Systems Medicine Lecture Notes

Uri Alon (Spring 2019)

https://youtu.be/BPPAEY44H-E

Lecture 9

Evolutionary tradeoffs, division of labor in tissues, and Universal cancer tasks

[Jean Hauser, Pablo Szekely, Noam Bar, Anat Zimmer, Hila Sheftel, Carlos Caldas, Uri Alon]

Introduction:

Cancer is an example of evolution within the body. Cells mutate, cease to be good citizens and start developing a new growing tissue. They keep mutating to adapt to their environment and escape the immune system.

Each cancer is a unique disease with its own mutations. Drugs that work well in one patient work less well in others. How can we make sense of this diversity? We will see that despite its messiness and scariness, cancer shows some understandable patterns based on the theory of evolution towards multiple simultaneous objectives.

Cancer arises from a series of mutations

Cancers are a growth of mutant cells that invades nearby tissues, and form new colonies in other tissues called **metastases**. Once metastases form, they are very difficult to treat with current approches; most metastatic cancers are unfortunately lethal.

Cancers are distinct from more benign growths of cells that don't produce metastases. An example of such growth are the hyper-sensing mutants we discussed in lecture 3: such mutants can grow to make, for example, thyroid nodules which secrete thyroid. The cells keep their original identity – they function according to their perceived signal. They do not colonize new tissues and make metastases. In contrast, cancer cells, such as thyroid cancer cells, typically lose most of their identity, stop secreting thyroid hormone, and invade other tissues, leading to a lethal disease.

Cancers typically occur in a multi-step process due to a series of mutations. A good example is colon cancer. Let's first talk about the normal colon tissue (Fig 9.1). The colon, the lower part of the intestine, acts to import nutrients and secrete anti-bacterial peptides. To increase its surface area, the intestine is made of columns of cells called villi (villus in singular). Villi have stem cells on the bottom, S, which differentiate to form the D cells that make up the walls of the villus. The D cell get pushed up and are removed at the top of the villus, such that each D cell lasts about 3 days. A chemical signal called wnt is secreted at the bottom and its concentration reduces toward the top of the villus. Wnt acts as a "height" signal (Fig 9.1). High wnt levels at the bottom make S cells divided to form more S cells, and at higher levels differentiated to make D cells. Thus, high wnt is a "stemness" signal. Note that the D cells have multiple tasks - to pump in sugars, fats and secrete peptides.

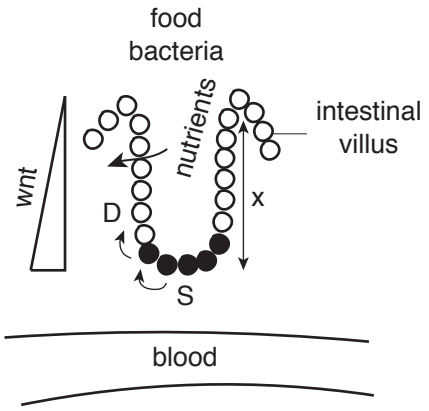

Figure 9.1

Colon cancer is quite common, with a prevalence of around 1%, due in part to the huge number of colon cells (about 100g or 10^{11}), and their fast turnover time: there are many divisions of S cells and hence many mutations.

The early steps in colon cancer are relatively easy to view because of endoscopy, in which a microscope is inserted into the colon. Bert Vogelstein used endoscopy to define what is now called the "**Vogelgram"** (Fig. 9.2): in terms of tissue shapes (histology), normal tissue becomes a growth that bends out to form a polyp, which then becomes a mass of ugly looking cells, and finally breaks into the neighboring tissue and into the blood stream to form metastases. This process takes years.

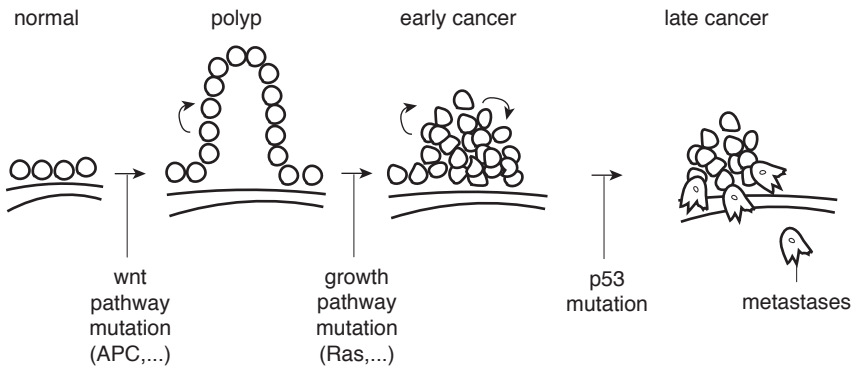

Figure 9.2

Transitions between stages are due to specific mutations. These mutations cause a gene to form a mutant protein that is over- or under-active.

The first mutation is usually in the pathway sensing wnt (APC or other mutations). It makes S cells stop differentiating and proliferate more. This forms the polyp. More cell divisions mean more mutations. The next stages are mutations that activate pathways for cell division (Ras or other mutations). The next steps include a step that makes the cell increase its own mutation rate! This is usually in the gene called the "guardian of the genome", p53. The normal function of p53 is to makes cells kill themselves or turn senescent if they are damaged. With mutant p53, cells make a lot more mutations and DNA rearrangements and survive. Thus, cells now explore more of mutation space, and can adapt to more environments. The cells that grow faster multiply and pass their genomes to their daughter cells.

We can ask about the mutations in the cancer in its original primary site and in its metastases. Work on sequencing the DNA of multiple tumors from the same patient that all arose from the same primary tumor show a tree of mutations (figure 9.3). The original cancer has some "trunk mutation" (blue line in Fig 9.3) and then the different metastases have different combinations of mutation. From the end-point tumors you can infer the evolutionary tree of mutations (Fig 9.3). Some of these mutations do not drive the cancer growth but are instead random "**passenger**" mutations. Other mutations "**driver mutations**" help the cancer adapt to its new niches and changing situation. Telling passengers from drivers is an important question for cancer research, because you can design drugs to target the driver mutant proteins and help to slow cancer.

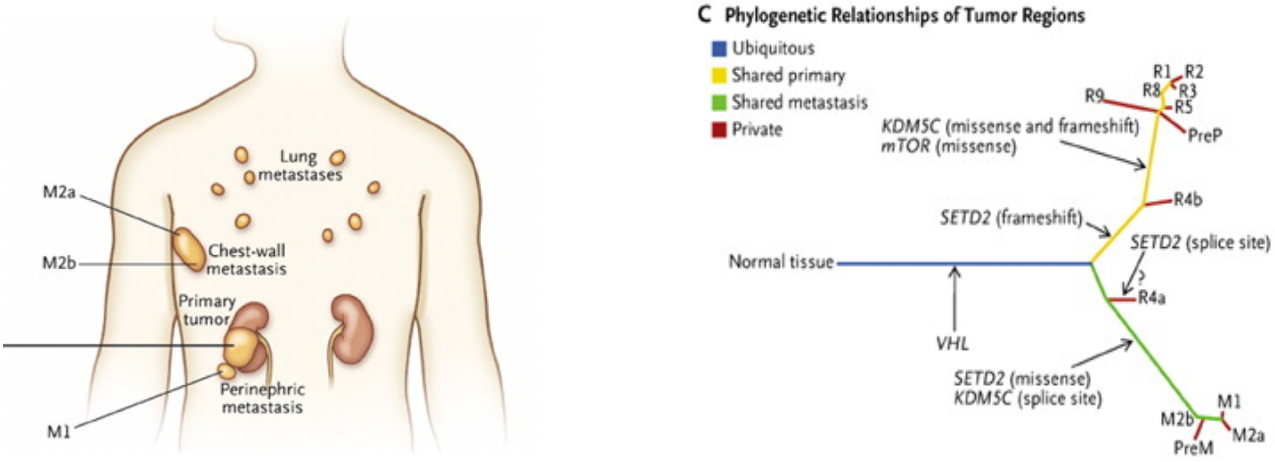

Figure 9.3

Cancer depends on a microenvironment

As the cancer grows, it organizes the tissue around it to form a **micro-environment** by making normal cells work for the cancer. The cancer microenvironment is similar to the hot fibrosis of the previous lecture (Figure 9.4). It is rich with collagen, macrophages and myofibroblasts --called tumor associated macrophages (TAMs) and cancer associated fibroblasts (CAFs).

