green ellipses), which may express any of five cell-surface receptor genes (small ellipses) that may be targeted selectively by modular treatments (the blue ellipses). If one targets TEK, cells A and B will be killed (upper right). If one targets {APP, KDR, MET}, all cells will be killed (lower left). However, if, instead, one would target {CR2, TEK} then all cells will be killed with two targets instead of three (lower right), providing a better, optimal solution. (B) Cohort target size (CTS) (averaged over 20 subsamples) as a function of the number of cells sampled for four single-cell data sets. lb denotes the lower bound on the proportion of tumor cells to be killed; we depict this function for two lb values, our baseline value (0.8) and a more stringent one (0.9).

We focus on a few key research questions. How many targets are needed to kill most cells of a tumor and what is the tradeoff between cancer cells killed and non-cancer cells spared? To formalize these and other questions as combinatorial optimization hitting set problems, we define the following parameters and baseline values and explore how the answers vary as functions of these parameters: We specify a lower bound on the fraction of tumor cells that should be killed, lb, which ranges from 0 to 1. Similarly, we define an upper bound on the fraction of non-tumor cells killed, ub, which also ranges from 0 to 1. Our baseline settings are $lb = 0.8$ and $ub = 0.1$.To represent the observed non-linear relationship between the number of binding peptides on the drug-delivery molecule and the probability of entering the cell via RME (Martinez-Veracoechea and Frenkel 2011), we introduce an additional parameter r. The expression ratio r defines which cells are killed, as follows. Denote the mean expression of a gene g in *non-cancer* cells that have non-zero expression by $\mathrm{E}(g)$. A given cell is considered killed if gene g is targeted and its expression level in that cell is at least $r \times E(g)$. Higher values of r model more selective killing. In addition to enabling the study of treatment combinations across a broad range of treatment stringencies, having r as a modifiable parameter anticipates that in the future one could experimentally tune the overexpression level at which the treatment enters cells and kills it in a gene-specific manner (cf. Delaney et al. 2019). For most of our analyses, the expression ratio r is varied from 1.5 to 3.0, with a baseline of 2.0, based on experiments in the lab of N.A and related to combinatorial chemistry modeling (Martinez-Veracoechea and Frenkel 2011).

Given these definitions, we solve the following combinatorial optimization hitting set problem (see Methods for more details): Given an input of a single cell transcriptomics sample of non-tumor and tumor cells for each patient in a cohort of multiple patients, bounds ub and lb, ratio r, and a set of up to 1269 target genes (later we present results for a specific subset of 58 such genes), we seek to find three key outputs: **(a)** the size of the *global minimum-size hitting set* for all patients such that the subset of genes targeted *for each patient is also of minimum size*, subject to the constraints that in each patient, at least lb proportion of cancer cells are killed by the optimal hitting set and at most ub proportion of non-tumor cells are killed by the hitting set **(b)** the size of the combination of targets that 'hit' each patient, termed the minimum *individual hitting set (IHS) size*, and finally **(c)** the identities of the gene targets that compose the optimal global and individual target sets. In non-technical terms, the *global minimum-size hitting set (GHS)* denotes the set of all the genes targeted across all patients, which we term the *cohort target set (CTS)* and similarly the IHS denotes the *individual target set* (*ITS*). This optimum hitting set problem with constraints *can be solved to optimality using integer linear programming (ILP)* (Methods).

Of note, this formulation is "personalized" as *each patient receives the minimum possible number of treatments*. The global optimization comes into play when there are multiple solutions of the same size to treat a patient. For example, suppose we have two patients such that patient A could be treated by targeting either {*EGFR*, *FGFR2*} or {*MET*, *FGFR2*} and patient B could be treated by targeting either {*EGFR*, *CD44*} or {*ANPEP*, *CD44*}. Then we prefer the global hitting set {*EGFR*, *FGFR2*, *CD44*} of size 3 and we treat patient A by targeting {*EGFR*, *FGFR2*} and patient B by targeting {*EGFR*, *CD44*}.

As the number of cells per patient varies by three orders of magnitude across data sets, we use random sampling to obtain hitting set instances of comparable sizes (Methods). We report that sampling hundreds of cells from the tumor is sufficient to get enough data to represent all cells. In most of the experiments shown, the number of cells sampled, which we denote by c, was 250, which is a compromise between the low number of cells available in some data sets and sampling enough cells to represent intra-tumor heterogeneity. As shown in (Figure 1B and Supplemental Materials 1, Figures S1-S4), 250 cells are roughly sufficient for CTS size to plateau for our baseline parameter settings described in the next subsection for most data sets, while 500-750 cells is more appropriate for more extreme values on some data sets. For the colorectal cancer data set only, we sampled 100 cells since 250 cells are not available.

Cohort and Individual Target Set Sizes as Functions of Tumor and Non-Tumor Killing Goals

Given the single cell tumor datasets and the ILP optimization framework described above, we first studied how do the resulting optimal cohort target set (CTS) vary as a function of the parameters defining the optimization objectives. Figures 2 and S5-S9 in Supplemental Materials 3 show spline-interpolated landscapes of CTS sizes varying *ub*, *lb*, and r around the baseline values. We show separate plots for $r = 1.5, 2.0, 2.5, 3.0$. Individual values on the z axis are not necessarily integers because each value represents the mean of 20 replicates of sampling c (usually 250; for colorectal cancer (GSE81861) only, $c = 100$)) cells (Figure 1B). The CTS sizes for melanoma are largest due to the larger number of patients in that data set.

Encouragingly, for most data sets and parameter settings, we see that the optimal CTS sizes are in the single digits. However, in several data sets, we observe a sharp increase in CTS size for lb values above 0.8 and/or ub is set below 0.1, with a more pronounced effect of varying lb. This transition is more discernable at the lowest value of r (1.5), probably because when r is lower it becomes harder to find genes that are individually selective in killing tumor cells and sparing non-tumor cells. The observed transition in CTS sizes occurs robustly for a broad range of thresholds for filtering out low expressing cells when preprocessing the data (Supplemental Materials 4, Figures S10-S12).

