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

Saba Ahmadi[1,7^], Pattara Sukprasert[2,8^], Rahulsimham Vegesna[3], Sanju Sinha[3], Fiorella Schischlik[3], Natalie Artzi[4,5,6], Samir Khuller[2,8], Alejandro A. Schäffer[3*], Eytan Ruppin[3*]

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

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

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

[6] Institute for Medical Engineering and Science, MIT, Cambridge, MA 02139 USA

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

[8] Part of this research was 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. 15-C1, Bethesda, MD 20892 USA

Keywords: targeted cancer therapy, combination therapy, personalized medicine, combinatorial optimization, hitting set, single-cell transcriptomics

Abbreviations:

CTS: cohort target set, synonym of global hitting set

GEO: Gene Expression Omnibus

GHS: global hitting set, synonym of cohort target set

GTEx: Genotype-Tissue Expression (project or consortium)

HPA: Human Protein Atlas

HUGO: Human Genome Organization

IHS: individual hitting set, synonym of individual target set

ILP: integer linear programming

ITS: individual target set

lb: lower bound on fraction of tumor cells killed

RME: receptor-mediated endocytosis

TPM: transcripts per million

ub: upper bound on fraction of non-tumor cells killed

Abstract

The availability of single-cell transcriptomics data opens new opportunities for rational design of combination cancer treatments in a systematic manner. Mining such data, we employed combinatorial optimization techniques to explore the landscape of optimal combination therapies in solid tumors, including brain, head and neck, melanoma, lung, breast and colon cancers. We assume that each individual therapy can target any one of 1269 genes encoding cell surface receptors, which may be targets of CAR-T, conjugated antibodies or coated nanoparticle therapies. In most cancer types, personalized combinations composed of at most four targets are sufficient to kill at least 80% of the tumor cells while killing at most 10% of the non-tumor cells in each patient. The number of distinct targets needed to do that for all patients in 8 of the 9 cohorts we studied is at most 11, while one larger melanoma cohort requires over 30 distinct targets. Further requiring that the target genes be lowly expressed across many different healthy tissues uncovers qualitatively similar trends. However, as one requires either more stringent killing thresholds or more stringent sparing of non-cancerous tissues beyond these baseline values, the number of targets needed rises rapidly. Emerging promising targets include the gene *PTPRZ1*, which is frequently found in the optimal combinations for brain and head and neck cancers, and *EGFR*, a recurring target in multiple tumor types. In sum, this is the first systematic single-cell based characterization of the landscape of combinatorial receptor-mediated cancer treatments, identifying promising targets for future development.

Introduction

Personalized oncology offers hope that each patient's cancer can be treated based on its genomic characteristics[1,2]. Several trials have suggested that it is possible to collect genomics data fast enough to inform treatment decisions[3-5]. Meta-analysis of Phase I clinical trials completed during 2011-2013 showed that overall, trials that used molecular biomarker information to influence treatment plans gave better results than trials that did not[6]. However, most precision oncology treatments utilize only one or two medicines, and resistant clones frequently emerge, emphasizing the need to deliver personalized medicine as multiple agents combined[6-11]. Important opportunities to combine systems biology and design of nanomaterials have been recognized to deliver medicines in combination to overcome drug resistance and combine biological effects[12].

Here, we propose and rigorously study a new conceptual framework for designing future precision oncology treatments. It is motivated by the growing recognition that tumors typically have considerable intra-tumor heterogeneity (ITH)[13,14] and thus need to be targeted with a combination of medicines such that as many as possible tumor cells are hit by at least one medicine. Our analysis is based on two recently emerging technologies: (1) the advancement of single-cell transcriptomics and proteomics measurements from patients' tumors, which is anticipated to gradually enter into clinical use[15], and (2) the introduction of "modular" treatments that target specific overexpressed genes/proteins to recognize cells in a specific manner and then use either the T cell immune response or a lethal toxin to kill the tumor cells preferentially.

Based on these two foundations, we formulate and systematically answer two basic questions. First, how many targeted treatments are needed to selectively kill most tumor cells while sparing most of the non-tumor cells in a given patient? And second, given a cohort of patients to treat, 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 encode receptors on the cell surface, as these may be precisely targeted by any one of at least six technologies: e.g., by CAR-T therapy[16], immunotoxins ligated to antibodies[17-18], immunotoxins ligated to mimicking

peptides[19], conventional chemotherapy ligated to nanoparticles[20], degraders associated with ubiquitin E3 ligases[21] and designed ankyrin repeat proteins (DARPins)[22,23]. These treatments are all "modular", including one part that specifically targets the tumor cell via one gene/protein and another part, the cytotoxic mechanism that kills the cells.

Two recent genome-wide analyses of modular therapies have focused on CAR-T therapy[24,25], so we focus first on this technology to put our work in context. In the original formulation, CAR-T therapy used one cell surface target that marks the cells of interest, such as CD19 as a marker for B cells. To date, CAR-T therapy has been effective in achieving remissions for some blood cancers[16,26], but less effective for solid tumors. MacKay et al.[25] focused primarily on single targets and looked at combinations of two targets and did all analysis *in silico*. Dannenfelser et al.[24] focused on predicting combinations of two and three targets and did most of their work *in silico*, with *in vitro* validation of two high-scoring predicted combinations in renal cancer. Importantly, these studies have analyzed *bulk tumor and normal expression* data to identify likely targets. Here we present the first analysis that aims to identify modular targets based on the analysis of *tumor single-cell transcriptomics*. This enables to study the research questions at a higher resolution but presents new analytical challenges that need to be addressed.

Two related difficulties with CAR-T therapy are i) toxicity to non-cancer cells[27,28] and ii) difficulty in finding single targets that are sufficiently selective[25]. To address the toxicity problem, MacKay et al.[25] selected 533 targets that had low expression in most tissues in the Genotype-Tissue Expression (GTEx) data; however, their analysis did not require that the targets are cell surface proteins. We proceed in a stepwise manner; we start with a formal analysis of a space of 1269 candidate cell surface receptors. Then, we proceed to add a low-expression requirement like that of MacKay et al.[25] and parameterized by a transcripts per million (TPM) expression threshold. For completeness, we also tested their set of 533 genes.

To address the selectivity problem, various groups have engineered composite forms of CAR-T treatments that implement Boolean AND, OR, and NOT gates that have been tested for combinations of up to three target proteins[29-33]. Both MacKay et al. and Dannenfelser et al. presented *in silico* methods focusing on AND gates and pairs or trios of targets; Dannenfelser et al. analyzed 2538 likely cell surface proteins that are not necessarily receptors. We have chosen

to focus on the simpler logical OR construction because that can be achieved not only by CAR-T technology[30,31], but can also be implemented via other modular treatment technologies by combining multiple single-target treatments, assuming that the composite treatment kills a cell if any one of the single treatments kill the cell. Conceptually, such a logical OR combination treatment can still achieve selectivity by choosing targets, each of which is expressed on a much higher proportion of cancer cells than non-cancer cells. One of our key contributions is to show that by using techniques from combinatorial optimization, one can find such effective combinations involving a large number of targets, while previous studies were limited to at most three targets.

Beyond CAR-T, our analysis applies to several additional types of modular treatment technologies that rely instead on receptor-mediated endocytosis (RME) delivering a toxin via a targeted receptor to enter the cell[34,35]. Like CAR-T, these RME-based technologies do not downregulate the target receptor. For RME technologies and other technologies that work intracellularly, we anticipate combining modular treatments from one technology such that *all treatments use the same toxin or mechanism of cell killing*, thereby mitigating the need to test for interaction effects between pairs of different treatments.

To address these research questions, we designed and implemented a computational approach named MadHitter (after the Mad Hatter from *Alice in Wonderland*) to identify optimal precision combination treatments that target membrane receptors (**Figure 1, A-C**). We define three key parameters related to the stringency of killing the tumor and protecting the non-tumor cells and explore how the optimal treatments vary with those parameters (**Figure 1B, C**). Solving this problem is analogous to solving the classical "hitting set problem" in combinatorial algorithms[36], which is formally defined in the Methods (see also **Supplementary Materials 1**). Unlike the previous studies on CAR-T targets, we define the problem in a personalized manner, intending that each patient will get optimal treatments for her or his tumor from among a collection of available treatments.