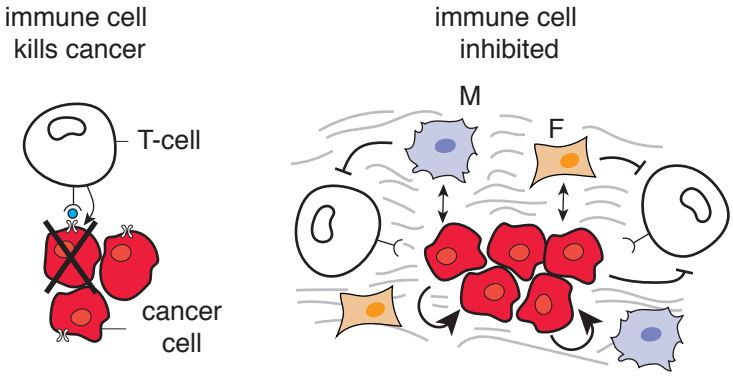

Figure 9.4

The microenvironment helps protect the cancer from the immune system. Indeed, we get micro-cancers all the time, but they are eliminated by the innate immune system like NK cells, and the adaptive immune system like T-cells that sense the mutant proteins made by the cancer cells (neoantigens). But the tumor microenvironment somehow downregulates these immune functions protecting the cancer. There seems to be threshold size, above about one million cancer cells, in which cancer is no longer eliminated by the immune system, due to its self-created environment. Instead, the growing cancer reaches co-existence with the body for a while.

Indeed, recent advances, called **cancer immunotherapy**, prevent immune system to be lulled to sleep by the cancer, and help the immune system attack the cancer cells. Immunotherapy works by inhibiting the checkpoints that shut off the immune system such as receptors on T-cells that inhibit their activity. Unleashing T-cells causes side effects of auto-immune diseases, mainly the endocrine autoimmune diseases we discussed in lecture 3. It is mysterious why immunotherapy currently works only for a few types of cancer (melanoma, lung) and not for others, and only for a small percentage of patients.

As cancer grows it lacks oxygen and needs to call in new blood vessels and to rearrange the extracellular matrix so it can grow and expand within its fibrosis-like environment. The microenvironment also helps cancer arrange this vasculature.

Cancer incidence rises with age

The incidence of most cancers rises exponentially with age, and drops at very old ages. An early theory for this rise with age was the **multiple-hit theory**, in which the chance to get k mutations in the same cell goes as time to the power k, t^k. Such a power law looks similar to the exponential rise in incidence for at least part of the range. Indeed, childhood cancers usually arise because one of the mutations is already present in the fertilized egg, and so the mutation is present in all cells of the body, increasing

greatly the chances of getting all k mutations in the same cell. Such mutations are **genetic risk** factors for cancer. A well-known example is mutations that reduce DNA repair in cells, like BRCA1, that increases risk of breast cancer to about 50-70%.

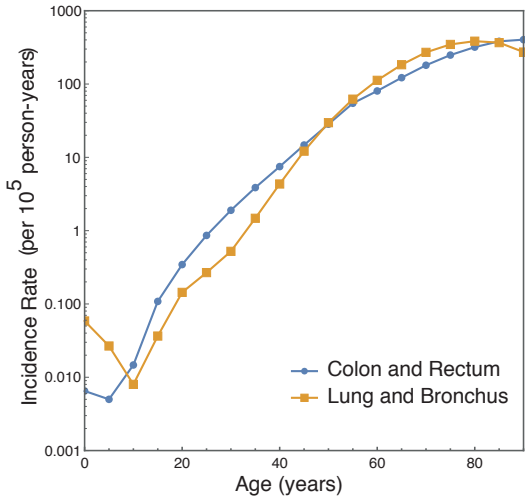

Figure 9.5

Modern theories note that even cancers based on a single mutation, like a leukemia called CML, have an exponential rise with age. Thus, multiple hits do not explain all of the age incidence. It is likely that changes in cancer removal by the immune system play a role: as we discussed in lecture 5, rising senescent cell numbers (SnCs) saturate the immune system that removes them; this part of the immune system, like NK cells, also has a role in removing cancer, and thus preoccupation with SnC makes cancer more likely to cross the threshold and establish itself. The function of the adaptive immune system such as T-cells also declines with age. Furthermore, rising inflammation with age due to SnC secretions helps cancers get the fibrosis-like environment they need- inflammation often is a precursor to cancer (e.g. colitis in the gut, inflammation in the pancreas). Thus, a threshold-crossing, first-passage-time mechanism similar to lecture 7 might explain part of the age dependence of cancer.

Along with genetic factors, there are **environmental factors** such as UV in the skin, smoking in the lungs, and toxins in the colon. These factors increase the chance of mutations and lead to more cancer in that organ. Factors that cause inflammation like acid reflux in the throat or liver inflammation and fibrosis also increase the risk of cancer.

Cancer can arise in different tissues- lung, breast, colon and skin being among the most prevalent. Cancers share certain **hallmarks of cancer**: they have dysregulated growth, they need to deal with the immune system, and they need to call in blood vessels, and so on. Thus, cancers need to multi-task: each different niche in the body and each stage over time requires different tasks at different degrees of importance.

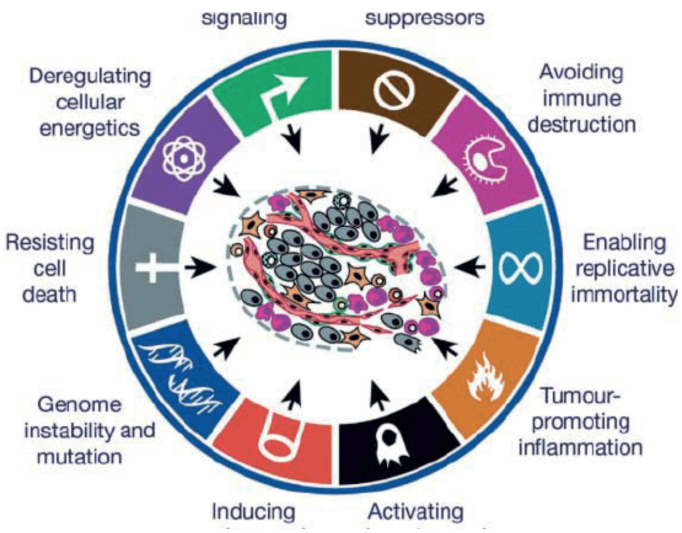

Figure 9.6

We want to address several questions:

Can we make sense of the variation between tumors?

Why do some tumors respond well to some drugs and not others?

What do driver mutations do, and how can we tell them apart from passenger mutations?

To do so, we will take a look at cancer from the point of view of evolution inside the body. We will first consider healthy tissue and its multiple tasks, and then return to cancer.

Tissues require performance at several tasks

Let's consider a healthy tissue such as the intestinal villus of Fig 9.1. It has multiple tasks. The D cells (enterocytes) need to pump in different nutrients, to secrete different antimicrobial peptides and so on.

Each D cell expresses genes to make proteins such as the pumps, the ribosomes that make new proteins, and so on. A cells' gene expression is thus described by 20,000 numbers, the expression of gene 1, gene 2, and so on, for all of the 20,000 genes in the human genome (about two thirds of these numbers are zero in a typical cell, for example many gene are expressed only in brain). The cell can be described therefore as is a vector G in gene-expression space that has 20,000 dimensions. In fact, experiments called single-cell mRNA sequencing can measure mRNA from each individual cell and produce this kind of high dimensional data. We can plot each

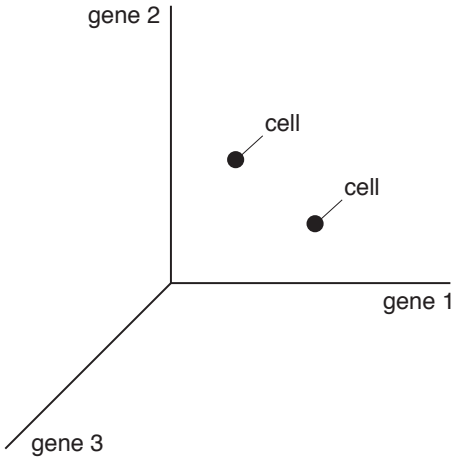

Figure 9.7

cell as a point in gene-expression space. Fig 9.7 shows the vectors of two cells, using three of these dimensions.

Let's consider one task of the cells, let's say to pump in sugars. Each task has a **performance function**, for example the number of sugar molecules pumped per unit time. Performance is a function of the cells gene expression, P(G). There is a gene expression profile G* that is optimal for this task, and maximizes P(G). This optimal vector has lots of sugar pumps, a few ribosomes to make more pumps, and so on. Performance decreases if we change gene expression and move away from G* (Fig 9.8). For example, making too few pumps reduces pumping rate, whereas making too many pumps does not leave enough gene expression for ribosomes to make essential proteins for cell survival, and so on. We call G* the **archetype** for the task – the optimal gene expression. We can plot contours of performance that drop away from G* (Fig 9.8). These contours are like a topographical map describing a hill whose peak is G*.