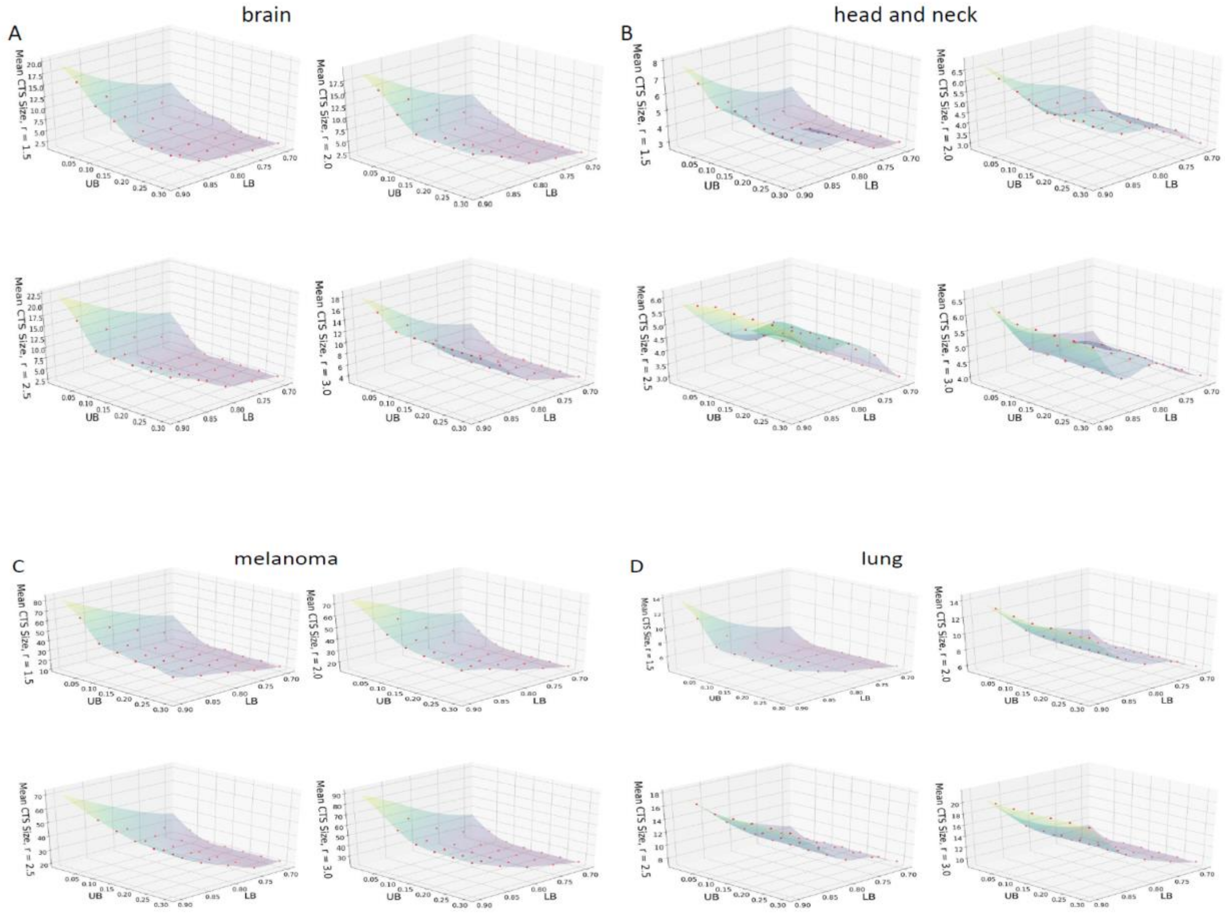

Figure 2. Surface plots showing how cohort target set size (CTS) varies as a function of lb, ub, r. For brain cancer (A) the mean values are in the range [2.1, 16.85]; for head and neck cancer (B) in [3,6.95]; for melanoma (C) in [12.8, 67.5]; for lung cancer (D) in [4.6, 20.5].

We then turned to examine what are the resulting *individual target set* (ITS) sizes that are obtained in the optimal combinations under the same conditions. In all data sets, the mean ITS sizes are in the single digits for most values of lb and ub. The distributions of ITS sizes are shown as bar graphs for four data sets and two combinations of (lb, ub) (Figure 3). Overall, the mean ITS sizes with the baseline parameter values ($r = 2.0, lb = 0.8, ub = 0.1$) range from 1.0 to 2.67 among the nine data sets (Supplemental Table S2), such that on average 3 targets per patient should suffice if enough single-target treatments are available in the cohort target set. However, there is considerable variability across patients. Evidently, as we make the treatment requirements more stringent (formally increasing lb from 0.8 to 0.9 and decreasing ub from 0.1 to 0.05), the variability in ITS size across patients becomes larger. Importantly, these bar plots provide rigorous quantifiable evidence grounding the prevailing observation that among tumors of the same type, some tumors are much harder to treat than others. Taken together, these results show that we can compute precise estimates of the number of targets needed for cohorts (tens) and individual patients (single digits usually) and that these estimates are sensitive to the killing stringency, especially when lb increases above 0.8. The landscape for more aggressive killing regimes, with values of lb up to 0.99 for the baseline $r = 2.0$ is portrayed in Figures S13-S14 in Supplemental Materials 5. For fixed $lb = 0.8, lb = 0.1$and varying r, smallest CTS sizes are typically obtained for r values close to 2.0, further motivating our choice of $r = 2$ as the default value (Supplemental Materials 6, Figures S15-S16, Table S1). Finally, we show that a ‘control’ heuristic algorithm searching for small and effective target combinations finds ITS sizes substantially larger than the optimal ITS sizes identified using our optimization algorithm. The greedy CTS size is greater than the ILP CTS size for eight out of nine data sets (Supplemental Materials 7 (Table S2), Methods).

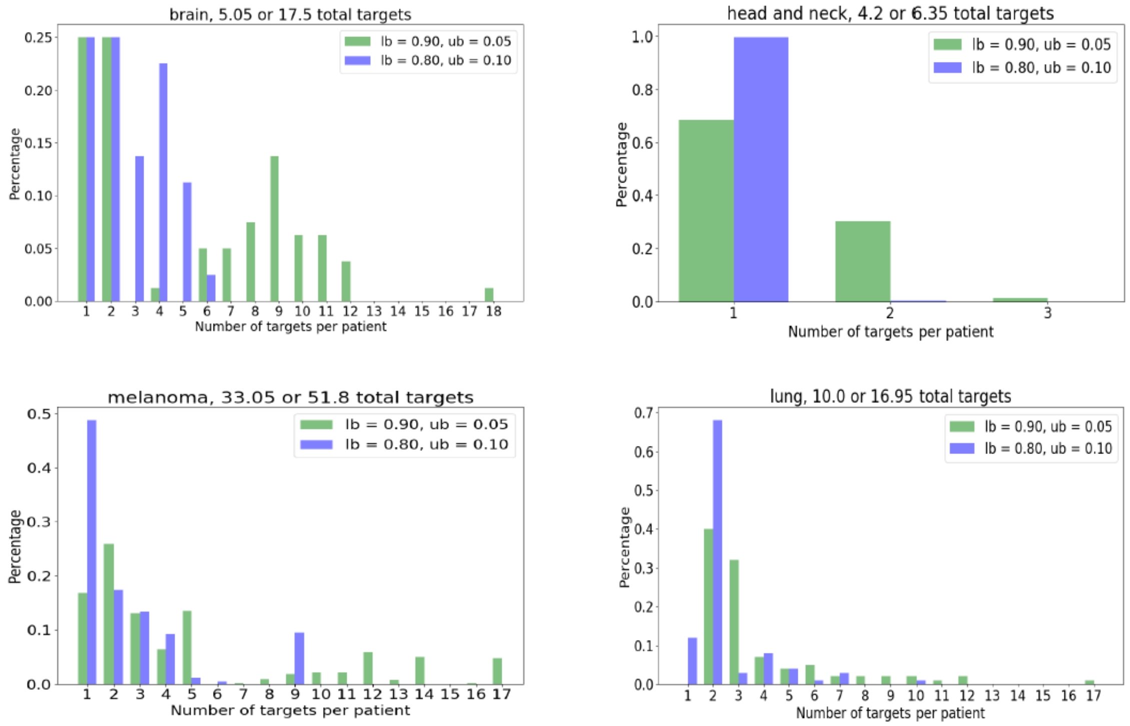

Figure 3. The distribution of ITS values across four different cancer types, at our baseline parameter setting (blue) and one more stringent parameter setting (green).

The Landscape of Combination Sizes Achievable with Currently Targetable Receptors

To get a more current view of combination treatments possible with targets that have already been validated, we conducted a literature search identifying 58 out of the 1269 genes for which ligand-mimicking peptides have been published and tested in *in vitro* models, usually cancer models (Methods; Tables 2 and 3). We asked whether we can find feasible optimal combinations in this case and if so, how do the optimal CTS sizes compare vs. those computed for all 1269 genes?

Computing the optimal CTS and ITS solutions for this basket of 58 targets, we found feasible solutions for six of the data sets across all parameter combinations we surveyed, but for three data sets, in numerous parameter combinations we could not find optimal solutions that satisfy the optimization constraints, especially for melanoma (Supplemental Materials 8). That is, the currently available targets do not allow one to design treatments that may achieve the specified selective killing objectives, underscoring the need to develop new targeted cancer therapies to make personalized medicine more effective for more patients. Overall, comparing the optimal solutions obtained with 58 targets to those we have obtained with the 1269 targets, three qualitatively different behaviors are observed (Figures S17-S20, Supplemental Materials 8): On some data sets, it is just a little bit more difficult to find optimal solutions with the 58-gene pool, while in others, the restriction to a smaller pool can be a severe constraint. In one data set (melanoma), the smaller basket of genes forces more patients to receive similar individual treatment sets and thereby reduces the size of the CTS. Unlike the CTS size, the ITS size must stay the same or increase when the pool of genes is reduced, by definition. Overall, the ITS sizes using the pool of 58 genes range from 1.16 to 4.0. Among the feasible instances (i.e., the cases we managed to solve), the average increases in the IHS sizes in the 58 genes space vs that of the 1269 case was quite moderate, ranging from 0.16 to 1.33.