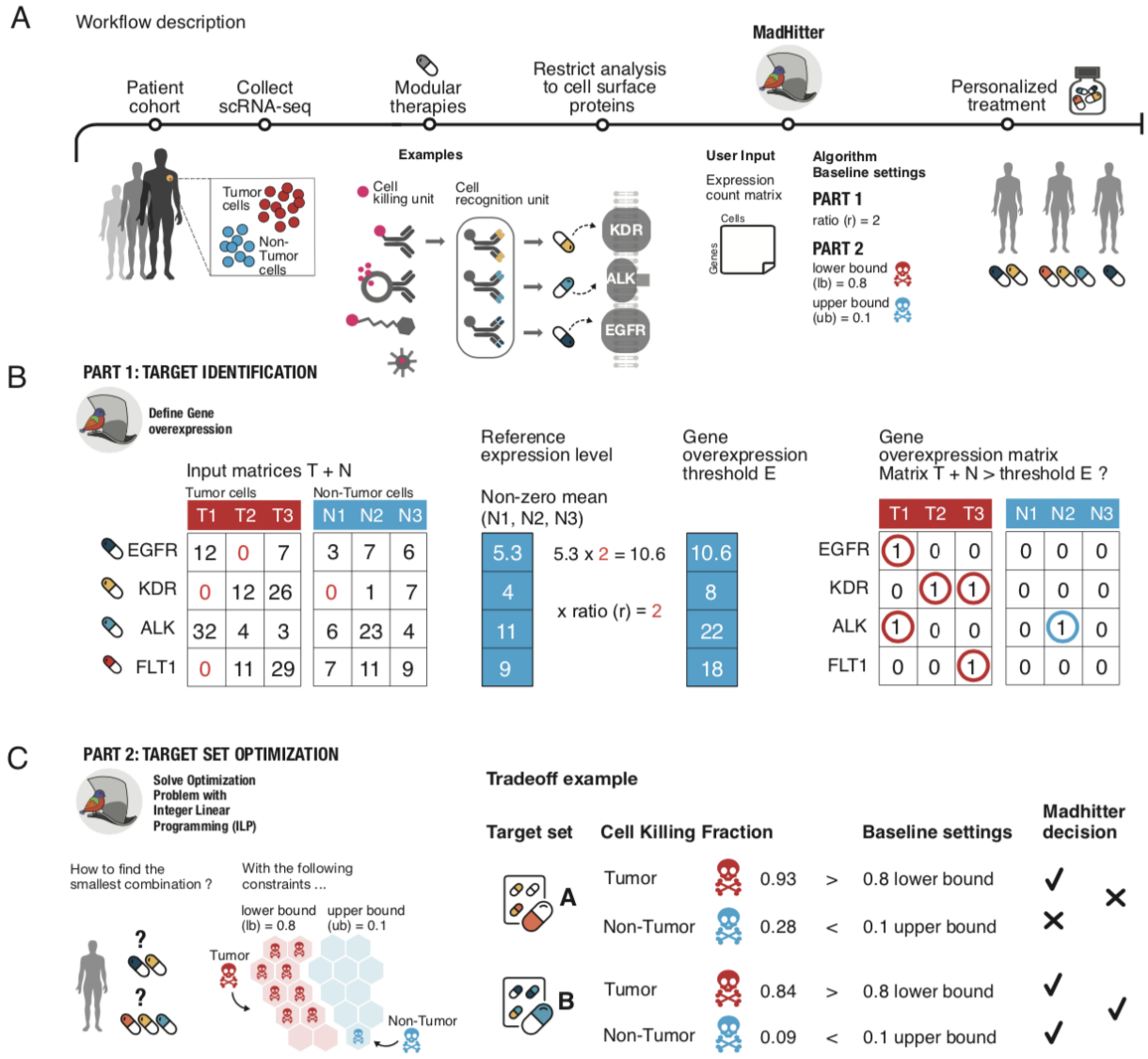

Figure 1. Conceptual schematic example of MadHitter analysis of single-cell data transcriptomics from three cancer patients. (A) A cohort of patients (three in this example) arrives for a study in which single-cell tumor microenvironment (TME) transcriptomics data are collected from each patient; the data are analyzed with MadHitter and each patient receives an optimal personalized combination of targeted therapies from a pre-specified set (pill bottle). MadHitter is aimed at optimizing combinations of targeted therapies that are *modular*, that is, having a recognition unit that is gene/protein-specific, and a joint killing subunit (similar for all gene targets). Icons of four such modular therapies are shown; for three of these, the target protein must be on the cell surface and for two it must be a receptor, so we focus our analyses on cell surface receptors. Three main algorithm parameters are denoted near the MadHitter icon in panel **A** and explained in the later panels. **(B)** The

single-cell TME data are represented in two matrices with the genes as rows and cells as columns, partitioned into tumor (T) and non-tumor (N) cells. The expression ratio *r* determines by how much a gene must be overexpressed for a cell to be considered as a targeted. A gene is considered 'overexpressed' in either a non-tumor cell or a tumor cell if its expression is at least *r* times the mean, reference level; e.g, the reference level for *FLT1* is (7+11+9)/3 = 9 and only cell T3 has *FLT1* expression above 9×2 = 18**.** The matrices on the right side show a Boolean representation of which targets kill which cells, based on the expression values presented in this toy problem in matrix B and taking *r*=2. Accordingly, the combination of EGFR and KDR would kill all tumor cells and would spare all non-tumor cells. **(C)** The main algorithm in MadHitter seeks a combination of targets that is as small as possible and would kill many tumor cells and few non-tumor cells, in a patient-specific manner. The lb and ub parameters are the lower bound on the fraction of tumor cells killed and the upper bound on the fraction of non-tumor cells whose killing is tolerated, respectively. Baseline settings used in our analyses are $r = 2, lb = 0.8$ and $ub = 0.1$, and are varied in some of the analyses. The right side of the panel shows a hypothetical example of the tradeoff between killing tumor cells and sparing non-tumor cells. While target set A could kill a larger fraction of tumor cells than target set B, MadHitter would select target set B since only it satisfies both our baseline settings and kills at most 0.1 fraction of the non-tumor cells.

Results

The Data and the Combinatorial Optimization Framework

We focused our analysis searching for optimal treatment combinations in nine single-cell RNAseq data sets that include tumor cells and non-tumor cells from at least three patients for that were publicly available at the onset of our investigation (Methods; **Table 1**). Those data sets include four brain cancer data sets and one each from head and neck, melanoma, lung, breast and colon cancers. Most analyses were done for all data sets, but for clarity of exposition, we focused in the main text analyses on four data sets from four different cancer types (brain, head and neck, melanoma, lung) that are larger than the other five and hence, make the optimization problems more challenging. Results on the other five data sets are provided in **Supplementary Materials 2**.

Analyzing separately each of these data sets, we ask how many targets are needed to kill most cells of a given tumor and what is the tradeoff between cancer cells killed and non-cancer cells spared? **Figure 2** shows a small schematic example in which there are alternative target sets of sizes two and three. One would prefer the target set of size two because the patients would need to receive only two distinct treatments rather than three treatments.

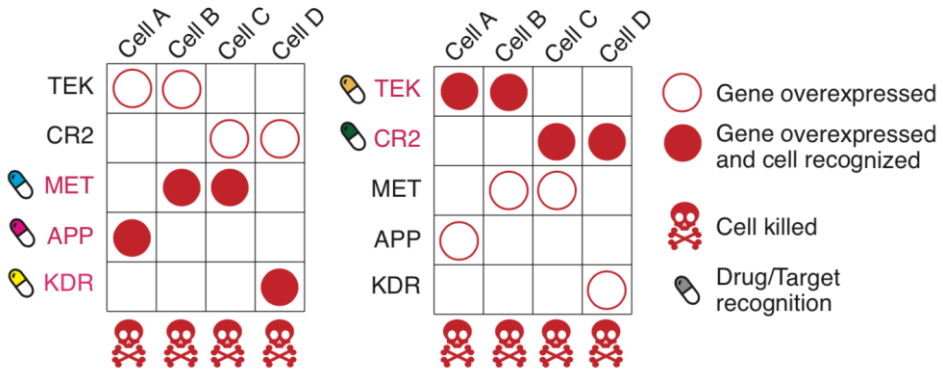

Figure 2. A schematic small example of killing a four tumor cells illustrating why choosing a minimum-size combination of targets may be non-trivial. The schematic tumor has four cancer cells (A, B, C, D in separate columns), which may express any of five cell-surface receptor genes (rows) that may be targeted selectively by modular treatments (pills). If one targets {APP, KDR, MET}, all cancer

cells will be killed (left panel). However, if, instead, one would target {CR2, TEK} then all cancer cells in the given example tumor will be killed with just two targets (right panel) instead of three, providing a smaller solution.

To formalize our questions as combinatorial optimization hitting set problems, we define the following parameters and baseline values and explore how the optimal 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 concept that only cells that overexpress the target, we introduce an additional parameter r. The expression ratio r defines which cells are killed, as follows (**Figure 1B**): 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 thus model more selective killing.

Having r as a modifiable parameter anticipates that in the future one could experimentally tune the overexpression level at which cell killing occurs[24]. In this respect, technologies that rely on RME to get a toxin into the cell are particularly tunable because there is known to be a non-linear relationship between the number of protein copies on the cell surface and the probability that RME occurs successfully[37]. In these technologies, the toxin or other therapy delivered by the modular treatment enters cells in a gene-specific manner[38], while CAR-T therapy activates T-cell killing against cells in a gene-specific manner[24,25]. 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[37]; in one analysis, we varied r up to 5.0 (**Supplementary Table S1**).

Given these definitions, we solve the following combinatorial optimization hitting set problem (Methods): 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 target genes, we seek to find a solution that finds a minimum-size combination of targets in each individual patient, while additionally minimizing the size of all targets given to the patients cohort. The latter is termed the *global minimum-size hitting set (GHS*) in computer science terminology or the *cohort target set (CTS)* in terminology specific to our problem, while the

optimal hitting set of genes targeting one patient is termed the *individual target set* (*ITS*). This optimum hitting set problem with constraints *can be solved to optimality using integer linear programming (ILP)* (Methods). We solve different optimization problem instances, each of which considers a different set of candidate target genes: 1269 genes encoding cell surface receptor proteins, and subset of 58 out of these 1269 genes that already have published ligand-mimicking peptides, and a nested collection of sets of 424-900 out of the 1269 genes that are lowly expressed below a series of decreasing gene expression thresholds[25]. From a computational standpoint, there is no inherent limit on the size of the candidate gene set.