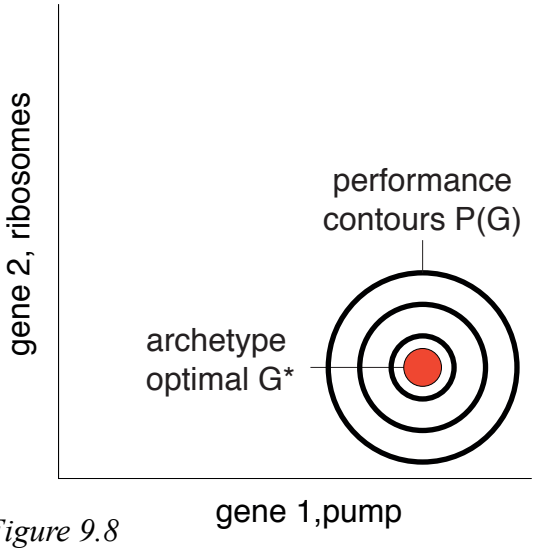

Figure 9.8

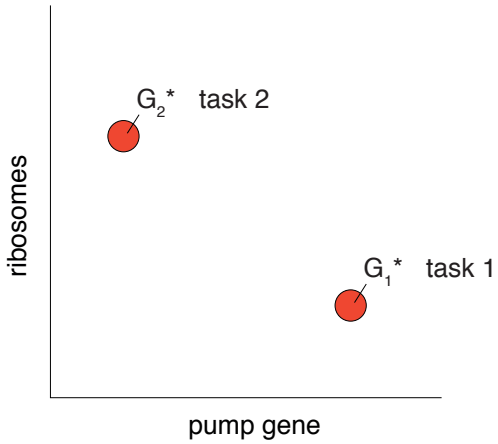

Figure 9.9

If that was the only tasks, all cells would have G=G*, at the top of the performance hill. Cells would from a cloud (due to noise) around G* in gene expression space.

What if the cells need to do two tasks, such as pumping in sugars and secreting peptides? Each task has its own performance function, which we call P1(G) and P2(G). Performance at task 1 is maximal at archetype G*1, and performance at task 2 at archetype G*2. For example, for making and secreting peptides one needs many ribosomes and few pumps. The archetypes are therefore at different points (Fig 9.9). Where would the cells be in gene-expression space? You can't do both tasks optimally with a finite amount of gene expression. If cells were animals, we might say that 'there is no animal that can fly like an eagle, run like a cheetah and swim like a dolphin'.

To resolve this dilemma, we need to think of **total tissue function**, which combines the performance at the two tasks. Since the tissue is made of many cells, the performance at each task is the sum of the efforts of all the cells (Fig 9.10). Since cells lie at different spatial positions in the tissue, we also need to take into account the spatial gradients along the tissue that can affect performance: some positions might have for example more oxygen because they are closer to arteries, other positions might make a certain function more important to the tissue and so on. Thus, each performance at taks i can be multiplied by a spatial factor $\phi_i(x_k)$, where x_k is the position of cell k. The better the task can be performed at that position, the higher phi.

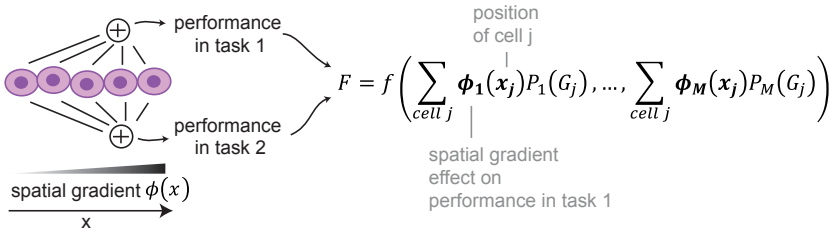

Figure 9.10

The total performance at task 1 from all cells is thus $S_1 = \sum_k \phi_1(x_k)P_1(G_k)$ where G_k is gene-expression of cell k. Similarly, the summed performance at task 2 is $S_2 = \sum_k \phi_2(x_k)P_2(G_k)$. Total fitness is a function of these summed performances

$$F = f(S_1, S_2)$$

The function f is not usually known- it depends on the situation: whether task 1 or task 2 is more important at a given time. The nice thing is that we don't need to know anything the exact form of the function f. One can be sure of one thing: is we improve performance at both tasks, we increase total fitness. Thus, f is an increasing function of S_1 and S_2.

We also don't need to know the exact functional forms of the performance function or of the spatial gradients phi(x) to make progress, as we will now see.

Optimal gene expression falls on a line for two tasks, a triangle for three tasks, a tetrahedron for four

We can finally ask- what is the optimal gene expression for cell k, Gk. To do this, we need to solve for the maximum of total tissue performance F, using the well-known condition for a maximum

$$\frac{dF}{dG_k} = 0$$

(the second derivative of F also needs to be negative). It turns out that there is an easy solution if we make an assumption about the performance functions. This assumption can be dropped with only minor effects on the solution. The assumption is that all of the performance functions decay with Euclidean distance from their maximum, the archetype G_i^*. Thus, $P_i(G_k) = P((G_k - G_i^*)^2)$. Another way of saying this is that the contours of the performance functions are circles, like in Figure 9.8.

If we make this assumption, we can find that the optimal solution is that all cells to fall on the line between the two archetypes in gene-expression space. To see this, imagine that one of the cells has gene expression G_k that is off the line, at point B (Fig 9.11). The performance of that cell at task 1 is determined by the distance to archetype 1, and likewise for task 2. By geometry, there is a point A on the line that is closer to both archetypes. That point therefore has higher performance in both tasks. Therefore, total function F would improve if we move the cell to the line. We can thus erase point B – it not part of the best solution.

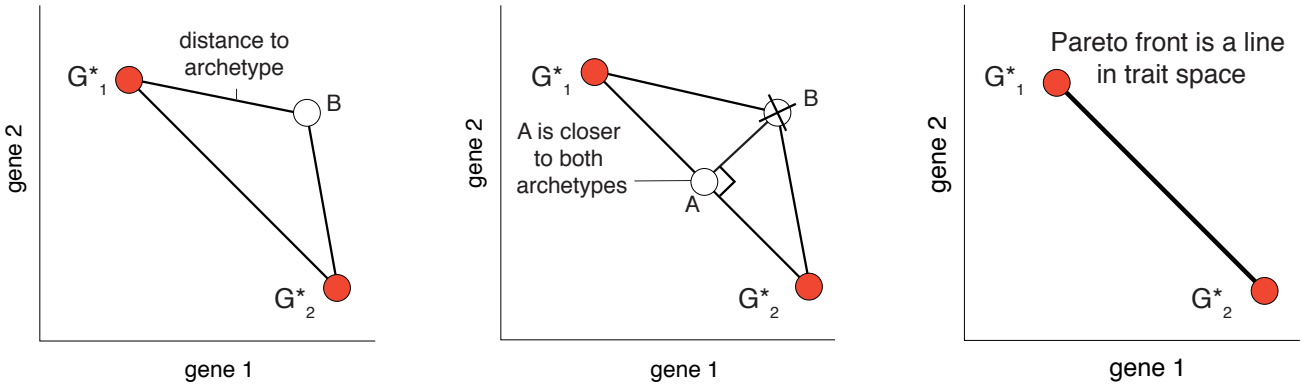

Figure 9.11

Now, there was nothing special about point B, so we can erase all points except for the line segment between the archetypes. This is where the optimal solutions lie. It is called the **Pareto front** in optimization theory.

If we relax the assumption that performances drop with Euclidean distance from the archetype, we typically get slightly curved lines (except for some pathological cases).

What if we measure 20,000 genes and not only two? The result is still a line in 20,000 dimensions. We can view it by plotting any two genes, because a projection of a line on a plane is still a line (Fig 9.12a).

What if we have three tasks? A similar argument shows that all cells must lie within a triangle, whose vertices are the three archetypes. Thus, gene expression data should fall on a plane, and within that plane on a triangle (fig 9.12b). Four tasks lead to a tetrahedron (Fig 9.12c). For a mathematical proof see solved exercise 1.

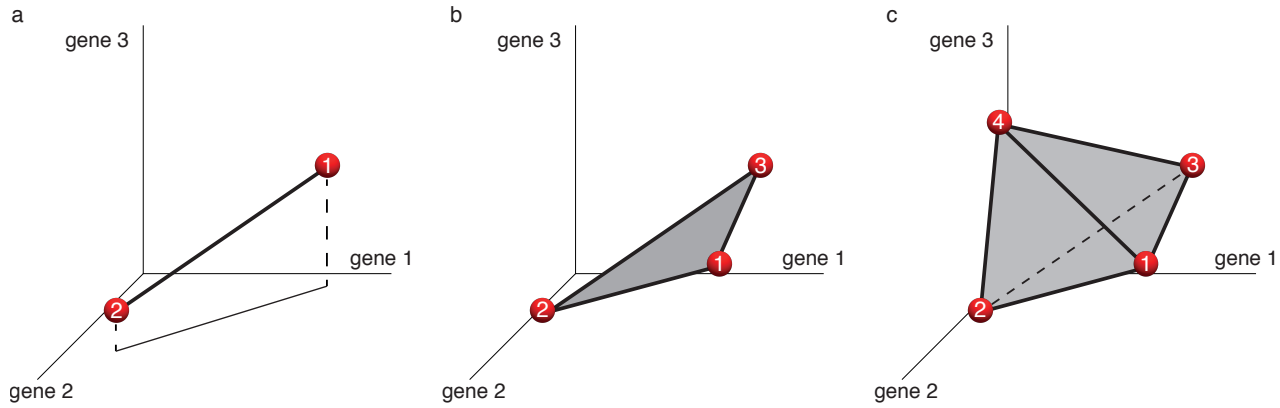

Figure 9.12

This approach can be used to infer the number of tasks. If you find that data falls, for example, on a plane (much of the variance, for example, is explained by two principle components), and within that plane on a triangle, you can starts to guess that there are three tasks. But his it is still not enough- perhaps the triangle is due to other reasons. The theory predicts that cells near each vertex should be specialists at a specific task (Fig 9.13). Cells in the middle should be generalists. This approach, called **Pareto Task inference (ParTI)** thus allows one to infer directly from the data how many tasks there are, and what the tasks might be- according to the genes expressed by the cells near each archetype [Shoval 2012,Hart 2015, Korem 2016, Adler 2019].