Optimal Fairness-Based Combination Therapies for a Given Cohort of Patients

Until now we have adhered to a *patient-centered approach* that aims to find the minimum-size ITS for each patient, first and foremost. We now study a different, *cohort-centered approach*, where given a cohort of patients, we seek to minimize the total size of the overall CTS size, while allowing for some increase in the ITS sizes. The key question is how much larger are the resulting ITS sizes if we optimize for the cohort (CTS size) rather than the individuals (ITS size)? This scenario is an example of 'fairness' problems (Supplemental Materials 1), where we seek potentially sub-optimal individual solutions that are beneficial for the entire community from a social or economic perspective. Here, the potential benefit is economic since running a basket trial would be less expensive if one reduces the size of the basket of available treatments (Figure 4).

We formalized the 'fair CTS problem' by adding a parameter α as the limit on the excess number of targets selected for any individual patient, compared to the number selected in the individual-based approach (formally, the latter corresponds to setting $\alpha = 0$). We formulated and solved via ILP this fair hitting set problem for up to 1269 possible targets on all nine data sets (Methods). We fixed $r = 2$ and $ub = 0.1$ while varying α and lb. Figures S21-S26 in Supplemental Materials 9 show the optimal CTS and ITS sizes w.r.t. $\alpha = 0, \ldots, 5$.

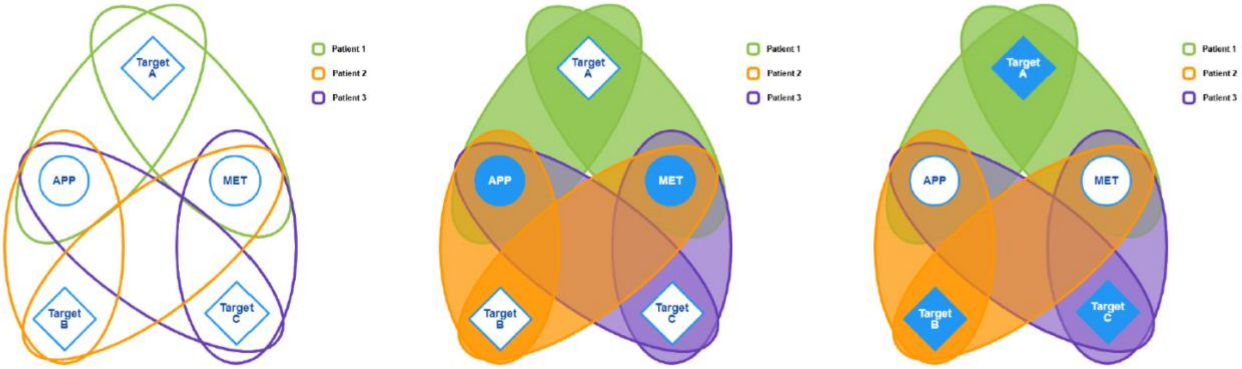

Figure 4. A schematic example showing the concept of *fairness*. Each patient has two cells (portrayed in a different color for each patient, left panel). Employing the individual-based optimizing objective, each patient could be treated by an IHS of size 1 by targeting the receptors A, B, C (middle panel), but this would result in an optimal CTS of size 3. Instead, employing a fairness-based optimization objective would result in a smaller CTS of size 2 ({APP, MET}, right panel) that could be used to treat each patient. This solution thus beneficially reduces the overall cohort target set size, but at the cost of increasing the size of each ITS from 1 to 2 (thus having an unfairness value $\alpha = 1$ because the worst difference among all patients is that a patient receives 1 more treatment than necessary).

For 8 out of 9 data sets, we encouragingly find that the unfairness *α is bounded by a constant of* 3, i.e, we cannot find smaller CTS by allowing α to increase above 3. For the largest data set (melanoma), the CTS is close to smallest possible at $\alpha = 3$. As we show in Supplemental Materials 9, empirically, even if one requires lower α values, then as those approach 0 the size of the fairness-based CTS grows fairly moderately and remains in the lower double digits, and the mean size of the number of treatments given to each patient (their ITS) is overall < 5. While, theoretically, we show that one can design instances for which α would need to be at least $\sqrt{n} - 1$to get a CTS (global hitting set) of size less than n (Supplemental Materials 9), in practice, fairness-based treatment strategies would be a reasonable and worthy economic option to consider in the future.

The Landscape of Gene Targets Appearing Frequently in Optimal Solutions

Examining the optimal combinations of targets that we identified across the different data sets can also give information regarding the cell surface receptors that merit more experimental effort in designing new targeted treatments. The results obtained for each dataset, at our baseline parameter setting, for $\alpha = 0{,}1{\ldots}{,}5$ are displayed in Tables S3-S5 in Supplemental Materials 10.

We further constructed co-occurrence networks in which edges connect targets that frequently co-occur in optimal solutions, analyzing the 58-gene and the 1269-gene solution spaces (Figure 5, Methods). Notably, with 1269 targets the resulting network is quite dense but for 58 targets, the resulting network is sparse where, remarkably, most of the frequent pairs include *EGFR* and one of the cadherin genes, *CDH1* and *CDH2*. More details can be found in Supplemental Materials 10.

Strikingly, one gene, *PTPRZ1* (protein tyrosine phosphatase receptor zeta 1), appears far more frequently than others, especially in the brain cancer data sets. This finding coincides with previous reports that *PTPRZ1* is overexpressed in glioblastoma (GBM) (Müller et al. 2003; Ulbricht et al. 2003). Notably, various cell line studies and mouse studies have shown that inhibiting PTPRZ1, for example by shRNAs, can slow glioblastoma tumor growth and migration (Ulbricht et al. 2006; Bourgonje et al. 2014). In the four brain cancer data sets, *PTPRZ1* is expressed selectively above the baseline $r = 2.0$ in 0.99 (GSE89567), 0.84 (GSE70630), 0.96 (GSE102130) and 0.27 (GSE84465) proportion of cells as an unweighted average of all the patients in each cohort.

Among the more widely studied genes, the oncogene *EGFR* stands out both because of its frequency in the 58-target network and via the gene/protein network tools provided along with the STRING database (Szklarczyk et al. 2019) to perform several types of gene set and pathway enrichment analyses. Among the 30 genes in the 1269-gene set most commonly in optimal solutions, there six kinases (out of 88 total human transmembrane kinases with a catalytic domain, p < 1e-6), namely {*EGFR, EPHB4, ERBB3, FGFR1, INSR, NTRK2*} and two phosphatases {*PTPRJ, PTPRZ1*}. The KEGG pathways most significantly enriched, all at FDR < 0.005, are (“proteoglycans in cancer”) represented by {*CD44, EGFR, ERBB3, FGFR1, PLAUR*}, (“adherens junction”) represented by {*EGFR, FGFR1, INSR, PTPRJ*}, and (“calcium signaling pathway") represented by {*EDNRB*, *EGFR*, *ERBB3*, *GRPR*, *P2RX6*}. The one gene in the intersection of all these pathways and functions is *EGFR*.

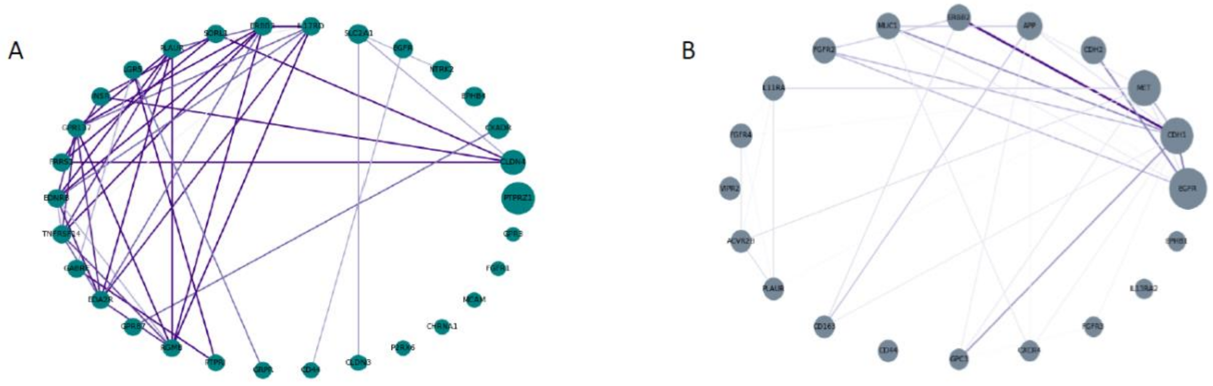

Figure 5. Network illustrations of the genes occurring most frequently (and their pairwise co-occurrence) in optimal target combinations. (A) The 30 most common target genes out of 1269 encoding cell-surface receptors and (B) the 20 most common target genes among the genes encoding 58 currently targetable receptors. There is an edge between two genes if they occur together in at least 70% (A) or 40% (B) respectively of the maximum possible. Node diameters reflect the frequencies of occurrence of each individual gene in optimal combinations. Edges are thicker and darker when the frequency of co-occurrence is higher.