Our formulation is personalized as each patient receives the minimum possible number of treatments. The global optimization comes into play only 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 CTS {*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 and yet adequately capture tumor heterogeneity. We found 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 500. In some smaller data sets, we had to sample smaller numbers of cells (Methods). As shown in (**Supplementary Materials 2**, **Figures S1-S2**), 500 cells, when available, are roughly sufficient for CTS size to plateau for our baseline parameter settings, $lb = 0.8, ub = 0.1, r = 2.0$. For each individual within a data set, we performed independent sampling of c cells 20 times and their results were summarized.

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

Given the single-cell tumor data sets and the ILP optimization framework described above, we first studied how the resulting optimal cohort target set (CTS) may vary as a function of the parameters defining the optimization objectives in different cancer types. **Figures 3 and S3-S7 in Supplementary Materials 3** show heatmaps of CTS sizes when varying lb, ub, and r around the baseline values of 0.8, 0.1, and 2.0, respectively. The CTS sizes for melanoma were largest, partly due to the larger number of patients in that data set (**Table 1**). Indeed, as we sampled subsets of 5 or 10 patients uniformly and observed that the mean CTS sizes grew from 7.9 (5 patient subsets) to 12.3 (10 patient subsets) to 31.0 (all patients, as shown in **Figure 3**).

Encouragingly, for most data sets and parameter settings, the optimal CTS sizes are in the single digits. However, in several data sets, we observe a sharp increase in CTS size as lb values are increased above 0.8 and/or as the ub is decreased 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 (**Supplementary Figures S3-S7**). The qualitative transition observed in CTS sizes occurs robustly regardless of the threshold for filtering out low expressing cells when preprocessing the data (**Supplementary Materials 4, Figures S8-S10**).

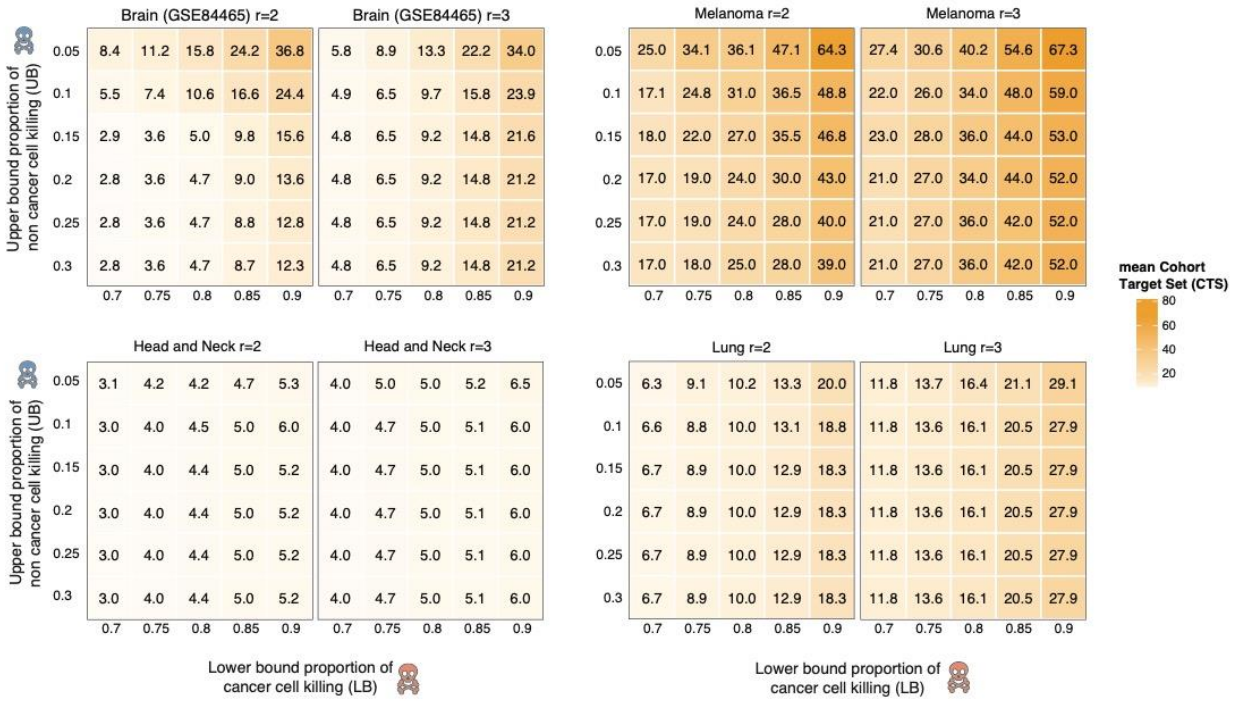

Figure 3. Heat maps showing how the cohort target set size (CTS) varies as a function of lb, ub, r and across data sets. For each plot the x-axis and y-axis represent *lb* and *ub* parameter values, respectively. The scale on the right shows the cohort target set sizes by color scale. We show separate plots for $r = 2.0, 3.0$ here and a larger set {1.5, 2.0, 2.5, 3.0} in **Supplementary Materials 3**. Individual values are not necessarily integers because each value represents the mean of 20 replicates of sampling c (500 for each of the data sets shown here) cells (**Figure S1**).

We next examined what are the resulting *individual target set* (ITS) sizes 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 for four data sets and two combinations of (lb, ub) (**Figure 4**) and for additional data sets in **Supplementary Materials 5, Figure S11**. Overall, the mean ITS sizes with the baseline parameter values ($r = 2.0, lb = 0.8, ub = 0.1$) range from 1.0 to 3.91 among the nine data sets studied (**Supplementary Table S2**); on average 4 targets per patient should hence 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 (by increasing lb from 0.8 to 0.9 and decreasing ub from 0.1 to 0.05), the variability in ITS size across patients became

larger. Importantly, this analysis provides rigorous quantifiable evidence grounding the prevailing observation that among tumors of the same type, some individual tumors may be 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 (in the tens) and individual patients (in the single digits usually) and that these estimates are sensitive to the killing stringency, especially when the lb increases above 0.8. The variation for more aggressive killing regimes, with values of lb up to 0.99 for the baseline $r = 2.0$ is displayed in **Figures S12-S13 in Supplementary Materials 6**. For fixed $lb = 0.8, ub = 0.1$ 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 (**Supplementary Materials 7, Figures S14-S15, Supplementary Table S1**). Finally, we show that, as expected, a 'control' greedy heuristic algorithm searching for small and effective target combinations finds ITS sizes substantially larger than the optimal ITS sizes identified using our optimization algorithm (**Figure 4**). The greedy CTS size is greater than the ILP optimal CTS size for eight out of nine data sets (**Table S2** in **Supplementary Materials 8,** Methods).

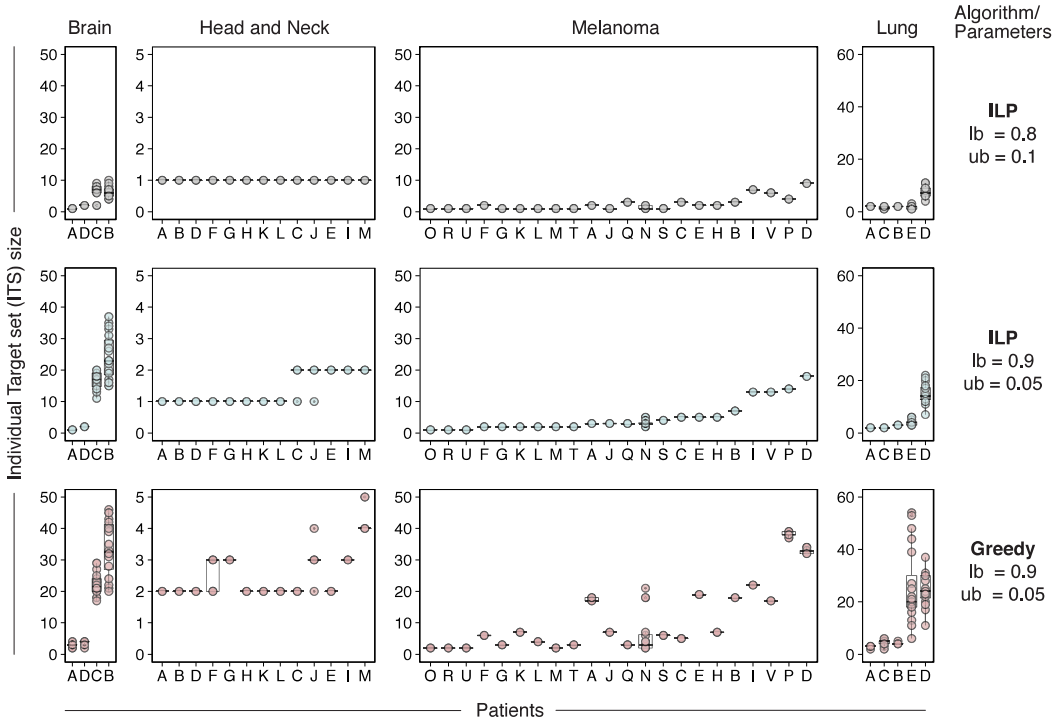

Figure 4. The distribution of optimal and greedy individual treatment combination sizes (ITS) values in four different cancer types. We study both our baseline parameter setting (upper row panels) and a markedly more stringent one (middle row plots). For the more stringent parameter setting, we compare the ITS sizes obtained using MadHitter (middle row plots) and a greedy algorithm that tries to add pairs of genes at a time (bottom row plots). In each plot, the patients are sorted from left to right according to their mean ITS values in the optimal stringent regime. Additional comparisons between ITS sizes at different parameter settings can be found in **Supplementary Materials 5**. Description of the greedy algorithm and more comparisons between the optimal and greedy algorithms are provided in **Supplementary Materials 8**.