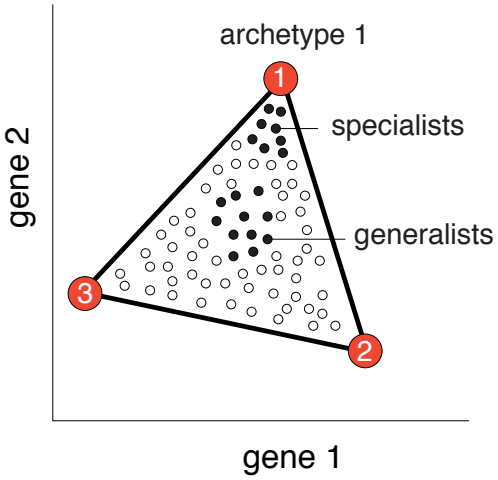

Figure 9.13

There are ParTI algorithms that can detect the data dimensionality, and the best fit polyhedron and its statistical significance (the chance that random data would give such a polyhedron). The software can thus tell the number of potential tasks, and provide clues to what the tasks are by the genes expressed in cells near each archetype. See an R version at https://github.com/vitkl/ParetoTI.

This theory does not tell us where in the polyhedra (line, triangle tetrahedron, etc.) the cells should fall. In some cases, if the tasks are incompatible, the cells will only fall at the vertices. We only have specialists. In this case, a tissue would produce different cell types, with mutually exclusive gene expression. Gene expression space shows distinct well separated clusters, one for each cell type (that is in fact how cell types are now defined). In other cases, we will see a continuum within a single cell type, with both specialist cells and generalist cells.

Intestinal enterocytes fall on a one-dimensional continuum with three tasks

Let's look at data for individual intestinal cells, collected by Ido Amit and Shalev Itzkovitz and studied by Miri Adler and colleagues (Adler et al 2019). The gene expression of individual mouse enterocytes falls on a plane, and within the plane on a V-shaped line. This suggests a triangle with three vertices, and hence three tasks. The cells fall in a continuum along two edges, from task 1 to 2 and from 2 to 3 (Fig 9.14, plotted in the space of the first two principal components of gene expression).

The three tasks are (1) cell adhesion and pumping in fats (lipids), (2) pumping in sugars (carbohydrates) and amino acids, and (3) defense- secreting anti-bacterial peptides.

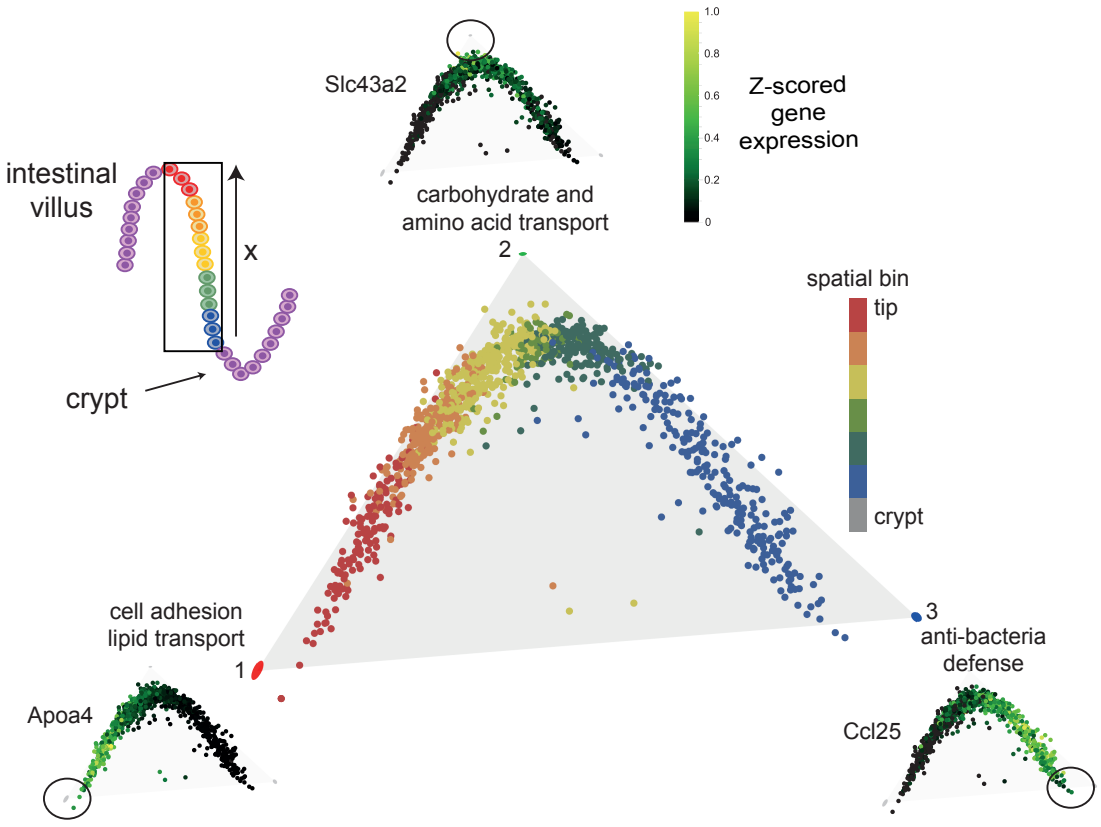

Figure 9.14

Itzkovitz and Amit also mapped where each of the cells lies along the intestinal villus. It turns out that the cells that do task 1 (adhesion/lipids) are at the top of the villus, task 2 (sugars/amino acids) lie in the middle, and task 3 specialists (defense) at the bottom. This arrangement makes sense: cell adhesion is important at the top where the cells fall off, dealing with lipids requires less oxygen than dealing with sugars, and there is less oxygen at the tip of the villus than in the middle. Bacterial defense is most important at the bottom, in order to defend the stem cells S. Thus, the space-dependent factors we mentioned above, $\phi_i(x)$ come into play. The optimal division of labor is to do each task at the positions where it is done best, or where it is most important to the tissue.

The broken-line-shaped continuum is a reflection of the fact that this tissue is effectively one-dimensional: the height along the tissue x. Such broken lines in gene-expression space are found also when there is a temporal development (e.g. going for stem cells to a series of differentiated cells). Here time t is the one-dimensional parameter that affects ϕ_i. Algorithms can generate a "pseudo-time" based on such data.

Liver cells show a continuum inside a tetrahedron

What about other polyhedra? A nice example is found in the liver, a classic multi-tasking organ. The liver is made of hepatocyte cells. These cells have dozens of tasks. Hepatocytes make glucose in several ways, store glucose, metabolize lipids, secrete many blood proteins, vitamins and hormones, detoxify the blood, turn ammonia to urea, produce antioxidants like glutathione …phew… the list goes on. How do they manage these tasks?