Discussion

In this multi-disciplinary study, we harnessed techniques from combinatorial optimization to analyze publicly available single-cell tumor transcriptomics data to chart the landscape of future personalized combinations that are based on 'modular' therapies. We showed that a modest number of modular drugs that target different overexpressed receptors but kill cells with the same toxin may suffice to kill most cells of a solid tumor, while sparing most of the non-cancerous cells (Figures 2 and 3). Remarkably, we found that if one designs the optimal set of treatments for an entire cohort adopting a fairness-based policy, then the size of the projected treatment combinations for individual patients are not much larger than the optimal solutions that would have been assigned for these patients based on an individual-centric policy (Supplemental Materials 10).

We compared gene expression levels between non-cancer cells and cancer cells sampled from the same patient, which avoids inter-patient expression variability (Seoane et al. 2014).

However, we did little to account for “expression dropout” beyond the normalization performed by the providers of the data sets did, aiming to preserve the public data as it was submitted to GEO or Array Express. To achieve some uniformity, we only added a step to filter low expressing cells because some data sets had already been filtered in this way. We surmise that our results about the sizes of optimal TS should be viewed as *estimated upper bounds* that are likely to decrease if the dropout rate decreases or if cells expressing few genes are eliminated from the analysis more stringently. Finally, a more general future analysis should ideally include single cell transcriptomics data from normal, non-cancer cells, from all accessible tissues in the human atlas (not just the tumor microenvironment of a given individual), aiming to avoid targeting genes that are overexpressed in organs unrelated to the cancer site.

Even though the combinatorial optimization problems solved here are in the worst-case exponentially hard (NP-complete (Karp, 1972) in computer science terminology), the actual instances that arise from the single-cell data could be either formally solved to optimality or shown to be infeasible with modern optimization software. Of note, Delaney et al. (2019), have recently formalized a related optimization problem in analysis of single-cell clustered data for immunology. Their optimization problem is also NP-complete in the worst case but they could solve it for solution sets of up to size four using heuristic methods.

We additionally studied the precision combination landscape when the pool of targets consists of the 58 cell surface receptors that have proven ligand-mimicking peptides (Gray and Brown 2014; Liu et al. 2017) (Tables 2 and 3). The sizes of the optimal cohort treatment sets are not much larger and at times smaller than when considering all 1269 cell surface receptor genes. This finding reassuringly indicates that biochemists developing ligand-mimicking peptides have been astute about choosing ligands whose receptors are often overexpressed in tumors. However, we did find some data sets and instances for which optimal solution could not be found, which indicates that additional targeted treatments for solid tumors need to be developed. Functional enrichment analysis of genes commonly occurring in the optimal target sets reinforced the central role of the widely-studied oncogene *EGFR* and other transmembrane kinases, while also suggesting that the less-studied phosphatase *PTPRZ1* is a useful target, especially in brain cancer.

In summary, this multi-disciplinary study is the first to harness combinatorial optimization tools to analyze emerging single-cell data to portray the landscape of personalized combinations cancer medicine. Our findings uncover promising membranal targets for the development of future oncology drugs that may serve to optimize the treatment of cancer patient cohorts in several cancer types. The approach presented and the accompanying software, termed MadHitter, after the Mad Hatter from Alice in Wonderland (see Methods), can be readily applied to analyze additional cancer data sets as they become available.

Methods

Data Sets

We retrieved and organized data sets from NCBI's Gene Expression Omnibus (GEO, Clough and Barrett 2016) and Ensembl's ArrayExpress (Kolesnikov et al. 2015) and the Broad Institute's Single Cell Portal (https://portals.broadinstitute.org/single_cell). Nine data sets had sufficient tumor and non-tumor cells and were used in this study; an additional five data sets had sufficient tumor cells only and were used in testing early versions of MadHitter. Suitable data sets were identified by searching scRNASeqDB (Cao et al. 2017), CancerSea (Yuan et al. 2019), GEO, ArrayExpress, Google Scholar, and the 10x Genomics list of publications (https://www.10xgenomics.com/resources/publications/). We required that the data contains measurements of RNA expression on single cells from human primary solid tumors of at least two patients and the metadata are consistent with the primary data. We are grateful to several of the data depositing authors of data sets for resolving metadata inconsistencies by e-mail correspondence and by sending additional files not available at GEO or ArrayExpress. We excluded blood cancers and data sets with single patients. When it was easily possible to separate cancer cells from non-cancer cells of a similar type, we did so.

The main task in organizing each data set was to separate the cells from each sample or each patient into one or more single files. Representations of the expression as binary, as read counts, or as normalized quantities such as transcripts per million (TPM) were retained from the original data. When the data set included cell type assignments, we retained those to classify cells as "cancer" or "non-cancer", except in the data set of (Karaayvaz et al. 2018) where it was necessary to reapply filters described in the paper to exclude cells expressing few genes and to identify likely cancer and likely non-cancer cells. If cell types were not distinguished, all cells were treated as cancer cells. To achieve partial consistency in the genes included, we filtered data sets to include only those gene labels recognized as valid by the HUGO Gene Nomenclature Committee (http://genenames.org), but otherwise we retained whatever recognized genes that the data submitters chose to include. After filtering out the non-HUGO genes, but before reducing the set of genes to 1269 or 58, we filtered out cells as follows. Some data sets came with low expressing cells filtered out. To achieve some homogeneity, we filtered out any cells expressing fewer than 10% of all genes before we reduced the number of genes. In Supplemental Materials 4, we tested the robustness of this 10% threshold. Finally, we retained either all available genes from among our set of 1269 genes encoding cell-surface receptors.

Table 1. Summary descriptions of single-cell data sets from solid tumors used either for analysis (9) or preliminary testing (5 additional). Data sets are ordered so that those from the same or similar tumor types are on consecutive rows. The first 13 data sets were obtained either from GEO or the Broad Institute Single Cell Portal, but the GEO code is shown. The data set on the last row was obtained from ArrayExpress. In some data sets that have both cancer and non-cancer cells, there may be samples for which only one type or the other is provided. Hence, the numbers in parentheses in the third and fourth columns may differ. Data set GSE115978 (Jerby-Arnon et al. 2018) supersedes and partly subsumes GSE7256 (Tirosh et al. 2016b)

Data set code	Cancer type(s)	Cancer cells(samples)	Non-cancer cells (samples)	Clinical follow-up	Reference(s)
GSE75688	Breast	441(11)	--	Metastasis or not	Chung et al. 2017
GSE118389	Breast	804(6)	314(6)	Metastasis or not	Karaayvaz et al. 2018
GSE89567	brain(glioma)	5097(10)	1146(9)	No	Venteicher et al. 2017
GSE103224	brain(glioma)	23793(8)	--	No	Yuan et al. 2018
GSE70630	brain(glioma)	4044(6)	303(6)	No	Tirosh et al. 2016a

GSE57872	brain(glioma)	440(6)	--	No	Patel et al. 2014
GSE102130	brain(6 glioma and 3 glioblastoma)	2858(9)	94(5)	No	Filbin et al. 2018
GSE84465	brain(glioblastoma)	1091(4)	651(4)	No	Darmanis et al. 2017
GSE81861	colorectal	272(10)	160(6)	No	Li et al. 2017
GSE103322	head and neck	2093(13)	3197(15)	No	Puram et al 2017
GSE115978	melanoma	2018(23)	4334(32)	Yes, immuno-therapy	Jerby-Arnon et al. 2018; Tirosh et al. 2016b
GSE118828	ovarian	1415(11 primary) 973 (5 metastasis)	578(2)	No	Shih et al. 2018
GSE67980	prostate	124(21)	--	Metastasis or not	Miyamoto et al. 2015
E-MTAB-6149	lung	7351(5)	2730(5)	No	Lambrechts et al. 2018

Definition of the Minimum Hitting Set Problem and Feasibility

One of Karp's original NP-complete problems is called "hitting set" and is defined as follows (Karp, 1972). Let U be a finite universal set of elements. Let $S_1, S_2, S_3, \dots, S_k$ be subsets of U**.** Is there a small subset $H \subseteq U$ such that for $i = 1, 2, 3, \dots, k$**,** $S_i \cap H$ is non-empty. In our setting, U is the set of target genes and the subsets $S_i$are the single cells. In (Gainer-Dewar and Vera-Lincona 2017), numerous applications for hitting set and the closely related problems of subset cover and dominating set are described; in addition, practical algorithms for hitting set are compared on real and synthetic data.