The Landscape of Combinations Achievable with Receptors Currently Targetable by Published Ligand-Mimicking Peptides

To get a view of the combination treatments that are possible with receptor targets for which there are already existing modular targeting reagents, we conducted a literature search identifying 58 out of the 1269 genes with published ligand-mimicking peptides that have been already tested in *in vitro* models, usually cancer models (Methods; **Tables 3 and 4**). We asked whether we could find feasible optimal combinations in this case and if so, how do the optimal CTS and ITS sizes compare vs. those computed for all 1269 genes?

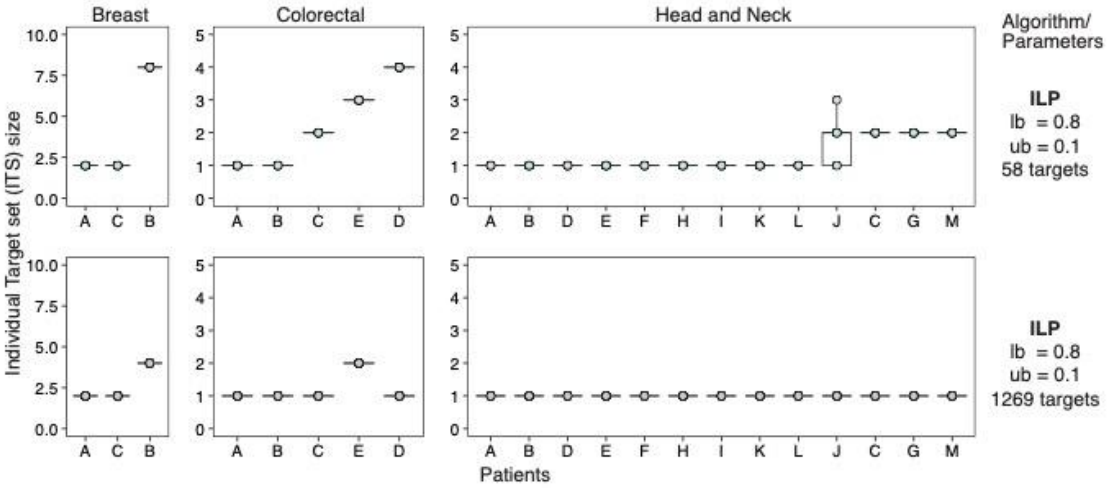

Figure 5. Comparison of Individual Target Set Sizes with 1269 or 58 targets for three out of the six data sets that have feasible solutions. We attempted to find feasible solutions for all patients using 58 cell surface receptors that have published ligand-mimicking peptides that have been tested *in vitro* or in pre-clinical models. There are feasible solutions for all patients in six data sets, but not for the brain (GSE84465), melanoma (GSE115978), and lung (E-MTAB-6149), which were displayed in previous figures. Instead, we show here results for breast and colorectal cancers, for which other analyses, such as those in **Figures 3** and **4**, are in the Supplementary Materials. Some of the optimal solutions obtained on the 58-receptors restricted set are of the same size to those obtained on the whole receptors set and some are larger.

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 and

three of these six are illustrated for each patient in **Figure 5**. However, for three data sets, in numerous parameter combinations we could not find optimal solutions that satisfy the optimization constraints (**Supplementary Materials 9, Figures S16-S18**). 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 (**Supplementary Materials 9, Figures S16-S18**): (1) In some datasets, it is just a little bit more difficult to find optimal ITS and CTS solutions with the 58-gene pool, while in others, the restriction to a smaller pool can be a severe constraint making the optimization problem infeasible. (2) The smaller basket of gene targets may force more patients to receive similar individual treatment sets and thereby reduces the size of the CTS. (3) Unlike the CTS size, the ITS size must stay the same or increase when the pool of genes is reduced, because we find the optimal ITS size for each patient. Overall, the average ITS sizes across each cohort using the pool of 58 genes for baseline settings range from 1.16 to 4.0. Among cases that have any solution, the average increases in the ITS sizes at baseline settings in the 58 genes case vs. that of the 1269 case were 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 minimizing the cohort (CTS size), rather than the individuals (ITS size)? This challenge is motivated by a 'fairness' perspective (**Supplementary Materials 1**), where we seek solutions that are beneficial for the entire community from a social or economic perspective (in terms of CTS size) even if they are potentially sub-optimal at the individual level (in terms of ITS sizes). 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 6A-B**).

We formalized this 'fair CTS problem' by adding a *cost parameter* α that specifies the limit on the excess number of (ITS) targets selected for any individual patient, compared to the number selected in the individual-based approach that was studied up until now (formally, the latter corresponds to setting $\alpha = 0$). We formulated and solved via ILP this fair CTS 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. **Figure 6C** and **Figures S19-S23** in **Supplementary Materials 10** show the optimal CTS and ITS sizes for $\alpha = 0, \ldots, 5$.

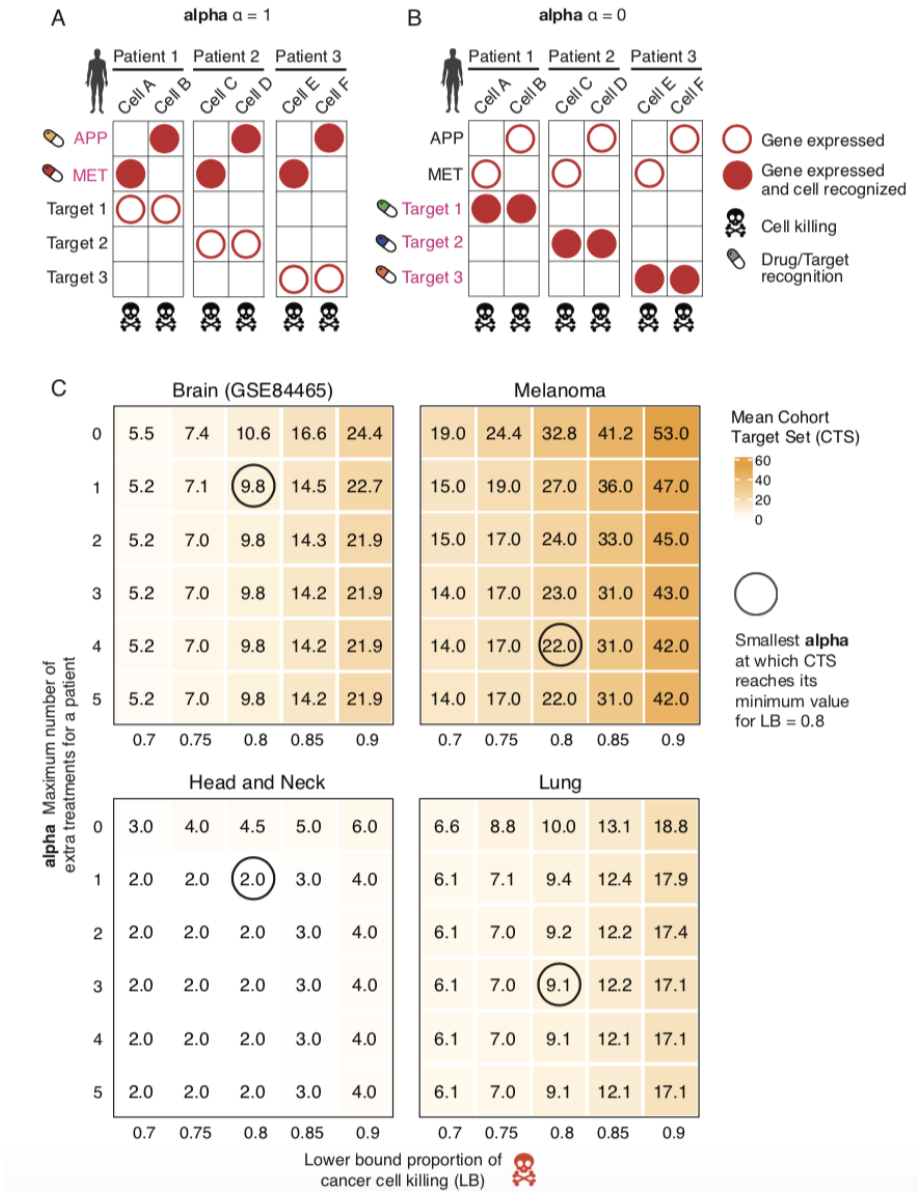

Figure 6. A schematic example demonstrating the rationale and workings of *fairness*-based solutions. (A, B) Let us assume that each of three patients has two tumor cells (columns), each displaying five membrane receptors that are highly expressed only on the tumor cells and not on the non-tumor ones (rows). If we target {APP, MET} (panel **A,** $\alpha = 1$) in all patients, then this achieves a CTS size of 2, which is the minimum possible. Employing the original individual-based optimizing objective, each patient could instead be treated by an ITS of size 1 by targeting the distinct receptors called Target 1 (specific to Patient 1), Target 2 and Target 3, respectively, but this would result in an optimal CTS of size 3 (panel **B,** $\alpha = 0$). The solution in panel A has an unfairness value $\alpha = 1$ because the worst difference among all patients is that a patient

receives 1 more treatment than necessary. (**C**) Heatmaps showing how the CTS size varies as α increases (y-axis), starting from its baseline value of 0 where each patient is assigned a minimum-sizes individual treatment set (top row). The lower bound on tumor cells killed (x-axis) is also varied while the upper bound on non-tumor cells killed is kept fixed at 0.1. We are particularly interested in finding the smallest value on the y-axis at which the CTS size reaches its minimum value, which is circled for the baseline $lb = 0.8$, because this bounds the tradeoff between the achievable reduction in the number of targets needed to treat the whole cohort and the number of extra targets above the ITS minimum that any patient might need to receive.