Itzkovtiz and Amit analyzed individual hepatocytes and, with Adler et al, found that these cells fall in a tetrahedron in gene expression space (Fig 9.15). The four tasks are making blood proteins, making glutathione,

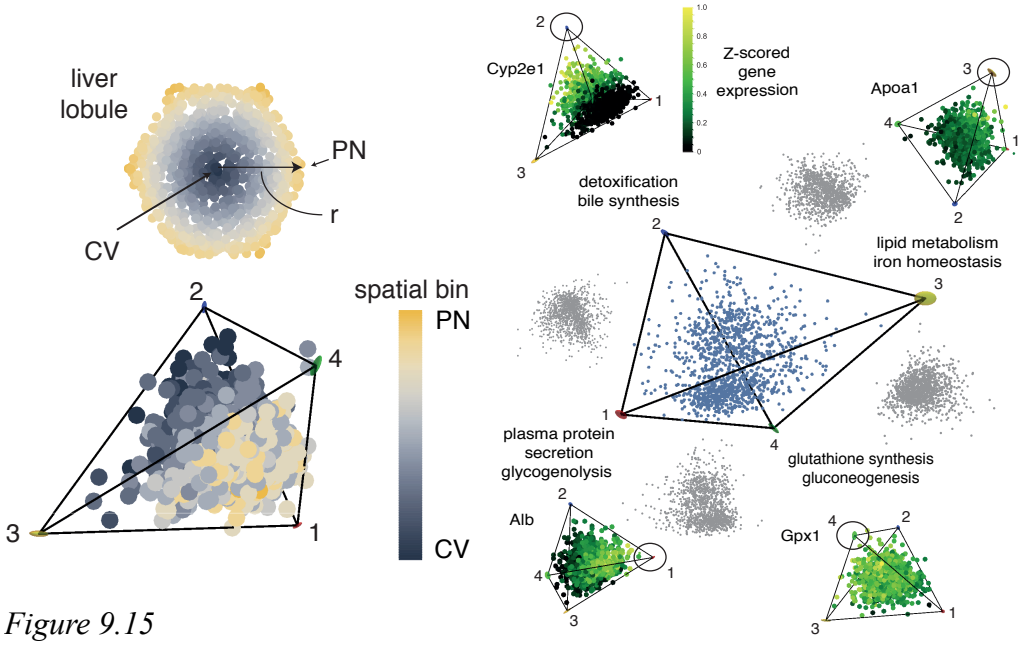

Figure 9.15

detoxifying blood, and surprisingly, iron homeostasis/lipid metabolism. Each archetype has additional secondary tasks. Many cells in the middle of the tetrahedron are generalists that multi-task.

We can ask about how these tasks are arranged in space along the tissue. The liver is made of repeating hexagonal columns called liver lobules. A cross-section shows the hexagon which is about 15 cells across, with blood input at the periphery and blood output in the center the central vein (Fig 9.16). The oxygen-requiring tasks of synthesis of proteins and antioxidant are done in the oxygen rich periphery; detoxification is done at the oxygen-poor center. The remaining iron/lipid ask is done in the middle regions of the lobule. These middle-region localizations may reflect an prediction of the theory that a task with no spatial preference (constant phi(x)) will be relegated to areas which are bad for the other tasks. Real-estate is thus optimally divided between the tasks.

The 3D-nature of the filled tetrahedron suggests that there are gradients in 3D in this tissue, depending on the three spatial dimensions. An exact account of these gradients remains to be discovered.

Many types of cancers fall on polyhedra in gene-expression space

We now return to cancer. We might expect that multi-task evolution theory may apply to cancer because cancer is a case of intense evolution inside the body that can play out over years, with generation times of cells that can be on the order of days (Driessens *et al.*, 2012; Gillies, Verduzco and Gatenby, 2012; Labi and Erlacher, 2015; Lan *et al.*, 2017). Furthermore, cancer cell growth and survival is conditioned on fulfilling multiple tasks, including growth, stress resistance, interaction with the immune system and so forth, as described by the concept of hallmarks of cancer (Hanahan and Weinberg, 2011) . Each task requires a different profile of gene expression – ribosomes for growth and stress proteins for survival. Presumably no tumor can be optimal at all tasks at once, because cells can only make a limited amount of protein per unit biomass, and proteins for different functions can interfere with each other. Thus, cell communities that optimally manage the trade-off relevant for their particular niche in the body will outgrow and out-survive cells that are suboptimal.

To test such ideas, it is useful that cancer research benefits from excellent public datasets. One such resource is TCGA which has gene-expression data from many different types of cancer (colon, liver lung etc.), and for hundreds of tumors from each type of cancer. Here, each tumor was ground up and the data is the averaged over all cells in the tumor, unlike the single-cell data we just discussed. Single cell data is rapidly becoming available also for cancers. Jean Hauser applied ParTi to cancer data to search for task and tradeoffs.

Out of the 15 cancer types that have at least 250 primary tumor samples in TCGA, about half the show no convincing polyhedra. These cancer types include kidney renal clear-cell carcinoma and ovarian cancer. The data looks like a cloud in gene expression space. This could be due to several reasons: there might be only one major task. There might be too many tasks and thus too many vertices to resolve. There could be hidden heterogeneity not currently understood (eg a mix of several types of tumors), or it could be that the theory does not apply.

The other half (8/15) are well described by polyhedra with 3-5 archetypes (Fig 9.16). Statistical tests suggest that these polyhedra fit the data much better than they do shuffled data (Fig 9.16, inset shows that real data is fit by a smaller triangle than shuffled data). These cancer type include breast, colon, thyroid, bladder, low-grade glioma, liver, lung and head&neck cancer.

The tumors thus fall on a plane for the case of triangles, on a 3D subspace in the case of tetrahedra, and so on. When I say 'fall', I mean that these low dimensional subspaces explain a surprisingly high amount of variation from the full 20,000-dimensional space- typically 20-40% of the variation for these cancers. Data is thus flattened like a pancake along, say a plane, but still with some width sticking out of the plane. When projected on the plane, their distribution fills out a shape with sharp corners. The corners are the archetype gene–expression profiles.

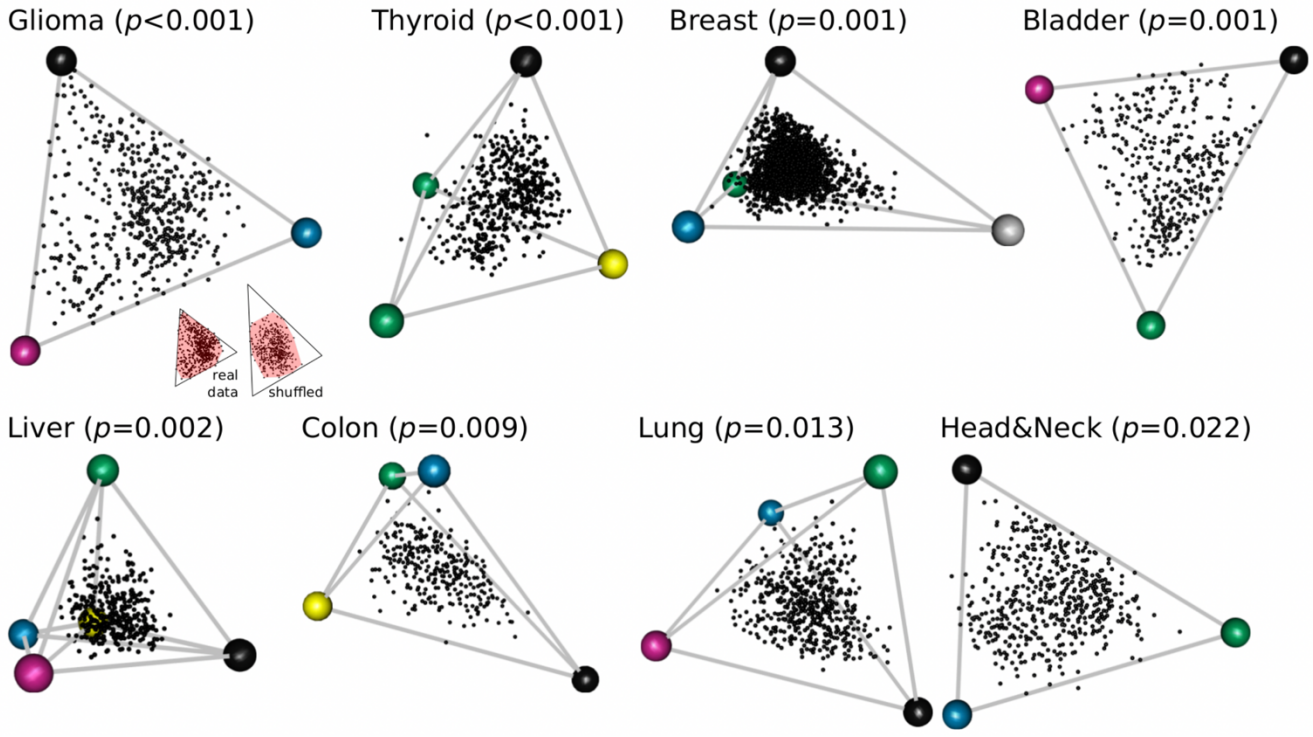

Figure 9.16

There are five universal cancer tasks shared by many cancers

It turns out that the different cancer types (liver, lung etc) have archetypes that are similar in terms of gene-expression. This raises the possibility that that tumors from different cancer types face similar trade-offs between similar tasks.

In order to test this requires comparing cancers from different tissues. This is made difficult by the fact that the tissue-type effects gene expression strongly. Each tissue has its own specific genes and its own biases (more or less ribosomes etc.) due to its different functions. To correct for this, one can pool the 3180 primary tumors from the different cancer types, after correcting gene expression profiles for tissue identity. This correction is done by normalizing each sample by the mean expression profile of its cancer type. This normalization is done by dividing the expressing of each gene by its average over all tumors form the same tissue. Typically, we plot gene expression in log space, since genes have widely different gene expression and log space makes it easier to compare them. In log space, this normalization means that you subtract the mean gene expression vector for each cancer type.