Among the applications of hitting set and closely related NP-complete problems in biology and biochemistry are stability analysis of metabolic networks (Haedlicke and Klamt 2011; Haus et al. 2011; Jarrah et al. 2007; Klamt and Gilles 2004; Trinh et al. 2009), identification of critical paths in gene signaling and regulatory networks (Ideker 2000; Wang and Albert 2011; and Zvedei-Oancea et al. 2005) and selection of a set of drugs to treat cell lines

(Vazquez 2009; Mellor et al. 2010) or single patients (Vera-Lincona et al. 2013; Pang et al. 2014). More information about related work can be found in Supplemental Materials 1.

Two different difficulties arising in problems such as hitting set are that 1) an instance may be *infeasible* meaning that there does not exist a solution satisfying all constraints and 2) an instance may be *intractable* meaning that in the time available, one cannot either i) prove that the instance is infeasible or ii) find an optimal feasible solution. *All instances of minimum hitting set that we considered were tractable* on the NIH Biowulf system. Many instances were provably infeasible; in almost all cases. we did not plot the infeasible parameter combinations. However, in Figure 3, the instance for the melanoma data set with the more stringent parameters was infeasible because of only one patient sample, so we omitted that patient for both parameter settings in Figure 3.

Lists of Target Genes

We are interested in the set of genes G that i) have the encoded protein expressed on the cell surface and ii) for which some biochemistry lab has found a small peptide (i.e. amino acid sequences of 5-30 amino acids) that can attach itself to the target protein and get inside the cell carrying a tiny cargo of a toxic drug that will kill the cell and iii) encode proteins that are receptors. The third condition is needed because many proteins that reside on the cell surface are not receptors that can undergo RME. The first condition can be reliably tested using a recently published list of 2799 genes encoding human predicted cell surface proteins (Bausch-Fluck et al. 2018). For the second condition, we found two review articles in the chemistry literature (Gray and Brown 2014; Liu et al. 2017) that effectively meet this condition. Intersecting the lists meeting each condition currently gives us 38 genes/proteins that could be targeted (Table 2).

We expect that the true list is larger due to more of the small mimicking peptides being developed recently for a variety of applications. Importantly, many of the genes/proteins on our list are overexpressed in various types of cancer and in some cases are targeted by gene-targeted chemotherapeutic drugs, which may nevertheless have difficulty penetrating the tumor and may cause side effects or lead to drug resistance. To enforce the third condition that the gene encodes a receptor, we took the union of the 38 genes in Table 2 and any genes meeting condition i) such that the work "receptor" is in the gene name. This union yields a list of 1269 genes that we consider as candidates to target. However, most of the data sets listed in Table 1 had expression data on 1200-1220 of these genes because the list of 1269 includes many olfactory receptor genes that may be omitted from standard genome-wide expression experiments. Among the 38 genes in Table 2, 13/14 data sets have all 38 genes, but GSE57872 was substantially filtered and has only 10/38 genes; since GSE57872 lacks non-tumor cells, we did not use this data set in any analyses shown.

Because the latter review (Liu et al. 2017) was published in 2017, we expected that there are now additional genes for which ligand-mimicking peptides are known. We found 20 additional genes and those are listed in Table 3. Thus, our hitting set analyses restricted to genes with known ligand-mimicking peptides use $58 = 38 + 20$ targets.

Table 2. Single proteins that can be targeted by peptides based on (Gray and Brown 2014; Liu et al. 2017) and are expressed on the cell surface (Bausch-Fluck et al. 2018). For easier correspondence with the gene expression data, the entries are listed in alphabetical order by gene symbol. In this table, we follow the clinical genetics formatting convention that proteins are in Roman and gene symbols are in *italics*.

Protein	Gene Symbol
APN/CD13	*ANPEP*
APP	*APP*
PD-L1	*CD274*
CD44	*CD44*
P32/gC1qR	*CD93*
E-cadherin	*CDH1*
N-cadherin	*CDH2*
CD21	*CR2*
EGFR	*EGFR*
Epha2	*EPHA2*
EphB4	*EPHB4*
HER2	*ERBB2*
FGFR1	*FGFR1*
FGFR2	*FGFR2*
FGFR3	*FGFR3*
FGFR4	*FGFR4*

VEGFR1	*FLT1*
VEGFR3	*FLT4*
PSMA	*FOLH1*
GPC3	*GPC3*
IL-10RA	*IL10RA*
IL-11Rα	*IL11RA*
IL-13Rα2	*IL13RA2*
IL-6Rα	*IL6R*
GP130	*IL6ST*
VEGFR2	*KDR*
MUC18	*MCAM*
Met	*MET*
MMP9	*MMP9*
Thomsen-Friedenreich carbohydrate antigen	*MUC1*
NRP-1	*NRP1*
PDGFRβ	*PDGFRB*
CD133	*PROM1*
PTPRJ	*PTPRJ*
HSPG	*SDC2*
E-selectin	*SELE*
Tie2	*TEK*
VPAC1	*VIPR1*

Table 3. Single proteins that can be targeted by ligand-mimicking peptides but are not included in the two principal reviews that we consulted (Gray and Brown 2014; Liu et al. 2017) and are among 1269 cell surface receptors (Bausch-Fluck et al. 2018). Since the evidence that these 20 genes have ligand-mimicking peptides is scattered in the literature, we include at least one PubMed ID of a paper describing a suitable peptide.

Protein	Gene Symbol	At Least One PubMed ID
ActRIIB	*ACVR2B*	28955765
CD163	*CD163*	27563889
CXCR4	*CXCR4*	19482312, 22523575
ephrin A4	*EPHA4*	15681844, 22523575
ephrin B1	*EPHB1*	15722342, 22523575

ephrin B2	*EPHB2*	15722342, 22523575
ephrin B3	*EPHB3*	15722342, 22523575
gonadotrophin releasing hormone receptor	*GNRHR*	20814857, 22523575
G Protein coupled receptor 55	*GPR55*	28029647
bombesin receptor 2	*GRPR*	20814857, 22523575
IL4 receptor	*IL4R*	19012727
low density lipoprotein receptor	*LDLR*	27656777
leptin receptor	*LEPR*	19233229, 26265355
LRP1	*LRP1*	29090274
melanocortin 1 receptor	*MC1R*	22964391
melanocortin 4 receptor	*MC4R*	17591746
CD206	*MRC1*	30768279
urokinase plasminogen activator receptor	*PLAUR*	25080049
neurokinin-1 receptor	*TACR1*	29498264
VPAC2	*VIPR2*	30077368

Basic Hitting Set Formulation

Given a collection $S = \{S_1, S_2, S_3, \ldots\}$ of subsets of a set U, the hitting set problem is to find the smallest subset $H \subseteq U$that intersects every set in S. The hitting set problem is equivalent to the set cover problem and hence is NP-complete. The following ILP formulates the hitting set problem:

$$min \sum_{g \in U} x(g)$$

$$\textstyle\sum_{g \in C} \; x(g) \; \geq 1 \quad \forall C \in S \quad (1)$$

In this formulation, there is a binary variable $x(g)$for each element $g \in U$that denotes whether the element g is selected or not. Constraint (**1)** makes sure that from each set $S_i$in S, at least one element is selected.