For 8 out of 9 data sets, we encouragingly find that the unfairness cost parameter α *is bounded by a constant of 3;* i.e., it is sufficient to increase α by no more than 3 to obtain the smallest CTS sizes in the optimally fair solutions. For the largest data set (melanoma), $\alpha = 4$. As we show in **Supplementary Materials 10**, 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. Theoretically, we show that one can design instances for which α would need to be at least $\sqrt{n} - 1$ to get a CTS of size less than the overall number of targets n (**Supplementary Materials 10**). However, in practice, we find that given the current tumor single-cell expression data, fairness-based treatment strategies are likely to be a reasonable economic option in the future.

The Landscape of Optimal Solutions Targeting Receptors that are Lowly Expressed Across Many Healthy Tissues

We turn to examine the space of optimal solutions when restricting the set of eligible surface receptor gene targets to those that have lower expression across many noncancerous human tissues (Methods), aiming to mitigate potential damage to tissues unrelated to the tumor site. To this end, we selected subsets of the 1269 cell surface receptor targets in which the genes have overall lower expression across multiple normal tissues, by mining GTEx and the Human Protein Atlas (HPA) (Methods). Varying the *selectivity expression thresholds* (expressed in transcripts per million (TPM)) used to filter out genes whose mean expression across the normal adult tissues is above values of 10, 5, 2, 1, 0.5, and 0.25 (i.e., employing more and more extensive

filtering as this threshold is decreased), decreases the size of the target cell surface receptor gene list by more than half (**Table 2**).

As shown in **Figures 7A, B (and Supplementary Figures S24-S26)**, MadHitter identifies very different cohort target sets (which are larger than the original optimal solutions, as expected) as the TPM selectivity threshold value is decreased. Furthermore, different ITS instances may become infeasible (**Supplementary Figure S27**). At an individual patient level, using lower selectivity threshold levels, which leads to a smaller space of membrane receptors to choose from, also leads to increased mean ITS sizes (**Supplementary Figures S28, S29**). Across the nine data sets, the selectivity threshold at which the CTS problem became infeasible varied (**Supplementary Figure S27**). The differences observed could be the result of expression heterogeneity of the cancer, number of patients within the data set, size of target gene set, lack of expression of available gene targets and other unknown factors. In the future, further experimentation is required to identify tissue-specific optimal gene expression thresholds that will minimize side effects while allowing cancer cells to be killed by combinations of targeted therapies. Finally, for completeness, we also tested MadHitter on the set of 533 lowly expressed genes suggested by MacKay et al.[25] All instances with default setting of r, lb, ub have feasible solutions for all patients. Mean ITS sizes are below 4 for eight of nine data sets, but close to 10 for the brain cancer data set GSE84465. More details can be found in **Supplementary Materials 11** and **Table S3**.

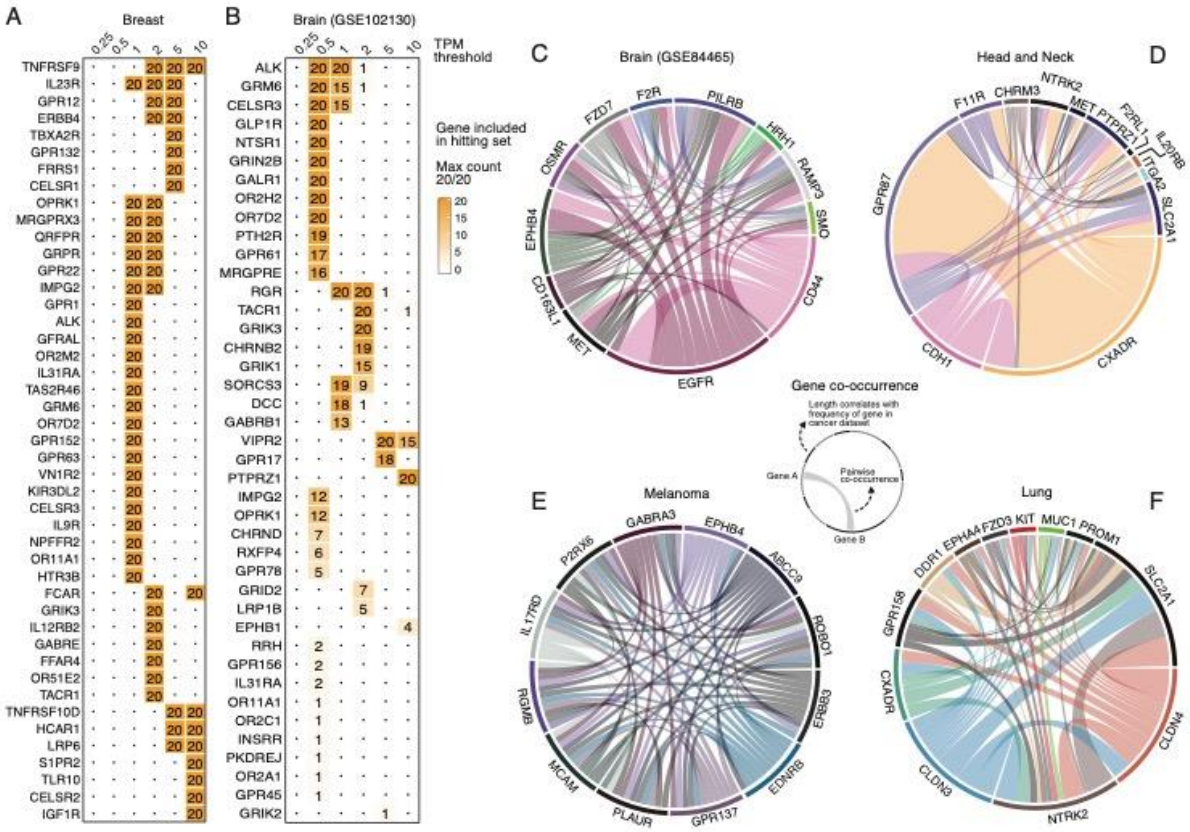

Figure 7. Variation in the CTS size and composition as function of the magnitude of filtering of genes expressed in noncancerous human tissues, for different tumor types. (A-B) The number of times a gene (cell-surface receptor) is included in the CTS (out of 20 replicates, which is therefore the Max count in panels A-B), where each column presents the CTS solutions when the input target genes sets are filtered using a specific TPM filtering threshold (Methods), for (**A)** a breast cancer and **(B)** brain cancer. These data sets were selected due to their relatively small cohort target set sizes, permitting their visualization. **(C-F)** CircOs plots of the genes occurring most frequently in optimal CTS solutions (length of arc along the circumference) and their pairwise co-occurrence (thickness of the connecting edge) for the four main cancer types, in our original target space of 1269 encoding cell-surface receptors. For each data set, we sampled up to 50 optimal CTS solutions. Network representations of the 12 most common target genes out of 1269 encoding cell-surface receptors (with greater than 5% frequency of occurrence) are represented in a cancer specific manner for **(C)** brain cancer, **(D)** head and neck cancer, (only seven genes have a frequency of 5% or more across optimal

solutions), **(E)** melanoma, and **(F)** lung cancer. Genes and connections have distinct colors for improved visibility.

Key Targets Composing Optimal Solutions Across the Space of 1269 membrane receptors

To identify the genes that occur most often in optimal solutions for our baseline settings, since there may be multiple distinct optimal solutions composed of different target genes, we sampled up to 50 optimal solutions for each optimization instance solved and recorded how often each gene occurs and how often each pair of genes occur together (Methods). We analyzed and visualized these gene (co-)occurrences in three ways. First, we constructed co-occurrence circus plots in which arcs around the circle represent frequently occurring genes and edges connect targets that frequently co-occur in optimal CTS solutions. **Figure 7C-F** shows the co-occurrence visualizations for optimal CTS solutions obtained with the original, unfiltered target space of 1269 genes and in baseline parameter settings. The genes frequently occurring in optimal solutions are quite specific and distinct between different cancer types. In melanoma, the edges form a clique-like network because virtually all optimal solutions include the same clique of 12 genes (**Figure 7E**). The head and neck cancer data set has only one commonly co-occurring pair {*GPR87, CXADR*} (**Figure 7D**). Of the cancer types not depicted in **Figure 7**, the breast cancer data set has a commonly co-occurring set of size 4, {*CLDN4, INSR, P2RY8, SORL*}, and the colorectal cancer data set has a different commonly co-occurring set of size 4, {*GABRE, GPRR, LGR5, PTPRJ*} (data not shown).

We next tabulated sums of how often each gene occurred in optimal solutions for all nine data sets (**Supplementary Materials 12, Tables S4, S5 and S6**), obtained when solving for either 58 gene targets or 1269 gene targets. Strikingly, one gene, *PTPRZ1* (protein tyrosine phosphatase receptor zeta 1), appears far more frequently than others, especially in three brain cancer data sets (GSE70630, GSE89567, GSE102130, **Supplementary Table S6**). *PTPRZ1* also occurs commonly in optimal solutions for the head and neck cancer data set (**Figure 7D**). The brain cancer finding coincides with previous reports that *PTPRZ1* is overexpressed in glioblastoma (GBM)[39,40]. *PTPRZ1* also forms a fusion with the nearby oncogene *MET* in some brain tumors that have an overexpression of the fused *MET*[41]. Notably, various cell line studies

and mouse studies have shown that inhibiting PTPRZ1, for example by shRNAs, can slow glioblastoma tumor growth and migration[42,43]. There have been some attempts to inhibit PTPRZ1 pharmacologically in brain cancer and other brain disorders[44,45]. 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 in each cohort. The much lower relative level of *PTPRZ1* expression in GSE84465 is likely due to the heterogeneity of brain cancer types in this data set[46]. Among the 58 genes with known ligand-mimicking peptides, *EGFR* stands out as most common in optimal solutions (**Supplementary Table S4**). Even when all 1269 genes are available, *EGFR* is most commonly selected for the brain cancer data set (GSE84465) in which *PTPRZ1* is not as highly overexpressed (**Figure 7C**).