After this correction, the tumors fall in a continuum inside a polyhedron bounded by five archetypes (9.17a). Tumors from different cancer types are spread widely within the polyhedron and are found close

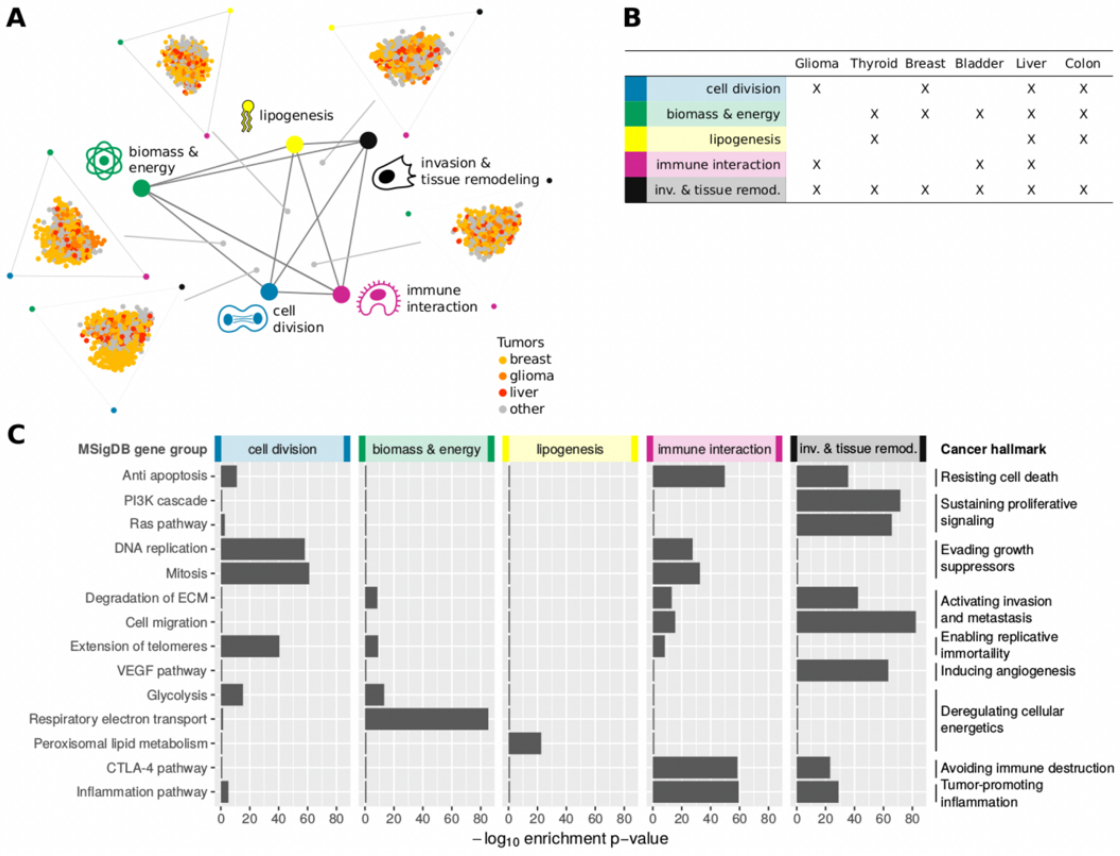

Figure 9.17

to 3-5 of the archetypes depending on the cancer type (9.17b). Tumors are not found in the immediate vicinity of archetypes for statistical and biological reasons.

The tasks performed by these five universal archetypes are revealed by determining the pathways and functional gene groups that are expressed highest in the tumors closest to a given archetype (using software that groups genes into biological functions, such as MSigDB *(Subramanian et al., 2005)*). One can also determine which clinical properties are frequent among the tumors closest to a given archetype compared to other tumors in the dataset.

There are clear tasks for each of the five archetypes (Figure 19.17c). The five tasks are: cell division, biomass/energy production, lipogenesis (making cell membranes), immune-interaction and invasion/tissue remodeling in which the cells dissolve their surrounding matrix and start to move to colonize other tissues. The tasks match the hallmarks of cancer defined by Hanahan and Weinberg (9.17C). A given hallmark can contribute to one or more tasks.

There seems to be a temporal sequence, in which early tumors (Stage II tumors) are found near the cell division, biomass/energy and lipogenesis archetypes whereas late-stage tumors (Stage III tumors) are found near the immune-interaction and invasion/tissue remodeling archetypes. This temporal sequence may stem from changes in the trade-offs during tumor progression, with growth/proliferation being more important during early stages and survival during later stages. Finally, tumors from patients with higher number of invaded lymph nodes (are sign of movement away from the primary cancer site towards making metastases) are found near the invasion/tissue remodeling archetype.

Not all tasks are universal: Breast cancer shows three universal archetypes (division, invasion, biomass & energy) and a fourth one that is tissue specific (Her-2 positive tumors).

Specialist tumors are sensitive to drugs that interfere with their task

If indeed tumors close to an archetype are specialists at a task, one can reason that these tumors will be most sensitive to drugs which specifically disrupt that task. To test this, we can use the blessing of public datasets from large experiments that asked how cells from many different tumors are effected by many different drugs. In these experiments, a single cell from a tumor is grown in plates to make a cell-line. The cell line is divided into many plates, and in each plate the cell-line is are challenged by a different drug. The cells growth and survival are monitored. You can take the gene expression of these cells-lines from different tumors, and project it on the tumor tetrahedron.

Indeed, the cells whose expression is closest to an archetype are most sensitive to drugs that disrupt that task (Fig 9.18). Cell lines closest to the cell-division archetype are sensitive to ixabepilone which stabilizes microtubules, and thus targets mitosis (Fig. 2G). Cell lines near the biomass&energy archetype are most sensitive to drugs which inhibit mTOR (Fig. 2H), a controller of cell growth *(Laplante and Sabatini, 2012)*. Similarly, cell lines close to the breast-cancer-specific HER2 archetype (tumors that over-expresses the erbB-2 receptor) are sensitive to Herceptin, an erbB-2 inhibitor and so on. This differential sensitivity to drugs supports the hypothesis that tumors close to archetypes are task specialists.

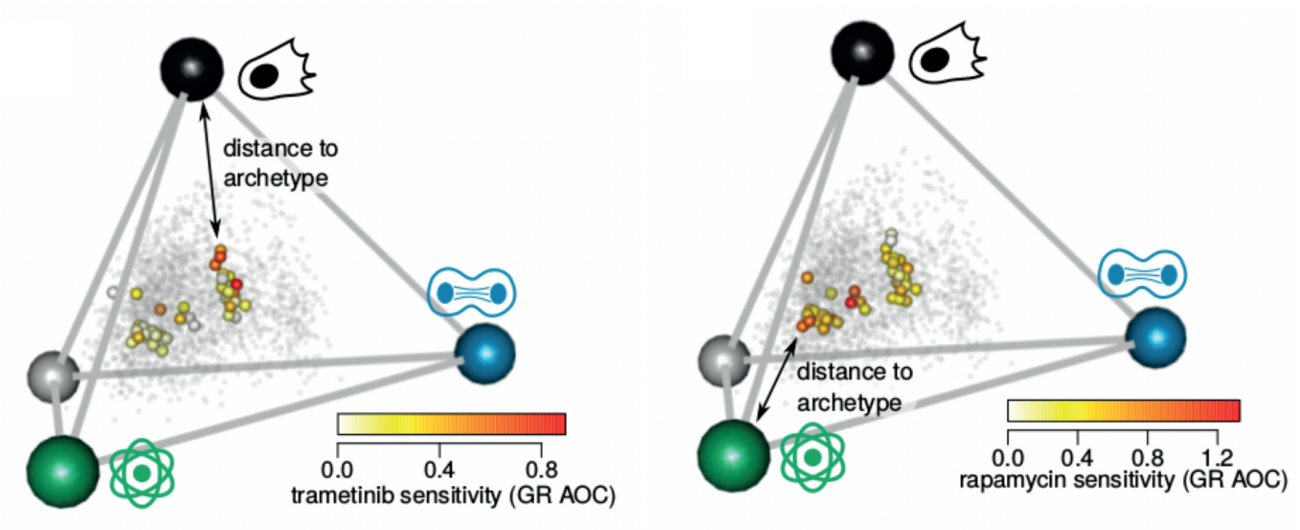

Figure 9.18

Driver mutations are like knobs that push gene expression to specialize in specific tasks

So far, we looked at gene expression, tasks and drug sensitivity. What about the diversity of mutations between different tumors? How do genetic alterations in tumors fit into the trade-off picture? Genetic alterations are mutations, deletions of whole pieces form the DNA, duplications of pieces and so on. We will call all of these changes in the DNA 'mutations' for simplicity.