For any data set of tumor cells, we begin with the model that we specify a set of genes that can be targeted, and that is U. Each cell is represented by the subset of genes in Uwhose expression is greater than zero. In biological terms, a cell is killed (hit) if it expresses at any level on one of the genes that is selected to be a target (i.e., in the optimal hitting set) in the treatment. In this initial formulation, all tumor cells are combined as if they come from one patient because we model that the treatment goal is to kill (hit) all tumor cells (all subsets). In a later subsection, we consider a fair version of this problem, taking into account that each patient is part of a cohort. Before that, we model the oncologist's intuition that we want to target genes that are overexpressed in the tumor.

Combining Data on Tumor Cells and Non-Tumor Cells

To make the hitting set formulation more realistic, we would likely model that a cell (set) is killed (hit) only if one of its targets is overexpressed compared to typical expression in non-cancer cells. Such modeling can be applied in the nine single-cell data sets that have data on non-cancer cells to reflect the principle that we would like the treatment to kill the tumor cells and spare the non-tumor cells.

Let NT be the set of non-tumor cells. For each gene g, define its average expression $E(g)$ as the arithmetic mean among all the non-zero values of the expression level of $g$and cells in NT. The zeroes are ignored because many of these likely represent dropouts in the expression measurement. Following the design of experiments in the lab of N. A., we define an expression ratio threshold factor r whose baseline value is 2.0. We adjust the formulation of the previous subsection, so that the set representing a cell (in the tumor cell set) contains only those genes g such that the expression of g is greater than $r \times E(g)$instead of greater than zero. We keep the objective function limited to the tumor cells, but we also store a set to represent each non-tumor cell, and we tabulate which non-tumor cells (sets) would be killed (hit) because for at least one of the genes in the optimal hitting set, the expression of that gene in that non-tumor cell exceeds the threshold $r \times E(g)$. We add parameters $lb$and ub each in the range [0,1] and representing respectively a lower bound on the proportion of tumor cells killed and an upper bound on the proportion of non-tumor cells killed. The parameters lb, ub are used only in two constraints, and we do not favor optimal solutions that kill more tumor cells or fewer non-tumor cells, so long as the solutions obey the constraints.

Sampling Process to Generate Replicates of Data Sets

As shown in Table 1, the number of cells available in the different single-cell data sets varies by three orders of magnitude; to enable us to compare the findings across different datasets and cancer types on more equal footing, we employed sampling from the larger sets to reduce this difference to one order of magnitude. This goes along with the data collection process in the real world as we might get measurements from different samples at different times. Suppose for a data set we have n genes, m_t tumor cells and m_n non-tumor cells (which is zero in 5 of the 14 data sets). We might want to come up with a data set of $m_t' + m_n' \quad \leq m'$ tumor cells. We select a set of random m' cells uniformly from a set of all cells. Then we extract tumor cells m_t' and non-tumor cells $m_n'$from these m' cells. For each desired data set size, we might repeat this process many times so as to minimize the chance of getting a very noisy data set (e.g., a data set with very small number of tumor cells). We call each generated data set of a certain size a *replicate*.

Fair Hitting Set for a Multi-Patient Cohort

We want to formulate an integer linear program that selects a set of genes S^* from available genes in such a way that, for each patient, there exists a hitting set $H_i^{S^*} \subseteq S^*$of a relative small size (compared to the optimal hitting set of that patient alone which is denoted by $H(i)$).

Let $U = \{g_1, g_2, ..., g_{|U|}\}$ be the set of genes. There are $\boldsymbol{n}$ patients. For the i^{th} patient, we denote by $S_{P(i)}$, the set of tumor cells related to patient *i*. For each tumor cell $C \in S_{P(i)}$, we describe it as a set of genes which is known to be targetable to cell C. That is, $g \in C$ if and only if a drug containing g can target the cell C.

In the ILP, there is a variable $x(g)$ corresponding to each gene $g \in U$ that shows whether the gene g is selected or not. There is a variable $x(g, P(i))$ which shows whether a gene g is selected in the hitting set of patient $P(i)$. The objective function is to minimize the total number of genes selected, subject to having a hitting set of size at most $H(i) + \alpha$ for patient *P(i)* where $1 \leq i \leq n$.

Constraint **(3)** ensures that, for patient $P(i)$,we do not select any gene g that are not selected in the global set.

Constraint **(4)** ensures all the sets corresponding to tumor cells of patient $P(i)$ are hit.

$$min \sum_{g \in U} \ x(g) \qquad (1)$$

$$\sum_{g \in S_{P(i)}} \ x(g, P(i)) \ \leq H(i) + \alpha \quad \forall i \quad (2)$$

$$x(g, P(i)) \ \leq \ x(g) \quad \forall i \forall g \ \in U \qquad (3)$$

$$\sum_{g \in C} \ x(g, P(i)) \ \geq 1 \quad \forall i \forall C \ \in S_{P(i)} \qquad (4)$$

Parameterization of Fair Hitting Set

In the Fair Hitting Set ILP shown above, we give more preference towards minimizing number of genes needed in the global hitting set. However, we do not take into account the number of non-tumor cells killed. Killing (covering) too many non-tumor cells potentially hurts patients. In order to avoid that, we add an additional constraint to both the ILP for the local instances and the global instance. Intuitively, for patient $P(i)$, given an upper bound of the portion of the non-tumor cell killed UB, we want to find the smallest hitting set $H(i)$ with the following properties:

1. $H(i)$ covers all the tumor cells.
2. $H(i)$ covers at most $UB * |NT_{P(i)}|$ where $NT_{P(i)}$ is the set of non-tumor cells known for patient $P(i)$.

The ILP can be formulated as follows:

$$min \sum_{g \in U} \ x(g) \qquad (1)$$

$$\sum_{g \in C} \ x(g) \ \geq 1 \quad \forall C \ \in S_{P(i)} \qquad (2)$$

$$y(C) \ \geq \max_{g \in C} x(g) \quad \forall C \ \in NT_{P(i)} \qquad (3)$$

$$\sum_C \ y(C) \ \leq UB * |NT_{P(i)}| \ \ \forall C \ \in NT_{P(i)} \qquad (4)$$

With this formulation, the existence of a feasible solution is not guaranteed. However, covering all tumor cells might not always be necessary either. This statement can be justified as (1) measuring data is not always accurate, and some tumor cells could be missing and (2) in some cases, it might be possible to handle uncovered tumor cells using different methods. Hence, we add another parameter LB to let us model this scenario. In the high-level, this is the ratio of the tumor cells we want to cover. The ILP can be formulated as follows:

$$min \ \sum_{g \in U} \ x(g) \qquad (1)$$

$$\sum_C \ y(C) \ \geq LB * |S_{P(i)}| \ \ \forall C \ \in S_{P(i)} \qquad (2)$$

$$y(C) \ \geq \max_{g \in C} x(g) \ \ \forall C \ \in S_{P(i)} \cup NT_{P(i)} \qquad (3)$$

$$\sum_C \ y(C) \ \leq UB * |NT_{P(i)}| \ \ \forall C \ \in NT_{P(i)} \qquad (4)$$

Notice that the constraint (2) here is different from the one above as we only care about the total number of tumor cells covered.

Even with both UB and LB, the feasibility of the ILP is still not guaranteed. However, modeling the ILP in this way allows us to parameterize the ILP for various other scenarios of interest. While the two ILPs above are designed for one patient, one can extend these ILPs for multi-patient cohort.

$$min \ \sum_{g \in U} \ x(g) \qquad (1)$$

$$\sum_{g \in C} \ x(g, P(i)) \ \leq H(i) + \alpha \quad \forall i \forall C \in S_{P(i)} \quad (2)$$

$$x(g, P(i)) \ \leq \ x(g) \quad \forall i, g \ \in U \qquad (3)$$

$$y(C, P_{P(i)}) \geq \max_{g \ \in C} x(g, P(i)) \ \forall i \forall C \ \in S_{P(i)} \quad (4)$$

$$(4) \sum_C \ y(C, P_{P(i)}) \geq LB * |S_{P(i)}| \ \forall i \quad (5)$$

$$\sum_C \ y(C, P_{P(i)}) \leq UB * |NT_{P(i)}| \ \forall i \quad (6)$$

Implementation Note and Software Availability

We implemented in Python 3 the above fair hitting set formulations, with the expression ratio r as an option when non-tumor cells are available. The parameters α, LB, UB can be set by the user in the command line. To solve the ILPs to optimality we used the SCIP library and its Python interface (Achterberg 2009). The software package is called MadHitter.