PTPRZ1 was the fifth most frequently occurring gene in optimal solutions for the head and neck cancer data set (GSE103322). The two most common genes by a large margin are *CXADR* and *GPR87*. *CXADR* has been studied primarily by virologists and immunologists because it encodes a receptor for cocksackieviruses and adenoviruses[47]. In one breast cancer study, CXADR was found to play a role in regulating PTEN in the AKT pathway, but CXADR was underexpressed in breast cancer[48] whereas it is overexpressed in the head and neck cancer data we analyzed. *GPR87* is a rarely studied G protein-coupled receptor with an unknown natural ligand[49]. In the context of cancer, GPR87 has previously been reported as overexpressed in several tumor types including lung and liver[49] and its overexpression may play an oncogenic role via either the p53 pathway[50] the NFκB pathway[51] or other pathways.

Finally, we analyzed the set of genes in optimal solutions via the STRING database and associated tools[52] to perform several types of gene set and pathway enrichment analyses. **Figures S30-S33 (Supplementary Materials 12)** show STRING-derived protein-protein interaction networks for the 25 most common genes in the same four data for which we showed co-occurrence graphs in **Figure 7C-F**. Again, *EGFR* stands out as being a highly connected protein node in the solution networks for both the brain cancer and head and neck cancer data sets. Among the 30 genes in the 1269-gene set most commonly in optimal solutions (**Supplementary Table S5**), there are six kinases (out of 88 total human transmembrane kinases with a catalytic domain, STRING gene set enrichment $p < 1e - 6$), 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*.

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, including CAR-T therapy. We showed that, for most tumors we studied, four modular medications targeting different overexpressed receptors may suffice to selectively kill most tumor cells, while sparing most of the non-cancerous cells (**Figures 3** and **4** and **Table S2**). For the more restricted sets of low-expression genes[25] or the 58 receptors with validated ligand-mimicking peptides (**Tables 3** and **4**), some patients do not have feasible solutions, especially as we reduce the TPM expression used for filtering the gene set to avoid targeting non-cancerous tissues. These findings indicate, on one hand, that researchers designing ligand-mimicking peptides have been astute in choosing targets relevant to cancer. On the other hand, these results suggest that there is a need for extending the set of cell surface receptors that can be targeted to enter tumor cells with ligated chemotherapy agents.

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 at most 3 targets larger, and in most data sets at most 1 target receptor larger than the optimal solutions that would have been assigned for these patients based on an individual-centric policy (**Figure 6, Supplementary Materials 10**). This suggests that the concern that the personalized treatment for any individual will be suboptimal solely because that individual happens to have registered for a cohort trial appears to be tightly bounded.

Like the study of MacKay et al.[25], our study is a conceptual computational investigation. We studied nine data single-cell expression data sets for the first time, but it would be helpful to

analyze more and larger data sets in the future. Even among four data sets of the same (brain) cancer type, we observed considerable variability in CTS and ITS sizes. Since our approach is general and the software is freely available as source code, other researchers can test the method on new data sets or add new variables, constraints, and optimality criteria. These investigations may of course lead other to further improve our method and broaden its applicability. In future work, we plan to apply our approach to study ways for selectively killing specific populations of immune cells, such as myeloid-derived suppressor cells, because they inhibit tumor killing, while sparing most other non-cancer cells.

We compared gene expression levels between non-cancer cells and cancer cells sampled from the same patient, which avoids inter-patient expression variability[53]. However, we did little to account for "expression dropout" beyond the normalization performed by the providers of the data sets, aiming to preserve the public data as it was submitted to GEO or Array Express. To achieve some uniformity and also to take as cautious an approach as possible, we added a step to filter low expressing cells because some data sets had already been filtered in this way. One could instead apply imputation methods such as MAGIC[54] or scImpute[55] to infer denser gene expression matrices and then apply our method to the adjusted input data. Therefore, our results about the sizes of optimal ITS 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. Another limitation of our method is that we viewed the measured gene expression as being valid over all time, even though gene expression is known to be a stochastic process[56]. In the future, we would like to extend our approach to use the single-cell data to infer how stochastic is the expression of each gene[57] and to prefer targets whose expression is more stable.

Even though the combinatorial optimization problems solved here are in the worst-case exponentially hard (NP-complete[36] 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., have recently formalized a related optimization problem in analysis of single-cell clustered data for immunology[38]. Their optimization problem is also NP-complete in the worst case and they could solve sets of up to size four using heuristic methods[38]. We have shown that the optimal ILP solutions we obtained

are often substantially smaller than solutions obtained via a greedy heuristic (**Figure 4**, **Supplementary Materials 8** including **Table S2**).

On the cautionary side, experiments with target gene sets that were further filtered by low expression in normal tissues showed that the individual target set problem can become infeasible in many instances. Even when the instance remained feasible, optimal cohort treatment set sizes increased rapidly as the expression levels allowed decreased (**Figure 7**), pointing to potential inherent limitations of applying such combination approaches to patients in the clinic and the need to carefully monitor their putative safety and toxicity in future applications. Finally, 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. We also found that that the less-studied phosphatase *PTPRZ1* is a useful target, especially in brain cancer.

In summary, this study is the first to harness combinatorial optimization tools to analyze emerging single-cell data to portray the landscape of feasible personalized combinations in cancer medicine. Our findings uncover promising membranal targets for the development of future oncology medicines that may serve to optimize the treatment of cancer patient cohorts in several cancer types. The MadHitter approach presented and the accompanying software made public can be readily applied to address additional fundamental related research questions and 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)[58] and Ensembl's ArrayExpess[59] 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[60], CancerSea[61], GEO, ArrayExpress, Google Scholar, and the 10x Genomics list of publications (https://www.10xgenomics.com/resources/publications/). We required that each

data set contain 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.[62] 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 900 or 424 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 **Supplementary Materials 4**, we tested the robustness of this 10% threshold. Finally, we retained either all available genes from among either our set of 1269 genes encoding cell-surface receptors that met additional criteria on low expression or available ligand-mimicking peptides.

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[63] supersedes and partly subsumes GSE7256[64].

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	65
GSE118389	Breast	804(6)	314(6)	Metastasis or not	62
GSE89567	Brain (glioma)	5097(10)	1146(9)	No	66
GSE103224	Brain (glioma)	23793(8)	--	No	61
GSE70630	Brain (glioma)	4044(6)	303(6)	No	67
GSE57872	Brain (glioma)	440(6)	--	No	68
GSE102130	Brain (6 glioma and 3 glioblastoma)	2858(9)	94(5)	No	69
GSE84465	Brain (glioblastoma)	1091(4)	651(4)	No	46
GSE81861	Colorectal	272(10)	160(6)	No	70
GSE103322	Head and Neck	2093(13)	3197(15)	No	71
GSE115978	Melanoma	2018(23)	4334(32)	Yes, immuno-therapy	63,64
GSE118828	Ovarian	1415(11 primary) 973 (5 metastasis)	578(2)	No	72
GSE67980	Prostate	124(21)	--	Metastasis or not	73
E-MTAB-6149	Lung	7351(5)	2730(5)	No	74

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 data sets 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, and m cells comprising tumor cells and non-tumor cells. We want to select a subset of $m' < m$ cells. We select a set of m' cells uniformly at random without replacement from among all cells. Then we partition the selected cells into $m_t{}'$ tumor cells and m_n' non-tumor cells to define one *replicate*. In most of the computational experiments shown we used 20 replicates and we report either the arithmetic mean or entire distribution of quantities such as the CTS size.

Considering a previously defined set of target genes and of HPA gene expression across different normal tissues

The general aim of our methods is to target the cancer cells while sparing the adjacent non-cancer cells as much as possible. A related concern is that genes within the target set could be expressed at high levels in other normal tissues that are not part of the non-cancer cells from the tumor microenvironment included in the input data sets. One way to address this problem is to identify genes that have low expression in the majority of the tissues and to use them to obtain a target set. This approach has been pioneered in a recent paper on selecting gene targets suitable for CAR-T therapy[25]. The authors selected 533 candidate genes that they judged could be reasonable targets for CAR-T. They made this selection based on expression data from the Human Protein Atlas[75] and the Genotype-Tissue Expression consortium (GTEx)[76], which have expression information from multiple tissues which was used to identify low expressed target genes.

McKay et al.[25] used a threshold of 15 TPM units of expression (written in their work as $\log2(\text{TPM}+1) \leq 4$), but they allowed a small number of tissues to exceed this threshold. Instead, we used quantitative levels of expression for finer granularity in analysis, as described in the next subsection. One clinical difference is that we looked only at adult tissues because we are

analyzing adult tumors, while CAR-T therapy can be used for either childhood or adult tumors. The reason to focus on cell-surface receptors, as suggested by Dannenfelser et al.[24], is that CAR-T therapy requires a cell-surface target that may or may not be a receptor, antibody technologies require a cell surface receptor, and the ligand-mimicking peptide nanotechnology that we summarized in the Introduction also requires cell surface receptor targets.