The effects of a given mutation in a cancer of a certain type can be described by the **effect vector** of a mutation. You take all tumors with that mutation, compute their mean gene expression G_{mut} , and subtract the mean of expression all tumors without the mutation G_{no} . This vector, $E = G_{mut} - G_{no}$ describes how the mutation affects gene expression on average (schematically shown in Fig. 9.19).

One can now compare this effect vector with the polyhedron for the cancer type. There are two possible situations: the effect vector can align with the polyhedron, or instead can point away from the polyhedron. To visualize this, if the front were a triangle on a plane, the effect vector could lie on the same plane (have a small angle with the plane), or could point away from the plane (have a large angle) (Fig 9.19). Importantly, shuffled controls, in which the mutation data is shuffled between cancers, typically have effect vectors that point away (angle=60º-80º), because the polyhedron explains only a fraction (20-40%) of the variation in the data.

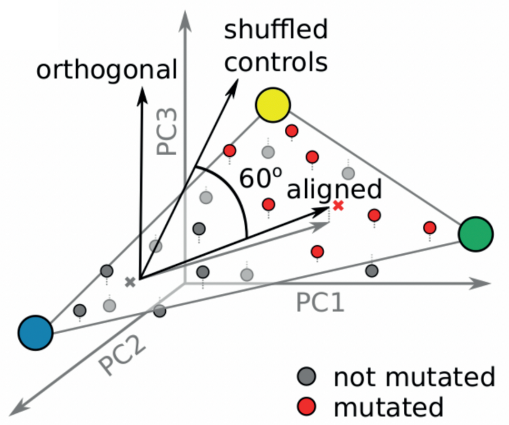

Figure 9.19

Strikingly, the effect vectors of the best-known driver mutations align with the polyhedron much more closely than expected from shuffled data in glioma, thyroid cancer, breast cancer, bladder cancer and colon cancer (Fig 9.20). Drivers are much more aligned than non-driver cancer genes and passenger mutations. This alignment suggests that driver mutations are involved closely with the tasks. We can then ask- how do they relate to the tasks?

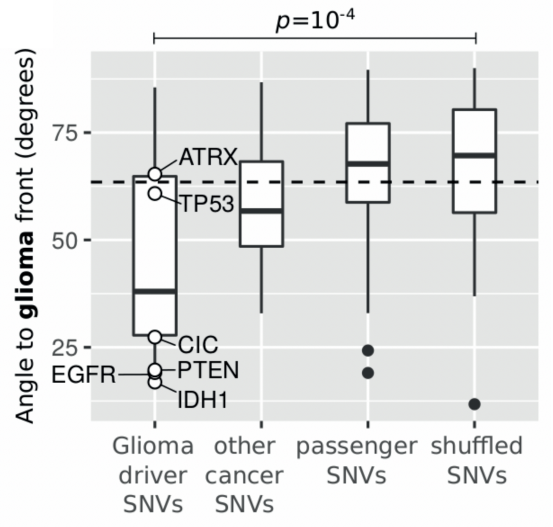

Figure 9.20

Interestingly, driver mutations move gene expression towards specific archetypes (fig 9.21). For example, *IDH1*, a strong driver in glioma, shifts gene expression towards the cell-division archetype (Fig. 9.21). In breast cancer, the common *P53* mutation

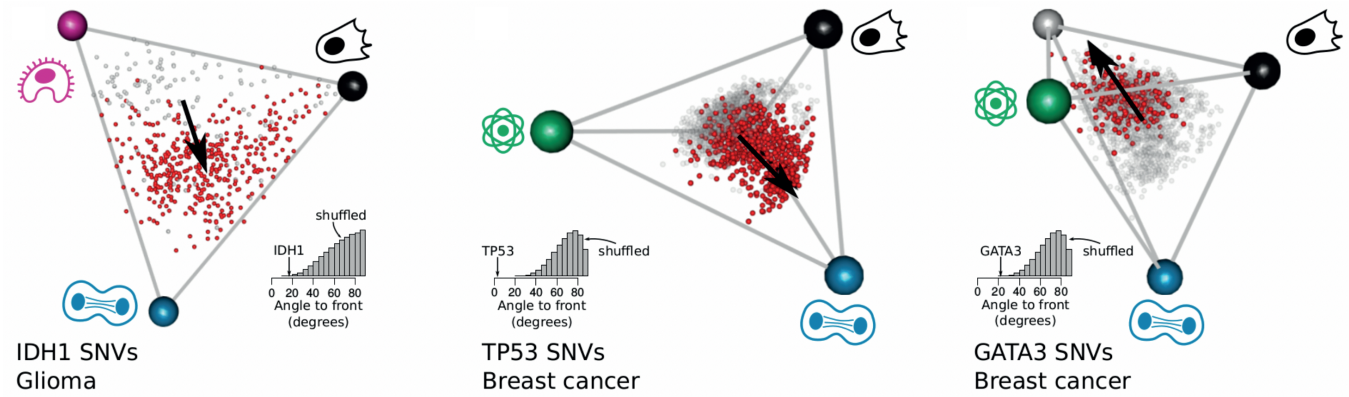

Figure 9.21

is the most aligned with the front. It points directly towards one archetype, cell division (angle to archetype = 18°). Mutations in *P53* in breast cancer and *IDH1* in glioma thus coordinate gene expression towards specializing in the cell-division task. Another breast cancer driver, *GATA3*, shifts gene expression towards the face defined by the lipogenic, *HER2* and invasion/tissue remodeling tasks and away from the cell-division archetype. The same conclusion is found for all 21 drivers with significant alignment to the polyhedra.

Thus, aligned driver mutations can be interpreted as knobs that tune gene expression towards some tasks and away from others. Although the tasks are universal, the drivers that shift gene expression toward each task are often tissue-specific.

Summary

Evolution under multiple tasks offers a framework for understanding division of labor between cells in healthy tissues. It also offers a way to understand the diversity between tumors. Tumor gene expression lies in a continuum in a polyhedron whose vertices are archetypal expression programs for universal cancer tasks that recur in different cancer types. Tumors can be specialists at a task or generalists: specialists have gene expression close to a vertex and generalists lie in the middle of the polyhedron. Specialists in a task seem to be more sensitive to drugs that disrupt that task. Driver mutations are like knobs that tune gene expression towards specialization in specific tasks. This framework offers a way to understand tumor variation in terms of task specialization.

Exercise 1:

Prove that if fitness is an increasing function of k performance functions in gene space, and that each performance function $i = 1 \dots k$ has a maximum at a point called archetype G_i, and that performance drops with Euclidean distance from the archetype d_i, then the point of maximum fitness is found inside the polytope defined by the k archetypes (Shoval , 2012).

Solution:

Each phenotype is described by a vector of genes $\vec{G}$ (for convenience we will drop the vector sign from now on). Fitness F is an increasing function of the performance at the k different tasks, $F(G) = F\big(P_1(G), P_2(G), \dots, P_k(G)\big)$. Each performance function P_i has a maximum at archetype i, G_i, and performance decreases with Euclidean distance from the archetype $P_i(G) = P_i(||G - G_i||)$. We will show that the optimal phenotype (the phenotype that maximizes F) is a

(i) a weighted average of the archetypes.

(ii) Weights are positive and sum to one.

which means that the optimal phenotype is inside the polytope defined by the k archetypes. Another way to say this is that the optimal phenotypes are convex combinations of the archetypes.

The optimal phenotype maximizes fitness, $dF/dG = 0$. Let's denote the distance from the archetype of task i by $d_i = ||G - G_i||$, so that $dd_i/dG = 2(G - G_i)$. Using the chain rule,

$$\frac{dF}{dG} = \sum_i \frac{dF}{dP_i}\frac{dP_i}{dd_i} 2(G - G_i) = 0$$

Thus, the optimal G that solves this equation, G_{opt} , is a weighted average of the archetypes

$$G_{opt} = \sum_i \theta_i G_i$$

with weights

$$\theta_i = \frac{\frac{dF}{dP_i}\frac{dP_i}{dd_i}}{\sum_j \frac{dF}{dP_j}\frac{dP_j}{dd_j}}$$

where all derivatives are at G_{opt}. Note that the the weights sum to one, $\sum \theta_i = 1$. The weights are positive $\theta_i > 0$ because F increases with performances (so that $\frac{dF}{dP_i} > 0$) and performance decreases with distance from the archetype ($\frac{dP_i}{dd_i} < 0$) so that all terms have a negative sign which cancels out. Hence

(i) The optimal phenotype is a weighted average (convex combination) of the archetypes

$$G_{opt} = \sum \theta_i G_i$$

(ii) The weights are positive and sum to one ($\theta_i > 0$, $\sum \theta_i = 1$). Another way to say this is that the k maxima of the k performance functions define a $k - 1$ dimensional shape in gene space (the convex hull of those k points), and optimal phenotypes are trapped within inside this shape. For two tasks, $k = 2$, this shape is a line segment:

$$G = \theta G_1 + (1 - \theta) G_2, 0 \le \theta \le 1.$$

References

Adler, M. *et al.* (2019) 'Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type Article Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type', *Cell Systems*. Elsevier Inc., 8(1), pp. 43-52.e5. doi: 10.1016/j.cels.2018.12.008.