The software is available on GitHub at https://github.com/ruppinlab/madhitter

Acknowledgements

This research is supported in part by the Intramural Research program of the National Institutes of Health, National Cancer Institute. This research is supported in part by the University of Maryland Year of Data Science Program and in part by start-up funds from Northwestern University to support the research of S.K. This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov). Thanks to E. Michael Gertz for technical assistance with SCIP and Biowulf. Thanks to Allon Wagner, Keren Yizhak, and Sushant Patkar for assistance in identifying and retrieving suitable single-cell RNAseq data sets. Thanks to Leandro Hermida and Sanju Sinha for technical advice.

References

Achterberg T. SCIP: Solving constraint integer programs. Mathematical Programming Computation 2009; 1(1), 1-41.

Arnedos M, Vielh P, Soria JC, Andre F. The genetic complexity of common cancers and the promise of personalized medicine: is there any hope? Journal of Pathology 2014; 232(2): 274-282.

Bausch-Fluck D, Goldmann U, Müller S, van Oostrum M, Müller M, Schubert OT, Wollscheid B. The in silico human surfaceome. Proceedings National Academy of Sciences Sci USA. 2018; 115: E10988-E10997.

Bjorn MJ, Ring D, Frankel A. Evaluation of monoclonal antibodies for the development of breast cancer immunotoxins. Cancer Research 1985; 45(3): 1214-1221.

Bourgonje AM, Navis AC, Schepens JT, Verrijp K, Hovestad L, Hilhorst R, Harroch S, Wesseling P, Leenders WP, Hendriks WJ. Intracellular and extracellular domains of protein tyrosine phosphatase PTPRZ-B differentially regulate glioma cell growth and motility. Oncotarget 2014; 5(18):8690-8702.

Byron SA, Tran NL, Halperin RF, Phillips JJ, Kuhn JG, de Groot JF, Colman H, Ligon KL, Wen PY, Cloughesy TF, Mellinghoff IK, Butowski NA, Taylor JW, Clarke JL, Chang SM, Berger MS, Molinaro AM, Maggiora GM, Peng S, Nasser S, Liang WS, Trent JM, Berens ME, Carpten JD, Craig DW, Prados MD. Prospective feasibility trial for genomics-informed treatment in recurrent and progressive glioblastoma. Clinical Cancer Research 2018; 24(2):295-305.

Cao Y, Zhu J, Jia P, Zhao Z. scRNASeqDB: A database for RNA-seq based gene expression profiles in human single cells. Genes 2017; 8(12):368.

Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, Kan Z, Han W, Park WY. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nature Communications 2017; 8:15081.

Clough E, Barrett T. The Gene Expression Omnibus Database. Methods of Molecular Biology 2016; 1418:93-110.

Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, Samghababi P, Zhang Y, Neff N, Kowarsky M, Caneda C, Li G, Chang SD, Connolly ID, Li Y, Barres BA, Gephart MH, Quake SR. Single-cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Reports 2017; 21(5):1399-1410.

Delaney C, Schnell A, Cammarata LV, Yao-Smith A, Regev A, Kuchroo VK, Singer M. Combinatorial prediction of marker panels from single-cell transcriptomic data. Molecular Systems Biology 2019; 15(10):e9005.

Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, Neftel C, Frank N, Pelton K, Hebert CM, Haberler C, Yizhak K, Gojo J, Egervari K, Mount C, van Galen P, Bonal DM, Nguyen QD, Beck A, Sinai C, Czech T, Dorfer C, Goumnerova L, Lavarino C, Carcaboso AM, Mora J, Mylvaganam R, Luo CC, Peyrl A, Popović M, Azizi A, Batchelor TT, Frosch MP, Martinez-Lage M, Kieran MW, Bandopadhayay P, Beroukhim R, Fritsch G, Getz G, Rozenblatt-Rosen O, Wucherpfennig KW, Louis DN, Monje M, Slavc I, Ligon KL, Golub TR, Regev A, Bernstein BE, Suvà ML. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science 2018; 360(6386):331-335.

Fisher SL, Phillips AJ. Targeted protein degradation and the enzymology of degraders. Current Opinion in Chemical Biology 2018; 44:47-55.

Gainer-Dewar A., Vera-Lincona P. The minimal hitting set generation problem: Algorithms and computation. SIAM Journal on Discrete Mathematics 2017; 31:63-100.

Gray BP, Brown KC. Combinatorial peptide libraries: mining for cell-binding peptides. Chemical Reviews 2014; 114(2):1020-1081.

Haedlicke O, Klamt S. Computing complex metabolic intervention strategies using constrained minimal cut sets. Metabolic Engineering 2011; 13:204-213.

Ideker T. Discovery of regulatory interactions through perturbation: inference and experimental design. Pacific Symposium on Biocomputing 2000; 5:302-313.

Jameson GS, Petricoin EF, Sachdev J, Liotta LA, Loesch DM, Anthony SP, Chadha MK, Wulfkuhle JD, Gallagher RI, Reeder KA, Pierobon M, Fulk MR, Cantafio NA, Dunetz B, Mikrut WD, Von Hoff DD, Robert NJ. A pilot study utilizing multi-omic molecular profiling to find potential targets and select individualized treatments for patients with previously treated metastatic breast cancer. Breast Cancer Research and Treatment 2014; 147(3):579-588.

Jarrah AS, Laubenbacher R, Stigler B, Stillman M. Reverse-engineering of polynomial dynamical systems. Advances in Applied Mathematics 2007; 39:477-489.

Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, Leeson R, Kanodia A, Mei S, Lin JR, Wang S, Rabasha B, Liu D, Zhang G, Margolais C, Ashenberg O, Ott PA, Buchbinder EI, Haq R, Hodi FS, Boland GM, Sullivan RJ, Frederick DT, Miao B, Moll T, Flaherty KT, Herlyn M, Jenkins RW, Thummalapalli R, Kowalczyk MS, Cañadas I, Schilling B, Cartwright ANR, Luoma AM, Malu S, Hwu P, Bernatchez C, Forget MA, Barbie DA, Shalek AK, Tirosh I, Sorger PK, Wucherpfennig K, Van Allen EM, Schadendorf D, Johnson BE, Rotem A, Rozenblatt-Rosen O, Garraway LA, Yoon CH, Izar B, Regev A. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell 2018; 175(4):984-997.e24

Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nature Communications 2018 9(1):3588.

Karp R. M. Reducibility among combinatorial problems. In Complexity of Computer Computations, Plenum Press, New York, 1972, pp. 85-103.

Klamt S., Gilles ED. Minimal cut sets in biochemical reaction networks. Bioinformatics 2004; 20:226-234.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. ArrayExpress update--simplifying data submissions. Nucleic Acids Research 2015; 43(Database Issue): D1113-D1116.

Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, Bassez A, Decaluwé H, Pircher A, Van den Eynde K, Weynand B, Verbeken E, De Leyn P, Liston A, Vansteenkiste J, Carmeliet P, Aerts S, Thienpont B. Phenotype molding of stromal cells in the lung tumor microenvironment. Nature Medicine 2018; 24(8):1277-1289.

Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, Wong M, Choi PJ, Wee LJK, Hillmer AM, Tan IB, Robson P, Prabhakar S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature Genetics 2017; 49(5):708-718.

Liu R, Li X, Xiao W, Lam KS. Tumor-targeting peptides from combinatorial libraries. Advances in Drug Delivery Reviews 2017; 110-111:13-37.

Martinez-Veracoechea, FJ and Frenkel, D. Designing super selectivity in multivalent nano-particle binding. Proceedings of the National Academy of Sciences USA 2011; 108, 10963-10968.

Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. Biochimica et Biophysica Acta 2010; 1805(1): 105-117.

McGranahan N, and Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell 2015; 27(1):15-26.

Mellor D, Prieto E, Mathieson L, Moscato P. A kernelisation approach for multiple d-hitting set and its application in optimal multi-drug therapeutic combinations. PLoS ONE 2010; 5(10):e13055.

Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, Desai R, Fox DB, Brannigan BW, Trautwein J, Arora KS, Desai N, Dahl DM, Sequist LV, Smith MR, Kapur R, Wu CL, Shioda T, Ramaswamy S, Ting DT, Toner M, Maheswaran S, Haber DA. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science 2015; 349:1351-1356.

Müller S, Kunkel P, Lamszus K, Ulbricht U, Lorente GA, Nelson AM, von Schack D, Chin DJ, Lohr SC, Westphal M, Melcher T. A role for receptor tyrosine phosphatase zeta in glioma cell migration. Oncogene 2003; 22(43):661-668.

Nikanjam M, Liu S, Yang J, Kurzrock R. Dosing Three-Drug Combinations That Include Targeted Anti-Cancer Agents: Analysis of 37,763 Patients. Oncologist. 2017 22(5):576-584.

Pang K, Wan YW, Choi WT, Donehower LA, Sun JC, Pant D, Liu ZD. Combinatorial therapy discovery using mixed integer linear programming. Bioinformatics 2014; 30:1456-1463.

Pastan I, Willingham MC, FitzGerald DJP. Immunotoxins. Cell 1986; 47(5): 641-648.

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014; 344(6190):1396-1401.

Plückthun A. Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy. Annual Review of Pharmacology and Toxicology 2015; 55:489-511.

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, Deschler DG, Varvares MA, Mylvaganam R, Rozenblatt-Rosen O, Rocco JW, Faquin WC, Lin DT, Regev A, Bernstein BE. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell 2017; 171(7): 1611-1624.e24.

Rebollo J, Sureda M, Martinez EM, Fernández-Morejón FJ, Farré J, Muñoz V, Fernández-Latorre F, Manzano RG, Brugarolas A. Gene expression profiling of tumors from heavily pretreated patients with metastatic cancer for the selection of therapy: A pilot study. American Journal of Clinical Oncology 2017; 40(2):140-145.

Saulnier Sholler GL, Bond JP, Bergendahl G, Dutta A, Dragon J, Neville K, Ferguson W, Roberts W, Eslin D, Kraveka J, Kaplan J, Mitchell D, Parikh N, Merchant M, Ashikaga T, Hanna G, Lescault PJ, Siniard A, Corneveaux J, Huentelman M, Trent J. Feasibility of implementing molecular-guided therapy for the treatment of patients with relapsed or refractory neuroblastoma. Cancer Medicine 2015; 4(6):871-86.

Schwaederle M, Zhao M, Lee JJ, Lazar V, Leyland-Jones B, Schilsky RL, Mendelsohn J, Kurzrock R. Association of biomarker-based treatment strategies with response rates and progression-free survival in refractory malignant neoplasms: A meta-analysis. JAMA Oncology 2016; 2(11):1452-1459.

Seoane J, De Mattos-Arruda L. The challenge of intratumour heterogeneity in precision medicine. Journal of Internal Medicine 2014; 276(1):41-51.

Shih AJ, Menzin A, Whyte J, Lovecchio J, Liew A, Khalili H, Bhuiya T, Gregersen PK, Lee AT. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. PLoS One 2018; 13(11): e0208778.

Sicklick JK, Kato S, Okamura R, Schwaederle M, Hahn ME, Williams CB, De P, Krie A, Piccioni DE, Miller VA, Ross JS, Benson A, Webster J, Stephens PJ, Lee JJ, Fanta PT, Lippman SM, Leyland-Jones B, Kurzrock R. Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. Nature Medicine 2019; 25(5):744-750.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research 2019; 47(D1): D607-D613.

Sokolova EA, Shilova ON, Kiseleva DV, Schulga AA, Balalaeva IV, Deyev SM. HER2-specific targeted toxin DARPin-LoPE: Immunogenicity and antitumor effect on intraperitoneal ovarian cancer xenograft model. Intrnational Journal of Molecular Sciences 2019 20(10). pii: E2399.

Sureda M, Rebollo J, Martínez-Navarro EM, Fernández-Morejón FJ, Farré J, Muñoz V, Bretcha-Boix P, Duarte M, Manzano RG, Crespo A, Del Carmen Redal M, Valenzuela B, Brugarolas A. Determining personalized treatment by gene expression profiling in metastatic breast carcinoma patients: a pilot study. Clinical and Translational Oncology 2018; 20(6):785-793.

Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, Neftel C, Desai N, Nyman J, Izar B, Luo CC, Francis JM, Patel AA, Onozato ML, Riggi N, Livak KJ, Gennert D, Satija R, Nahed BV, Curry WT, Martuza RL, Mylvaganam R, Iafrate AJ, Frosch MP, Golub TR, Rivera MN, Getz G, Rozenblatt-Rosen O, Cahill DP, Monje M, Bernstein BE, Louis DN, Regev A, Suvà ML. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016a; 539(7628):309-313.

Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin JR, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CG, Kazer SW, Gaillard A, Kolb KE, Villani AC, Johannessen CM, Andreev AY, Van Allen EM, Bertagnolli M, Sorger PK, Sullivan RJ, Flaherty KT, Frederick DT, Jané-Valbuena J, Yoon CH, Rozenblatt-Rosen O, Shalek AK, Regev A, Garraway LA. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016b; 352(6282):189-196.

Tortorella S, Karagiannis TC. Transferrin receptor-mediated endocytosis: a useful target for cancer therapy. Journal of Membrane Biology 2014; 247(4):291-307.

Trinh CT, Wlaschin A, Srienc F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Applied Microbiology and Biotechnology 2009; 81:813-826.

Ulbricht U, Brockmann MA, Aigner A, Eckerich C, Müller S, Fillbrandt R, Westphal M, Lamszus K. Expression and function of the receptor protein tyrosine phosphatase zeta and its ligand pleiotrophin in human astrocytomas. Journal of Neuropathology and Experimental Neurology 2003; 62(12):1265-1275.

Ulbricht, U., Eckerich, C., Fillbrandt, R., Westphal, M. & Lamszus, K. RNA interference targeting protein tyrosine phosphatase zeta/receptor-type protein tyrosine phosphatase beta suppresses glioblastoma growth in vitro and in vivo. J Neurochem 98, 1497–1506, 2006.

Vazquez A. Optimal drug combinations and minimal hitting sets. BMC Systems Biology 2009; 3:81.

Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, Hovestadt V, Escalante LE, Shaw ML, Rodman C, Gillespie SM, Dionne D, Luo CC, Ravichandran H, Mylvaganam R, Mount C, Onozato ML, Nahed BV, Wakimoto H, Curry WT, Iafrate AJ, Rivera MN, Frosch MP, Golub TR, Brastianos PK, Getz G, Patel AP, Monje M, Cahill DP, Rozenblatt-Rosen O, Louis DN, Bernstein BE, Regev A, Suvà ML. Science 2017; 355(6332): pii:eaai8478.

Vera-Licona P, Bonnet E, Brillot E, Zinovyev A. OCSANA: optimal combinations of interventions from network analysis. Bioinformatics 2013; 29:1571-1573.

Von Hoff DD, Stephenson JJ Jr, Rosen P, Loesch DM, Borad MJ, Anthony S, Jameson G, Brown S, Cantafio N, Richards DA, Fitch TR, Wasserman E, Fernandez C, Green S, Sutherland W, Bittner M, Alarcon A, Mallery D, Penny R. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. Journal of Clinical Oncology 2010; 28(33):4877-4883.

Wang RS, Albert R. Elementary signaling modes predict the essentiality of signal transduction network components. BMC Systems Biology 2011; 5:44.

Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, Xu L, Luo T, Yan H, Shi ZLA, Zhao T, Xiao Y, Li X. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Research 2019; 47(Database Issue): D900-D908.

Yuan J, Levitin HM, Frattini V, Bush EC, Boyett DM, Samanamud J, Ceccarelli M, Dovas A, Zanazzi G, Canoll P, Bruce JN, Lasorella A, Iavarone A, Sims PA. Single-cell transcriptome analysis of lineage diversity in high-grade glioma. Genome Medicine 2018; 10(1);57.

Zvedei-Oancea I; Schuster S. A theoretical framework for detecting signal transfer routes in signaling networks. Computers & Chemical Engineering 2005; 29:597-617.