Construction of target gene sets that are lowly expressed in normal tissues

To analyze the tissue specificity of the 1269 candidate target genes, the RNAseq based multiple tissue expression data was obtained from the Human Protein Atlas (HPA) database (https://www.proteinatlas.org/about/download ; Date: May 29, 2020). The HPA database includes expression values (in units of transcripts per million (TPM)) for 37 tissues from HPA (rna_tissue_hpa.tsv.zip)[75] and 36 tissues from the Genotype-Tissue Expression consortium (rna_tissue_gtex.tsv.zip)[76]. Next, to identify target genes with low or no expression within majority of adult human tissues, for the 1269 candidate genes we identified genes whose average expression across tissues is below certain threshold value (0.25, 0.5, 1, 2, 5, and 10 TPM) in both HPA and GTEx data sets. Using the intersection of low expression candidate genes from HPA and GTEx data sets, we generated lists of high confidence targets. The size of the resulting high confidence target genes varied from 424 (average expression less than 0.25 TPM) to 900 (average expression across tissue less than 10 TPM) genes (**Table 2**). While the total number of genes decreases slowly, the decrease is much steeper if one excludes olfactory receptors and taste receptors (**Table 2**). These sensory receptors are not typically considered as cancer targets, although a few of these receptors are selected in optimal target sets when there are few alternatives (**Figure 7**). MadHitter was run on all nine data sets using the expression information from the high confidence gene lists.

Table 2: Size of high confidence target gene sets for different thresholds.

Thresholds expression across	Size of gene set	No of genes which are NOT olfactory	No of genes with ligand mimicking

tissues (TPM)		(OR*) and taste receptors (TAS*)	peptides (intersection with Tables 3 and 4)
0.25	424	97	1
0.5	494	141	3
1	547	187	3
2	632	269	6
5	762	398	10
10	900	536	19

Assembling Lists of Membrane 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[77]; we reduced the list to 1269 by requiring that the proteins be receptors, which is necessary for RME-based therapies but not for CAR-T therpy[24]. For condition ii), we found two review articles in the chemistry literature[19-20] that list targets effectively meeting this condition. Intersecting the lists meeting conditions i) and ii) gave us 38 genes/proteins that could be targeted (**Table 3**).

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 3**, 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[20] 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 4**. Thus, our target set analyses restricted to genes with known ligand-mimicking peptides use $58 = 38 + 20$ targets.

Table 3. Single proteins that can be targeted by peptides based on references 18, 19 and are expressed on the cell surface[77]. 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 4. Single proteins that can be targeted by ligand-mimicking peptides but are not included in the two principal reviews that we consulted[19-20] and are among 1269 cell surface receptors[77].

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

Definition of the Minimum Hitting Set Problem and Solution Feasibility

One of Karp's original NP-complete problems is called "hitting set" and is defined as follows[36]. Let U be a finite universal set of elements. Let $S_1, S_2, S_3, \ldots, S_k$ be subsets of U**.** Is there a small subset $H \subseteq U$ such that for $i = 1, 2, 3, \ldots, k$**,** $S_i \cap H$ is non-empty. In our setting, U is the set of target genes and the subsets $S_i$are the single cells. In reference 78, 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[79-83], identification of critical paths in gene signaling and regulatory networks[84-86] and selection of a set of drugs to treat cell lines[87-88] or single patients[89-90]. More information about related work can be found in **Supplementary 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 4**, 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 4**.

Basic Optimal Target 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 this target set problem:

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

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

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 U whose 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 target 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 target 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.

The Fair Cohort Target Set Problem 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 an individual target set $H_i^{S^*} \subseteq S^*$of a relative small size (compared to the optimal ITS 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 n patients. For the ith 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 target set of patient $P(i)$. The objective function is to minimize the total number of genes selected, subject to having a target 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.

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

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

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

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

Parameterization of the Fair Cohort Target Set Problem

In the Fair Cohort Target Set ILP shown above, we give more preference towards minimizing number of genes needed in the CTS. 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 cohort target set $H(i)$ with the following properties:

1. $H(i)$ covers all the tumor cells of patient $P(i)$.
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 number of non-tumor cells killed is represented by the variable y.

The ILP can be formulated as follows:

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

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

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

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

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.

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

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

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

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

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

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

Implementation Note, Accounting for Multiple Optima and Software Availability

We implemented in Python 3 the above fair cohort target 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 usually used the SCIP library and its Python interface[891]. To obtain multiple optimal solutions of equal size we used the Gurobi library (https://www.gurobi.com) and its python interface. When evaluating multiple optima, for all feasible instances, we sampled 50 optimal solutions that may or may not be distinct, using the Gurobi function select_solution(). To determine how often each gene or pair of genes occur in

optimal solutions, we computed the arithmetic mean of gene frequencies and gene pair frequencies over all sampled optimal solutions.

The software package is called MadHitter. The main program is called hitting_set.py. We include in MadHitter a separate program to sample cells and generate replicates, called sample_columns.py. So long as one seeks only single optimal solutions for each instance, exactly one of SCIP and Gurobi is sufficient to use MadHitter. We verified that SCIP and Gurobi give optimal solutions of the same size. If one wants to sample multiple optima, this can be done only with the Gurobi library. The choice between SCIP and Gurobi and the number of optima to sample are controlled by command-line parameters use_gurobi and num_sol, respectively. The MadHitter 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. This research is supported in part by start-up funds from Northwestern University and a research award from Amazon 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, Gurobi, 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 for technical advice.

Competing Interests

The authors declare that they have no competing interests.

References

1. Von Hoff D.D., et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. *J. Clin. Oncol.* **28**(33), 4877-4883 (2010).

2. Schütte M., et al. Cancer precision medicine: Why more is more and DNA is not enough. *Pub. Health Genomics* **20**(2):70-80 (2017).
3. Jameson, G.S., et al. 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 Res. Treat.* **147**(3), 579-588 (2014).
4. Saulnier Sholler, G.L., et al. Feasibility of implementing molecular-guided therapy for the treatment of patients with relapsed or refractory neuroblastoma. *Cancer Med.* **4**(6):871-886 (2015).
5. Byron, S.A., et al. Prospective feasibility trial for genomics-informed treatment in recurrent and progressive glioblastoma. *Clin. Cancer Res.* **24**(2), 295-305 (2018).
6. Schwaederle, M., et al. Association of biomarker-based treatment strategies with response rates and progression-free survival in refractory malignant neoplasms: A meta-analysis. *JAMA Oncology* **2**(11), 1452-1459 (2016).
7. Arnedos, M., Vielh, P., Soria, J.C. & Andre, F. The genetic complexity of common cancers and the promise of personalized medicine: is there any hope? *J. Pathol.* **232**(2): 274-282 (2014).
8. Nikanjam, M., Liu, S., Yang, J. & Kurzrock, R. Dosing three-drug combinations that include targeted anti-cancer agents: Analysis of 37,763 patients. *Oncologist* **22**(5), 576-584 (2017).
9. Rebollo, J. et al. Gene expression profiling of tumors from heavily pretreated patients with metastatic cancer for the selection of therapy: A pilot study. *Am. J. Clin. Oncol.* **40**(2), 140-145 (2017).
10. Sureda, M., et al. Determining personalized treatment by gene expression profiling in metastatic breast carcinoma patients: a pilot study. *Clin. Trans. Oncol.* **20**(6), 785-793 (2018).
11. Sicklick, J.K., et al. Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. *Nat. Med.* **25**(5), 744-750 (2019).
12. Joo, J.I., et al. Realizing cancer precision medicine by integrating systems biology and nanomaterial engineering. *Adv. Mater.* **32**(35):e1906783 (2020).
13. Marusyk, A. & Polyak, K. Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta* **1805**(1), 105-117 (2010).
14. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27**(1):15-26 (2015).
15. Yofe, I., Dahan, R. & Amit I. Single-cell genomic approaches for developing the next generation of immunotherapies. *Nat. Med.* **26**(2), 171-177 (2020).
16. Neelapu, S.S., et al. Axicabatagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma. *New Engl. J. Med.* **377**(26):2531-2544, 2017.
17. Bjorn, M.J., Ring, D. & Frankel, A. Evaluation of monoclonal antibodies for the development of breast cancer immunotoxins. *Cancer Res.* **45**(3), 1214-1221 (1985).
18. Pastan, I., Willingham, M.C. & FitzGerald D.J.P. Immunotoxins. *Cell* **47**(5), 641-648 (1986).
19. Gray, B.P. & Brown, K.C. Combinatorial peptide libraries: mining for cell-binding peptides. *Chem. Rev.* **114**(2), 1020-1081 (2014).
20. Liu, R., Li, X., Xiao, W. & Lam, K.S. Tumor-targeting peptides from combinatorial libraries. *Adv. Drug Delivery Rev.* **110-111**, 13-37 (2017).
21. Fisher, S.L. & Phillips, A.J. Targeted protein degradation and the enzymology of degraders. *Curr. Opin. Chem. Biol.* **44**, 47-55 (2018).