Aktipis, C. A. *et al.* (2013) 'Life history trade-offs in cancer evolution', *Nature Reviews Cancer*. Nature Research, 13(12), pp. 883–892. doi: 10.1038/nrc3606.

Balaban, N. Q. *et al.* (2004) 'Bacterial persistence as a phenotypic switch.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 305(5690), pp. 1622–5. doi: 10.1126/science.1099390.

Barretina, J. *et al.* (2012) 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*, 483(7391), pp. 603–307. doi: 10.1038/nature11003.

Ciriello, G. *et al.* (2013) 'Emerging landscape of oncogenic signatures across human cancers', *Nature Genetics*. Nature Research, 45(10), pp. 1127–1133. doi: 10.1038/ng.2762.

Curtis, C. *et al.* (2012) 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.', *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 486(7403), pp. 346–52. doi: 10.1038/nature10983.

Dawson, S.-J. J. *et al.* (2013) 'A new genome-driven integrated classification of breast cancer and its implications.', *The EMBO journal*. Nature Publishing Group, 32(5), pp. 617–28. doi: 10.1038/emboj.2013.19.

Driessens, G. *et al.* (2012) 'Defining the mode of tumour growth by clonal analysis', *Nature*. Nature Publishing Group, 488(7412), pp. 527–530. doi: 10.1038/nature11344.

Evdokimova, V. *et al.* (2009) 'Reduced proliferation and enhanced migration: two sides of the same coin? Molecular mechanisms of metastatic progression by YB-1.', *Cell cycle*, 8(18), pp. 2901–6. doi: 10.4161/cc.8.18.9537.

Futuyma, D. J. and Moreno, G. (1988) 'The Evolution of Ecological Specialization', *Annual Review of Ecology and Systematics*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 19(1), pp. 207–233. doi: 10.1146/annurev.es.19.110188.001231.

Gillies, R. J. *et al.* (2018) 'Eco-evolutionary causes and consequences of temporal changes in intratumoural blood flow', *Nature Reviews Cancer*. Nature Publishing Group, p. 1. doi: 10.1038/s41568-018-0030-7.

Gillies, R. J., Verduzco, D. and Gatenby, R. A. (2012) 'Evolutionary dynamics of carcinogenesis and why targeted therapy does not work', *Nature Reviews Cancer*. Nature Publishing Group, 12(7), pp. 487–493. doi: 10.1038/nrc3298.

Gonzalez-Perez, A. *et al.* (2013) 'IntOGen-mutations identifies cancer drivers across tumor types', *Nature Methods*. Nature Publishing Group, 10(11), pp. 1081–1082. doi: 10.1038/nmeth.2642.

Hafner, M. *et al.* (2016) 'Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs', *Nature Methods*. Nature Research, 13(6), pp. 521–527. doi: 10.1038/nmeth.3853.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: the next generation.', *Cell*. Elsevier, 144(5), pp. 646–74. doi: 10.1016/j.cell.2011.02.013.

Hart, Y. *et al.* (2015) 'Inferring biological tasks using Pareto analysis of high-dimensional data.', *Nature methods*. Nature Publishing Group, 12(3), pp. 233–235. doi: 10.1038/nmeth.3254.

Hatzikirou, H. *et al.* (2012) '"Go or grow": The key to the emergence of invasion in tumour progression?', *Mathematical Medicine and Biology*, 29(1), pp. 49–65. doi: 10.1093/imammb/dqq011.

Heiser, L. M. *et al.* (2012) 'Subtype and pathway specific responses to anticancer compounds in breast cancer.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 109(8), pp. 2724–9. doi: 10.1073/pnas.1018854108.

Iorio, F. *et al.* (2016) 'A Landscape of Pharmacogenomic Interactions in Cancer', *Cell*, 166(3), pp. 740–754. doi: 10.1016/j.cell.2016.06.017.

Kim, C. *et al.* (2018) 'Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing', *Cell*. Elsevier Inc., 173(4), pp. 1–15. doi: 10.1016/j.cell.2018.03.041.

Korem, Y. *et al.* (2015) 'Geometry of the Gene Expression Space of Individual Cells', *PLOS Computational Biology*, 11(7), p. e1004224. doi: 10.1371/journal.pcbi.1004224.

Labi, V. and Erlacher, M. (2015) 'How cell death shapes cancer.', *Cell death & disease*, 6(3), p. e1675. doi: 10.1038/cddis.2015.20.

Lan, X. *et al.* (2017) 'Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy', *Nature*. Nature Publishing Group, 549(7671), pp. 227–232. doi: 10.1038/nature23666.

Laplante, M. and Sabatini, D. M. (2012) 'mTOR Signaling in Growth Control and Disease', *Cell*, 149(2), pp. 274–293. doi: 10.1016/j.cell.2012.03.017.

McLendon, R. *et al.* (2008) 'Comprehensive genomic characterization defines human glioblastoma genes and core pathways', *Nature*. Nature Publishing Group, 455(7216), pp. 1061–1068. doi: 10.1038/nature07385.

Merlo, L. M. F. *et al.* (2006) 'Cancer as an evolutionary and ecological process.', *Nature reviews. Cancer*, 6(12), pp. 924–35. doi: 10.1038/nrc2013.

Nik-Zainal, S. *et al.* (2016) 'Landscape of somatic mutations in 560 breast cancer whole-genome sequences', *Nature*. Nature Publishing Group, 534(7605), pp. 1–20. doi: 10.1038/nature17676.

Nowell, P. C. (1976) 'The clonal evolution of tumor cell populations.', *Science (New York, N.Y.)*, 194(4260), pp. 23–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/959840 (Accessed: 24 July 2017).

Pemovska, T. *et al.* (2013) 'Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia.', *Cancer discovery*. American Association for Cancer Research, 3(12), pp. 1416–29. doi: 10.1158/2159-8290.CD-13-0350.

Pereira, B. *et al.* (2016) 'The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes', *Nature Communications*, 7(May), p. 11479. doi: 10.1038/ncomms11479.

Reiter, J. G. *et al.* (2017) 'Reconstructing metastatic seeding patterns of human cancers', *Nature Communications*. Nature Publishing Group, 8, p. 14114. doi: 10.1038/ncomms14114.

Santarius, T. *et al.* (2010) 'A census of amplified and overexpressed human cancer genes.', *Nature reviews. Cancer*. Nature Publishing Group, 10(1), pp. 59–64. doi: 10.1038/nrc2771.

Schluter, D. (1996) 'Adaptive radiation along genetic lines of least resistance', *Evolution*, 50(5), pp. 1766–1774. doi: 10.1111/j.1558-5646.1996.tb03563.x.

Scott, M. *et al.* (2010) 'Interdependence of Cell Growth and Gene Expression: Origins and Consequences', *Science*, 330(6007), pp. 1099–1102. doi: 10.1126/science.1192588.

Sheftel, H. *et al.* (2018) 'Evolutionary trade-offs and the structure of polymorphisms.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 373(1747), p. 20170105. doi: 10.1098/rstb.2017.0105.

Shen, R., Olshen, A. B. and Ladanyi, M. (2009) 'Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis', *Bioinformatics*. Oxford University Press, 25(22), pp. 2906–2912. doi: 10.1093/bioinformatics/btp543.

Shoval, O. *et al.* (2012) 'Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space', *Science*, 336(6085), pp. 1157–1160. doi: 10.1126/science.1217405.

Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(43), pp. 15545–50. doi: 10.1073/pnas.0506580102.

Sullivan, R. and Graham, C. H. (2007) 'Hypoxia-driven selection of the metastatic phenotype', *Cancer and Metastasis Reviews*, 26(2), pp. 319–331. doi: 10.1007/s10555-007-9062-2.

Szekely, P. *et al.* (2015) 'The Mass-Longevity Triangle: Pareto Optimality and the Geometry of Life-History Trait Space', *PLoS Computational Biology*, 11(10), pp. 1–19. doi: 10.1371/journal.pcbi.1004524.

Tendler, A., Mayo, A. and Alon, U. (2015) 'Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells.', *BMC systems biology*, 9(1), p. 12. doi: 10.1186/s12918-015-0149-z.

Tirosh, I. *et al.* (2016) 'Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 352(6282), pp. 189–96. doi: 10.1126/science.aad0501.

Tsai, J. H. and Yang, J. (2013) 'Epithelial-mesenchymal plasticity in carcinoma metastasis', *Genes & Development*, 27(20), pp. 2192–2206. doi: 10.1101/gad.225334.113.

Vogelstein, B. *et al.* (2013) 'Cancer Genome Landscapes', *Science*, 339(6127). Available at: http://science.sciencemag.org.ezproxy.weizmann.ac.il/content/339/6127/1546.long (Accessed: 8 April 2017).

Wang, Y. K. *et al.* (2017) 'Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes', *Nature Genetics*, 49(6), pp. 856–865. doi: 10.1038/ng.3849.