22. Plückthun A. Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy. *Annu. Rev. Pharmacol. Toxicol.* **55**, 489-511 (2015).
23. Sokolova, E.A., et al. HER2-specific targeted toxin DARPin-LoPE: Immunogenicity and antitumor effect on intraperitoneal ovarian cancer xenograft model. *Int. J. Mol. Sci.* **20**(10). pii: E2399 (2019).
24. Dannenfelser, R., et al., Discriminatory power of combinatorial antigen recognition in cancer T cell therapies. *Cell Syst.* **11**(3):215-228 (2020).
25. MacKay, M., et al. The therapeutic landscape for cells engineered with chimeric antigen receptors. *Nat. Biotech.* **38**(2), 233-244 (2020).
26. Maude, S.L., et al. Tisagenlecleusel in children and young adults with B-cell lymphoblastic leukemia. *New Engl. J. Med.* **378**(5):439-448 (2018).
27. Lamers, C.H., et al. Treatment of metastatic renal cell carcinoma with CAIX CAR-engineered T cells: clinical evaluation and management of on-target toxicity. *Mol. Ther.* **21**(4):904-912 (2013).
28. Thistlethwaite, F.C., et al. The clinical efficacy of first-generation carcinoembryonic antigen (CEACAM5)-specific CAR T cells is limited by poor persistence and transient pre-conditioning-dependent respiratory toxicity. *Cancer Immunol. Immunother.* **66**(11):1425-1436 (2017).
29. Fedorov, V.D., Themeli, M., Sadelain, M. PD-1- and CTLA-4-based inhibitory chimeric antigen receptors (iCARs) divert off-target immunotherapy responses. *Sci. Transl. Med.* **5**(215):215ra172 (2013).
30. Grada, Z., et al. TanCAR: A novel bispecific chimeric antigen receptor for cancer immunotherapy. *Mol. Ther. Nucl. Acids* **2**:e105 (2013).
31. Hegde, M., et al. Combinational targeting offsets antigen escape and enhances effector function of adoptively transferred T cells in glioblastoma. *Mol. Ther.* **21**(11):2087-2101 (2013).
32. Roybal, K.T., et al. Engineering T cells with customized therapeutic response using synthetic Notch receptors. *Cell* **167**(2):419-43.e16 (2016).
33. Williams, J.Z., et al. Precise T cell recognition programs designed by transcriptionally linking multiple receptors. *Science* **370**(6520):1099-1104 (2020).
34. Říhová, B. Receptor-mediated targeted drug or toxin delivery. *Adv. Drug Deliv. Rev.* **29**(3), 273-289 (1998).
35. Tortorella, S. & Karagiannis, T.C. Transferrin receptor-mediated endocytosis: a useful target for cancer therapy. *J. Membr. Biol.* **247**(4), 291-307 (2014).
36. Karp, R.M. Reducibility among combinatorial problems. In *Complexity of Computer Computation*s, pp. 85-103 (Plenum Press, New York, 1972).
37. Martinez-Veracoechea, F.J. & Frenkel, D. Designing super selectivity in multivalent nano-particle binding. *Proc. Natl Acad. Sci. USA* **108**, 10963-10968 (2011).
38. Delaney, C., et al. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Mol. Syst. Biol.* **15**(10), e9005 (2019).
39. Müller, S., et al. A role for receptor tyrosine phosphatase zeta in glioma cell migration. *Oncogene* **22**(43), 661-668 (2003).
40. Ulbricht, U. et al. Expression and function of the receptor protein tyrosine phosphatase zeta and its ligand pleiotrophin in human astrocytomas. *J. Neuropathol. Exp. Neurol.* **62**(12), 1265-1275 (2003).

41. Chen, H.M., et al. Enhanced expression and phosphorylation of the MET oncoprotein by glioma-specific PTPRZ1-MET fusions. *FEBS Lett.* **589**(13):1437-1443 (2015).
42. 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).
43. Bourgonje. A.M., et al. Intracellular and extracellular domains of protein tyrosine phosphatase PTPRZ-B differentially regulate glioma cell growth and motility. *Oncotarget* **5**(18), 8690-8702 (2014).
44. Fujikawa, A., et al. Targeting PTPRZ inhibits stem cell-like properties and tumorigenicity in glioblastoma cells. *Sci. Rep.* **7**:5609 (2017).
45. Pastor, M., et al. Development of inhibitors of receptor protein tyrosine phosphatase β/ζ (PTPRZ1) as candidates for CNS disorders. *Eur. J. Med. Chem.* **144**:318-329 (2018).
46. Darmanis, S., et al. Single-cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21**, 1399-1410 (2017).
47. Bergelson, J.M., et al. Isolation of a common receptor for coxsackie B viruses and adenoviruses 2 and 5. *Science* **275**(5304): 1320-1323 (1997).
48. Nilchian, A., et al. CXADR-mediated formation of an AKT inhibitory signalosome at tight junctions controls epithelial-mesenchymal plasticity in breast cancer. *Cancer Res.* **79**(1):47-60 (2019).
49. Arfelt K.N., et al. Signaling via G proteins mediates tumorigenic effects of GPR87. *Cell. Signal.* **30**:9-18 (2017).
50. Zhang, Y., Qian, Y., Lu, W., Chen, X. The G protein-coupled receptor 87 is necessary for p53-dependent cell survival in response to genotoxic stress. *Cancer Res.* **69**(15):6049-6056 (2009).
51. Wang, L., et al. Overexpression of G protein-coupled receptor GPR87 promotes pancreatic cancer aggressiveness and activates NF-κB signaling pathway. *Mol. Cancer* **16**:61 (2017).
52. Szklarczyk, D., et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019; **47**(D1), D607-D613 (2019).
53. Seoane, J. & De Mattos-Arruda L. The challenge of intratumour heterogeneity in precision medicine. *J. Intern. Med.* **276**(1), 41-51 (2014).
54. Van Dijk D, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**(3):716-729.e27 (2018).
55. Li, W.V. & Li, J.J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Comm.* **9**, 997 (2018).
56. Raj, A. & van Oudenaarden, A. Stochastic gene expression and its consequences. *Cell* **135**(2): 216-226 (2008).
57. Kim, J.Y. & Marioni, J.C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**, R7 (2013).
58. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Meth. Mol. Biol.* **1418**, 93-110 (2016).
59. Kolesnikov, N., et al. ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.* **43**(Database Issue), D1113-D1116 (2015).
60. Cao, Y., Zhu, J., Jia, P. & Zhao Z. scRNASeqDB: A database for RNA-seq based gene expression profiles in human single cells. *Genes* **8**(12), 368 (2017).

61. Yuan J, et al. Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med.* **10**(1), 57 (2018).
62. Karaayvaz, M., et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Comm.* **9**(1), 3588 (2018).
63. Jerby-Arnon, L., et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell* **175**(4):984-997.e24 (2018).
64. Tirosh, I., et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**(6282), 189-196 (2016).
65. Chung, W., et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Comm.* **8**,15081 (2017).
66. Venteicher, A.S., et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**(6332): pii:eaai8478 (2017).
67. Tirosh, I., et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**(7628), 309-313 (2016a).
68. Patel, A.P., et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**(6190), 1396-1401 (2014).
69. Filbin, M.G., et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**(6386), 331-335 (2018).
70. Li, H., et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**(5), 708-718 (2017).
71. Puram, S.V., et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**(7), 1611-1624.e24 (2017).
72. Shih, A.J., et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. *PLoS ONE* **13**(11), e0208778 (2018).
73. Miyamoto, D.T., et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **349**, 1351-1356 (2015).
74. Lambrechts, D., et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277-1289 (2018).
75. Uhlén, M., et al. Tissue-based map of the human proteome. Science 347(6220), 1260419 (2015).
76. The GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat. Genet. 45(6), 580-585 (2013).
77. Bausch-Fluck, D., et al. The in silico human surfaceome. Proc. Natl. Acad. Sci USA 115, E10988-E10997 (2018).
78. Gainer-Dewar, A. & Vera-Lincona, P. The minimal hitting set generation problem: Algorithms and computation. *SIAM J. Discr. Math.* **31**, 63-100 (2017).
79. Haedlicke, O. & Klamt, S. Computing complex metabolic intervention strategies using constrained minimal cut sets. *Metabolic Eng.* **13**, 204-213 (2011).
80. Haus, U.-U., Klamt, S. & Stephen, T. Computing knock-out strategies in metabolic networks. *J. Comput. Biol.* **15**(3), 259-268 (2008).
81. Jarrah, A.S., Laubenbacher, R., Stigler, B. & Stillman, M. Reverse-engineering of polynomial dynamical systems. *Adv. Appl. Math.* **39**, 477-489 (2007).
82. Klamt, S. & Gilles, E.D. Minimal cut sets in biochemical reaction networks. *Bioinformatics* **20**, 226-234 (2004).

83. Trinh, C.T., Wlaschin, A. & Srienc F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl. Microbiol. Biotech.* **81**, 813-826 (2009).
84. Ideker, T. Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.* **5**, 302-313 (2000).
85. Wang, R.S. & Albert, R. Elementary signaling modes predict the essentiality of signal transduction network components. *BMC Syst. Biol.* **5**, 44 (2011).
86. Zvedei-Oancea, I. & Schuster, S. A theoretical framework for detecting signal transfer routes in signaling networks. *Comput. Chem. Engineer.* **29**, 597-617 (2005).
87. Vazquez A. Optimal drug combinations and minimal hitting sets. *BMC Syst. Biol.*, **3**, 81 (2009).
88. 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* **5**(10), e13055 (2010).
89. Vera-Licona, P., Bonnet, E., Brillot, E & Zinovyev, A. OCSANA: optimal combinations of interventions from network analysis. *Bioinformatics* **29**, 1571-1573 (2013).
90. Pang, K., et al. Combinatorial therapy discovery using mixed integer linear programming. *Bioinformatics* **30**, 1456-1463 (2014).
91. Achterberg, T. SCIP: Solving constraint integer programs. *Math. Program. Comput.* **1**(1), 1-41 (2009).